# Typological profiles -- L1 varieties

Benedikt Szmrecsanyi
University of Freiburg

Acknowledgments

I am grateful to Christoph Wolk for creating the network representation in Figure 4, and for valuable feedback. The usual disclaimers apply.

## 1. Introduction

This chapter is concerned with the morphosyntactic profiles of the 30 L1 varieties in the WAVE survey. Compared to the morphosyntax survey (see Kortmann and Szmrecsanyi 2004) coming with the *Handbook of Varieties of English* (Kortmann et al. 2004) (henceforth: the HVE survey), the most obvious advancement of the WAVE survey is that it has a much broader variety and feature coverage. In addition, as for L1 varieties of English specifically the new WAVE survey also systematically distinguishes between TRADITIONAL L1 VARIETIES, and HIGH-CONTACT L1 VARIETIES. This distinction builds on the TRUE TYPOLOGICAL SPLIT hypothesis formulated by Trudgill (2009) (see also Trudgill 2011), according to which traditional (or: low-contact) varieties are characterized by COMPLEXIFICATION (more irregularity, less transparency) while high-contact varieties are characterized by SIMPLIFICATION (less irregularity, more transparency), thanks to past and present adult language acquisition. In this spirit, the WAVE survey defines the 10 traditional L1 varieties in the sample (O&SE, ScE, North, SW, EA, SE, NfldE, OzE, AppE, SEAmE) as "regional dialects (non-standard varieties) which are long-established mother tongue varieties and which are characterized by a low degree of contact with other dialects and languages" (http://www.ewave-atlas.org). By contrast, the 20 high-contact L1 varieties covered in the survey (IrE, ManxE, WelE, ChIsE, CollAmE, UAAVE, RAAVE, EAAVE, BahE, LibSE, WhZimE, WhSAfE, CollSgE, AusE, AusVE, AbE, NZE, FlkE, StHE, TdCE) are defined as varieties "characterized by a high degree of contact between different dialects of English and/or between English and other languages" (http://www.ewave-atlas.org).

All quantitative statements in this chapter are based on a binary distinction between varieties that attest a particular feature (however pervasively or sporadically), and those that do not. This is another way of saying that we conflate the ratings 'A' ("feature is pervasive or obligatory"), 'B' ("feature is neither pervasive nor extremely rare"), and 'C' ("feature exists, but is extremely rare") in the survey into an 'attested' category. The other values ("attested absence", "not applicable", "don't know") are merged into a 'not attested' category. This distinction is admittedly somewhat simplistic but ensures compatibility with analytical work based on the 2004 HVE survey (for example, Szmrecsanyi and Kortmann 2009a, 2009b), as well as with other synopses in the present volume.

This chapter is structured as follows. In Section 2, we explore particularly frequent and rare features in L1 varieties of English, and we also identify features that set apart traditional L1 varieties from high-contact L1 varieties of English. In Section 3, we discuss aggregate similarities and differences between L1 varieties of English from a bird's eye perspective, taking into account variability in all 235 features covered in the WAVE survey. Section 4 offers some concluding remarks

## 2. Feature synopsis

We thus begin by adopting a feature-centered perspective. What are the most and least frequent features attested in L1 varieties of English? What are the features that set apart traditional L1 varieties of English from high-contact L1 varieties of English?

### 2.1 Most frequent features

Table 1 lists those 8 features in the WAVE survey that are attested (as either 'A', 'B', or 'C') in at least 28 of the 30 L1 varieties of English covered (i.e. in over 90%). We note, first, that this list is basically the second coming of Kortmann and Szmrecsanyi's (2004: Table 23) similar list based on the HVE survey.

| **feature #** | **description** | **domain** | **# attested** | **exceptions** |
|---|---|---|---|---|
| F7 | *Me* instead of *I* in coordinate subjects | Pronouns | 30 | – |
| F147 | *Was* for conditional *were* | Verb Morphology | 30 | – |
| F221 | Other adverbs have the same form as adjectives | Adverbs & Prepositions | 30 | – |
| F159 | *Never* as preverbal past tense negator | Negation | 29 | O&SE (D) |
| F172 | Existential / presentational *there's/there is/there was* with plural subjects | Agreement | 29 | CollSgE (D) |
| F8 | *Myself/meself* instead of *I* in coordinate subjects | Pronouns | 28 | CollSgE (D)<br>AbE (D) |
| F34 | Forms or phrases for the second person plural pronoun other than *you* | Pronouns | 28 | O&SE (D)<br>SE (D) |
| F220 | Degree modifier adverbs have the same form as adjectives | Adverbs & Prepositions | 28 | LibSE (D)<br>EA (D) |

Table 1: Most frequent features in L1 varieties of English (attested in at least 28 of 30 varieties in the survey).

