Parameters of morphosyntactic variation in World Englishes: prospects and limitations of searching for universals

Bernd Kortmann and Benedikt Szmrecsanyi
Freiburg Institute for Advanced Studies (FRIAS)

1. Introduction

Anyone interested in linguistic typology, on the one hand, and the partnership between cross-linguistic and language-internal variation (or: macro- and microparametric variation), on the other hand, spontaneously sympathizes with the idea of searching for universals. Thus the idea of searching for universals in vernaculars, of English and other languages, as suggested by Jack Chambers in various papers (2000, 2001, 2003, 2004), also has something immediately appealing about it. However, in Filppula, Klemola, and Paulasto (to appear 2009) and in some papers in the present volume, Chambers' notion of *vernacular universals* has given rise to a controversy, with often quite critical evaluations of the concept pointing to fundamental problems or even apparent flaws. It is against the backdrop of this critical discourse that the present paper needs to be seen. Its aims are twofold: it tries to (a) contribute to an objectification of vernacular universals, thus somewhat relativising the predominantly critical opinions which have been voiced concerning this notion in the recent literature and giving them their appropriate place, and (b) point out the advantages and value this notion may have for dialectological and variationist studies and, more broadly, for language typology. The focus in doing so will solely be the morphosyntax of non-standard varieties of English around the world (L1, L2, English-based Pidgins and Creoles).

In section 2 some pros and cons of Chambers' notion will be weighed against each other, both in the light of rather general considerations and of recent and ongoing empirical research by ourselves. The ultimate idea this section will lead up to is that there are two lines of understanding vernacular universals, which as a result inform two strands of research into large-scale morphosyntactic patterns across a large number of varieties of

English, and that these, in turn, may help to give Chambers' notion some more credit than it has received in recent accounts.

While section 2 is largely concerned with individual or sets of two or three morphosyntactic features (i.e. the outcome of what we suggest to call *feature-based approaches*), sections 3 and 4 will present some results from our current research on large-scale cross-varietal patterns in entire coding strategies of varieties of English around the world (i.e. *strategy-based approaches*). Our overall argument in this paper will be that the notion of vernacular universals will be most useful (and least controversial) if we do not explore the alleged universality of individual features (in the sense of absolute universals), but rather the universality of "conspiracies of morphosyntactic features" and strategies for coding certain types of grammatical information which can be identified for individual types of varieties in any language (e.g. L1 as opposed to L2 varieties, low-contact vs. high-contact varieties, spontaneous spoken vs. written varieties).

In section 5, finally, we will sketch how the approach we are suggesting for determining different degrees of syntheticity and analyticity in the coding of grammatical categories among (spoken and written) varieties of English also holds potential for language typology. The systematic difference in preferred grammar coding strategies (synthetic vs. analytic) which can be identified for spontaneous spoken and written varieties in English respectively also seems to hold, as pilot studies suggest, across languages. This again would argue in favour of looking at entire coding strategies and large feature bundles rather than isolated morphosyntactic features when trying to determine candidates for structural properties which are universal in spontaneous spoken (including vernacular) language.

2. Weighing the pros and cons of vernacular universals

Chambers' well-known claim concerning the socio-dialectological notion of vernacular universals (or: vernacular roots) is that these universals comprise "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" (2004: 128). He goes on to argue that the putative ubiquity of such features, not just in varieties of English but in vernaculars of all languages, is unlikely to be due to sociolinguistic diffusion. Rather, they must be "primitive features of vernacular dialects" (Chambers 2003: 243), that is, unlearned and thus innate. Chambers (2004: 129) lists the following four morphosyntactic candidates for vernacular universals:

1. conjugation regularization, or levelling of irregular verb forms: e.g. *John seen the eclipse*, *Mary heared the good news*;
2. default singulars, or subject-verb nonconcord: e.g. *They was the last ones*;
3. multiple negation, or negative concord: e.g. *I don't/ain't know nothing*;
4. copula absence, or copula deletion: e.g. *She smart*, *We going as soon as possible*.

