Available online at www.sciencedirect.com
ScienceDirect
Lingua 119 (2009) 1643–1663
Lingua
www.elsevier.com/locate/lingua

The morphosyntax of varieties of English worldwide: A quantitative perspective

Benedikt Szmrecsanyi *, Bernd Kortmann

Freiburg Institute for Advanced Studies (FRIAS), Albertstr. 19, 79104 Freiburg, Germany

Received 5 August 2006; received in revised form 21 September 2007; accepted 21 September 2007
Available online 3 December 2008

Abstract

What are the large-scale patterns and generalizations that emerge when investigating morphosyntactic variation in World Englishes from a bird's eye perspective? To address this question, this study draws on the questionnaire-based morphosyntactic database of the *Handbook of Varieties of English*, utilizing a number of quantitative analysis techniques (frequency and correlation measures, multidimensional scaling, cluster analysis, and principal component analysis). We demonstrate (i) that the database yields a number of generalizations and implicational tendencies relating to vernacular angloversals and universals of New Englishes, (ii) that there is a surprisingly consistent typological division between English L1 vernaculars, on the one hand, and English-based pidgins and creoles on the other hand, and (iii) that World Englishes can, on aggregate, be seen to vary along two major dimensions which we interpret as being indicative of morphosyntactic complexity and analyticity. In conclusion, we offer that the *Handbook*'s morphosyntactic database presents some interesting methodological challenges to dialectology and dialectometry.

Keywords: Variation; World Englishes; Dialectology; Dialectometry; Typology; Creolistics

1. Introduction

Much as in other dialectology literatures, it would be hard to argue that there is a shortage of phonological and lexicological surveys of non-standard dialects and varieties of English. By stark contrast, non-standard English morphology and syntax have only lately received the attention they deserve, spurred by the advent of data sources specifically designed to strengthen research on morphosyntactic variation in English. Thanks to this recent availability of adequate databases, then, dialectologists, sociolinguists, and typologists have begun to ask new and exciting questions about varieties and dialects of English.

It is precisely this strand of research into non-standard English morphosyntax that the present study will seek to engage in. Our evidence comes from the most comprehensive database of non-standard English morphosyntax to date, *viz.* the morphosyntactic survey in the recent *Handbook of Varieties of English* (Kortmann et al., 2004). This survey constitutes a database sampling 46 varieties of English around the globe with regard to 76 non-standard morphosyntactic features. One interesting aspect of this database is that it is rather less homogeneous than most

* Corresponding author. Tel.: +49 761 203 97387; fax: +49 761 203 97420.
E-mail addresses: bszm@frias.uni-freiburg.de (B. Szmrecsanyi), bernd.kortmann@anglistik.uni-freiburg.de (B. Kortmann).

0024-3841/$ – see front matter
doi:10.1016/j.lingua.2007.09.016

databases analyzed in many previous quantitative investigations into language-internal variation: not only does the survey sample native vernaculars and traditional dialects of English, as is standard in dialectology and dialectometry, but it also spans English as Second Language varieties and English-based pidgins and creoles. It is this kind of comprehensiveness that enables the analyst to remedy a shortcoming of much previous research into non-standard English grammar, that is, a narrow focus on one particular phenomenon in one particular dialect or dialect area on the basis of a limited data base. Methodologically, this study is in keeping with the Freiburg program of combining functional Greenbergian typology with dialectology, such that the observable patterns of cross-dialectal and cross-varietal variation are analyzed and interpreted in terms of the same methodological and interpretational apparatus familiar from the typological study of large-scale cross-linguistic variation (see Kortmann, 2004 for a collection of papers in this spirit).

Our overall concern in this paper, then, is with large-scale morphosyntactic patterns and variation in World Englishes. More specifically, we seek, first, to sketch how a typological angle can help uncover a number of instructive generalizations, and second, we propose to show that a lot can be learned about language-internal variation by complementing qualitative inquiry with large-scale quantitative analysis of comparative data.

This study is going to be structured as follows. Section 2 will introduce our database. In section 3, we shall discuss some generalizations that can be uncovered using some relatively straightforward frequency and correlation measures. Section 4 will rely on multidimensional scaling to map the range of variation in varieties of English. This will be complemented, in section 5, by a more fine-grained cluster analysis of patterns and groupings across varieties of English. In section 6, we will utilize principal component analysis to uncover major dimensions of linguistic variance in the database. Section 7 will recapitulate this study's major findings and sketch some directions for future research.[1]

2. The morphosyntactic survey in the *Handbook of Varieties of English*

As indicated above, our inquiry empirically rests on the *Handbook of Varieties of English* (Kortmann et al., 2004), the most comprehensive reference work to date on the phonology (Volume 1) and morphosyntax (Volume 2) of varieties of English around the globe. The *Handbook* contains survey articles by some 100 specialists, covering some 60 almost exclusively non-standard varieties or groups of varieties: native vernaculars (henceforth: L1), which include all main national standard varieties, such as New Zealand English; distinctive ethnic, regional, and social varieties, e.g., African American Vernacular English; English-based pidgins and creoles, such as Tok Pisin; and the major English as a Second Language (henceforth: L2) varieties, e.g., Malaysian English. Crucially, all of the varieties surveyed are spoken.

What will take center-stage in the present study is the morphosyntactic survey of the multimedia reference tool, available on CD-ROM and online (http://www.mouton-online.com), accompanying the *Handbook* (cf. Kortmann and Szmrecsanyi, 2004).[2] This first-ever comprehensive survey of non-standard English morphosyntax was conducted in a very simple way: 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, then, the contributors were asked to specify into which of the following three categories the relevant feature in the relevant variety, or set of closely related varieties, fell:

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)[3]

Kortmann and 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 40 *Handbook* authors provided us with data on 46 non-standard varieties of English, which amounts to more than 85% of all non-standard varieties covered in the morphosyntax chapters of the *Handbook*. More information on the varieties sampled is given in Table 1. As can be seen, all seven anglophone world regions (British Isles, America, Caribbean, Australia, Pacific, Asia, Africa), as well

[1] On a technical note, we used SPSS 13.0 for most advanced statistical analyses. The measures for correlational and implicational tendencies to be presented in section 3 were computed using Practical Extraction and Report Language (Perl) scripts which are available from the authors on request.

[2] Note that there is also a phonological survey (cf. Schneider, 2004), which will, however, not be subject to discussion in this paper.

[3] Category C thus equates evidence of absence ('does not exist') with absence of evidence ('is not documented'). Consider here that the great majority of the varieties documented in the survey are comparatively well-researched: thus, to all intents and purposes, if a feature is not documented in a given variety, it is certainly not 'pervasive' (category A) and moreover unlikely to be salient enough to license a 'B' rating.

Table 1
Varieties of English sampled according to variety type.

Variety type	Variety
L1 varieties	Orkney and Shetland, Scottish E (ScE), Irish E (IrE), Welsh E (WelE), East Anglia, North, Southwest and Southeast of England, Colloquial American E (CollAmE), Isolated Southeast US E (IsSE), Appalachian E (AppE), Ozarks E (OzE), Newfoundland E (NfldE), Urban African-American Vernacular E (Urban AAVE), Earlier African-American Vernacular E (Earlier AAVE), Colloquial Australian E (CollAusE), Australian Vernacular E (AusVE), Norfolk, regional New Zealand E (NZE), White South African E (WhSAfE)
L2 varieties	Chicano E (ChcE), Fiji English, Standard Ghanaian E (GhE), Cameroon E (CamE), East African E (EAfE), Indian South African E (InSAfE), Black South African E (BlSAfE), Butler E (ButlE), Pakistan E (PakE), Singapore E (SgE), Malaysian E (MalE)
English-based pidgins and creoles	Gullah, Suriname Creoles (SurCs), Belizean Creole (BelC), Tobagonian/Trinidadian Creole (Tob/TrnC), Bahamian E (BahE), Jamaican Creole (JamC), Bislama, Solomon Islands Pidgin (SolP), Tok Pisin (TP), Hawaiian Creole (HawC), Aboriginal E (AbE), Australian Creoles (AusC), Ghanaian Pidgin E (GhP), Nigerian Pidgin E (NigP), Cameroon Pidgin E (CamP)

Fig. 1. Varieties sampled in the morphosyntactic database of the *Handbook*.

as a fair mix of L1 varieties, L2 varieties, and pidgins/creoles, are represented in the survey.[4] Fig. 1 visualizes the geographical distribution of varieties of English in the survey.

The feature catalogue in its entirety is provided in Appendix A to this study. The features are numbered from 1 to 76 (for easy reference in later parts of the paper, the feature numbers will be referred to in square brackets, e.g., [37]) and provided with the short definitions and illustrations given as input to the *Handbook* contributors serving as informants. As has been said above, they include all major phenomena discussed in previous survey articles on grammatical properties of (individual groups of) non-standard varieties of English, with a slight bias towards features observed in L1 varieties. The 76 features fall into 11 groups corresponding to the following broad areas of morphosyntax, each of which we illustrate with one example feature:

Pronouns: for example, [3] special forms or phrases for the second person plural pronoun (e.g., *youse, y'all, aay', yufela, you together, all of you, you ones/'uns, you guys, you people*, etc.);
Noun phrase: for example, [15] group plurals (e.g., *That President has two Secretary of States);*

[4] Note that America is used as a shorthand for North America, Caribbean as a shorthand for the Caribbean, Central and South America, and Asia as a shorthand for South and Southeast Asia.