### 2.2 Features rare and very rare in L1 varieties

In what follows we discuss features that are rare in L1 varieties, such that they are attested only in two of 30 L1 varieties ("*rara*"; see Table 2), and features that are very rare, such that they are attested only once in 30 L1 varieties ("*rarissima*"; see Table 3).

What is it that one should expect to find here? From reading the literature on sociolinguistically conditioned complexity variation (e.g. Szmrecsanyi and Kortmann 2009c; Trudgill 2009), one may conjecture that *rara* and *rarissima* should be found primarily in traditional, low-contact varieties, which supposedly have a knack for useless complexification. This hypothesis relies on the implicit assumption that *rara* and *rarissima* are prime candidates for non-functional “historical baggage” (Trudgill 1999: 149), “ornamental elaboration” (McWhorter 2001: 132) or “baroque accretion[s]” (McWhorter 2001: 126). After all, if *rara* and *rarissima* – or so the argument goes – weren't baroque accretions without a clear functional bonus, more varieties should have them. The WAVE survey, however, does not corroborate this neat hypothesis, plausible as it may seem. The top attestors of *rara* and *rarissima* in Tables 2 and 3 are actually CollSgE and AbE, two high-contact varieties *par excellence*. Further *rara* and *rarissima* hotspots include RAAVE, LibSE, and TdCE. Meanwhile, the only traditional low-contact varieties that actually do attest *rara* and *rarissima at all* are NfldE and SEAmE and the North of England. We conclude that *rara* and *rarissima* do not seem to speak to the issue of sociolinguistic typology.

| **feature #** | **description** | **domain** | **attested in …** |
|---|---|---|---|
| F6 | Generalized third person singular pronoun: object pronouns | Pronouns | AbE<br>NfldE |
| F87 | Attributive adjectival modifiers follow head noun | Noun Phrase | ManxE<br>AbE |
| F94 | Progressive marker *stap* or *stay* | Tense & Aspect | UAAVE,<br>RAAVE |
| F98 | *After*-perfect | Tense & Aspect | IrE<br>NfdlE |
| F107 | Completive/perfect marker *slam* | Tense & Aspect | RAAVE<br>SEAmE |
| F110 | *Finish*-derived completive markers | Tense & Aspect | LibSE<br>CollSgE |
| F137 | Special inflected forms of *do* | Verb Morphology | WelE<br>North |
| F143 | Transitive verb suffix *-em/-im/-um* | Verb Morphology | AbE<br>TdCE |
| F148 | Serial verbs: *give* = 'to, for' | Verb Morphology | RAAVE<br>CollSgE |
| F149 | Serial verbs: *go* = 'movement away from' | Verb Morphology | CollSgE<br>AbE |
| F153 | *Give* passive: NP1 (patient) + *give* + NP2 (agent) + V | Voice | CollSgE<br>TdCE |
| F162 | *No more/nomo* as negative existential marker | Negation | RAAVE<br>AbE |
| F230 | Doubly filled COMP-position with *wh*-words | Discourse & Word Order | IrE<br>TdCE |

Table 2: *Rara* (features attested twice in 30 L1 varieties).

| **feature #** | **description** | **domain** | **attested in …** |
|---|---|---|---|
| F36 | Distinct forms for inclusive/exclusive first person non-singular | Pronouns | AbE |
| F37 | More number distinctions in personal pronouns than simply singular vs. plural | Pronouns | AbE |
| F73 | Existential construction to express possessive | Noun Phrase | ManxE |
| F108 | *Ever* as marker of experiential perfect | Tense & Aspect | CollSgE |
| F115 | Volition-based future markers other than *will* | Tense & Aspect | StHE |
| F141 | Other forms/phrases for copula 'be': before locatives | Verb Morphology | BahE |
| F150 | Serial verbs: *come* = 'movement towards' | Verb Morphology | CollSgE |
| F167 | Fronted invariant tag | Negation | LibSE |
| F168 | Special negative verbs in imperatives | Negation | WelE |
| F195 | Postposed *one* as sole relativizer | relativization | CollSgE |
| F196 | Correlative constructions | relativization | SEAmE |
| F206 | Existentials with forms of *have* | Complementation | LibSE |

Table 3: *Rarissima* (features attested once in 30 L1 varieties).