Below we will address some pros and, especially, cons that have been voiced in the recent literature, and we will seek to put these in perspective. Criticism has been levelled at Chambers' notion of vernacular universals primarily from two angles: language typology and socio-/variationist linguistics. Let us first examine the problems which typologists (may) have with this notion.

2.1. Some general considerations

From a typological angle there are at least five points worth making:

(i) The fundamental problem typologists have with the notion of vernacular universals is that a special status is claimed for universals of vernaculars (as opposed to universals of non-vernacular varieties, especially written standard varieties of languages). Rather, the candidates for vernacular universals should form a proper subset of those universals (be they

statistical or not, implicational or not) which have been identified in decades of solid, methodologically increasingly refined typological research exploring large-scale cross-linguistic variation of individual parameters of variation. Universals proper, the argument basically runs, apply to all languages and all their varieties equally, thus there is no need for postulating an extra-set of vernacular universals.

In our view, there is a lot to be said in favour of this line of argumentation. At the same time, however, it needs to be acknowledged that typological studies of the last 40 years have not really covered all the linguistic variation existing on the globe. Especially for languages with a long literary tradition (such as the synchronically and historically extensively well-described major European languages), there has been a bias towards the written standard varieties. Spontaneous spoken varieties have largely, until recently almost entirely, been left out of consideration.[1] Judged on a global scale, this leads to the following major methodological apples-and-oranges problem in language typology, which as yet has hardly been acknowledged as such by its practitioners. For the majority of languages in the world, especially for those lacking a literary tradition or even a codified written representation, spoken language data form the basis for typological descriptions and generalizations, while for most European languages written data taken from the standard varieties form the object of typological study. So linguistic typology would definitely benefit from the systematic inclusion of vernaculars, at least for languages whose written standards have exclusively been taken to represent them in, for example, such magnificent typological undertakings as the *World Atlas of Language Structures* (*WALS*; Haspelmath et al. 2005). This would offer the opportunity to make the methodology of typology and the resulting typological claims (e.g. when formulating generalizations in the form of universals) more watertight. As has variously been argued and shown for English, for example, Standard British English is the "odd man out" in several domains of morphosyntax (e.g. negation, agreement, relativisation, reflexives) compared with the vast majority of (standard and) non-standard varieties of English (cf. e.g. Kortmann et al. 2005). Thus taking Standard British English as "the" representative of English in typological studies may crucially distort the picture. For relative clauses, for instance, the most frequent strategy in non-standard varieties of English spoken in the British Isles is the relative particle strategy (notably *that* and *what* as in *The man that/what came in*) and not the relative pronoun strategy, as claimed in the *WALS*. For putting the typological picture right it is especially non-standard features with a wide areal and/or social reach which will be most useful (cf.

Auer 2004), i.e. exactly those which qualify as candidates for Chambers' vernacular universals. Moreover, such candidates for vernacular universalhood may well turn out to be veritable candidates for universals proper (like the prototypical Greenbergian universal, i.e. (statistical) implicational universals).

(ii) Typologists criticizing the notion of vernacular universals should also acknowledge the following: (a) So far, it has been mainstream typological practice not to include pidgins and creoles in their samples. Only now is there a research initiative headed by the Max-Planck Institute of Evolutionary Typology at Leipzig (Germany) which is working on a World Atlas of Pidgin and Creole Structures, following the model of the Leipzig-made *WALS*, and which ultimately pursues the question whether pidgins and creoles represent a language type of its own; (b) more generally, with its strict focus on the structure of genetically, areally, and historically unrelated languages, linguistic typology has largely left out of account the role of language contact in the formation of individual languages, and even more so the role of socially driven diffusion of individual language structures and linguistic types across languages (cf. the illuminating paper by Bisang 2004). By contrast, anyone interested in putting to test and possibly giving (more) substance to the notion of vernacular universals is bound *not* to neglect aspects such as those in (a) and (b).

