Uncovering the big picture: measuring the typological relatedness of varieties of English

Benedikt Szmrecsanyi (KU Leuven)

This contribution surveys various large-scale quantitative techniques that have been utilized in the literature on varieties and dialects of English to determine their typological relatedness: (a) aggregative measures of distance or similarity, based on atlas or survey data; (b) Typological Profiling, a technique that draws on naturalistic text corpora to calculate usage- and frequency-based measures of grammatical analyticity and syntheticity; (c) a corpus-based method, inspired by work in information theory, that is designed to map out varieties based on how they differ in terms of language/dialect complexity: and (d) an approach to calculate distances between varieties as a function of the extent to which grammatical variation patterns in usage data are dissimilar.

1. Introduction

Determining relatedness between languages, dialects, and varieties is needless to say one of the core tasks in typology and dialect typology. In cross-linguistic typology we find essentially two ways of determining relatedness: top-down approaches, where relatedness is established on the basis of pre-defined categories or types; and bottom-up approaches, in which types or patterns emerge from comparatively unstructured data. Greenberg (1960), for example, utilizes a top-down approach to calculate various indices that operationalize pre-defined, Sapir-inspired parameters (e.g. synthesis, agglutination, suffixing, etc.). Conversely, Cysouw (2013) adopts among other things a bottom-up perspective to explore the strength of the correlation between geographical distance and typological distance, based on information from the World Atlas of Language Structure (WALS) (see Cysouw 2013: Fig. 3).

Against this backdrop, this chapter canvasses various large-scale quantitative approaches, both top-down and bottom-up, that have been employed in the English dialect typology literature to investigate the typological relatedness of varieties of English. All of these approaches can in principle also be utilized to investigate the dialect typology of other languages, such as Spanish. In this spirit, Section 2 reviews bottom-up work that calculates aggregate measures of distance or similarity, based on atlas or survey data. Section 3 is concerned with Typological Profiling, a technique inspired by work in quantitative morphological typology that draws on naturalistic text corpora to calculate usage- and frequency-based measures of grammatical analyticity and syntheticity. Section 4 reviews work that maps out varieties in terms of language complexity, which is a fashionable research topic in contemporary typology and sociolinguistics. Section 5 discusses how we may make use of variationist measures to calculate distances between varieties. Section 6 offers some concluding remarks.

2. Aggregate measures of distance or similarity

Bottom-up feature aggregation is often based on surveys and/or dialect atlases, in both crosslinguistic typology and in dialectology. The data source taking center stage to furnish

the case study in this section is the morphosyntax survey that accompanies the *Handbook of Varieties of English* (see Kortmann & Szmrecsanyi 2004 and http://www.varieties.mouton-content.com/). This survey was conducted by compiling a catalogue of 76 non-standard features, and subsequently the authors of the chapters in the morphosyntax volume of the *Handbook* were asked to specify into which of the following three categories the features fall in the relevant variety:

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

The survey covers 46 vernacular varieties of English around the world.[1]

Szmrecsanyi and Kortmann (2009a) present an analysis of the survey that seeks, among other things, to measure and visually depict linguistic distances between the varieties of English covered in the survey, for the sake of identifying a signal of typological relatedness. This objective is accomplished as follows: to facilitate statistical processing, Szmrecsanyi and Kortmann first combine the A and B ratings into an "attested" category, while C counts as "unattested". Subsequently, to calculate distances the study marshals the well-known squared Euclidean (a.k.a. Manhattan) distance measure, which calculates the distance between any two varieties as the number of feature classifications with regard to which the two varieties differ. We exemplify by taking a look at the first 10 features in the survey and two varieties, Irish English and Singapore English (Table 1).

	Irish English (IrE)	Singapore English (SgE)
[1] *them* instead of demonstrative *those*	attested	not attested
[2] *me* instead of possessive *my*	attested	not attested
[3] special forms or phrases for 2nd ps pl pronoun	attested	not attested
[4] regularized reflexives-paradigm	attested	not attested
[5] object pronoun forms serving as base for reflexives	attested	not attested
[6] lack of number distinction in reflexives	attested	attested
[7] *she*/*her* used for inanimate referents	attested	not attested
[8] generic *he*/*his* for all genders	not attested	not attested
[9] *myself*/*meself* in a non-reflexive function	attested	attested
[10] *me* instead of *I* in coordinate subjects	attested	attested

Table 10.1. Feature classifications (first 10 features only) in the morphosyntax survey accompanying the *Handbook of Varieties of English* (Kortmann et al. 2004) for Irish English and Singapore English.

According to Table 1, Irish English and Singapore English share four feature classifications ([6]: attested in both; [8] not attested in either; [9]: attested in both; [10] attested in both),

[1] There is also an updated version, the "Electronic World Atlas of Varieties of English" (eWAVE) (see Kortmann & Lunkenheimer 2013).

and do not agree with regard to the remaining six features (for example, *them* instead of demonstrative *those* is attested in Irish English but not in Singapore English). The distance between Irish English and Singapore English thus comes out as 6 squared Euclidean distance points. Repeating this exercise considering all 76 features in the essay as well as all 46 × 45/2 = 1,035 unique variety pairings (given that the survey covers 46 varieties) yields as its end product a so-called distance matrix (of dimensionality 46 × 46) that specifies pairwise linguistic distances between all varieties covered in the survey.

High-dimensional distance matrices like this are hard to grasp via eye-balling, so Szmrecsanyi and Kortmann (2009a) turn to Multidimensional Scaling (MDS) (Kruskal & Wish 1978) to visually depict linguistic similarity/distance relationships between the varieties under study. MDS is a well-known dimension reduction technique that translates distances between objects (in the case at hand, squared Euclidean linguistic distances between varieties of English) in high-dimensional space into a lower-dimensional representation that can be visually depicted in plots such as the plot in Figure 1: the data points in this plot are the varieties of English subject to study, and distances between the varieties in the plot are proportional to the linguistic distance between the varieties. In other words, proximity in the plot indicates linguistic similarity and typological relatedness.

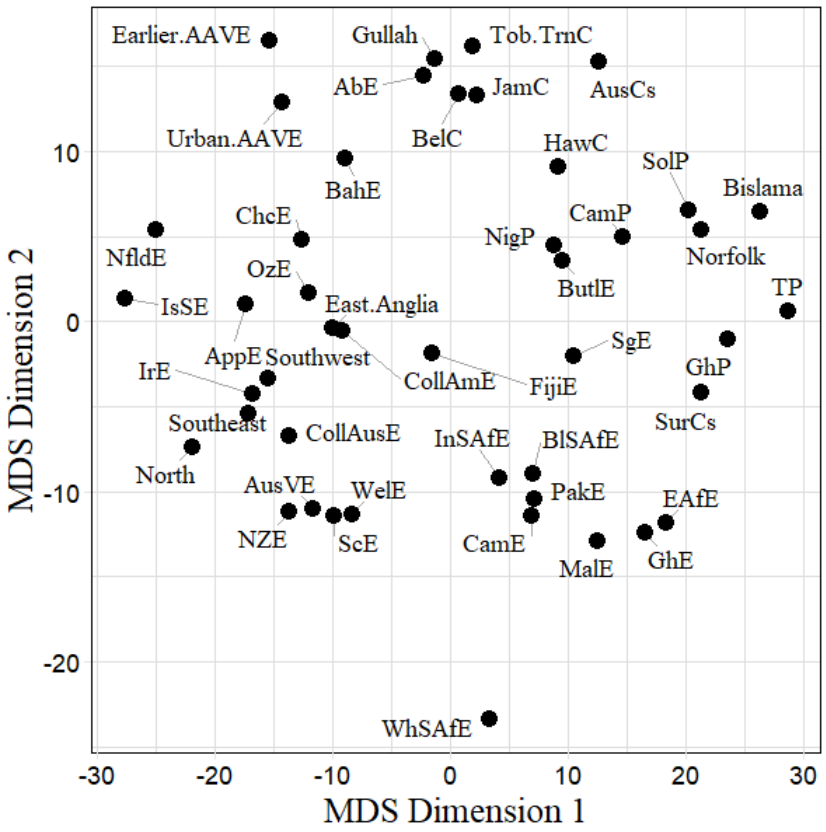

Figure 10.1. Metric Multidimensional Scaling (MDS) map, based on the morphosyntax survey in the *Handbook of Varieties of English* (see Kortmann & Szmrecsanyi 2004). Proximity between varieties in the plot is proportional to their aggregate morphosyntactic similarity.[2]