## 2.3 Features absent from L1 varieties

For the sake of completeness, Table 4 reports five features that are entirely absent from the L1 varieties covered in the WAVE survey. It should surprise no one that these features typically have a creole feel to them, such as F17 creation of possessive pronouns with prefix *fi-* + personal pronoun, as in *fi-mi* ‘my’), or F152 (serial verbs: constructions with 4 or more verbs, as in *Agnes ron komot go lef in mama na makit*).

| feature # | description | domain |
|---|---|---|
| F17 | Creation of possessive pronouns with prefix *fi-* +personal pronoun | Pronouns |
| F18 | Subject pronoun forms as (modifying) possessive pronouns: first person singular | Pronouns |
| F40 | Plural forms of interrogative pronouns: reduplication | Pronouns |
| F152 | Serial verbs: constructions with 4 or more verbs | Verb Morphology |
| F199 | Reduced relative phrases preceding head-noun | relativization |

Table 4: Features not attested in 30 L1 varieties.

## 2.2. Low-contact versus high-contact L1 varieties

We now move on to a discussion of the features that differentiate well between high-contact L1 varieties and traditional, low-contact L1 varieties. To this end, we examine

- the top ten features which are, in comparison to high-contact varieties, conspicuously frequent in traditional, low-contact varieties (Table 5)
- the top ten features which vis-à-vis traditional, low-contact varieties are fairly overrepresented in high-contact varieties (Table 6).

To illustrate: it turns out (see the first entry in Table 5) that F188 (Relativizer *at*, as in *This is the man* at *painted my house*) is attested in only 15% of all high-contact varieties in the survey. At the same time, 70% of all traditional, low-contact varieties have this feature. F188 is thus highly characteristic of traditional L1 varieties. By contrast, F132 (Zero past tense forms of regular verbs, as in *I walk* ‘I walked’) is attested in only 10% of traditional L1 varieties but in 57% of all high-contact varieties (see the first entry in Table 6). This is another way of saying that F132 is diagnostic of high-contact L1 varieties.

Again, we begin by drawing on sociolinguistic theory to inform expectations. Following Trudgill's (2009) "true typological split" hypothesis, we would expect to see many complexifying features in Table 5, because traditional L1 varieties are presumably complexifying. On the other hand, there should be many simplifying features in Table 6, since high-contact L1 varieties are supposedly simplifying. Inevitable subjectivity in rating individual features one way or the other notwithstanding, this hypothesis seems to be borne out by and large. In Table 5, we find a good deal of features that will strike many as a tad ornamental and somewhat 'quirky' – and thus, as complex and/or as complexifying. Consider F181, which among other things captures the Northern Subject Rule (Agreement sensitive to subject type, as in *birds sings* vs. *they sing*), or F2 (*He*/*him* used for inanimate referents, as in *I bet thee cansn' climb he* [= a tree]), or F144 (Use of *gotten* and *got* with distinct meanings (dynamic vs. static), as in *They've gotten a new car* [‘have received’] vs. *They've got a new car* [‘possess’]).

| feature # | Description | domain | % attestations in high-contact varieties | % attestations in traditional varieties |
|---|---|---|---|---|
| F188 | Relativizer *at* | Relativization | 15% | 70% |
| F181 | Agreement sensitive to subject type | Agreement | 25% | 70% |
| F35 | Forms or phrases for the second person singular pronoun other than *you* | Pronouns | 20% | 60% |
| F2 | *He*/*him* used for inanimate referents | Pronouns | 25% | 60% |
| F232 | Either order of objects in double object constructions (if both objects are pronominal) | Discourse & Word Order | 25% | 60% |
| F187 | Relativizer as | relativization | 5% | 40% |
| F202 | Unsplit *for to* in infinitival purpose clauses | Complementation | 45% | 80% |
| F32 | Distinction between emphatic vs. non-emphatic forms of pronouns | Pronouns | 0% | 30% |
| F144 | Use of *gotten* and *got* with distinct meanings (dynamic vs. static) | Verb Morphology | 20% | 50% |
| F96 | *There* with past participle in resultative contexts | Noun Phrase | 60% | 90% |

Table 5: Top 10 features whose presence is most characteristic of traditional varieties (ordered by per cent point differential of attestation).

At the same time, the features in Table 6 – which is concerned with features particularly characteristic of high-contact L1 varieties – have a simple or simplifying flavor to them, according to customary complexity notions. Specifically, Table 6 covers an abundance of null and deletion phenomena that do away with overt contrasts and markers (often inflectional) that are obligatory in Standard English. Consider, for example, F132 (Zero past tense forms of regular verbs), or F58 (Plural marking generally optional: for nouns with non-human referents, as in *The tree-Ø don't grow very tall up there*), or copula deletion phenomena (F176 – F178). By way of an interim summary, we thus stress that the difference between traditional L1 varieties and high-contact L1 varieties is in line with sociolinguistic theory (Trudgill 2009, 2011).