(iii) Typologists discussing and trying to evaluate Chambers' notion in an open-minded and objective manner should not restrict vernacular universals to absolute universals, but should, for example, also inquire into the possibility of identifying implicational universals (or rather implicational tendencies). No typologist seriously expects to find, across languages, absolute universals in grammar beyond that mere handful mentioned as likely candidates in textbooks. So why should the fact that a given candidate for a vernacular universal does not occur in 100% of all vernaculars in a given language, let alone across languages, be turned against the notion of a vernacular universal? Even if this may be a problem for Chambers himself, since – as someone thinking of universals in a formalist sense, i.e. as part of UG, or outgrowths of the bioprogram – he formulates his criteria for vernacular universals in a most sweeping way, for a typologist it would be amazing already if a given feature can be observed in, say, 70% or more of the languages of the world.

From the point of view of the typologist investigating the world-wide spread of morphosyntactic features in varieties of English (in our

study, 76 morphosyntactic features in 46 varieties of English; for details, see section 3), it therefore (a) does not come as a surprise that not a single feature is found in all varieties, and (b) it is rather astonishing to learn that the five most widely found features of English morphosyntax worldwide occur in 80–89% of all varieties of English. By contrast, all four of Chambers' candidates for vernacular universals (or rather *angloversals*) are documented in fewer varieties (70–78%). Even multiple negation is found in no more 76% of the 46 varieties. The lesson, then, to be learnt from these findings is that we should interpret Chambers' vernacular universals as having a very wide (areal and/or social) reach in individual languages, and possibly across languages, but not as absolute universals. 100% scores are possible only if we consider subsets of non-standard varieties. Thus it is possible to identify morphosyntactic features, including all of Chambers' four top candidates for vernacular universals, which are part of all non-standard varieties of North America (US and Canada; something we have elsewhere suggested to call *areoversals*; cf. Szmrecsanyi and Kortmann to appear 2009b), or in the vast majority of individual variety types (e.g. in L2 varieties of English, so-called *varioversals*[2]). We will return to the importance of variety types repeatedly from here on.

(iv) Implicational universals are the prototypical type of universals in functional typology. So when putting to test the notion of vernacular universals, it should be explored for them, too, whether biconditional implications (alternatively: *equivalences*) or one-way implications (also known as *preferences*) can be identified. On the basis of the 46 non-standard varieties of English we investigated, such dependencies can clearly be confirmed. Here are a few examples (for details cf. Szmrecsanyi and Kortmann to appear 2009b,c). With respect to *biconditional implications*, the bulk of varieties of English have both *ain't* as the negated form of *be* and *ain't* as the negated form of *have*, or they have none of these options. In a similar vein, non-coordinated subject pronoun forms in object function (*Did you get **he** out of bed?*) and non-coordinated object pronoun forms in subject function (***Us** say 'er's dry*) tend to go together, as is the case for habitual *do* (*He does catch fish every day*) and *do* as an unstressed tense and aspect marker (*The man what did say that…*). Likewise, if a variety exhibits the relative particle *at*, it will have the relative particle *as*, and vice versa. In short, pairings like the above are best seen as feature bundles instead of pairs of independent features. Such biconditional implications also depend on the variety type: so, for instance, among L1 varieties we find a correlation such that when in a variety past forms of

irregular verbs can replace participle forms, unmarked forms are also possible, and vice versa. Both among L2 varieties and among English-based pidgins and creoles (but not in L1 varieties), we observe that if a variety has one use of *ain't* (either for *be*, *have*, or as a generic negator), it will also have the other two uses of the form.

As regards *one-way implications*, any variety in our sample that attests *would* in *if*–clauses also displays loosening of the sequence of tense rule, but not necessarily vice versa. Moreover, a variety will not have the relative particle *as* (*There's the man **as** kicked me in the face*) unless that variety also has (i) the relative particle *what* (*There's the man **what** kicked me in the face*) and (ii) gapping or zero-relativisation in subject position (*There's the man ____ kicked me in the face*). And once again, distinguishing between variety types is instructive, as the following two examples show: an L1 variety, unlike L2 and Pidgin/Creole varieties, needs habitual *do* in its inventory in order to attest *do* as a tense and aspect marker, and in L2 varieties, the possibility of regularized reflexives-paradigms appears to necessitate generic *he/his* for all genders (as in *The car, he's broken*).