Let us summarize the main points in Szmrecsanyi and Kortmann (2009a: 8–10). What we see in Figure 1 is a three-way split: the English-based pidgin and creole languages (e.g. Tok Pisin [TP], Ghanaian Pidgin [GhP]) are located in the right half of the plot, while native L1 varieties (e.g. Newfoundland English [NfldE], dialects in the North of England [North]) are located in the left half of the diagram. In other words, the areas where one finds L1 varieties (left) and where one finds pidgins and creoles (right) do not tend to overlap, which is another way of saying that we are dealing with distinct types.

Sandwiched, as it were, in between the L1 variety cluster and pidgin/creole cluster in Figure 1 we find indigenized L2 varieties of English (e.g. Bahamian English [BahE], Fiji English [FijiE], Malaysian English [MalE], and so on). The interpretation, of course, is that indigenized L2 varieties form an intermediate type. That being said, there are some borderline cases: for example, Butler English (ButlE) and Singapore English (SgE) conspicuously border on the pidgin/creole area. The coordinates of Butler English, in point of fact, are consonant with Hosali's (2004: 1032) evaluation that it is hard to say whether ButlE should be considered a pidgin or an early fossilized interlanguage. As for SgE, there is controversy as to whether the variety should be understood in terms of a post-creole continuum or in terms of diglossia (Wee 2004: 1070). Note in this connection that colloquial Singapore English has been called a 'creoloid' or 'semi-pidgin' (Gil 2003: 469f).

Chicano English (ChcE), on the other hand, intrudes into the L1 cluster. In the *Handbook,* the label ChcE refers to ethnolects spoken by (i) Mexican Americans who acquired English as their first language; (ii) Mexican Americans who acquired English and Spanish simultaneously; and (iii) speakers who began to acquire English at school age (Bayley & Santa Ana 2004: 374). Given this heterogeneity, is it maybe not that surprising that ChcE is somewhat peripheral to the L2 cluster in Figure 1. The location of Orkney and Shetland English, close to the L2 cluster, is also curious. While there should be no doubt that the English spoken in Orkney and Shetland is an L1 vernacular, the dialect has an exceptional history, both linguistically and politically. Specifically, the variety attests many Scandinavian features that we do not find in other British varieties of English (Melchers 2004), which may be part of the explanation why Orkney and Shetland English is set apart from many other L1 varieties in Figure 1. Finally, we should add a word about Earlier African American English (Earlier AAVE) and (contemporary) Urban African American English (Urban AAVE), which Figure 1 locates on the periphery of the L1 cluster and in relative proximity to the L2 and pidgin/creole groups. For reasons of space we cannot review here the rich literature on the origins of AAVE. Suffice it to say that sociolinguists agree that the history of AAVE is unique: a contentious point is the question of whether AAVE has creole origins or English origins

[2] All calculations and analyses in this chapter were conducted using the statistical software package R (https://www.r-project.org/).

(Poplack 2000). The dataset under study suggests that both Earlier AAVE and Urban AAVE are fairly distant morphosyntactically from 'white' varieties of American English, but according to their location in Figure 1 they are even more distinct from English-based pidgin and creole languages.

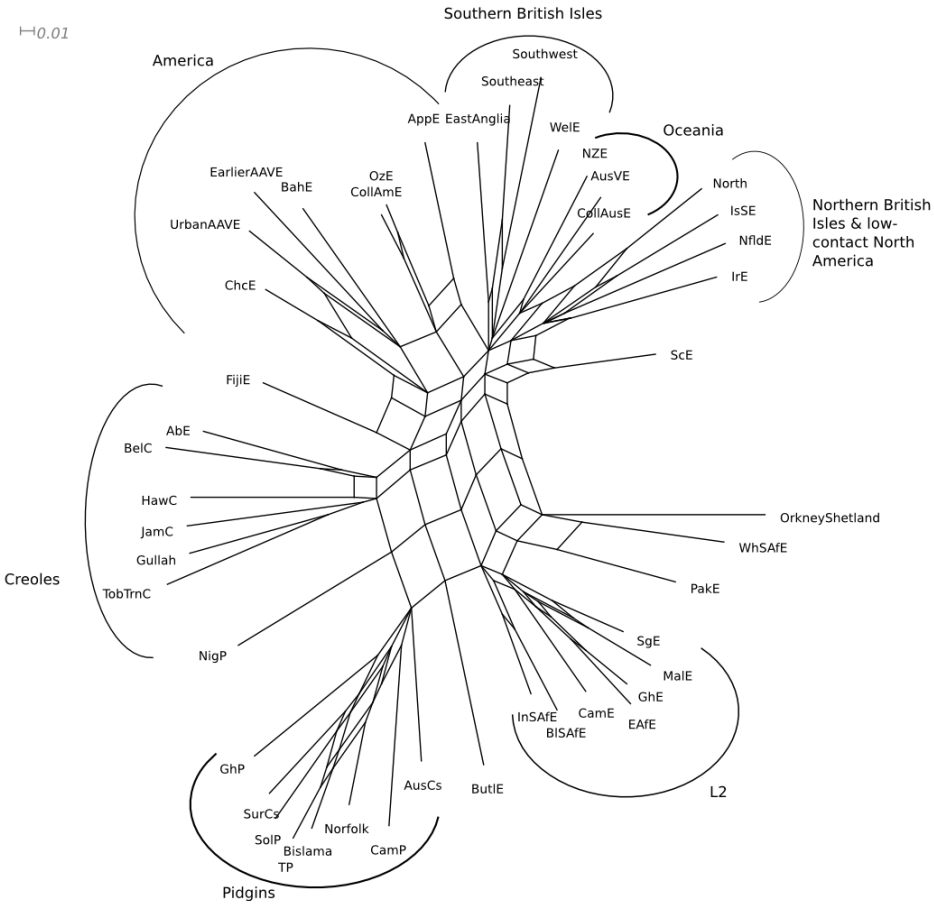

Figure 10.2. Visualizing aggregate similarities: NeighborNet diagram, based on the morphosyntax survey in the *Handbook of Varieties of English* (see Kortmann & Szmrecsanyi 2004). Internode distances (branch lengths) are proportional to cophenetic linguistic distances. Source: Wolk (2009: Figure 4.6).