| feature # | Description | domain | % attestations in high-contact varieties | % attestations in traditional varieties |
|---|---|---|---|---|
| F3 | Alternative forms/phrases for referential (non-dummy) *it* | Pronouns | 70% | 20% |
| F66 | Indefinite article *one/wan* | Noun Phrase | 55% | 10% |
| F132 | Zero past tense forms of regular verbs | Verb Morphology | 55% | 10% |
| F176 | Deletion of copula *be*: before NPs | Agreement | 45% | 0% |
| F177 | Deletion of copula *be*: before AdjPs | Agreement | 45% | 0% |
| F178 | Deletion of copula *be*: before locatives | Agreement | 45% | 0% |
| F174 | Deletion of auxiliary *be*: before progressive | Agreement | 60% | 20% |
| F21 | Subject pronoun forms as (modifying) possessive pronouns: third person plural | Pronouns | 50% | 10% |
| F38 | Specialized plural markers for pronouns | Pronouns | 50% | 10% |
| F58 | Plural marking generally optional: for nouns with non-human referents | Noun Phrase | 40% | 0% |

Table 6: Top 10 features whose presence is most characteristic of high-contact varieties (ordered by per cent point differential of attestation).

## 3. Similarities and distances between L1 varieties of English

In this section, we take a step back from occurrence likelihoods of individual morphosyntactic features, and focus instead on AGGREGATE morphosyntactic distances and similarities between L1 varieties of English. The subsequent discussion thus takes into account joint variability in all 235 features covered in the WAVE survey.

Technically, we calculate pairwise distances and similarities between varieties of English by using the well-known SQUARED EUCLIDEAN DISTANCE MEASURE, which in the case of the present dataset is fully proportional to the MANHATTAN DISTANCE MEASURE (see, for example, Aldenderfer and Blashfield 1984: 24–25). The squared Euclidean distance measure works in a very straightforward and intuitive way – it simply counts the number of different feature ratings. To illustrate: the two L1 varieties in the WAVE survey that are, on aggregate, most similar to each other are EA and FlkE. This is a variety pairing that scores a squared Euclidean distance value of 23 points. In other words, the two varieties disagree in regard to just 23 (of 235) feature classifications (recall, again, that we conflate the 'A', 'B', and 'C' ratings into a broad 'attested' category here). By contrast, the two L1 varieties in the survey that are most distant morphosyntactically are O&SE and RAAVE, which disagree in terms of no less than 143 feature classifications; hence, their squared Euclidean distance score is 143.

### 3.1. L1 varieties on a map: the geographic perspective

Is there a geolinguistic signal in the dataset? We initially address this question by projecting aggregate distances and similarities, according to the squared Euclidean distance measure, to geography using a so-called NETWORK MAP (Nerbonne and Heeringa 1997) in Figure 1. The projection visually depicts pairwise similarities between varieties by making link blueness inversely proportional to pairwise morphosyntactic distance. Therefore, in Figure 1 we observe a network of relatively strong morphosyntactic similarities between British L1 varieties of English, L1 varieties in North America, and L1 varieties in Australia and the Pacific region. FlkE also bears strong similarities with the aforementioned variety groups. On the other hand, all L1 varieties spoken in Africa, and CollSgE do not link particularly well to the other L1 varieties in the sample.