(v) We saw in (iii) that Chambers' candidates for vernacular universals do not even stand up to the status of vernacular angloversals, i.e. features found in 100% of all non-standard varieties of English. But even if we did find absolute, unrestricted angloversals, there is much room for doubt, to put it mildly, that counterparts of these vernacular universals for English were to be found in vernaculars in other languages, too, as Chambers (2004: 129) assumes:

> I have listed the vernacular universals with their English names and illustrated them with English examples. This is misleading. In so far as these processes arise naturally in pidgins, child language, vernaculars, and elsewhere, they are primitive features, not learned. As such they belong to the language faculty, the innate set of rules and representations that are the natural inheritance of every human being. They cannot be merely English. They must have counterparts in the other languages of the world that are demonstrably the outgrowths of the rules and representations in the bioprogram.

Considerable room for doubt there is at least as long as in searching for vernacular universals the focus is on individual, highly specific features

like those given by Chambers as promising candidates at the beginning of section 2. Take, for example, copula absence, as in *She___ smart* or *He __ a sailor*. Cross-linguistically, it turns out that zero copulas for such predicate nominals are impossible in the majority of languages (211 vs. 175 where it is possible) in the sample investigated by Stassen in *WALS* Map 120 (see online version). If zero copulas were part of the language faculty, there should be more languages that have them. Since this is evidently not the case, zero copulas cannot claim the status of "promising candidate for a vernacular universal". Something similar can be observed for subject-verb nonconcord (or default singulars). In order to exhibit subject-verb nonconcord in any non-vacuous way, a language needs structural means to display subject-verb concord (at least historically). Yet many languages – for instance, highly isolating languages such as Vietnamese – do not show agreement at all; so vernacular Vietnamese showing nonconcord is an utterly unremarkable fact. Along similar lines, the other of Chambers' candidates for vernacular universals, too, could rather easily be attacked from a typological angle (with the possible exception of multiple negation).

In short, what we mean to suggest here is that a feature like subject-verb nonconcord is rather typical of languages which, like English, have some inflectional morphology but are in the process of getting rid of what remains. But what is happening in non-standard varieties of English and, possibly, languages belonging to the same morphological type as English, does not necessarily apply to vernaculars of inflectional or agglutinating languages such as Finnish, Hungarian (which have a great deal of grammatical agreement), or Turkish. A feature such as subject-verb nonconcord may be better referred to as a vernacular *typoversal* – a feature, in other words, which is typical of vernaculars of inflecting languages. It is unlikely that the 'language faculty' would provide for special rules and representations applying to English-like vernaculars; zero copulas and default singulars are just too conditioned on the linguistic type of English to be cross-linguistically "universal". At the same time, notice that it is not only loss of agreement or loss of redundancy that we can observe in vernaculars. Individual vernaculars have, and can indeed be shown to currently develop, a more elaborate inflectional morphology or, for example, agreement system than the standard variety has (e.g. de Vogelaer et al. 2002, Wagner 2004, or Haser/Kortmann 2008). Given these observations, we have argued (Szmrecsanyi and Kortmann to appear 2009b) that a candidate feature for a vernacular universal should, at a minimum, fulfil the following criteria:

- The candidate feature should be attested in a vast majority of a given language's vernacular varieties.
- The candidate feature should not be patterned geographically or according to variety type (in the case of English: L1 (high- vs. low-contact), L2, or pidgin/creole).
- For the sake of cross-linguistic validity, the candidate feature should not be tied to a given language's typological make-up (inflectional, agglutinating, etc.).
- The candidate feature should be cross-linguistically attested in a significant number of the world's languages (especially among the vast number of languages without a literary tradition).