We observe that distance matrices of the sort that underpin the MDS plot in Figure 1 can also be used to calculate NeighborNet diagrams (Bryant & Moulton 2004). Originally developed in biometry and bioinformatics to represent uncertainty in phylogenies and reticulate effects such as genetic recombination, NeighborNet diagrams are now also popular in linguistics (e.g. McMahon et al. 2007). Wolk (2009) uses the technique to study the morphosyntax survey accompanying the *Handbook;* his diagram is reproduced here as Figure 2. Without insisting on a strictly phylogenetic interpretation, Figure 2 visualizes aggregate similarities and distances between the varieties covered in the survey, much like the MDS plot in Figure 1 does, and thus offers an alternative perspective on the dataset and the signals of typological relatedness that it provides. 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. By and large, the NeighborNet diagram in Figure 2 suggests the same typological split that also shines through in Figure 1: we tend to find L1 varieties of English in the upper half of the diagram, English-based pidgin and creole languages in the lower left-hand quadrant, and indigenized L2 varieties in the buffer zones between these groups. In addition, we find a number of areal groupings.

3. Typological profiling

The method discussed in the preceding section is bottom-up in nature because it does not define relevant dimensions of relatedness *a priori* (apart, of course, from relying on a pre-defined feature catalogue). Top-down methods of relatedness calculation, by contrast, start out from dimensions, categories etc. that are of core interest, and then determine relatedness between languages or varieties as a function of these dimensions/categories. We will now discuss in some detail one such method, namely the "Typological Profiling" method (Szmrecsanyi 2009; Szmrecsanyi & Kortmann 2011; see also Maitz & Németh 2014; Grandel 2017; Laitinen 2018) , which determines relatedness based on how analytic and synthetic the coding of grammatical information is, according to naturalistic corpus data.

The distinction between analytic and synthetic languages goes back to August Wilhelm von Schlegel (1818). Sapir (1921) subsequently proposed a number of other parameters along which languages should be categorized. Sapir's typology, in turn, influenced Greenberg, who in a seminal (1960) paper entitled "A Quantitative Approach to the Morphological Typology of Language'' demonstrated that prima facie abstract typological notions are amenable to sufficiently precise numerical measurements by calculating a number of indices on the basis of naturalistic data ("texts"). Thus Greenberg proposed indices of synthesis, agglutination, compounding, derivation, isolation, and so on. For example, Greenberg defined the gross inflectional index as the number of inflectional morphemes (nonconcordial or concordial) in a sample text divided by the total number of words in the sample. Greenberg's method is thus essentially an early exercise in corpus-based typology.

The "Typological Profiling" method uses Greenberg's method in revised form. *Formal grammatical syntheticity* is about coding strategies where grammatical information is encoded with bound grammatical markers. By contrast to Greenberg's method, what is measured, in a given textual sample, is not the number of inflectional morphemes per sample (which is what Greenberg's gross inflectional index measures), but the number of words in a sample that bear at least one bound grammatical marker. This is a rather subtle difference (at least when attention is restricted to English and morphologically similarly simple languages). Second and more importantly, Greenberg did not calculate an analyticity index, but this can be done by relating the number of synsemantic words in a given text to the total number of words in that text (Kasevič & Jachontov 1982: 37; see also Kelemen 1970: 62). *Formal grammatical analyticity* thus covers coding strategies where grammatical information is encoded with free grammatical markers defined as closed-class function words without lexical meaning.

With these definitions in place, indices measuring formal grammatical analyticity and formal grammatical syntheticity can be established fairly straightforwardly drawing on part-of-speech (POS) annotated corpora, such as the British National Corpus, which can be analyzed exhaustively. If corpora subject to study are not or cannot be POS annotated exhaustively,

the analyst may manually annotate random samples of word tokens. (see Szmrecsanyi 2009: fn 6 for details and for statistics on how robust findings deriving from 1,000 token random samples are).

Once annotated material is available, we may define – in line with the definitions presented above – four part-of-speech groups into which individual word tokens in the corpus texts may be grouped:

1. Analytic word tokens: complementizers (as in *he thinks <u>that</u> he will go*), coordinating conjunctions (*he sleeps <u>and</u> she reads*), determiners (*<u>the</u> house*), infinitive markers (*he needs <u>to</u> go*), modals (*he <u>can</u> go*), negators (*she will <u>not</u> leave*), existential markers (*<u>there</u> are many examples*), pronouns (*I, you, me*, …), prepositions (*of, in, at*, …), comparative and superlative quantifiers (*more* and *most*), and auxiliary verbs (e.g. *I <u>have</u> eaten lunch*).
2. Synthetic word tokens: verbal inflections (*I walk* > *he walk-s*), nominal inflections (*one dog* > *two dog-s*), and adjectival inflections (*small* > *small-er*/*small-est*). This category also includes allomorphies such as ablaut phenomena (*sing* > *sang*), i-mutation (*goose* > *geese*), and other non-regular yet clearly bound grammatical markers.
3. Simultaneously analytic and synthetic word tokens such as inflected auxiliary verbs, as in *he <u>has</u> eaten lunch*.
4. Purely lexical word tokens, such as singular nouns. This is a wastebasket category that is uninteresting for present purposes.

Subsequently, based on this categorization the analyst may calculate typological indices[3] as follows:

- an *adjusted analyticity index*, which is calculated as the ratio of the number of free grammatical markers (i.e. function words) in a text to the total number of content words (i.e. the total number of orthographically transcribed words in the text minus function words) in the text, normalized to a sample size of 1,000 tokens.
- an *adjusted syntheticity index*, which is analogously calculated as the ratio of the number of words in a text that bear a bound grammatical marker to the total number of content words (i.e. the total number of orthographically transcribed words in the text minus function words) in the sample text, normalized to a sample size of 1,000 tokens.

[3] This exercise in index calculation is essentially the method proposed by Greenberg (1960). One key ingredient in this methodology is the reliance on orthographically transcribed "words" (instead of e.g. grams) as the basic unit of analysis. This reliance is not unproblematic: Haspelmath and Michaelis (2017), for example, point out that we cannot, in a cross-linguistically responsible fashion, reliably define the difference between words and affixes, and that therefore the concept of the "word" is unreliable. Hence, they conclude that the distinction between analytic and synthetic can only be made drawing on diachronic evidence – in fact, they suggest that "if we are dealing with a language whose history is totally unknown, we cannot classify it as synthetic or analytic" (p. 8). The spirit of the Typological Profiling method is less pessimistic, however, and prefers (minor) imperfections arising from defining "words" to principled agnosticism.