Tense and aspect: for example, [23] habitual *do* (e.g., *He does catch fish pretty*);
Modal verbs: for example, [34] double modals (e.g., *I tell you what we might should do);*
Verb morphology: for example, [41] *a*-prefixing on ing-forms (e.g., *They wasn't a-doin' nothin' wrong*);
Adverbs: for example, [42] adverbs (other than degree modifiers) have same form as adjectives (e.g., *He treated her wrong right from the start*);
Negation: for example, [44] multiple negation/negative concord (e.g., *He won't do no harm*);
Agreement: for example, [57] deletion of *be* (e.g., *She ___ smart*);
Relativization: for example, [66] gapping or zero-relativization in subject position (e.g., *The man ___ lives there is a nice chap*);
Complementation: for example, [71] as what/than what *in comparative clauses (e.g., It's harder than what you think it is*);
Discourse organization and word order: for example, [76] *like* as a quotative particle (e.g., *And she was like, what do you mean?*).

In summary, the database to be analyzed in this study is a matrix of 46 non-standard varieties of English, each characterized by a vector of 76 tripartite (A–B–C) variables describing the status of the features constituting our survey in the respective variety.

3. The distribution of non-standard morphosyntactic features: some generalizations

What can be learned about recurrent patterns of morphosyntactic variation by considering a number of fairly straightforward frequency and correlation measures? This section will explore some generalizations about the distribution of non-standard features in our database. Kortmann and Szmrecsanyi (2004) present an in-depth discussion of a number of salient yet comparatively simple frequency distributions in the Handbook's database. Let us begin by reiterating some of the key findings in regard to non-implicational frequency patterns. The *Handbook* 's comprehensive database effectively advertizes itself to flesh out empirically two related notions, VERNACULAR UNIVERSALS and UNIVERSALS OF NEW ENGLISHES. Vernacular universals, according to Jack Chambers (cf. Chambers, 2001, 2003, 2004), are ''a small number of phonological and grammatical processes [that] recur in vernaculars wherever they are spoken . . . not only in working class and rural vernaculars, but also in . . . pidgins, creoles and interlanguage varieties'' (Chambers, 2004:128). UNIVERSALS OF NEW ENGLISHES, on the other hand, is a term coined by Christian Mair to denote the set of joint tendencies observable in the course of the standardization of postcolonial varieties of English which cannot be explained historically or genetically (Mair, 2003:84; also cf. Sand, 2004, 2005; Simo Bobda, 2000).[5] As for likely morphosyntactic candidates for vernacular universals, Chambers (2004:129) lists the following four (indicated in square brackets are the features in our 76-features catalogue which correspond most closely to the four morphosyntactic processes named by Chambers): (i) conjugation regularization, or leveling of irregular verb forms: *John seen the eclipse, Mary heared the good news* [36–39]; (ii) default singulars, or subject–verb non-concord: *They was the last ones* [55, 59; marginally 53 and 54]; (iii) *multiple negation, or negative concord*: *He hasn't got no money* [44]; and (iv) copula absence, or copula deletion: *She smart, We going as soon as possible* [57; possibly 58, 73]. How frequent, then, are these features in vernacular varieties of English sampled in our database? It turns out that they are certainly rather frequent: multiple negation ([44]), for instance, is attested (as either pervasive, or existing but rare) in 76% of the surveyed varieties, regularization of irregular verb paradigms ([36]) in 70%. It should be added, too, that in Englishes spoken in the Americas, these features are even more widespread. Yet, on a global scale it would be hard to call these features universal, with none of them being attested in more than 80% of the varieties of English worldwide. What is more, there are features in our survey which are considerably more pervasive (though note that no feature is actually attested in each and every variety in our survey). The frontrunners, then, include the following non-standard phenomena:

- [74] lack of inversion in main clause *yes/no* questions, as in *You get the point?* (attested in 89% of the varieties in the *Handbook*);
- [10] *me* instead of *I* in coordinate subjects, as in *My brother and me were late for school* (87%);

[5] UNIVERSALS OF NEW ENGLISHES, as a notion, is subject to some terminological variance: Mair (2003) also uses the term 'angloversals' (a usage from which we deliberately deviate), and Simo Bobda (2000) adopts the label 'New Englishisms'.

- [49] *never* as preverbal past tense negator, as in *He never came* (87%);
- [42] adverbs same form as adjectives, as in *He treated her wrong right from the start* (85%).

We suggest that these features be best labeled candidates for VERNACULAR ANGLOVERSALS, that is, features which tend to recur in varieties of English, be they L1 vernaculars, L2 vernaculars, or pidgins and creoles. Indeed, more cross-linguistic research into vernacular varieties of other languages might have to be carried out before we can make empirically robust statements about vernacular universals.

In a similar vein, precisely because the database samples L1 varieties, L2 varieties, and pidgins and creoles, we might also attempt to shed light on the notion of UNIVERSALS OF NEW ENGLISHES (cf. Mair, 2003; Sand, 2004, 2005; Simo Bobda, 2000). One way to approach this issue is to define, first, as *pervasive* any feature that is attested in at least 75% of any given set of varieties. In a second step, the analyst can then explore which features are pervasive in both L2 varieties and in pidgins and creoles, but not in L1 vernaculars. This criterion yields exactly three non-standard phenomena:

- [40] zero past tense forms of regular verbs, as in *I walk* for *I walked* (attested in 91% of L2 varieties and 87% of pidgins/creoles, but only in 30% of L1 varieties);
- [73] lack of inversion/lack of auxiliaries in *wh*-questions, as in *What you doing?* (attested in 75% of L2 varieties and 100% of pidgins/creoles, but only in 65% of L1 varieties);
- [6] lack of number distinction in reflexives, as in *They saw it* themself (attested in 75% of L2 varieties and 80% of pidgins/creoles, but only in 50% of L1 varieties).

Mair (2003 fn.4) explicitly states that universals of New Englishes may be the result of learning strategies of non-native speakers, in other words, properties typical of L2 varieties. Observe, along these lines, that all of the above features can indeed be argued to simplify, or even do away with rules that hold in standard English. [40] gets rid of past tense formation, [73] deletes a word order rule of standard English, and [6] in effect regularizes the morphology of reflexives.[6]

We will now turn to a number of correlational tendencies and implications between individual non-standard features in our survey. Analysts of cross-linguistic variation make a distinction between two types of correlational tendencies, BICONDITIONAL IMPLICATIONS, also known as EQUIVALENCES (for instance, 'if in a language the genitive follows the noun, then the complement follows the adposition, and vice versa'; cf. Greenberg, 1963), and ONE-WAY IMPLICATIONS, also known as PREFERENCES(for instance, 'if a language has a marked singular, it has also a marked plural, but not necessarily vice versa'; cf. Greenberg, 1966). Crucially, for most analysts the correlations need not be absolute or perfect, but can have exceptions.

In our database of 76 features (hence, out of $76 \times 75/2 = 2850$ potential candidate pairings), there are no perfect biconditional implications. However, there are 361 biconditional implications with a felicity of at least 70% (that is, which hold true in at least 70% of the varieties in the sample), and 24 biconditional implications with a felicity of at least 85%. Some of the most felicitous tendencies are given in the upper half of Table 2. So, for instance, 94% of the varieties sampled either have both *ain't* as the negated form of *be* (as in *They're all in there, ain't they?*) and *ain't* as the negated form of *have* (as in *I ain't had a look at them yet*), or they have none of these things. This relationship – which is consonant with many dialect descriptions (cf., for instance, Anderwald, 2003:149–150) – can be neatly visualized by a tetrachoric table, where unpredicted cells are shaded in gray:

		ain't* as negated form of *have	
		attested	not attested
ain't* as negated form of *be	attested	AusVE, ...	(IrE, WelE)
	not attested	(NZE)	ScE, ...

Thus, Australian Vernacular English is a typical variety in that it exhibits both *ain't* as the negated form of *be* and *ain't* as the negated form of *have*. Scottish English also conforms with the biconditional implication in that it has none of

[6] It should be added here that pidgins and creoles do not tend to have many reflexives anyway.

Table 2
Some biconditional implications exhibited in the database (the occurrence/non-occurrence of feature 1 is conditioned on the occurrence/non-occurrence of feature 2, and vice versa). The designation of pair parts as 'feature 1' and 'feature 2', respectively, is arbitrary. Percentage of varieties where implication holds is given in the rightmost column.

Feature 1	Feature 2	% of varieties
[45] *ain't* as the negated form of *be*	[46] *ain't* as the negated form of *have*	94
[12] non-coordinated subject pronoun forms in object function	[13] non-coordinated object pronoun forms in subject function	89
[23] habitual *do*	[27] *do* as a tense and aspect marker	89
[63] relative particle *as*	[64] relative particle *at*	89
[15] group plurals	[16] group genitives	85
[21] wider range of uses of the Progressive	[24] non-standard habitual markers other than *do*	26
[21] wider range of uses of the Progressive	[29] past tense/anterior marker *been*	26

these features. New Zealand English, Irish English, and Welsh English, by contrast, are the three exceptions to the statistical tendency, each attesting only one of the features, but not the other.

Moving down in Table 2, non-coordinated subject pronoun forms in object function (e.g., *You did get he out of bed in the middle of the night*) and non-coordinated object pronoun forms in subject function (e.g., *Us say 'er's dry*) also tend to go together, as do habitual *do* (e.g., *He does catch fish pretty*) and *do* as a tense and aspect marker (e.g., *This man what do own this*). The same is true for the relative particle *at* (e.g., *This is the man at painted my house*) and the relative particle *as* (e.g., *He was a chap as got a living anyhow*), and for group plurals (e.g., *That President has two Secretary of States*) and group genitives (e.g., *The man I met's girlfriend is a real beauty*). In other words, pairings like these are best seen as feature bundles instead of pairs of independent features.