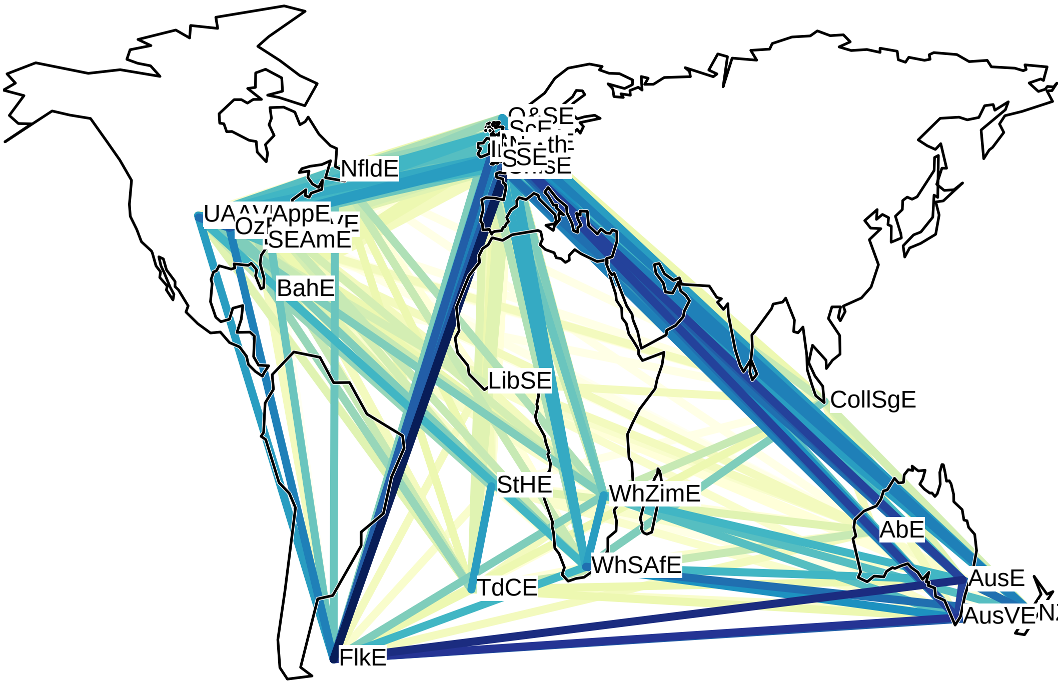


Figure 1. Visualizing aggregate similarities: Network map. Link blueness is proportional to pairwise morphosyntactic similarity (distance limit: 15,000km).

According to the FUNDAMENTAL DIALECTOLOGY PRINCIPLE (Nerbonne and Kleiweg 2007: 154), geographic proximity between dialects should predict dialectal similarity between dialects. Needless to say, the dataset that we analyze here is not a classically dialectological dataset – the L1 varieties it describes are neither traditional dialects, nor are they confined to a particular language area. Still, we are interested in the extent to which geographic proximity predicts morphosyntactic similarity in this dataset. In this spirit, Figure 2 locates all 30 × 29/2 = 435 unique variety pairings in the dataset on a two-dimensional plane which plots morphosyntactic distance (in squared Euclidean distance points) on the *y*-axis and geographic distance (in km) on the *x*-axis (which is log-scaled). For example, we have seen that EA and FlkE are close morphosyntactically, yet of course the two varieties are fairly distant geographically (some 13,000km, to be precise). This is why we find this particular pairing in the lower right corner of Figure 2. In all, it is amply clear that while there is a slight positive correlation between morphosyntactic and geographic distance, there are many exceptions (and the SE-FlkE pairing is just one of them). In quantitative terms, the correlation between morphosyntactic and geographic distance comes out as $r = .226$ ($p < .001$), which means that geographic distance explains a meager 5.1% ($R^2 = 0.051$) of the morphosyntactic variability in the dataset.

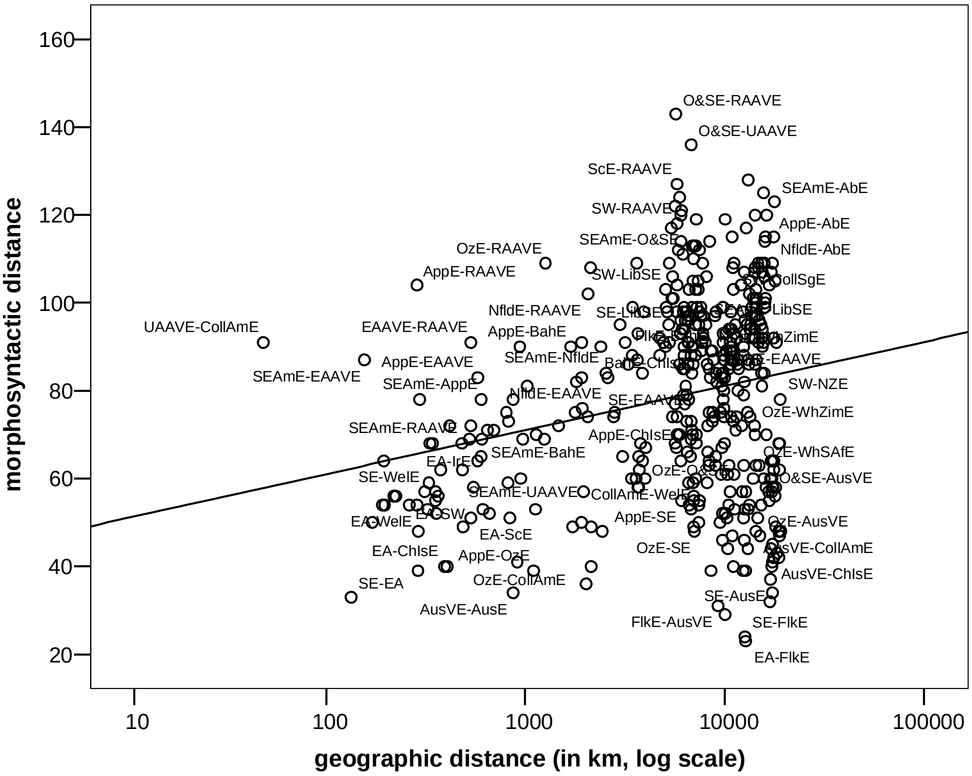


Figure 2. Scatterplot, morphosyntactic distance (squared Euclidean) versus geographic distances (in km). Each dot represents one variety pairing (selected labels displayed only).