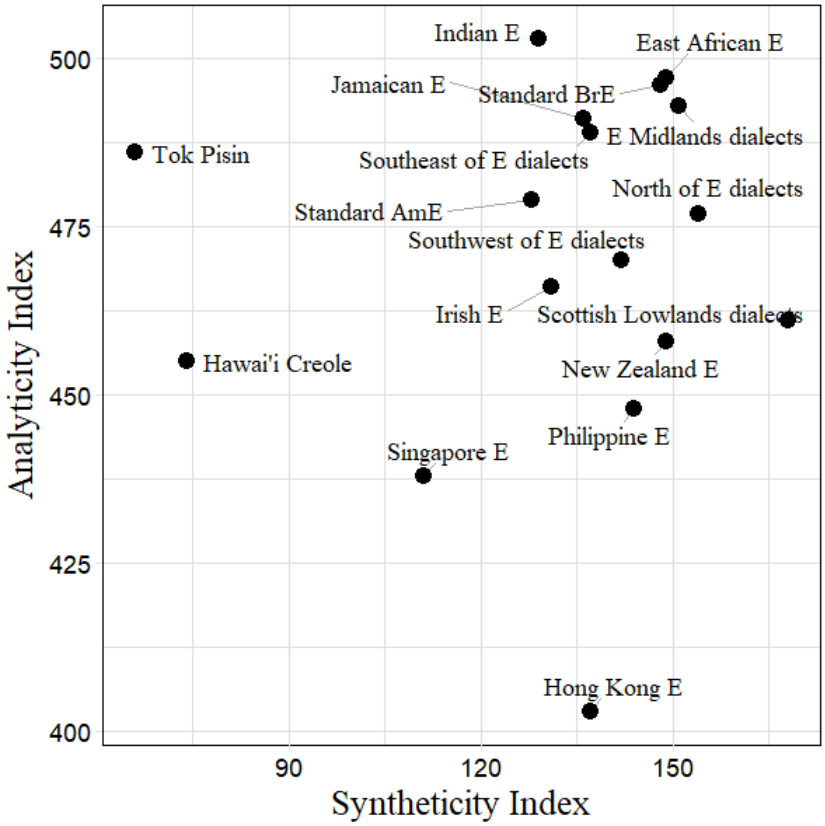

Figure 10.3. Tok Pisin and Hawai'i Creole vis-à-vis varieties of English: total number of analytic types against total number of synthetic types (adapted from Siegel et al. 2014: Figure 2)

But the crucial point here is that the index-based Typological profiling method can be used to assess the typological relatedness of varieties and languages. For example, Siegel et al. (2014) test the claim that creole languages are more analytic than other languages and varieties. They specifically investigate the coding of grammatical information (free vs. bound morphemes) in two English-lexified creoles (Tok Pisin and Hawai'i Creole) and – for benchmarking purposes – in a number of rural dialects of British English, non-native indigenized L2 varieties of English, transplanted L1 varieties of English around the world, and language-shift varieties. Using the Typological profiling method as discussed above, Siegel et al. show that, indeed, Tok Pisin and Hawai'i Creole are significantly less synthetic than other varieties of English in that they exhibit a greater ratio of analytic versus synthetic structures. That being said, the data suggest that Tok Pisin and Hawai'i Creole are not necessarily more analytic (in absolute terms) than indigenized or native varieties of English. Figure 3 illustrates this point by locating varieties in a two-dimensional syntheticity-analyticity space. Both Tok Pisin and Hawai'i Creole, in the left half of the diagram, use synthetic markers less often than other varieties of English, but they do exhibit a similar degree of analyticity compared to most other varieties in the sample. We also see that L1 varieties and traditional British dialects are more analytic than some L2 varieties (e.g. Singapore English, and especially Hong Kong English) in which zero marking is relatively frequent (see Kortmann & Szmrecsanyi 2011: 275 for discussion). With regard to distances and typological relatedness, then, Figure 3 can be interpreted as follows: adopting a top-down perspective in terms of grammatical analyticity and syntheticity, most L1 and L2 varieties of English are indeed quite close

linguistically, as the dense cluster in the upper-right hand quadrant of Figure 3 indicates. Tok Pisin and Hawai'i Creole are quite distant from this cluster on the syntheticity dimension, while Singapore English and Hong Kong English differ typologically along the analyticity dimension.

4. Typological relatedness and language complexity

Language complexity is currently a very fashionable research topic, particularly in sociolinguistic and typological circles. For most of the twentieth century, theoretical linguists were essentially in denial that complexity differentials exist. Recent years, however, have seen massive research activity on complexity and its measurement. One of the primers was a lead article in the journal *Linguistic Typology* in which John McWhorter suggested that creole languages tend to have simpler grammars than older languages, "by virtue of the fact that they were born as pidgins, and thus stripped of almost all features unnecessary to communication" (McWhorter 2001: 125; absence of "features unnecessary to communication" is, in a nutshell, McWhorter's definition of complexity). In other words, creole languages are typologically similar among other things thanks to the fact that their grammars eschew needless complexity on all linguistic levels.

An important theme in the literature is the extent to which processes of complexification or simplification have language-external triggers. Research in this vein typically assumes that languages are complex adaptive systems (Beckner et al. 2009) whose complexity profiles adapt to the communicative, cultural, and cognitive needs of their speakers (Bentz & Winter 2013: 19). Hence language structure and its complexity is seen as a function of social and sociohistorical factors. For example, it has been argued that a history of language contact and concomitant adult SLA (or: imperfect learning) triggers simplification (e.g. Bentz & Berdicevskis 2016; Szmrecsanyi & Kortmann 2009b). Conversely, complexification is thought to occur in contact scenarios that involve childhood bilingualism (Trudgill 2011: 42) and in the absence of contact (Nichols 2013; Wray & Grace 2007); for a view that compares child and adult bilingualism and possible outcomes (simplification, complexification), see Kerswill et al. (2013). History aside, it is well known that complexity can be a function of the speech situation – registers differ in terms of the extent and type(s) of complexity they exhibit (e.g. Biber, Gray & Poonpon 2011; Szmrecsanyi 2009).

So language and dialect complexity advertises itself as a diagnostic of sociolinguistic and/or typological relatedness. To gauge the complexity of languages and dialects, many analysts draw on system-based, langue-focused, top-down defined metrics. Take, for example, Nichols (2013), a study that investigates complexity differentials in 26 Nakh-Daghestanian languages spoken in the eastern half of the Great Caucasus range. Nichols is interested in two types of complexity: grammatical inventory size and opacity. Inventory size is defined "as the number of contrastive elements in a system or subsystem" (e.g. number of inflectional categories marked on the verb, number of major word orders). Opacity is defined as "the number of steps, processes, mergers, syncretisms, allomorphies, etc. standing between the underlying category or form and its surface exponent" (Nichols 2013: 44). Against this backdrop, Nichols compiles a catalogue of features relating to inventory size

and opacity. She then assesses each language under study with regard to these features, and subsequently calculates inventory size and opacity scores. An analysis of these scores reveals that language complexity is among other things a function of altitude, because altitude correlates with the sociolinguistic status – specifically: isolation and lack of contact – of languages.

Studies such as Nichols (2013), which rely on a top-down definition of system-oriented complexity concepts (consider that e.g. the number of inflectional categories is information that would be provided in reference grammars, but not in corpora), are extremely valuable and have substantially added to our knowledge of language complexity. Recent research, however, increasingly takes a bottom-up and more text- and/or usage-oriented approach to language complexity. *Kolmogorov complexity,* the complexity measure that we will be discussing in what follows, exemplifies this trend. As a measure that brings in information theory (Shannon 1948), Kolmogorov complexity is unsupervised, holistic, and radically text-based: the construct defines the complexity of a string or text as the length of the shortest possible description of that string or text. Juola (1998; 2008) was the first to utilize Kolmogorov complexity in the realm of language complexity research. His idea was that texts count as linguistically simple or complex to the extent that they can or cannot be predicted from previously seen texts. It is clear that Kolmogorov complexity is entirely agnostic about form-meaning relationships and such things; what is measured is text-based linguistic surface complexity/redundancy (see Ehret 2016 for extended discussion). Kolmogorov complexity can be conveniently approximated using off-the-shelf file compression programs, because these use adaptive entropy estimation, which in turn is a good approximation of Kolmogorov complexity (Juola 1998). More specifically, file compression programs compress text strings by describing new strings on the basis of previously seen and memorized (sub-)strings so that the amount of information and redundancy in a given string can be measured (Juola 2008: 93).