2.2. The sociolinguistic and variationist perspective

It may be recalled that Chambers, in his account of vernacular universals, draws a major divide between Standard English(es), on the one hand, and non-standard varieties of English, ranging from traditional dialects to pidgins and creoles, on the other hand. It is exactly this divide, along with the notion of vernacular universals in general, which Trudgill takes issue with in his contribution ("Vernacular universals and the sociolinguistic typology of English dialects") to the vernacular universals debate in the volume by Filppula/Klemola/Paulasto (to appear 2009). According to Trudgill, not enough vernacular universals have been found to make the concept fruitful and, more fundamentally, the "true typological split" among varieties of English lies elsewhere, not between vernacular and non-vernacular varieties. The major split – according to him -- rather lies between high-contact and low-contact varieties of English. The former include (a) koineised non-standard urban varieties in the British Isles and colonial varieties of North America, Australasia and South Africa; (b) indigenized non-native L2 varieties like Indian English or Nigerian English; (c) shift varieties like Irish English and Welsh English; (d) English-based pidgins and creoles (as extreme cases resulting from language contact); and, notably, (e) Standard English(es). As low-contact varieties Trudgill identifies the traditional dialects of English, located largely in the British Isles, but also including Appalachian English or Newfoundland English. For the purposes of the present paper, the crucial point of Trudgill's account is this: he argues that the grammars of high-contact varieties are characterized by processes of simplification as

opposed to processes of complexification to be observed in the grammars of low-contact varieties.

Three points follow from Trudgill's "true typological split" among varieties of English and the way in which he motivates it. First, he attributes the central role in his typology to language contact (and along with it adult language acquisition), highlighting its impact on the structural properties of languages and their varieties (for similar views cf., for example, the papers by Siemund and Winford in Filppula, Klemola and Paulasto, to appear 2009). Second, implicitly, there is an even more fundamental split underlying Trudgill's division of varieties of English (and essentially, given a similar socio-cultural background, the varieties of any language) into high- and low-contact varieties, namely the one between spontaneous spoken and written varieties – a point always worth remembering in dialectology, variationist studies, and especially in language typology. Third, Trudgill doesn't base his claim on individual morphosyntactic features but on what we prefer to call overall coding strategies (e.g. inflectional coding of grammatical information as a complex(ifying) strategy and analytic, or even zero, marking as simplifying strategies).

Trudgill's hypothesis concerning the significance of variety types for such overall grammar coding strategies as simplicity/simplification and complexity/complexification will be tested – and ultimately confirmed – in section 4. Systematic differences in the grammar coding strategies used in spontaneous spoken as opposed to written registers in English (and other languages) will be tracked in section 5. Before, however, it will be demonstrated in section 3 that even on a more general level, i.e. when not specifically exploring different degrees of morphosyntactic complexity and simplicity, the morphosyntax of non-standard varieties of English around the world clearly patterns according to variety type. Moreover, it will be shown that the clustering of varieties according to variety type has more explanatory power than geography.

3. Strategy-based approaches I: the significance of variety types

In this section, we will briefly demonstrate that in accounting for shared morphosyntactic features and, indeed, feature bundles and entire coding strategies, it is the *variety type* (L1 vs indigenized non-native L2 vs. pidgins and creoles) which is of towering importance. The following is a brief sketch of previous research (cf. Szmrecsanyi to appear 2009, Szmrecsanyi

and Kortmann to appear 2009b,c) where we used Principal Component Analysis to see the wood for the trees in the survey data we had collected in Kortmann and Szmrecsanyi (2004). Let us first review the nature of the survey data and the composition of our sample of varieties.

The source for our survey data, i.e. the classic data type in typological and dialectological research, is what we have informally come to call *The World Atlas of Morphosyntactic Variation in English*, i.e. the survey of morphosyntactic features underlying the interactive maps on the CD-ROM accompanying the *Handbook of Varieties of English* (Kortmann and Schneider 2004) and subjected to a first close examination in Kortmann and Szmrecsanyi (2004). For this survey, material was collected from (often native-speaker) experts on 76 non-standard morphosyntactic features from 46 (exclusively spoken) non-standard varieties of English around the world (for details of the survey procedure cf. Kortmann and Szmrecsanyi 2004: 1142–1145). The features in the survey are numbered from 1 to 76 and cover 11 broad areas of morphosyntax: pronouns, the noun phrase, tense and aspect, modal verbs, verb morphology, adverbs, negation, agreement, relativisation, complementation, discourse organization and word order. This, for example, is the complete set of agreement features in the survey (including the feature numbering):