### 3.2. L1 varieties in a plane: Multidimensional Scaling

So the upshot is that the fundamental dialectology principle more or less fails to explain the distribution of morphosyntactic similarities in L1 varieties of English world-wide. To probe factors that may be more elucidative, we now turn to Figure 3, which uses MULTIDIMENSIONAL SCALING (Kruskal and Wish 1978) to locate L1 varieties in a two-dimensional plot. In this visualization, proximity between data points is proportional to aggregate morphosyntactic similarity. Thus, FlkE and EA end up close in the plot (because, as we have seen, their morphosyntactic profiles are very similar) while RAAVE and O&SE are far apart (as they should be, given how different the two varieties are morphosyntactically). Observe that Figure 3 does not yield striking regional clusters, which confirms our earlier finding that geography is a weak predictor of morphosyntactic similarities. But be that as it may, Figure 3 suggests three generalizations about L1 varieties of English from a bird's eye perspective,

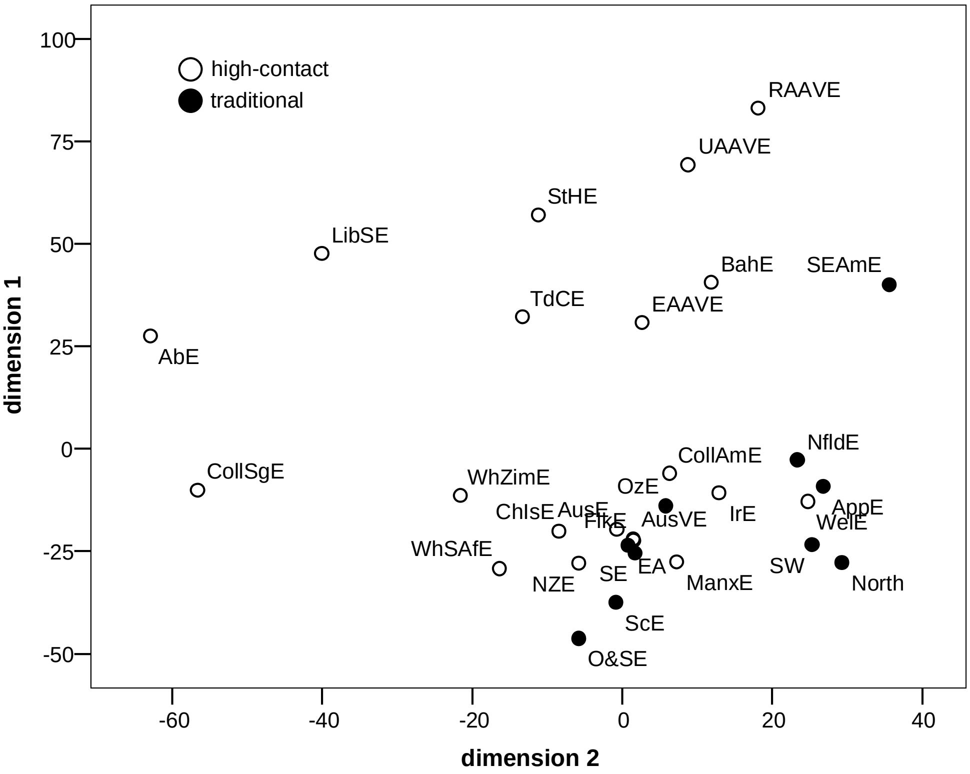
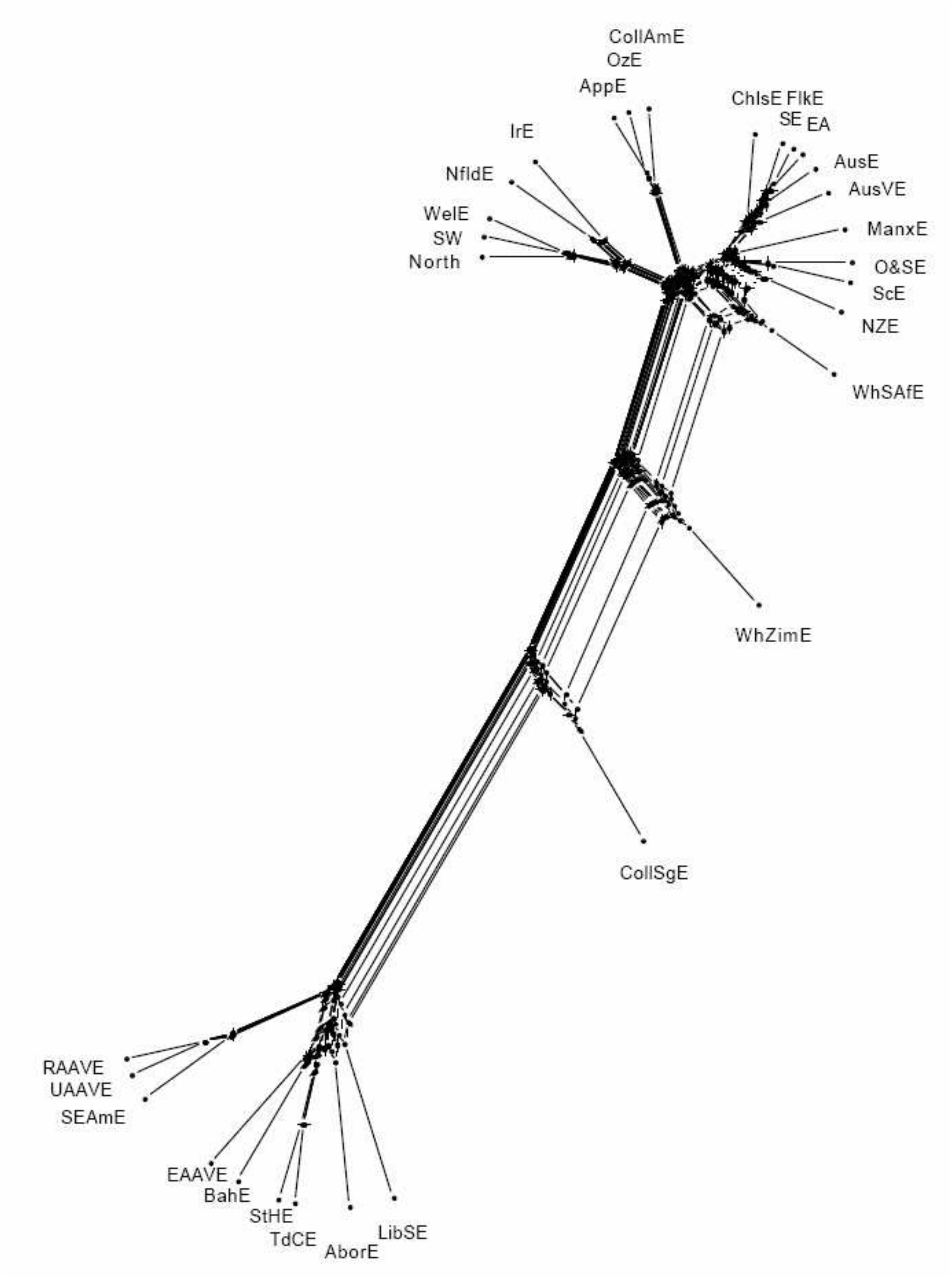