With an interest in both cross-linguistic and language-internal (cross-dialectal) complexity variation and the patterns of relatedness that it reveals, Ehret & Szmrecsanyi (2016) endeavor to investigate Kolmogorov complexity in a parallel text database containing translations of the Gospel of Mark into a number of languages (Esperanto, Finnish, French, German, Hungarian, Jamaican Patois, and Classical Latin), including (mostly historical) translations into English (from a West Saxon translation over the King James Bible to the English Standard Version, published in 2001). The summary that follows skips most of the technicalities. What is important is that, overall, Kolmogorov complexity of a text can be established using the following procedure: (1) feed corpus texts into a compression program such as *gzip* (the results to be reported below were obtained using *gzip* version 1.2.4), (2) note down file sizes before and after compression, (3) regress out the trivial correlation between the two measurements, (4) interpret the regression residuals (in bytes) as adjusted complexity scores: bigger adjusted complexity scores indicate more Kolmogorov complexity. Using this procedure, Ehret & Szmrecsanyi (2016) establish the ranking of Bible translations, from overall more complex to overall less complex, that is depicted in Table 2:

1. English-WestSaxon-10c

2. Hungarian
3. Finnish
4. Latin
5. German
6. French
7. Jamaican Patois
8. English-ESV-21c
9. English-Darby-19c
10. Esperanto
11. English-Webster's-19c
12. English-KingJames-17c
13. English-DouayRheims-16c
14. English-Wycliffe-14c
15. English-ASV-20c
16. English-Young's-19c
17. BasicEnglish-20c

Table 10.2. Overall Kolmogorov complexity, from most complex (rank 1) to least complex (rank 17 (adapted from Ehret & Szmrecsanyi 2016: Figure 1).

It appears, then, that the Kolmogorov approach ranks the complexity of the Bible texts in a way that seems to be compatible with what we think we know about the languages covered in the sample. The three most complex translations are the West Saxon, Hungarian, and Finnish texts. Jamaican Patois and Esperanto are medium-complex. Most translations into English (except for the West Saxon translation mentioned above) are below-average complex. The least complex data point in the sample is the Basic English translation of the Bible. Basic English is a simplified variety of English designed by Charles Kay Ogden as, among other things, an aid to facilitate the teaching of English as a foreign language (Ogden 1934). The ranking in Table 2 can, of course, be interpreted in terms of relatedness and distance: for example, West Saxon, Hungarian and Finish, three fairly inflectional languages, are typologically related and thus close; likewise, the post-West Saxon English data points are close thanks to their comparatively low overall complexity.

It is possible to add more granularity to the Kolmogorov measurements. Specifically, morphological and syntactic complexity may be measured by distorting the text samples prior to compression (see Juola 1998; Juola 2008 for pioneering work). In this spirit, Ehret & Szmrecsanyi (2016) also apply syntactic distortion to their dataset by deleting a random sample of 10% of all *word tokens* in each text file before applying the compression technique. This procedure leads to the disruption of word order regularities. The idea is that this disruption badly affects syntactically complex texts, i.e. texts with a comparatively fixed word order (such as e.g. Standard English texts -- see Dryer 2013), which compromises their compressibility. Languages with relatively free word order, on the other hand, are less affected, as they lack syntactic interdependencies that could be compromised in the first place. Comparatively bad compression ratios after syntactic distortion thus indicate comparatively high syntactic complexity, which means that high syntactic complexity

essentially entails word order rigidity (the logic here being that word order rigidity constrains word order choices, and hence complexifies production). In a similar fashion, Ehret & Szmrecsanyi (2016) measure Kolmogorov complexity at the morphological level: they delete at random 10% of all *characters* in each text file. This creates new word forms, which negatively affects the compressibility of morphologically simple texts which, on the whole, have fewer word forms than morphologically complex texts. By contrast, morphologically complex texts exhibit overall a relatively large amount of word forms in any case, and are thus not affected as much by this kind of distortion. Therefore, after distortion, comparatively bad compression ratios indicate low morphological complexity.

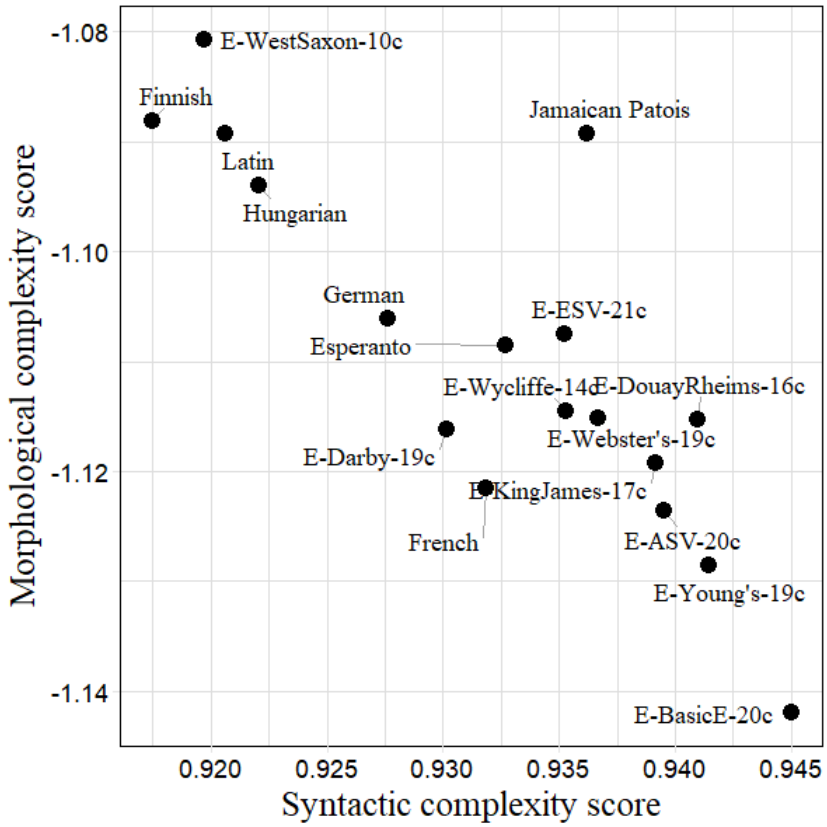

Figure 10.4. Morphological complexity (*y*-axis) by syntactic complexity (*x*-axis) (adapted from Ehret & Szmrecsanyi 2016: Figure 2).

The outcome of this exercise in distortion and disruption is visually depicted in Figure 4, which plots syntactic Kolmogorov complexity (on the horizontal axis) against morphological Kolmogorov complexity (on the vertical axis). The analysis suggests that the morphologically most complex languages in the sample are West Saxon, Finnish, and Latin (in that order) – a result that is certainly in line with what we know about these highly inflectional languages. The syntactically most complex data point (i.e. the text in which word order is most rigid) is the Bible in Basic English, which is a fairly consistent SVO variety. Overall, syntactic complexity trades off against morphological complexity: languages that tend to be syntactically complex are not that complex morphologically, and vice versa. With respect to English, most post-West Saxon translations tend toward the syntactic

complexity/morphological simplicity quadrant. And it is in this connection that we can see how the method can be used to determine typological relatedness: languages and varieties that share similar complexity profiles – potentially thanks to shared sociolinguistic ecologies – are located close to each other in Figure 4.