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,* or zero for *there*)
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*)

The 46 varieties are taken from all seven anglophone world regions (the British Isles, America, Caribbean, Australia, Pacific, South/Southeast Asia, Africa). Table 1 provides a breakdown of the 46 varieties by variety type (20 L1 varieties, 11 L2 varieties, 15 English-based pidgins and creoles),

which for the L1 varieties includes Trudgill's split between high-contact L1 varieties (12 out of 20) and low-contact varieties (8 out of 20; more on the high- vs. low-contact distinction in section 4.1):

Table 1: Varieties sampled in the *World Atlas* (Kortmann and Schneider 2004)

varieties	*variety type*
Orkney and Shetland, North, Southwest and Southeast of England, East Anglia, Isolated Southeast US E, Newfoundland E, Appalachian E	traditional L1
Scottish E, Irish E, Welsh E, Colloquial American E, Ozarks E, Urban African-American Vernacular E, Earlier African-American Vernacular E, Colloquial Australian E, Australian Vernacular E, Norfolk, regional New Zealand E, White South African E	high-contact L1
Chicano E, Fiji E, Standard Ghanaian E, Cameroon E, East African E, Indian South African E, Black South African E, Butler E, Pakistan E, Singapore E, Malaysian E	L2
Gullah, Suriname Creoles, Belizean Creole, Tobagonian/Trinidadian Creole, Bahamian E, Jamaican Creole, Bislama, Solomon Islands Pidgin, Tok Pisin, Hawaiian Creole, Aboriginal E, Australian Creoles, Ghanaian Pidgin E, Nigerian Pidgin E, Cameroon Pidgin E	P/C

When applying Principal Component Analysis (PCA) to this data set in order to explore the distribution of the 76 morphosyntactic features across the 46 varieties of English, we find that co-presence and co-absence patterns point to two highly explanatory dimensions (accounting for 38.4% of the total variance, which is quite remarkable). As can be seen in Figure 1, our pre-established 46 varieties cluster very nicely according to whether they are L1 varieties (represented by squares), L2 varieties (represented by triangles), or English-based pidgins and creoles (represented by circles) – and indeed better than geographically. Thus variety type turns out to be the

better predictor of overall similarity or distance between individual varieties than the world region where they are spoken.

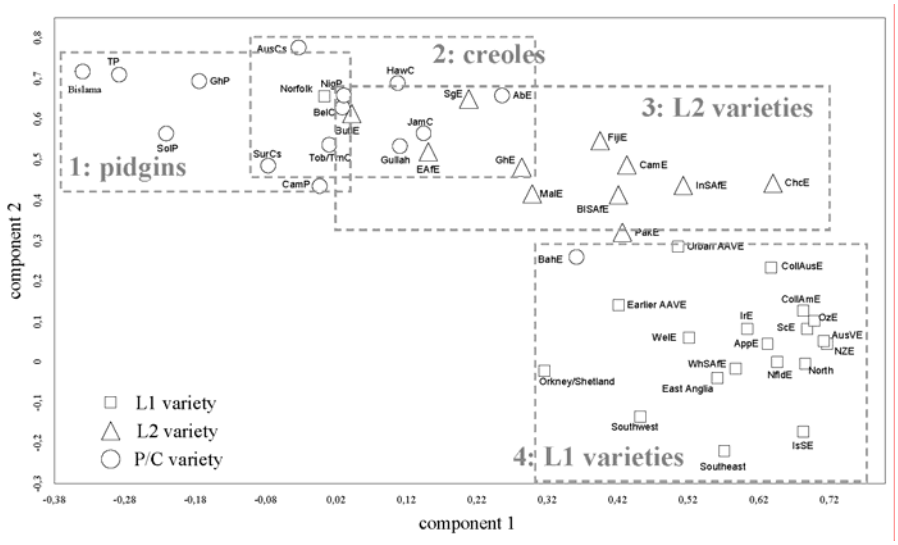

Figure 1: Visualization of principal components of variance in the 76 × 46 database. Dotted boxes indicate group memberships (cf. Szmrecsanyi and Kortmann to appear 2009b,c).