There are also features that have a marked tendency not to co-occur. In our database, some 70 pairs of features have a co-occurrence likelihood of less than 30%; some such pairings are given in the lower half of Table 2. Hence, only 26% of the varieties sampled have both a wider range of uses of the progressive (e.g., *I'm liking this, What are you wanting*) and non-standard habitual markers other than *do* or the past tense/anterior marker *been* (e.g., *I been cut the bread*). Hence, the tense-mood-aspect systems in about three quarters of the surveyed varieties do not allow for extended usages of the progressive if there is a set of specialized tense-mood-aspect markers available, and vice versa. Therefore, we are likely to be dealing here with a statistical tendency in the varieties sampled to avoid redundancy in their tense-mood-aspect systems.

Let us now embark on a brief exploration of some one-way implications. The varieties sampled in our database exhibit the surprisingly large number of 247 perfect (that is, 100% felicitous) bidirectional statistical correlations, out of $76 \times 75 = 5,700$ potential candidates; at the 85% felicity level, there are no less than 721 such correlations. Needless to say, a good deal of these *prima facie* implications turn out, upon statistical and/or substantial inspection, to be random and thus meaningless. Still, the database appears to contain a healthy number of substantial one-way implications: in an inevitably exploratory and eclectic fashion (a more detailed discussion of this matter is reserved for another occasion), Table 3 lists some perfect one-way implications in our database, indicating a measure of statistical robustness in the rightmost column.[7] Thus, for instance, any variety that attests *would* in *if*-clauses (e.g., *If I'd be you*, ...) will also display loosening of the sequence of tense rule (e.g., *I noticed the van I came in*), but not necessarily vice versa. The following tetrachoric table will illustrate (as before, unpredicted cells are shaded in gray):

		loosening of sequence of tense rule	
		attested	not attested
***would* in *if*-clauses**	attested	NfldE, ...	—
	not attested	East Anglia, ...	CollAmE, ...

[7] The specific statistical measure used is *Fisher's Exact test*; see Cysouw (2003:91–92) for a discussion of the merits of this testing method in typological analysis. We are aware that the *p*-values in Table 3 are not as good as they may seem: given that we tested 247 perfect implications, *p*-values greater than $0.05/247 = 0.0002$ (assuming an alpha value of 0.05 and $n = 247$ multiple comparisons) are, statistically speaking, not robust (cf. Abdi, 2007). Notice here, however, that in addition to statistical testing, we utilized substantive – that is, qualitative – evaluation as a second filter criterion, thus weeding out obviously spurious positives.

Table 3
Some perfect one-way implications exhibited in the database (if a variety has feature 1, it also has feature 2, but not necessarily vice versa). Rightmost column provides results of one-tailed Fisher's Exact tests on corresponding 2×2 tetrachoric tables.

Feature 1	Feature 2	Fisher's Exact test
[31] *would* in *if*-clauses	[30] loosening of sequence of tense rule	$p = 0.0003$
[47] *ain't* as generic negator before a main verb	[45] *ain't* as the negated form of *be*	$p = 0.0013$
[47] *ain't* as generic negator before a main verb	[46] *ain't* as the negated form of *have*	$p = 0.0008$
[47] *ain't* as generic negator before a main verb	[48] invariant *don't* for all persons in the present tense	$p = 0.0402$
[63] relative particle *as*	[61] relative particle *what*	$p = 0.0192$
[63] relative particle *as*	[62] relative particle *that* or *what* in non-restrictive contexts	$p = 0.0192$
[63] relative particle *as*	[66] gapping or zero-relativization in subject position	$p = 0.0387$

Newfoundland English has both [30] and [31]; Colloquial American English has neither; and East Anglia has [30] but not [31]. All of these configurations are predicted by the one-way implication. What is not predicted is a variety that has [31] (*would* in *if* -clauses) but not [30] (loosening of the sequence of tense rule), and indeed such a variety is not attested in our database.[8] A possible interpretation of this particular one-way implication seems to be that loosening of the sequence of tense rule triggers, or is a necessary precondition to, *would* in *if*-clauses; the phenomenon, in short, appears to be a specific manifestation of the loosening-of-the-sequence-of-tense-rule phenomenon.

In cross-linguistic typology, one-way implications are often interpreted in terms of historical evolution. This interpretation appears to be appropriate for the implications obtaining between *ain't* as generic negator before a main verb and (i) *ain't* as the negated form of *be*, (ii) *ain't* as the negated form of *have*, and (iii) invariant *don't* for all persons in the present tense. We had demonstrated before that, more often than not, there is an equivalence between *ain't* as the negated form of *be* and *ain't* as the negated form of *have*. What we see now is that we do not obtain *ain't* as generic negator before a main verb (e.g., *Something I ain't know about*) unless we also see the two more restricted uses of *ain't*, as well as invariant *don't* (e.g., *He don't like me*). This may well have a diachronic explanation such that a variety that displays *ain't* as a generic negator must have evolved the more specific uses of *ain't* and invariant *don't* first. This line of argument would square nicely with Anderwald (2003:48) who, in her book-length treatment of non-standard negation, argues that generic *ain't* (more specifically, *ain't* for *do*) is an extension of *ain't* for *be* and *have*, and that *ain't* is spreading (in Great Britain, at least) at the expense of *don't*.

The last bundle of implications in Table 3 falls into the domain of relativization strategies. It turns out that any given variety will not have the relative particle *as* (as in *He was a chap as got a living anyhow*) unless that variety has (i) the relative particle *what* (as in *This is the man what painted my house*), (ii) the relative particle *that* or *what* in non-restrictive contexts (as in *My daughter, that/what lives in London*, ...), and (iii) gapping or zero-relativization in subject position (as in *The man ___ lives there is a nice chap*). Elements of this quite restrictive implication are in accordance with what we know about the typology of relativization strategies in dialects of English: zero-relativization, for instance, is known to be a precondition to having *as* or *what* as relative markers (Herrmann, 2003:91). It is also known from corpus study that non-restrictive *what*, in particular, is common among broad dialects, in the same way that the relative particle *as* is a feature of broad dialect speakers (cf. Herrmann, 2003). The implication suggested in Table 3 between *what* and *as*, by contrast, is not expected and will have to be studied.

We hope that this section has succeeded in outlining the many ways in which comparatively straightforward analysis methods and typological generalization patterns can be brought to bear on a database mapping language-internal variation. This does not mean, of course, that the resulting findings (such as, for example, the implications and implicational constraints identified here) are necessarily typologically grounded. Whereas the granularity of typological implications is much coarser, implications capturing language-internal variation can be much more specific and go down to the lexical level (just take the restrictive implications for the relativizers *as*, *what*, and *that* above). In exceptional cases, it may be possible to find sound functional motivations for such rather ''surfacy'' findings involving individual grammatical markers which can be generalized across languages, but in the vast majority of cases this will not be possible

[8] The (purely statistical) null hypothesis, which follows from the overall occurrence likelihoods of the two features in the database and which is the basis of the significance test in Table 3, is that there should be approximately five entries in this cell.

Table 4
Percentage of shared classifications (overlapping features) between varieties of English. Top 5 and bottom 5 pairings.

Variety 1	Variety 2	% of feature classifications shared
CollAmE	OzE	90.8
Bislama	TP	89.5
InSAfE	BlSAfE	85.5
SolP	TP	82.9
SurCs	TP	81.6
IsSE	GhP	17.1
IsSE	TP	19.7
Southwest	NigP	21.1
Earlier AAVE	BelC	21.1
IsSE	Bislama	23.7

and should not even be attempted. The overall lesson to be learned from this is that implicational universals formulated in typological research should not only firmly be expected, but indeed need to stand the test against varieties of individual languages, whereas the reverse is not true for language-internal implicational generalizations.

4. Mapping similarities and dissimilarities in varieties of English: multidimensional scaling

This section will abandon the feature-centric perspective implicit in the previous section in favor of an approach that seeks to map varieties of English in terms of similarities and dissimilarities, a time-honored method in dialectology and dialectometry (see Goebl, 2004:26 for an evaluation, and Spruit, 2005 for an application to a syntactic database). Consider, in exactly this spirit, Table 4, which compares the five pairs of varieties in the survey which share most feature classifications in terms of the survey's tripartite A–B–C distinction to those five pairs which share, overall, the least amount of feature classifications (technically, Table 4 thus tallies overlapping features; cf. Séguy, 1971). The general pattern is very clear: the pairings that are most similar consist of varieties of the same type (L1, L2, or pidgins/creoles) and located in geographic proximity. So, for instance, Colloquial American English shares 90.8% of its feature classifications with Ozarks English; both varieties are North American L1 varieties. In a similar vein, Bislama shares 89.5% of its feature classifications with Tok Pisin, which, like the former, is a Pacific pidgin (though it is somewhat surprising, for a number of areal and historical reasons, that Tok Pisin shares more classifications with Bislama than with Solomon Island Pidgin). Conversely, those pairings which share a minimum number of feature classifications consist of varieties that are of a different type and that are from different anglophone world regions (e.g., Isolated Southeast American English, a L1 variety, and Ghanaian Pidgin English, an African pidgin, which share only 17.1% of their feature classifications).