5. Variation-based relatedness measures

The measure of typological distance and relatedness to be discussed in this section is top-down in nature and draws inspiration from variationist (socio)linguistics and probabilistic grammar research (Labov 1982; Tagliamonte 2001; Bresnan et al. 2007). The point of departure is the observation that variation is ubiquitous in language use, and thus crucial for understanding relatedness. In the empirical analysis to follow, we specifically restrict attention to onomasiological variation of the type that van Hout & Muysken (2016: 250) call "variability in the linguistic signal within a given language", also known as inherent variability (Labov 1969): we are only interested in patterns that are known to be variable for all members of the speech community, regardless of dialect background. On the theoretical plane, we will assume – in line with the basic tenets of usage-based linguistics – that grammar (including those parts of grammar that regulate variation patterns) is the "cognitive organization of one's experience with language" (Bybee 2006: 711). This organization includes experience about probabilistic constraints in linguistic choice making.

The analysis in this section is based on three variation phenomena (a.k.a. alternations) in the grammar of English: the particle placement alternation, as in (1); the genitive alternation, as in (2); and the dative alternation, as in (3).

(1) The particle placement alternation (Grafmiller & Szmrecsanyi in press)
   a. verb-object-particle order (V-DO-P)
   you can just [*cut*]$_{\text{verb}}$ [*the tops*]$_{\text{direct object}}$ [*off*]$_{\text{particle}}$ *and leave them.* [ICE-GB:S1A-007]
   b. verb-particle-object order (V-P-DO)
   [*Cut*]$_{\text{verb}}$ [*off*]$_{\text{particle}}$ [*the flowers*]$_{\text{direct object}}$ *as they fade*. [ICE-CAN:W2B-023]

(2) The genitive alternation (Heller 2018)
   a. the *s*-genitive
   [*Singapore*]$_{\text{possessor}}$'s [*small size*]$_{\text{possessum}}$ *meant it could be quick to respond to changes in economic conditions* [ICE-SIN:W2C-011]
   b. the *of*-genitive
   the [*size*]$_{\text{possessum}}$ *of* [*the eyes*]$_{\text{possessor}}$ *is to help them at night.* [ICE-GB:W2B-021]

(3) The dative alternation (Röthlisberger 2018)
   a. the ditransitive dative variant
   That will give [*the panel*]$_{\text{recipient}}$ [*a chance*]$_{\text{theme}}$ *to expand on what they've been saying.* [ICE-GB:S1B-036]

b. the prepositional dative variant
[...] *and that gives* [*a chance*]$_{\text{theme}}$ [*to Bhupathy*]$_{\text{recipient}}$ *to equalise the points at thirty all.* [ICE-IND:S2A-019]

The a. and b. variants in the above examples are all well-known "alternate ways of saying 'the same' thing" (Labov 1972: 188), where one variant can be paraphrased by the other without semantic change (see Lavandera 1978 for more discussion). All three alternations are subject to various, well-known probabilistic constraints on variation, such as e.g. the principle of end-weight (Behaghel 1909; Wasow 1997): long direct objects after transitive phrasal verbs tend to favor particle-object order (because the object is placed in final position), long possessors tend to favor the *of*-genitive (because the *of*-genitive places the possessor after the possessum), and long recipients tend to favor the prepositional dative (because the recipient is placed after the theme). We may now define the variational distance between varieties as being proportional to the extent to which constraints on variation, such as the principle of end-weight, have different effects and/or different effects strengths in different varieties.[4]

The steps necessary to conduct a variational distance/relatedness analysis in this spirit can be summarized as follows:

1. Define the variable(s) and identify grammatical variants in the corpus material.
2. Annotate each and every observed variant for a-priori defined constraints on variation (e.g. constituent length).
3. Based on the richly annotated datasets that step 2 generates, calculate multivariate models that predict variant choice.
4. Use model outputs to calculate variational distances between the varieties under study.

Steps 1-3 draw on run-of-the-mill methodology that has been customary in variationist (socio)linguistics and corpus-based variationist linguistics for decades. As to step 4, there are several ways to derive distance measures from variationist models. The following discussion presents a comparatively straightforward method to accomplish this task: the basic idea is to define a default variety (such as British English [BrE]) and to calculate a regression model that predicts choices in the default variety. Subsequently, determine the extent to which the default model generalizes to other varieties. Generalizability, in this approach, diagnoses relatedness.

We illustrate drawing on the well-known dative alternation, as in (3). The dataset to be used in this section (see Röthlisberger et al. 2017) derives from the *International Corpus of English* (ICE) and is annotated, among other things, for five major constraints on dative variation: the ratio between recipient and theme length, pronominality of the recipient, animacy of the recipient, definiteness of the theme, and genre. Based on this information, we fit a binary logistic regression model that describes how language users make dative choices in the British section of ICE. The model is summarized in table 3.

[4] See Szmrecsanyi, Grafmiller, and Rosseel (forthcoming) for a related approach.

```
Call:
glm(formula = f0, family = binomial, data = gb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3845  -0.4478  -0.1917   0.2292   3.1381

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.7420     0.2623 -10.452  < 2e-16 ***
z.logWeightRatio         4.2629     0.3655  11.663  < 2e-16 ***
RecPronnon-pron          0.8730     0.2907   3.003  0.00268 **
RecBinAnimacyinanimate   0.7289     0.2822   2.583  0.00978 **
ThemeDefinitenessdef     1.0579     0.2356   4.490 7.12e-06 ***
GenreCoarsemonologue    -0.1601     0.3319  -0.482  0.62949
GenreCoarsenon-printed   0.6553     0.3155   2.077  0.03778 *
GenreCoarseprinted      -0.4145     0.3128  -1.325  0.18511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1015.57  on 873  degrees of freedom
Residual deviance:  544.34  on 866  degrees of freedom
AIC: 560.34

Number of Fisher Scoring iterations: 6
```

Table 10.2. Binary logistic regression model predicting dative choices in ICE-GB. Predicted odds are for the prepositional dative. Predictive accuracy: 86.8% (baseline: 73.2%).

The signs of the coefficients (see column 'Estimate') are the theoretically expected ones: comparatively weighty recipients favor the prepositional dative, inanimate recipients favor the prepositional dative, and so on. In all, the model summarized in Table 3 correctly predicts 86.8% of all dative choices in the ICE-GB dataset, vis-à-vis a baseline of 73.2%, which is the percentage one would be able to predict successfully by categorically predicting the more frequent variant.

Next we use the regression model in Table 3, which is fine-tuned to dative variability in BrE, to predict choices in other varieties of English covered in the dataset: Irish English (IrE), Canadian English (CanE), New Zealand English (NZE), Jamaican English (JamE), Singapore English (SgE), Hong Kong English (HKE), Indian English (IndE), and Philippines English (PhiE). We specifically establish how much worse or better the BrE regression model predicts variation in these varieties by calculating the in- or decrease in predictive accuracy. To illustrate, the in- and decreases in predictive accuracy when the model in Table 3 is applied to the other varieties are as follows:

BrE	0 (default)
CanE	-0.31 per cent points
HKE	-3.95 per cent points
IndE	-5.38 per cent points
IrE	-2.87 per cent points
JamE	+1.49 per cent points
NZE	-2.15 per cent points

PhiE -0.92 per cent points
SgE +0.03 per cent points

Table 10.3. Decrease in model accuracy (in per cent points) when the binary logistic regression model predicting dative choices in ICE-GB (see Table 3) is used to predict dative outcomes in other varieties

So when the BrE model (which correctly predicts 86.8% of all outcomes) is applied to CanE data, accuracy decreases by 0.31 per cent points, if it is applied to HKE data, accuracy decreases by 3.95 per cent points, and so on. Crucially, the in- and decreases can be interpreted as distance measurements. The procedure outlined above is repeated for the genitive alternation (constraints: possessor animacy, final sibilancy, possessor length, possessum length, semantic relation) and the particle placement alternation (constraints: direct object length, semantics, presence of directional modifier, definiteness of direct object, phonology of final segment of verb form). In each case, we include in the models those five predictors with the greatest explanatory power. We thus obtain three predictive accuracy vectors, one per alternation under study. The vectors can be visually depicted in a three-dimensional plot as in Figure 5.