Figure 3: Visualizing aggregate similarities: MDS plot. Proximity in plot space is proportional to morphosyntactic similarity (MDS algorithm: Kruskal's method, *r* = .92).

First, there is a fairly impressive divide between traditional, low-contact varieties of English (black dots, clustered in the lower right quadrant, except for SEAmE, which is an outlier) and high-contact varieties of English. This pattern, of course, nicely dovetails with Trudgill's (2009, 2011) "true typological split" hypothesis. Second, traditional, low-contact L1 varieties form a tighter cluster than high-contact L1 varieties, which appear to be more of a mixed bag. Third (and relatedly), the reason for the high-contact L1 cluster's more extensive internal heterogeneity is that a number of high-contact L1 varieties (CollSgE, AbE, LibSE, the AAVEs, TdCE, StHE, and BahE – to be found in the top half of the plot) are noticeably different from other high-contact varieties of English. We conjecture that this is because CollSgE et al. are even more mixed and high-contact ('higher-contact', in other words) than more orthodox high-contact L1-varieties of English, such as CollAmE.

3.3. L1 varieties on a tree: a network diagram

An alternative way of visualizing aggregate similarities and distances in the dataset marshals NEIGHBORNET DIAGRAMS (Bryant and Moulton 2004). Originally developed in biometry and bioinformatics to represent uncertainty in phylogenies, and reticulate effects such as genetic recombination, NeighborNet diagrams are increasingly popular in linguistics (e.g. A. McMahon et al. 2007; Szmrecsanyi and Wolk 2011). Without insisting on a strictly phylogenetic interpretation, Figure 4 thus visualizes aggregate similarities and distances between L1 varieties of English.[1] Note that the diagram can be basically read like a family tree that is not rooted; branch lengths are proportional to linguistic distances. As in MDS plots, proximity in the plot broadly indicates morphosyntactic similarity, which is why FlkE and SE are depicted as closely related in Figure 4.