In short, the similarity indices in Table 4 give an impression of the similarities and dissimilarities exhibited in our database. For the remainder of this section, we will utilize MULTIDIMENSIONAL SCALING (henceforth: MDS), a considerably more refined set of statistical techniques, in order to uncover the hidden structure of variation in World Englishes (for an introduction to MDS, see Kruskal and Wish, 1978). This means that we will scale down the 76 original dimensions by which every variety in our database is characterized, which will make it possible to visualize the (dis-)similarities between the varieties in two-dimensional maps. The obvious advantage of such perceptual maps is that, because such visual representations use the straightforward concepts of space and distance, they can be interpreted fairly intuitively. Much as with geographic maps, the further two points are apart, the more dissimilar (in geographic terms, distant) they are. If two pairs of points are equally close or distant, the pairs of varieties they represent are equally (dis-)similar. We conducted the MDS analysis using the 'Alscal' algorithm implemented in SPSS, a classical MDS method that derives perceptual maps from Euclidian distances (cf. Young and Harris, 1992). The distance matrix was created from our 76×46 syntactic database. We settled on a two-dimensional MDS solution with a stress value of 0.19 and an R^2 value of 0.82, which means that 82% of the variance in the database is accounted for by the MDS model (we should add here that with a stress value of 0.13 and an R^2 value of 0.87, a three-dimensional solution would have had an only slightly better fit). The resulting visualization is given in Fig. 2.

To encourage confidence in Fig. 2, observe that the varieties in pairings in the upper half of Table 4 end up close in Fig. 2, as they should given how many feature classifications they share. Conversely, varieties in the pairings below the

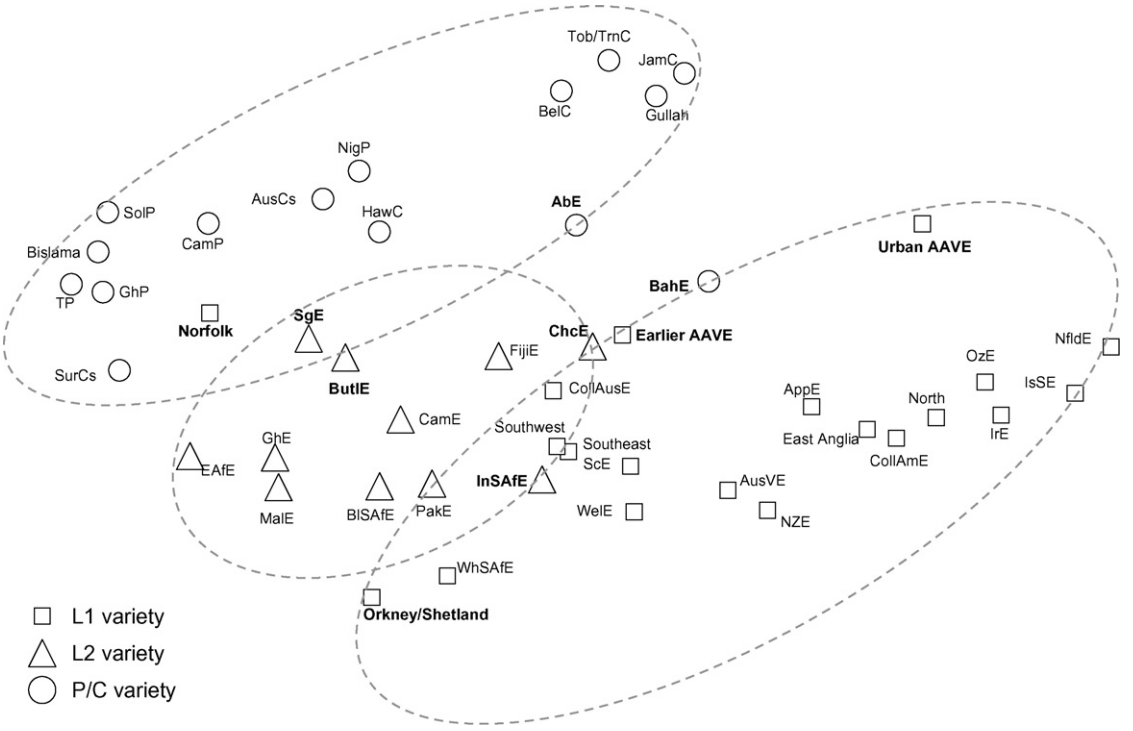

Fig. 2. MDS visualization based on the 76 × 46 database of varieties of English. Squares represent L1 varieties, triangles represent L2 varieties, circles represent English-based pidgins and creoles. Dotted ellipses group variety types.

dotted line in Table 4 are quite distant in Fig. 2– again, as they should.[9] As a general pattern, it is interesting that Fig. 2 groups varieties fairly consistently according to variety type (L1, L2, and pidgins/creoles; the dotted ellipses seek to impressionistically delineate the three variety types[10]). This means that the areas where one finds L1 varieties (in the lower half of Fig. 2) and where one finds pidgins and creoles (in the upper half of Fig. 2) tend not to overlap. Independent-samples *t*-tests on the MDS coordinates confirm that the visually salient spatial difference between pidgins/creoles and L1 varieties in Fig. 1 is robust statistically in both dimensions, horizontally ($t = 4.37$, $df = 33$, $p < 0.001$) and vertically ($t = -7.2$, $df = 33$, $p < 0.001$), as is the difference between L2 varieties and all other varieties (horizontally: $t = 3.1$, $df = 40.7$, $p = 0.004$; vertically: $t = 3.0$, $df = 38.8$, $p = 0.004$). This general pattern notwithstanding, there are two exceptions to the clear pattern of separation according to variety type: Norfolk, classified as a L1 variety in our survey, is located in the pidgin/creole area, and Bahamian English (BahE), classified as a pidgin/creole variety, ends up in the L1 group (although only marginally so). Now, the classification of Norfolk and BahE is actually not as clear-cut as the survey's classification (which was the product of lively discussions among the *Handbook*'s editors) might pretend: in the survey article on Norfolk, Mühlhäusler (2004:789) notes that this variety ''shares the characteristic of many creoles, koinés and mixed languages of not having a great deal of inflectional and derivational morphology'', and, as it were, Burridge (2004:1116) classifies Norfolk as a contact variety. Reaser and Torbert, on the other hand, devote a whole section of their survey article on BahE to the question as to why the linguistic status of the varieties covered under the label 'BahE' is a problematic one (2004: 392–392); it is well-known, too, that the settlement history of the Bahamas is rather unique.

Be that as it may, L1 varieties and pidgins and creoles are clearly demarcated in Fig. 2. They are 'linked', so to speak, by L2 varieties, the majority of which (e.g., Malaysian English, Ghanaian English, Cameroon English, . . .) are

[9] Still, it should be pointed out that Fig. 2 is not completely felicitous to Table 4 (for instance, the North of England is located even closer to Colloquial American English than Ozarks English). This is because MDS does not optimize over pairs of varieties only, but takes into account all 46 varieties in the dataset at once.

[10] Notice that these ellipses do not derive from statistical analysis but from visual inspection. As one reviewer remarked, they are based on – and are actually meant to emphasize – what we already know about the varieties in the survey, i.e. their variety type.

located in the area between L1 varieties and pidgins/creoles. There are some borderline cases, though: Butler English and Singapore English (SgE), for instance, border on the pidgin/creole area. The coordinates of Butler English, in point of fact, nicely square with Hosali's (2004:1032) evaluation that it is hard to say whether the variety should be considered a pidgin or an early fossilized interlanguage. As for SgE, it is unclear whether this variety should be understood in terms of a post-creole continuum or in terms of diglossia (cf. Wee, 2004:1070) – in the literature, colloquial Singapore English has actually been called a 'creoloid' or 'semi-pidgin' (Gil, 2003:469f). It is also interesting, along these lines, that Aboriginal English (AbE) is located closer to the L1 group than any other pidgin/creole apart from BahE. In fact, the status of AbE is not entirely uncontroversial either: Malcolm (2004:678), at any rate, notes that forms in AbE resemble acrolectal – rather than basilectal or mesolectal – forms in Australian Creoles, and Burridge (2004:1116) actually refers to AbE as a 'native' variety.

In Fig. 2, within the L1 circle (but at its periphery), the coordinates of the following varieties appear to us as deserving of some discussion:

- Indian South African English (InSAfE), which is classed as a L2 variety in our survey while clearly intruding into the L1 area in Fig. 2. Note that some analysts (for instance, Mesthrie, 2004:974) account for the variety in terms of a former L2 variety which has developed into a native L1 vernacular.
- Chicano English (ChcE) – a variety label which, in the *Handbook* , is actually a somewhat mixed bag, referring to the 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 (cf. Bayley and Santa Ana, 2004:374). Given that, it should surprise no one that ChcE is somewhat peripheral to the L2 group in Fig. 2.
- Orkney and Shetland English: while there should be no doubt that the English spoken in Orkney and Shetland is an L1 vernacular, the dialect (technically, a variety of Scots, though interestingly MDS does not locate the variety in particular proximity to Scottish English [ScE]) has an exceptional history, both linguistically and politically. In particular, it incorporates many Scandinavian features that are not attested in other British varieties of English (cf. Melchers, 2004). It is characteristics like these which appear to make the variety dissimilar from other L1 varieties in MDS analysis.
- Earlier African American English (Earlier AAVE) and (contemporary) Urban African American English (Urban AAVE) are transparently L1 ethnolects, yet MDS locates the two varieties on the periphery of the L1 cluster and in relative proximity to the L2 and pidgin/creole groups. Why is that? While this is not the place to even begin to review the sizable literature on AAVE, sociolinguists agree that the history of AAVE is unique. A contentious point, however, is the issue whether AAVE has creole origins (cf. Labov, 1972) or English origins (cf. Poplack, 2000). It is also unclear whether AAVE has been converging with white varieties of English, or whether there has been divergence (cf., for instance, Bailey and Maynor, 1989). Our MDS analysis – starkly inadequate as it is for offering any conclusive evidence relating to these issues – suggests that Earlier AAVE and Urban AAVE are fairly distant morphosyntactically, but that the two varieties of AAVE are equidistant from 'white' varieties of American English (though Earlier AAVE seems to be located somewhat closer to the pidgin/creole group in Fig. 2 than is Urban AAVE).