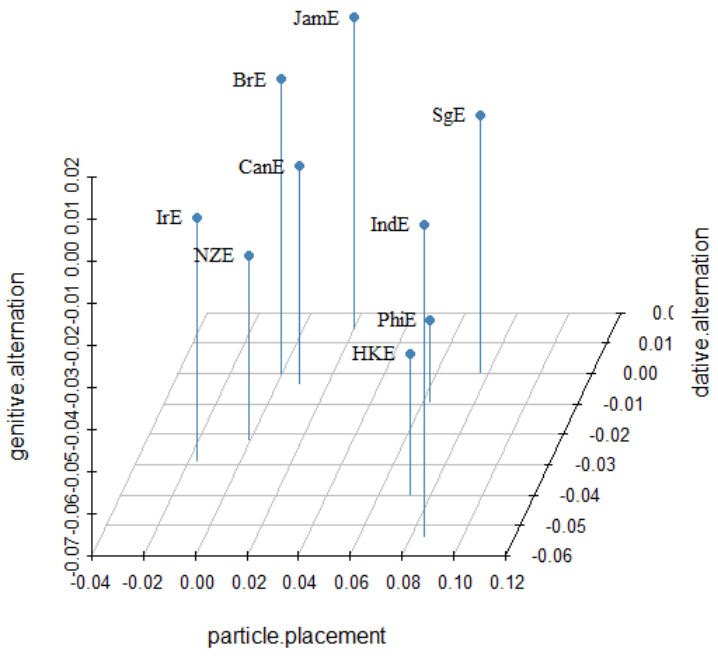

Figure 10.5. Variational distances between 9 varieties of English as sampled in the *International Corpus of English*. Distances between data points are proportional to the extent to which grammatical choice-making processes are different. Default variety: BrE.

Axes plot in- or decreases in predictive accuracy as a regression model trained on BrE data is used to predict grammatical variation in other varieties.

In Figure 5, each axis represents one of the variation phenomena under study, and the values plotted on the axis are predictive accuracy in- or decreases. We see a fairly clear split between L1 varieties (left half of plot) and indigenized L2 varieties of English (right half of plot) in the particle placement dimension. The genitive alternation axis pits PhiE and HKE, two Asian varieties, against all other varieties in the dataset. Some outliers notwithstanding, the dative alternation axis also tends to distinguish between Asian varieties of English (IndE, HKE) and non-Asian varieties (such as JamE at the back of the plot). As to variety clusters, we note that BrE and JamE are surprisingly close (i.e. related variationally), and so are IrE and NZE. In the L2 realm, HKE and IndE are located in proximity to each other. In the bird's eye perspective the variety that is most distinct from the default (i.e. BrE) variationally is IndE.

6. Conclusion

We have seen that there is a sizable literature in English linguistics on how to determine the relatedness and – more generally speaking – linguistic distance and/or similarity between varieties of English, inspired by work in crosslinguistic typology (see e.g. Nichols 1992; Bickel 2017). The present chapter has reviewed four techniques to accomplish this task: using aggregate measures to derive distances, in a bottom-up fashion, from survey data (Section 2); creating typological analyticity/syntheticity profiles, which is a top-down method (Section 3); establishing relatedness via complexity measurements, which can be accomplished in a bottom-up fashion using an information theory-inspired approach (Section 4); and calculating top-down variation-based relatedness measures (Section 5). It is clear that each of these methods highlights different aspects of linguistic reality, and none can plausibly claim to comprehensively cover the multiple ways in which varieties are (un)related to each other. This chapter has also taken the liberty to focus on grammar- and structure-focused methodologies -- due to space limitations, other linguistic levels have received short shrift. In particular, pronunciation and lexis are two linguistic levels that are in principle also amenable to an analysis of typological relatedness (see e.g. McMahon et al. 2007; Ruette, Ehret & Szmrecsanyi 2016).

As always, there are a number of directions for future research. For example, we currently have no good idea about the extent to which measurements of typological relatedness correspond to how language users perceive dialect and variety distances. While there is some pioneering work in the realm of perceptual dialectology (in the spirit of e.g. Niedzielski & Preston 1999), the World Englishes literature would profit from more systematic research that addresses this question. Second, of course, we will eventually need to know whether our measurement techniques yield the same patterns and splits (e.g. between L1 varieties and indigenized L2 varieties) in other languages with global reach, such as Spanish and French.

Acknowledgments

Thanks go to Melanie Röthlisberger for various calculations regarding variational distance. The usual disclaimers apply. A grant from the Research Foundation Flanders (FWO, grant # G.0C59.13N) is gratefully acknowledged.

References

Bayley, Robert & Otto Santa Ana. 2004. Chicano English: morphology and syntax. In Bernd Kortmann, Edgar Schneider, K. Burridge, R. Mesthrie & C. Upton (eds.), *A Handbook of Varieties of English*, vol. 2, 374–390. Berlin/New York: Mouton de Gruyter.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language Is a Complex Adaptive System: Position Paper. *Language Learning* 59. 1–26. doi:10.1111/j.1467-9922.2009.00533.x.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110–142.

Bentz, Christian & Aleksandrs Berdicevskis. 2016. Learning pressures reduce morphological complexity: linking corpus, computational and experimental evidence. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics (COLING 2016)*. Osaka, Japan.

Bentz, Christian & Bodo Winter. 2013. Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change* 3. 1–27. doi:10.1163/22105832-13030105.

Biber, Douglas, Bethany Gray & Kornwipa Poonpon. 2011. Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? *TESOL Quarterly* 45(1). 5–35. doi:10.5054/tq.2011.244483.

Bickel, Balthasar. 2017. Areas and Universals. In Raymond Hickey (ed.), *The Cambridge Handbook of Areal Linguistics*, 40–54. Cambridge: Cambridge University Press. doi:10.1017/9781107279872.004. https://www.cambridge.org/core/product/identifier/9781107279872%23CN-bp-3/type/book_part (19 October, 2018).

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the Dative Alternation. In Gerlof Boume, Irene Krämer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bryant, Davis & 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.

Bybee, Joan. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4). 711–733.

Cysouw, Michael. 2013. Disentangling geography from genealogy. In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in Language and Linguistics*. Berlin, Boston: DE GRUYTER. http://www.degruyter.com/view/books/9783110312027/9783110312027.21/9783110312027.21.xml (3 June, 2016).

Dryer, Matthew S. 2013. Order of Subject, Object and Verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/81.

Ehret, Katharina. 2016. An information-theoretic approach to language complexity: variation in naturalistic corpora. Albert-Ludwigs-Universität Freiburg.