[1] Technically, we first used clustering with noise (Nerbonne, Kleiweg, and Manni 2008) to enhance the robustness of the similarity signal before submitting the dataset to the NeighborNet analysis.

Figure 4. Visualizing aggregate similarities: NeighborNet diagram. Internode distances (branch lengths) are proportional to cophenetic linguistic distances.

The NeighborNet diagram in Figure 4 takes a slightly different perspective than the MDS plot in Figure 3. The most crucial split, according to NeighborNet, is between what we have called 'higher-contact' L1 varieties (at the bottom of the diagram) and the other L1 varieties in the sample (exceptions: SEAmE is grouped with the 'higher-contact' varieties, and CollSgE is an isolate). In the 'other' cluster to the top of the diagram, we find some regional sub-groupings, such as CollAmE/OzE/AppE (an American cluster), and WelE/SW/North (a British cluster). NfldE is nicely – given the history of the variety – positioned between IrE and the aforementioned British sub-grouping.

## 4. Conclusion

This chapter has sought to profile L1 varieties of English in the WAVE survey. We have seen that L1 varieties are characterized by the near-universalhood of features such as *me* instead of *I* in coordinate subjects, *was* for conditional *were*, and adverbs that have the same form as adjectives. The analysis has also demonstrated that what sets apart traditional, low-contact L1 varieties if English (such as O&SE) from high-contact L1 varieties (such as AusE) is the presence, in inventories of traditional L1 varieties, of features that can be considered complex or complexifying. Contrariwise, high-contact L1 varieties are characterized by features that are simplifying. We also endeavored to sketch the big picture, which is fueled by the calculation of linguistic distances between varieties based on joint variability of all 235 features in the WAVE survey. On aggregate, geography (specifically, geographic distances) turns out to be a weak predictor of overall morphosyntactic similarities between L1 varieties of English, accounting for no more than 5% of the linguistic variance. Instead, we diagnosed a typological split in line with Trudgill (2009, 2011) between traditional L1 varieties, high-contact L1 varieties, and what we have dubbed 'higher-contact' L1 varieties of English (such as the AAVE varieties).

## References

Aldenderfer, Mark S., and Roger K. Blashfield 1984 *Cluster Analysis*. Newbury Park, London, New Delhi: Sage Publications.

Bryant, Davis, and Vincent Moulton 2004 Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution* 21 (2): 255–265. doi:10.1093/molbev/msh018

Kortmann, Bernd, Edgar Schneider, Kate Burridge, Raj Mesthrie, and Clive Upton (eds.) 2004 *A Handbook of Varieties of English.*, Vols.1-2. Berlin, New York: Mouton de Gruyter.

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

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

McMahon, April, Paul Heggarty, Robert McMahon, and Warren Maguire 2007 The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics* 11 (01): 113. doi:10.1017/S1360674306002139

McWhorter, John 2001 The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125–166.

Nerbonne, John, and Wilbert Heeringa 1997 Measuring Dialect Distance Phonetically. In: John Coleman (ed.), *Workshop on Computational Phonology, Special Interest Group of the Association for Computational Linguistics*, 11–18. Madrid.

Nerbonne, John, and Peter Kleiweg 2007 Toward a Dialectological Yardstick. *Journal of Quantitative Linguistics* 14 (2): 148–166.

Nerbonne, John, Peter Kleiweg, and Franz Manni 2008 Projecting dialect differences to geography: bootstrapping clustering vs. clustering with noise. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, 647–654. Berlin: Springer.

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

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

Szmrecsanyi, Benedikt, and Bernd Kortmann 2009c Between simplification and complexification: non-standard varieties of English around the world. In: Geoffrey Sampson, David Gil, and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 64–79. Oxford: Oxford University Press.

Szmrecsanyi, Benedikt, and Christoph Wolk 2011 Holistic corpus-based dialectology. *Brazilian Journal of Applied Linguistics / Revista Brasileira de Linguística Aplicada* 11 (2): 561–592.

Trudgill, Peter 1999 Language contact and the function of linguistic gender. *Poznan Studies in Contemporary Linguistics* 35: 133–152.

Trudgill, Peter 2009 Vernacular universals and the sociolinguistic typology of English dialects. In: Marrku Filppula, Juhani Klemola, and Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, 302–329. London: Routledge.

Trudgill, Peter 2011 *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford; New York: Oxford University Press.