This significant result of our global study on English leads us to hypothesize that the importance of variety type in accounting for similarities and differences in the morphosyntax of non-standard varieties will be found confirmed when exploring morphosyntactic similarities and differences and, inspired or challenged by Chambers, looking for universal features in vernaculars of other languages, too. Moreover, we hold that the relevant feature bundles for each of the variety types (i.e. at its crudest: L1 vs. L2 vs. pidgins/creoles), regardless of what the individual features look like for a given language and its non-standard varieties, ultimately all "conspire" and jointly instantiate an overall coding strategy which is constitutive of the variety type at hand. These overall coding strategies can be seen to be instantiated by the two dimensions (components 1 and 2) that PCA yields in Figure 1. How, by the way, are these dimensions to be interpreted? Encouraged by an independent study (cf. Kortmann and Szmrecsanyi to appear 2009) in which we explore different degrees of

morphosyntactic complexity in non-standard varieties of English on the basis of exactly those survey data used for the PCA above, we are inclined to interpret component 1 as increasing degrees of L2-acquisition difficulty and component 2 as increasing degrees of transparency (i.e. regularity for synthetic markers of grammatical information).

Notice, now, that in a different large-scale cross-varietal study (Szmrecsanyi and Kortmann to appear 2009a), we explore two other global coding strategies (namely, analytic vs. synthetic coding of grammatical information) on the basis of naturalistic corpus data for a smaller set of spontaneous spoken varieties (all L1 or L2) – and again, variety type turns out to be the best predictor for a given variety's morphosyntactic profile. This particular study will be summarized in the following section.

4. Strategy-based approaches II: analyticity vs. syntheticity in varieties of English around the world

Our findings in this section are based on the dataset and method utilized in Szmrecsanyi and Kortmann (2009a). A total of three frequency-based metrics will be introduced here, which are applied to 15 varieties of English on the basis of naturalistic corpus data. The bulk of our corpus data stems from two major digitized speech corpora (the *Freiburg Corpus of English Dialects* (FRED; cf. Hernandez 2006, Anderwald and Wagner 2007, Kortmann and Wagner 2005) and the *International Corpus of English* (ICE; cf. Greenbaum 1996). From these two corpora we sampled 12 spoken varieties: two high-contact varieties (Welsh English and New Zealand English), five British low-contact varieties (Southeast England + East Anglia, Southwest England, English Midlands, English North, Scottish Highlands) and five L2 varieties (Hong Kong English, Philippine English, Singapore English, Indian English, Jamaican English). In addition, we used data from the *Northern Ireland Transcribed Corpus of Speech* (NITCS) to represent Northern Irish English, another high-contact variety. Purely for benchmarking purposes, we also included data from two high-contact *standard* varieties of British English (from the ICE-GB) and American English (from the *Corpus of Spoken American English*). Note that pidgins and creoles are not represented in these samples. Before turning to the frequency-based metrics, it is necessary to make clear (a) on what basis we have classified a given variety of English as belonging to one of three variety types high-contact L1, low-contact L1, and L2 (section 4.1), (b) what it is understood by *analyticity* and *syntheticity* in our study (section

4.2), and (c) how we calculate degrees of analyticity and syntheticity in our quantitative analysis (section 4.3).