Summing up, we have suggested in this section that MDS yields a statistical arrangement of variety types that is both meaningful from a linguistic point of view, and instructive in terms of visualizing the range of the observable variation. We believe we have also demonstrated that there is a marked divide between native English vernaculars, on the one hand, and English-based pidgins and creoles, on the other hand, with English as a Second Language varieties covering the middle ground.

5. Varieties of English: cluster analysis

We will now engage in a considerably more fine-grained analysis of clusterings and patterns across varieties of English in our database. To this purpose, we will conduct CLUSTER ANALYSIS, a set of techniques used to objectively group a large number of cases (in this study, varieties of English) into a smaller number of discrete and meaningful clusters on the basis of some sort of similarity – in our case, similarity of feature classifications. Cluster analysis is less visual than MDS but can establish patterns in a statistically somewhat more robust way. The method is widely used in disciplines such as biology, the social sciences (e.g., social network analysis), dialectology (cf., for instance, Shackleton, 2005), and dialectometry (cf., for example, Goebl, 2004; Nerbonne et al., 1999); for an introduction to the

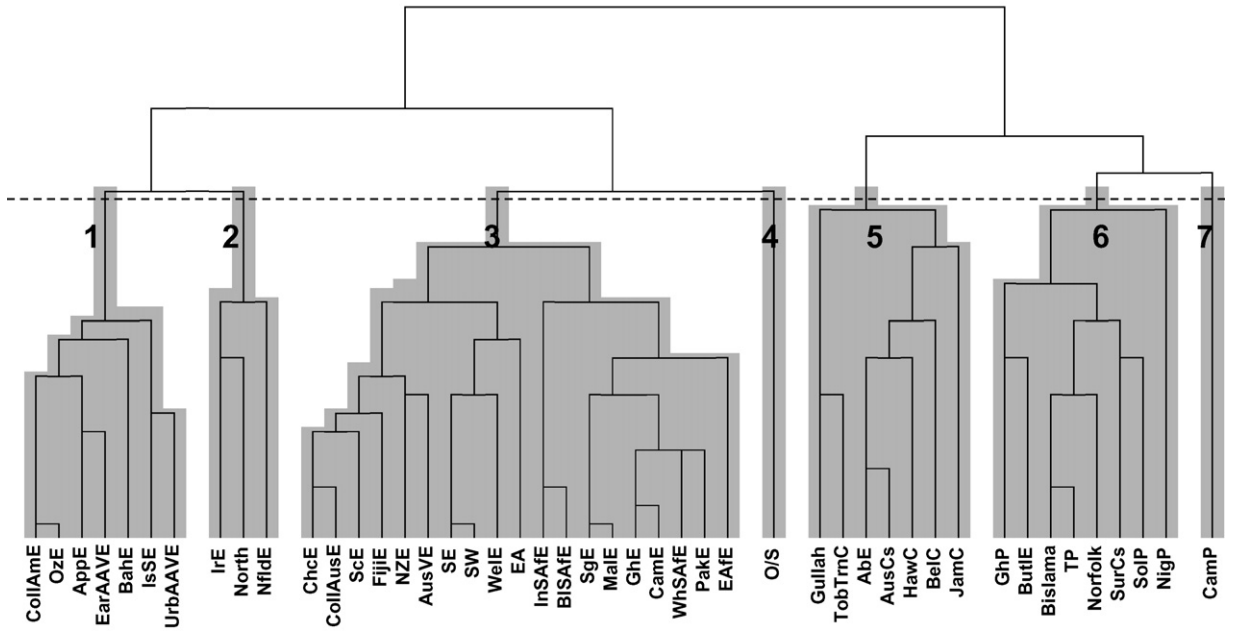

Fig. 3. Dendrogram derived from hierarchical agglomerative cluster analysis of the 76 × 46 database. Numbered clusters indicate a seven cluster solution.

technique from the social scientist's perspective, see Aldenderfer and Blashfield (1984). We used SPSS's hierarchical agglomerative clustering algorithm to partition the varieties in our database into clusters, relying on between-groups-linkage as clustering method and square Euclidean distance as interval measure.[11] We would like to caution readers that the clusters identified in this section, suggestive as they may be, are tentative (pending more rigorous future testing). The reason is that simple clustering of the type we perform here can be unstable unless resampling procedures or similar techniques are applied (cf. Nerbonne et al., 2007 for a discussion). Nonetheless, despite these caveats, we do believe that our preliminary findings (taken with a grain of salt) may be instructive in the context of the present paper.

Data clustering can be visually represented using tree diagrams, also know as DENDROGRAMS, where one finds individual varieties on the bottom and successively larger clusters as one moves up. Essentially, then, dendrograms work in much the same way as family trees.

The cluster analysis run on our database yields the dendrogram in Fig. 3. In a dendrogram, much as in a family tree, the closer to the bottom we find a split between a pair of varieties (or a pair of groups of varieties), the more closely related the varieties (or groups of varieties) are. Observe, along these lines, that Colloquial American English and Ozarks English, as well as Bislama and Tok Pisin, split relatively low in the dendrogram; it is not by accident that these two pairs are the frontrunners in the upper half of Table 4, which displays the percentage of shared variants between varieties of English. Let us now, step-by-step, work our way down from the top to the bottom of the dendrogram in Fig. 3.

- The most fundamental split occurs between English-based pidgins and creoles, to the right of the dendrogram, and other varieties of English, to the left of the dendrogram. Butler English, Bahamian English, and Norfolk are the only anomalies in that ButlE and Norfolk group with the pidgins/creoles and BahE groups with the non- pidgin/creole varieties (which, for the very reasons explicated in section 4 of this study, was to be expected). This fundamental split testifies, once again, to the paramount importance of variety type in our database (cf. Fig. 2).
- Next, the non-pidgin/creole branch, to the left of the dendrogram, regroups into an essentially American cluster, plus the North of England and Irish English, and a non-American cluster, spanning a number of both L1 and L2 varieties. At the same time, the pidgin/creole cluster splits up into a creole-only cluster (Gullah, Tob/TrnC, BelC, HawC, AbE,

[11] In plain English, this means that the clustering algorithm started out with individual varieties, merging them stepwise – when possible – with other varieties to arrive at successively larger clusters under the condition that cluster members be maximally similar.

Table 5
Hierarchical agglomerative cluster analysis. Group memberships for seven clusters solution.

Cluster	Variety
1	CollAmE, IsSE, OzE, AppE, Urban AAVE, Earlier AAVE, BahE
2	IrE, North, NfldE
3	ScE, East Anglia, WelE, Southeast, Southwest, ChcE, FijiE, NZE, AusVE, CollAusE, GhE, CamE, EAfE, WhSAfE, InSAfE, BlSAfE, PakE, SgE, MalE
4	Orkney/Shetland
5	Gullah, Tob/TrnC, BelC, HawC, AbE, AusCs, JamC
6	SurCs, SolP, Bislama, TP, Norfolk, GhP, ButlE, NigP
7	CamP

AusCs, JamC), and a cluster containing the English-based pidgins (SolP, CamP TP, Bislama, GhP, NigP) in addition to two English-based creoles (SurCs, Norfolk) and Butler English. Cluster analysis seems to suggest, therefore, that at this level the typological difference between English-based pidgins and English-based creoles is more marked than the difference between English L1 vernaculars and English L2 varieties.

- Subsequently, the database is split up into seven clusters, which are visualized in Fig. 3; the corresponding group memberships are also indicated in Table 5 for convenience. At this stage, Cameroon Pidgin English and Orkney and Shetland English are sufficiently distinct from the other varieties in the database to form clusters of their own (clusters 4 and 7). Since Orkney and Shetland English, in particular, is usually considered a dialect of Scots, this arrangement is interesting. Further, the algorithm now distinguishes an all-American cluster – CollAmE, IsSE, OzE, AppE, Urban AAVE, Earlier AAVE, BahE (cluster 1) – from a cluster containing Irish English, the North of England, and Newfoundland English (cluster 2). Given the large influx of Irish settlers into Newfoundland (cf. Clarke, 2004), it is not astonishing that IrE and NfldE end up in the same cluster, though it is noteworthy that the North of England, and not, say, England's Southwest (which contributed another large portion of settlers to Newfoundland) is being grouped with Newfoundland.

Some further instructive divisions further down the road in the dendrogram should also be addressed at this time. For example, cluster 3 in Table 5 is subject to a later split into an almost exclusive L2 cluster (InSAfE, BlSAfE, SgE, MalE, GhE, CamE, WhSAfE, PakE, EafE) and a cluster containing some L2 varieties (ChcE, FijiE) and almost all of the non-American L1 vernaculars (CollAusE, ScE, NZE, AusVE, SE, SW, WelE, EA). This latter cluster subsequently breaks up into a British-only L1 group (SE, SW, WelE, EA) and a strikingly heterogeneous party, from a historical and geographical perspective, comprising Chicano English, Colloquial Australian English, Scottish English, Fiji English, New Zealand English, and Australian Vernacular English. Further note that in the left branch of the dendrogram, Early African American Vernacular English and Urban African American Vernacular English are not exceedingly close – as a matter of fact, Urban AAVE forms a sub-cluster with Isolated Southeast US English. This observation might once again point to the considerable development that AAVE has undergone in its history. Last but not least, we should mention that in the right branch of the dendrogram a distinction emerges between Gullah and Tobagonian/Trinidadian Creole, for one thing, and a number of other English-based creoles (BelC, HawC, AbE, AusCs, JamC). In all, cluster analysis demonstrably yields groupings of varieties not unlike those established by more traditional approaches to dialectology. Yet, the dendrogram in Fig. 3 has considerably refined our earlier findings concerning similarities and dissimilarities between the varieties in our database. By way of recapitulation, the most fundamental split among varieties of English, from a bird's eye perspective, is that between English-based pidgins and creoles, on the one hand, and other varieties of English, that is, L1 and L2 varieties of English, on the other hand. Moreover, we have argued that this split overrides geographic patternings and appears to be somewhat more pivotal than the typological difference between L2 varieties of English and English L1 vernaculars.