Ehret, Katharina & Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler & Guido Seiler (eds.), *Complexity, Isolation, and Variation*, 71–94. Berlin, Boston: De Gruyter.

http://www.degruyter.com/view/books/9783110348965/9783110348965-004/9783110348965-004.xml (2 August, 2016).
Gil, David. 2003. English goes Asian: Number and (in)definiteness in the Singlish noun phrase. In Frans Plank (ed.), *Noun Phrase Structure in the Languages of Europe*, 467–514. Berlin/New York: Mouton de Gruyter.
Grafmiller, Jason & Benedikt Szmrecsanyi. in press. Mapping out particle placement in Englishes around the world. A case study in comparative sociolinguistic analysis. *Language Variation and Change*.
Grandel, Saskia. 2017. Morphosyntaktische Komplexität, Normativität und Sprachkontakt. Eine Projektskizze. *Zeitschrift für Dialektologie und Linguistik* 84(2–3). 152–177.
Greenberg, Joseph H. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26(3). 178–194.
Haspelmath, Martin & Susanne Maria Michaelis. 2017. Analytic and synthetic: Typological change in varieties of European languages. In Isabelle Buchstaller & Beat Siebenhaar (eds.), *Studies in Language Variation*, vol. 19, 3–22. Amsterdam: John Benjamins Publishing Company. doi:10.1075/silv.19.01has. https://benjamins.com/catalog/silv.19.01has (4 October, 2018).
Heller, Benedikt. 2018. Stability and Fluidity in Syntactic Variation World-Wide: The Genitive Alternation Across Varieties of English. Leuven: KU Leuven PhD dissertation.
Hosali, Priya. 2004. Butler English: morphology and syntax. In Bernd Kortmann, Edgar Schneider, K. Burridge, R. Mesthrie & C. Upton (eds.), *A Handbook of Varieties of English*, vol. 2, 1031–1044. Berlin/New York: Mouton de Gruyter.
Hout, Roeland van & Pieter Muysken. 2016. Taming Chaos. Chance and Variability in the Language Sciences. In Klaas Landsman & Ellen van Wolde (eds.), *The Challenge of Chance*, 249–266. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-26300-7_14 (13 December, 2016).
Juola, Patrick. 1998. Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics* 5(3). 206–213.
Juola, Patrick. 2008. Assessing linguistic complexity. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*, 89–108. Amsterdam, Philadelphia: Benjamins.
Kasevič, Vadim & Sergej E Jachontov (eds.). 1982. *Kvantitativnaja tipologija jazykov Azii i Afriki [A quantitative typology of Asian and African Languages]*. Leningrad.
Kelemen, J. 1970. Sprachtypologie und Sprachstatistik. In László Dezső & Peter Hajdú (eds.), *Theoretical problems of typology and the Northern Eurasian languages*, 53–63. Amsterdam: Gruener.
Kerswill, Paul, Jenny Cheshire, Susan Fox & Eivind Torgersen. 2013. English as a contact language: the role of children and adolescents. In Daniel Schreier & Marianne Hundt (eds.), *English as a contact language*, 258–282. Cambridge: Cambridge University Press.
Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2013. *eWAVE*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://ewave-atlas.org/.
Kortmann, Bernd, Edgar Schneider, Kate Burridge, Raj Mesthrie & Clive Upton (eds.). 2004. *A Handbook of Varieties of English*. Berlin: Mouton de Gruyter.
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.
Kortmann, Bernd & Benedikt Szmrecsanyi. 2011. Parameters of morphosyntactic variation in World Englishes: Prospects and limitations of searching for universals. In Peter Siemund (ed.), *Linguistic Universals and Language Variation*, 264–290. Berlin / New York: Mouton de Gruyter.
Kruskal, Joseph B & Myron Wish. 1978. *Multidimensional Scaling* (Quantitative Applications in the Social Sciences). Newbury Park, London, New Delhi: Sage Publications.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45. 715–762.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.
Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkiel (eds.), *Perspectives on Historical Linguistics*, 17–92. Amsterdam, Philadelphia: Benjamins.
Laitinen, Mikko. 2018. Placing ELF among the varieties of English: Some observation from typological profiling. In Sandra Deshors (ed.), *Modeling World Englishes in the 21st Century: Assessing the interplay of emancipation and globalization of ESL varieties*. Amsterdam: Benjamins.
Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7. 171–183.
Maitz, Péter & Attila Németh. 2014. Language Contact and Morphosyntactic Complexity: Evidence from German. *Journal of Germanic Linguistics* 26(01). 1–29. doi:10.1017/S1470542713000184.
McMahon, April, Paul Heggarty, Robert McMahon & Warren Maguire. 2007. The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics* 11(1). 113–142.
McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 6. 125–166.
Melchers, Gunnel. 2004. English spoken in Orkney and Shetland: morphology and syntax. In Bernd Kortmann, Edgar Schneider, K. Burridge, R. Mesthrie & C. Upton (eds.), *A Handbook of Varieties of English*, vol. 2, 34–46. Berlin/New York: Mouton de Gruyter.
Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.
Nichols, Johanna. 2013. The vertical archipelago: Adding the third dimension to linguistic geography. In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in language and linguistics: geographical, interactional, and cognitive perspectives*, 38–60. Berlin, New York: Walter de Gruyter.
Niedzielski, Nancy A. & Dennis Richard Preston. 1999. *Folk linguistics*. Berlin, New York: Mouton de Gruyter.
Ogden, C. K. 1934. *The system of Basic English*. New York: Harcourt.
Poplack, Shana (ed.). 2000. *The English history of African American English*. Oxford: Blackwell.
Röthlisberger, Melanie. 2018. Regional variation in probabilistic grammars: A multifactorial study of the English dative alternation. Leuven: KU Leuven PhD dissertation. https://lirias.kuleuven.be/handle/123456789/602938.
Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. to appear. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics*.
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. doi:10.1075/ijcl.21.1.03rue.
Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace and Company.
Schlegel, August Wilhelm von. 1818. *Observations sur la language et la littérature provençales*. Paris.
Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27. 379–423.
Siegel, Jeff, Benedikt Szmrecsanyi & Bernd Kortmann. 2014. Measuring analyticity and syntheticity in creoles. *Journal of Pidgin and Creole Languages* 29(1). 49–85. doi:10.1075/jpcl.29.1.02sie.
Szmrecsanyi, Benedikt. 2009. Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change* 21(03). 319–353. doi:10.1017/S0954394509990123.
Szmrecsanyi, Benedikt, Jason Grafmiller & Laura Rosseel. forthcoming. Variation-Based Distance and Similarity Modeling: a case study in World Englishes. *Frontiers*.
Szmrecsanyi, Benedikt & Bernd Kortmann. 2009a. The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua* 119(11). 1643–1663.

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

Szmrecsanyi, Benedikt & Bernd Kortmann. 2011. Typological profiling: learner Englishes versus indigenized L2 varieties of English. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, 167–187. Amsterdam, Philadelphia: Benjamins.

Tagliamonte, Sali. 2001. Comparative sociolinguistics. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *Handbook of Language Variation and Change*, 729–763. Malden and Oxford: Blackwell.

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

Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9. 81–105.

Wee, Lionel. 2004. Singapore English: morphology and syntax. In Bernd Kortmann, Edgar Schneider, K. Burridge, R. Mesthrie & C. Upton (eds.), *A Handbook of Varieties of English*, vol. 2, 1058–1072. Berlin/New York: Mouton de Gruyter.

Wolk, Christoph. 2009. Classifying Geographic Variation: Morphosyntax and Phonology. Freiburg: University of Freiburg MA Thesis.

Wray, Alison & George W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117(3). 543–578. doi:10.1016/j.lingua.2005.05.005.