6. Dimensions of morphosyntactic variance: principal component analysis

In this section, we shall explore what generalizations can be made about underlying dimensions of morphosyntactic variance in World Englishes, drawing on PRINCIPAL COMPONENT ANALYSIS of our database. Principal component analysis is a statistical technique that reduces a number of independent variables to a smaller number of hypothetical

constructs, or dimensions, known as COMPONENTS (or FACTORS), which should be assigned meaningful interpretations by the analyst. The basic idea in principal component analysis is that independent variables grouped in a component should be highly correlated to each other, but must not correlate substantially with independent variables in other components (Kim and Mueller, 1978 is a recommendable introduction to principal component analysis).

In some ways, principal component analysis is not dissimilar to MDS and cluster analysis: all three techniques seek to simplify a matrix of data. The crucial difference between principal component analysis and MDS is that while what is paramount in the latter is distances/dissimilarities between cases, what counts in principal component analysis is the arrangement of cases along – ideally – meaningful dimensions (or: components). The difference between principal component analysis and cluster analysis, in short, is that while the former groups variables, the latter groups cases. In linguistics, principal component analysis has been applied in studies of register variation (for instance, Biber, 1988), corpus linguistics (for instance, Keune et al., 2005), and dialectology (for instance, Shackleton, 2005). As Shackleton (2005:142) succinctly puts it, ''applied to a data set of linguistic features, principal component analysis may isolate sets of linguistic features that tend to occur together and not with other features.'' We will now probe where principal component analysis can take us in terms of our dataset: which bundles of morphosyntactic features account, in concert, for most variance between varieties of English? Treating varieties as variables and features as cases, we utilized principal component analysis (with Varimax rotation and Kaiser normalization) to extract two components for subsequent visualization. The first component accounts for 23.2% of the observable variance in the database, component 2 explains 15.2% of the variance; component 1 is thus statistically more robust than component 2. Fig. 4 visualizes the component scores for both principal components in a two-dimensional coordinate plane.

As for the linguistic interpretation of the two components, we offer that component 1 can be seen to display increased levels of MORPHOSYNTACTIC COMPLEXITY. While we are fully aware that complexity is a notoriously ill-defined notion, we set out to tentatively define morphosyntactic complexity as follows: A given variety *X* is morphosyntactically more complex than a given variety *X′* if variety *X* exhibits (i) a smaller amount of features that can be argued to simplify syntactic rules, (ii) a smaller amount of features that can be said to aid processing, and (iii) more features that are indicative of ''distinctions beyond communicative necessity'' (McWhorter, 2001:125; cf. also Szmrecsanyi and Kortmann, 2009 for an in-depth investigation of complexity differentials in varieties of English).

Further, we believe component 2 to indicate a given variety's degree of ANALYTICITY, a notion which we operationally define as bringing about a greater number of features that are rather autonomous – that is, invariable and periphrastic –

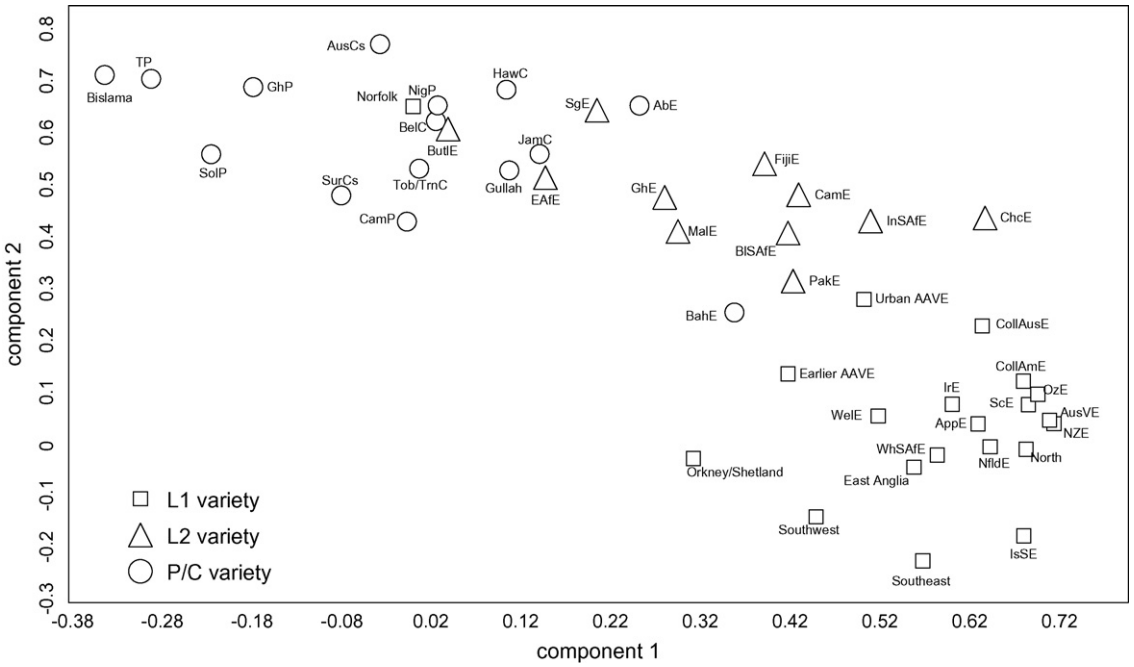

Fig. 4. Visualization of principal components of variance in the 76 × 46 database. Squares represent L1 varieties, triangles represent L2 varieties, circles represent English-based pidgins and creoles.

Table 6
Principal component analysis – selected component score coefficients for the first principal component.

Feature	Regression score
[19] double comparatives and superlatives	1.483
[69] inverted word order in indirect questions	1.419
[44] multiple negation/negative concord	1.231
[21] wider range of uses of the Progressive	1.050
[16] group genitives	0.594
[15] group plurals	0.387
[31] *would* in *if*-clauses	−0.727
[8] generic *he /his* for all genders	−1.274
[72] serial verbs	−1.643

Table 7
Principal component analysis – selected component score coefficients for the second principal component.

Feature	Regression score
[40] zero past tense forms of regular verbs	2.019
[53] invariant present tense forms due to zero marking for the third person singular	1.960
[49] *never* as preverbal past tense negator	1.388
[52] invariant non-concord tags	1.344
[50] *no* as preverbal negator	1.237
[29] past tense/anterior marker *been*	1.170
[15] group plurals	−0.617
[16] group genitives	−0.944
[51] *was–weren't* split	−1.223

in nature, much along the lines of Vincent's concise definition: "A Construction *C* is relatively more analytic than another construction *C′* having approximately the same grammatical content as *C* to the extent that the constituent elements of *C* show greater morphosyntactic and phonological autonomy than do those in *C′*" (1997:99).[12]

What is the evidence for this specific interpretation of the axes in Fig. 4? It is our sense that those features characteristic of varieties towards the right pole of the *x*-axis (that is, component 1) are, more often than not, indicative of increased levels of morphosyntactic complexity, and that many of those features that are characteristic of varieties scoring high on the *y*-axis (i.e., component 2) are rather analytic in nature. To illustrate, consider Tables 6 and 7, which display a choice of component score coefficients (as estimated in a regression analysis) associated with individual features in the first and second principal component. In Table 6, all the features that are positively correlated with component 1 can be argued to be more complex morphosyntactically. Thus, double comparatives and superlatives (e.g., *he's more taller*) are arguably more complex morphologically than 'simple' comparatives and superlatives (e.g., *he's taller*); inverted word order in indirect questions (e.g., *I'm wondering what are you gonna do?*) extends the inversion rule to an additional syntactic environment; multiple negation (e.g., *He won't do no harm*) adds an additional concord rule; wider ranges of uses of the Progressive (e.g., *what are you wanting*) can be possibly claimed to express distinctions beyond communicative necessity; and both group genitives (e.g., *the man I met's girlfriend*) and group plurals (e.g., *two secretary of states*) are likely to be more complex to parse as the plural/genitive markers have wider syntactic scope. Conversely, among the features negatively correlated with component 1, one tends to find phenomena such as *would* in *if*-clauses, a feature that – as we have seen in section 3– implicates a relaxation of the sequence of tenses rule. Likewise, generic *he/his* for all genders (e.g., *my car, he's broken*) does away with a gender distinction, and serial verbs are often considered a conjoined alternative to outright syntactic subordination (cf., for instance, Baker, 1989).

[12] It should be stressed that in a typological perspective, analyticity and complexity are two distinct notions. While analytic structures may sometimes be morphosyntactically less complex as well, this is not necessarily the case. Also, obviously one would not want to argue that Sinitic or Polynesian (analytic language par excellence) are inherently less complex than synthetic or inflectional languages.

Table 7 lists some of those features positively correlated with component 2, which we take to indicate analyticity. Zero past tense forms of regular verbs (e.g., *I walk* for *I walked*) is a feature that can be claimed to be fairly analytic in that it does away with overt inflectional marking. Invariant present tense forms (e.g., *he show up), never* as a preverbal past tense negator (e.g., *he never came*), invariant tags (e.g., *innit?*), *no* as a preverbal negator (e.g., *me no iit brekfus*), and the past tense/anterior marker *been* (e.g., *I been cut the bread*) all are features that replace concord structures with invariant, and hence more autonomous, material. Negatively correlated with component 1 are features such as group genitives and plurals, which strengthen inflectional, rather than analytic elements in the grammar of English, and the *was–weren't* split (e.g., *the boys was interested but Mary weren't*), a phenomenon which re-shuffles inflectional distinctions, as opposed to abandoning them. It is for these reasons, in short, that we believe the horizontal axis in Fig. 4(component 1) to indicate morphosyntactic complexity, and the vertical (component 2) axis to gauge analyticity.

Notice now that the plot in Fig. 4 locates variety types in fairly remote sectors of the diagram: English-based pidgins and creoles are located in the upper left-hand corner of the diagram, L2 varieties of English are to be found towards the upper right-hand corner of Fig. 4, and L1 vernaculars are clustered in the lower right-hand corner of the diagram. The outliers are the usual suspects familiar from previous sections in this study (i.e., Norfolk and BahE). A one-way analysis of variance (ANOVA) on the varieties' coordinates in Fig. 4 confirms this visual classification: there are significant differences between varieties of different types (conceptualized here in terms of a four-fold typology: L1s vs. L2s vs. creoles vs. pidgins) both in regard to mean scores on the horizontal dimension (i.e. component 1; $F = 38.45$, $df(n) = 45$, $p < 0.001$) and in regard to mean scores on the vertical dimension (i.e. component 2; $F = 39.03$, $df(n) = 45$, $p < 0.001$). With regard to morphosyntactic complexity (component 1), then, the hierarchy in (1) emerges:

(1) English-based pidgins < English-based creoles < L2 varieties of English < English L1 vernaculars

Is this hierarchy meaningful from a theoretical perspective? We offer that it is. As for (morpho-)syntactic complexity, we would like to draw attention to an ongoing debate in the literature about what (if anything) it is that sets apart creoles, as a synchronic class, from other languages. In exactly this connection, John McWhorter has argued that

> if all of the world's languages could be ranked on a scale of complexity, there would be a delineable subset beginning at the ''simplicity'' end and continuing towards the ''complexity'' one all of which were creoles. (2001:162)

This should be the case, according to McWhorter, because creole genesis tends to eliminate grammatical complexity and because not enough time has elapsed since this genesis to accumulate much historical complexification in the meantime (2001:155). Our modest empirical inquiry may thus well be seen to suggest that McWhorter's view is not implausible, at least as far as varieties of English and English-based creoles are concerned. Conflating the English-based pidgins in our survey with the English-based creoles – note that pidgins are not usually considered full-fledged languages in their own right due to their lack of grammatical structure (which is exactly why they pattern to the left of creoles in Fig. 4) – one finds that in Fig. 4 all of the leftmost varieties are English-based pidgins and creoles, with Bislama, Tok Pisin etc. constituting the most extreme cases. Moving rightwards, one then obtains overlap areas between pidgins, creoles, and L2 varieties of English: Nigerian Pidgin English, for example, scores as high on the complexity (i.e., horizontal) dimension as Belizean Creole and Butler English, an L2 variety. Even more rightwards, we find creoles, such as Aboriginal English, that are as complex as some L2 varieties, such as Ghanaian English. Finally, one gets a number of L2 varieties – Chicano English, Indian South African English, and so on – that are as complex as the average L1 vernacular, though there also appear to be L1 vernaculars (for instance, Orkney and Shetland English) that appear to be less complex morphosyntactically than some L2 varieties. Be that as it may, on aggregate the hierarchy in (1) clearly stands.

What can be said about the other dimension in Fig. 4? Analyticity (component 2) appears to yield the following hierarchy:

(2) English-based pidgins/English-based creoles/L2 varieties of English > English L1 vernaculars

This is another way of saying that in terms of the vertical axis (and thus, in terms of analyticity), there is not much variance between English-based pidgins, English-based creoles, and L2 varieties. Crucially, however, these three

groups score (the usual exceptions, e.g., BahE, apply) higher on the analyticity dimension than any L1 vernacular in the sample. It is well-known, in this regard, that creoles in general tend to be highly analytic if not isolating languages (cf. Schuchardt, 1979; Hagège, 1985; Mufwene, 1990; McWhorter, 1998). Similarly, the adult L2 acquisition process is known to be hostile toward inflectional morphology (cf., for instance, Klein and Perdue, 1997), a fact which might ultimately be responsible for the analytic/isolating nature of creole languages as well (cf. DeGraff, 1997). At any rate, it is reasonable that among the varieties sampled in our database, L1 vernaculars should be the least analytic ones. Within the L1 group, independent-samples *t*-tests on the coordinates of individual L1 vernaculars did not reveal any statistically significant patterns (in either dimension), though there seems to be a slight tendency for American varieties, on aggregate, to be a tad more complex ($p = 0.25$) and more analytic ($p = 0.15$) than most British varieties.

In sum, drawing on principal component analysis of our database, we have suggested in this section that varieties of English can be thought as varying along two fundamental dimensions, morphosyntactic complexity and analyticity. These dimensions yield hierarchies which arrange varieties of English, once again, according to variety type rather than according to other factors.

7. Conclusions

Drawing on a large morphosyntactic database comprehensively portraying almost four dozen varieties of English around the world – and thus doing away with the narrow focus in much previous dialectological and variationist research on one or two features in a handful (at most) of dialects or varieties – this study would seem to have offered a novel perspective on morphosyntactic variation in World Englishes. In conclusion, let us recapitulate what we believe are our core findings.

First, a large database such as ours offers a rich testing ground for, and a resource for the discovery of, (typological) generalizations. Along these lines, we have investigated the features in our database in regard to vernacular angloversals, universals of New Englishes, and implicational correlations. Second, we used multidimensional scaling to visualize the structure of variation across the Englishes in our database. We suggested, among other things, that there is a major typological division between English L1 vernaculars, on the one hand, and English-based pidgins and creoles on the other hand, whilst L2 varieties fairly consistently stick to the middle of the road. Third, we relied on cluster analysis to partition the varieties in our database into discrete clusters. This line of analysis confirmed, among other things, the important split between pidgins/creoles and other varieties in our database, and indicated that this divide is more pivotal than either geographical patterns or the difference between L1 varieties and L2 varieties. Lastly, we utilized principal component analysis to uncover the underlying dimensions of variance in our dataset. In a nutshell, we have offered that varieties of English can be thought of as varying along two major typological dimensions – morphosyntactic complexity and analyticity – and argued that once again, variety type is the best predictor of a given variety's location relative to these dimensions.

At this point, we would like to stress explicitly that it is not our intention to argue that the quantitative approach outlined in this study can ever replace in-depth qualitative dialectological study. What it can do, indeed, is complement qualitative inquiry. At the same time, the reader may rest assured that we are fully aware of the many ways in which our database, in its current form, is insufficient. Thus, as a prerequisite for more rigorous investigation, the feature catalogue on which the survey rests needs to be substantially expanded and refined, and more varieties of English need to be covered while relying on a wider circle of informants. Only then can one begin to deal with some of the questions we have raised. But however reductionist and simplistic our analysis may be, we believe that the benefits of being able to see the wood for the trees are worth some of the costs that this approach inevitably incurs.

Beyond the data source, what are some of the issues that future work on the global morphosyntax of English along the lines of the present study will need to resolve? For one thing, it should prove profitable to probe synchronic databases like ours in regard to genetic relationships and historical relatedness between varieties of the same language, in the spirit of, e.g., Dunn et al. (2005), Gray and Atkinson (2003), McMahon and McMahon (2006), and Ringe et al. (2002). Notice, for instance, that the dendrogram in Fig. 3 might lend itself to a more explicitly historical interpretation than we have offered in the present study. Secondly, it seems worth pointing out that some of our findings, which are admittedly conditioned on the comparatively heterogeneous nature of our database, challenge a time-honored principle, labeled the ''fundamental dialectology principle'' by Nerbonne and Kleiweg (2007), according to which geographically close varieties are more similar than distant ones. As we have demonstrated, this is not always the case as one leaves behind the department of geographically more or less contiguous native vernaculars and forays into the

realm of L2 varieties and contact languages without much of a shared history. If, as we have seen, geographic distance is not an exceedingly powerful predictor of variance in our database – if, that is, there is no close link between geography and typology – future study will want to explore other factors (difference in historical depth; immersion with speakers of other languages; difference in number of speakers; difference in per capita gross domestic product; and so on). In a similar vein, we will need to investigate ways in which the sophisticated visualization methods used in state-of-the-art dialectometry can be applied to data where, to reiterate, geographical contiguity does not necessarily obtain. And finally, future inquiry might want to address the extent to which a dataset such as the one analyzed here can yield sociolinguistic insights into ''the globalisation of vernacular variation'' (Meyerhoff and Niedzielski, 2003).

In all, we submit that coming to terms with a database like ours presents a challenge of sorts to traditional dialectology and also dialectometry, but one that will be well worth meeting.

Acknowledgements

We are grateful to the editor, John Nerbonne, and to three anonymous reviewers for many invaluable and truly helpful comments and suggestions. While the usual disclaimers apply, their input has made this a better paper.

Appendix A. The feature catalogue

Pronouns, pronoun exchange, pronominal gender

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, aay', yufela, you . . . together, all of you, you ones/'uns, you guys, you people,* etc.)
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, We like us town, Show us* 'me' *them boots, Us kids used to pinch the sweets like hell, Us'll do it*)
12. non-coordinated subject pronoun forms in object function (e.g., *You did get he out of bed in the middle of the night*)
13. non-coordinated object pronoun forms in subject function (e.g., *Us say 'er's dry*)

Noun phrase

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 (also known as 'group clitcs/inflections'; e.g., *The man I met's girlfriend is a real beauty*)
17. irregular use of articles (e.g., *Take them to market, I had nice garden, about a three fields, I had the toothache*, etc.)
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 kind a guy I know, in one of the most pretty sunsets*)

Verb phrase: Tense and aspect

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?, Some of us have been to New York years ago*)
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 don ate 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 'are standing' there you can see the flames)*
33. *after*-Perfect (e.g., *She's after selling the boat*)

Verb phrase: Modal verbs

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*)

Verb phrase: Verb morphology

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*)

Adverbs

42. adverbs (other than degree modifiers) derived from adjectives lack *–ly* (e.g., *He treated her wrong right from the start*)
43. degree modifier adverbs lack *–ly* (e.g., *That's real good*)

Negation

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?*)

Agreement

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*, or: *You was hungry but he was thirsty*)
60. Northern Subject Rule (e.g., *I sing vs. *I sings, Birds sings, I sing and dances*)

Relativization

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*)

Complementation

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., *We always had gutters in the winter time 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'*)

Discourse organization

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? Like for one round five quid, that was like three quid, like two-fifty each*)
76. like as a quotative particle (e.g., *And she was like, what do you mean?*)

References

Abdi, H., 2007. The Bonferroni and Sidak corrections for multiple comparisons. In: Salkind, N.J. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, pp. 103–107.

Aldenderfer, M.S., Blashfield, R.K., 1984. Cluster Analysis. In: Quantitative Applications in the Social Sciences, Sage Publications, Newbury Park, London, New Delhi.

Anderwald, L., 2003. Negation in Non-Standard British English: Gaps, Regularizations and Asymmetries. Routledge, London, New York.

Bailey, G., Maynor, N., 1989. The divergence controversy. American Speech 64, 12–39.

Baker, M., 1989. Object sharing and projection in serial verb constructions. Linguistic Inquiry 20, 513–553.

Bayley, R., Santa Ana, O., 2004. Chicano English: morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 374–390.

Biber, D., 1988. Variation across Speech and Writing. Cambridge University Press, Cambridge.

Burridge, K., 2004. Synopsis: morphological and syntactic variation in the Pacific and Australasia. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 1. Mouton de Gruyter, Berlin/New York, pp. 1116–1131.

Chambers, J.K., 2001. Vernacular universals. In: Fontana, J.M., McNally, L., Turell, M.T., Vallduv, E. (Eds.), ICLaVE 1: Proceedings of the First International Conference on Language Variation in Europe. Universitat Pompeu Fabra, Barcelona, pp. 52–60.

Chambers, J.K., 2003. Sociolinguistic Theory: Linguistic Variation and Its Social Implications. Blackwell, Oxford, Malden.

Chambers, J.K., 2004. Dynamic typology and vernacular universals. In: Kortmann, B. (Ed.), Dialectology meets Typology. Mouton de Gruyter, Berlin/New York, pp. 127–145.

Clarke, S., 2004. Newfoundland English: phonology. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 1. Mouton de Gruyter, Berlin/New York, pp. 366–382.

Cysouw, M., 2003. Against implicational universals. Linguistic Typology 7, 89–101.

DeGraff, M., 1997. Verb syntax in, and beyond, creolization. In: Haegeman, L. (Ed.), The New Comparative Syntax. Longman, London, pp. 64–94.

Dunn, M., Terrill, A., Reesink, G., Foley, R.A., Levinson, S.C., 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309, 2072–2075.

Gil, D., 2003. English goes Asian: Number and (in)definiteness in the Singlish noun phrase. In: Plank, F. (Ed.), Noun Phrase Structure in the Languages of Europe. Mouton de Gruyter, Berlin/New York, pp. 467–514.

Goebl, H., 2004. Sprache, Sprecher, und Raum: Eine kurze Darstellung der Dialektometrie. Das Fallbeispiel Frankreich. Mitteilungen der Österreichischen Geographischen Gesellschaft 146, 247–286.

Gray, R.D., Atkinson, Q.D., 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426, 435–439.

Greenberg, J.H., 1963. The Languages of Africa. Indiana University, Bloomington.

Greenberg, J.H., 1966. Language Universals, With Special Reference to Feature Hierarchies. Mouton, The Hague.

Hagège, C., 1985. L'homme de paroles. Fayard, Paris.

Herrmann, T., 2003. Relative clauses in dialects of English: a typological approach. PhD thesis, University of Freiburg.

Hosali, P., 2004. Butler English: morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 1031–1044.

Keune, K., Ernestus, M., van Hout, R., Baayen, H.R., 2005. Variation in Dutch: From written MOGELIJK to spoken MOK. Corpus Linguistics and Linguistic Theory 1, 183–223.

Kim, J.-O., Mueller, C.W., 1978. Introduction to Factor Analysis: What It Is and How To Do It. In: Quantitative Applications in the Social Sciences, Sage Publications, Newbury Park, London, New Delhi.

Klein, W., Perdue, C., 1997. The basic variety (or: Couldn't natural languages be much simpler?). Second Language Research 13, 301–347.

Kortmann, B. (Ed.), 2004. Dialectology Meets Typology: Dialect Grammar from a Cross-Linguistic Perspective. Mouton de Gruyter, Berlin/New York.

Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), 2004. A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York.

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

Kruskal, J.B., Wish, M., 1978. Multidimensional Scaling. In: Quantitative Applications in the Social Sciences, Sage Publications, Newbury Park, London, New Delhi.

Labov, W., 1972. Language in the Inner City. University of Philadelphia Press, Philadelphia.

Mair, C., 2003. Kreolismen und verbales Identitätsmanagement im geschriebenen jamaikanischen Englisch. In: Vogel, E., Napp, A., Lutterer, W. (Eds.), Zwischen Ausgrenzung und Hybridisierung. Ergon, Würzburg, pp. 79–96.

Malcolm, I.G., 2004. Australian Creoles and Aboriginal English: morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 657–681.

McMahon, A., McMahon, R., 2006. Why linguists don't do dates: evidence from Indo-European and Australian languages. In: Forster, P., Renfrew, C. (Eds.), Phylogenetic methods and the prehistory of languages. McDonald Institute for Archaeological Research, Cambridge, pp. 153–160.

McWhorter, J., 1998. Identifying the creole prototype: vindicating a typological class. Language 74, 788–818.

McWhorter, J., 2001. The world's simplest grammars are creole grammars. Linguistic Typology 6, 125–166.

Melchers, G., 2004. English spoken in Orkney and Shetland: morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 34–46.

Mesthrie, R., 2004. Indian South African English: morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 974–992.

Meyerhoff, M., Niedzielski, N., 2003. The globalisation of vernacular variation. Journal of Sociolinguistics 7, 534–555.

Mufwene, S., 1990. Transfer and the substrate hypothesis in creolistics. Studies in Second Language Acquisition 12, 1–23.

Mühlhäusler, P., 2004. Norfolk Island-Pitcairn English (Pitkern Norfolk): morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 789–801.

Nerbonne, J., Heeringa, W., Kleiweg, P., 1999. Edit distance and dialect proximity. In: Sankoff, D., Kruskal, J. (Eds.), Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. CSLI Press, Stanford, pp. v–xv.

Nerbonne, J., Kleiweg, P., 2007. Toward a dialectological yardstick. Journal of Quantitative Linguistics 14, 148–167.

Nerbonne, J., Kleiweg, P., Manni, F., Heeringa, P., 2007. Projecting dialect differences to geography: bootstrap clustering vs. noisy clustering. In: Preisach, C., Schmidt-Thieme, L., Burkhardt, H., Decker, R. (Eds.), Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society. Springer, Berlin.

Poplack, S. (Ed.), 2000. The English history of African American English. Blackwell, Oxford.

Reaser, J., Torbert, B., 2004. Bahamian English: morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 391–406.

Ringe, D., Warnow, T., Taylor, A., 2002. Indo-European and computational cladistics. Transactions of the Philological Society 100, 59–129.

Sand, A., 2004. Shared morpho-syntactic features of contact varieties: article use. World Englishes 23, 281–298.

Sand, A., 2005. The effects of language contact on the morpho-syntax of English. In: Moessner, L. (Ed.), Anglistentag 2004 Aachen – Proceedings. Wisenschaftlicher Verlag, Trier.

Schneider, E., 2004. Global synopsis: phonetic and phonological variation in English world-wide. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 1. Mouton de Gruyter, Berlin/New York, pp. 1111–1137.

Schuchardt, H., 1979. The Ethnography of Variation: Selected Writings on Pidgins and Creoles. Karoma, Ann Arbor, MI.

Séguy, J., 1971. La relation entre la distance spatiale et la distance lexicale. Revue de Linguistique Romane 35, 335–357.

Shackleton, R.G.J., 2005. English-American Speech Relationships: a quantitative approach. Journal of English Linguistics 33, 99–160.

Simo Bobda, A., 2000. Research on New Englishes: a critical review of some findings with a focus on Cameroon. Arbeiten aus Anglistik und Amerikanistik 25, 53–70.

Spruit, M.R., 2005. Classifying dutch dialects using a syntactic measure: the perceptual daan and blok dialect map revisited. Linguistics in the Netherlands 22, 179–190.

Szmrecsanyi, B., Kortmann, B., 2009. Between simplification and complexification: non-standard varieties of English around the world. In: Sampson, G., Gil, D., Trudgill, P. (Eds.), Language Complexity as an Evolving Variable. Oxford University Press, Oxford.

Vincent, N., 1997. Synthetic and analytic structures. In: Maiden, M. (Ed.), The Dialects of Italy. Routledge, London, pp. 99–105.

Wee, L., 2004. Singapore English: morphology and syntax. In: Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, vol. 2. Mouton de Gruyter, Berlin/New York, pp. 1058–1072.

Young, F., Harris, D., 1992. Multidimensional scaling. In: Noursis, M. (Ed.), SPSS for Windows: Professional Statistics. SPSS Inc., Chicago, pp. 155–222.