4.1 Classification of varieties

Our classification of varieties (in this sample and also in Table 1 above) into L2 varieties and, much more controversial, high- vs. low-contact L1 varieties rests on the following assumptions (on this issue, cf. also Szmrecsanyi to appear 2009). L2 varieties of English are non-native, indigenized varieties that do not have significant numbers of native speakers but that nonetheless have prestige and important normative status in certain political communities (e.g. Indian English or Jamaican English). As for L1 varieties, the distinction between high- and low-contact varieties goes back to Peter Trudgill (to appear 2009), who does not, however, provide hard-and-fast criteria for classifying a given L1 variety of English as either high- or low-contact. As outlined in section 2.2, Trudgill's idea, in a nutshell, is that contact implicates adult language learning, which in turn implicates simplification of grammars, especially by reduction of inflectional morphology (for a recent similar view, cf. Wunderlich 2008: 252). The resulting simplicity in different domains (see Trudgill to appear 2009 for an overview) is what sets apart high-contact from low-contact varieties as synchronic groups. In this spirit, we operationally define *high-contact L1 varieties* as varieties that fall into one of the following categories:[3]

Transplanted L1 Englishes or *colonial (standard) varieties* (cf. Mesthrie 2006: 382), i.e. varieties whose genesis is such that thanks to settlement colonization in the course of the past 400 years, settlers with diverse linguistic and/or dialectal backgrounds – with all the dialect and language contact that this implicates – formed new indigenized English dialects that have had native speakers from early on. Examples include New Zealand English and Australian English.

Language-shift Englishes, i.e. varieties "that develop when English replaces the erstwhile primary language(s) of a community" and that have "adult and child L1 and L2 speakers forming one speech community" (Mesthrie 2006: 383). We also include in this department what we call *shifted varieties*, i.e. varieties that used to

be genuine language-shift varieties within the past 400 years but which do not now have significant numbers of L2 speakers any more. A prime example is Irish English (cf. Mesthrie 2006: 383).

Standard varieties, such as Standard British English, the genesis of which, according to Trudgill (to appear 2009), always implicates a high degree of dialect contact.

Hence, high-contact L1 varieties, in our diction, are essentially 'New Englishes' (cf. Pride 1982; Platt *et al.* 1984) *minus* non-native, indigenized L2 varieties *minus* English-based pidgins and creoles *plus* standard varieties. Varieties that do not fall into one of the above categories will be considered *low-contact L1 dialects of English*, i.e. traditional non-transplanted regional dialects which are "long-established mother tongue varieties" (Trudgill to appear 2009).

This classification scheme gives rise to the following categorization of the 15 varieties studied in this part of the paper:

- *Non-native, indigenized ESL varieties*: Indian E; Jamaican E; Hong Kong E; Philippines E; and Singapore E.

- *High-contact L1 varieties of English*: New Zealand E (by virtue of being a transplanted English and a standard variety); Standard (colloquial) BrE + Standard (colloquial) AmE (by virtue of being standard varieties); Scottish Highlands E (by virtue of being a shifted or even shift variety, with English having been introduced not before the 18th century; cf. Mesthrie 2006: 388); and Welsh E (again, by virtue of being a shift variety, with English not really having spread before the 19th or even 20th century).

- *Low-contact L1 dialects of English*: the traditional dialects spoken in the Southwest of England; the traditional dialects spoken in the Southeast of England; the traditional dialects spoken in the English Midlands; and the traditional dialects spoken in the North of England.

4.2 Defining analyticity and syntheticity

The terms *analytic* and *synthetic* have a long history (cf. Schwegler 1990, chapter 1 for an excellent overview of the rich history of thought in this area). Unfortunately, both terms also "are used in widely different meanings by different linguists" (Anttila 1989: 315), which is why a concise definition is necessary at this point. Following Szmrecsanyi and Kortmann (to appear 2009a) and Szmrecsanyi (to appear 2009), we are interested, first, in the coding of grammatical information only, which is why lexical analyticity and syntheticity will not enter into consideration here. Second, our idea of grammatical analyticity and syntheticity is a strictly formal one (and not a semantic one), which roughly follows Danchev's notion that "[f]ormal analyticity [...] implies that the various meanings [...] of a given language unit are carried by [..] free morphemes, whereas formal syntheticity is […] characterized by the presence of [at least] one bound morpheme" (1992: 26). We thus operationally define grammatical analyticity and syntheticity as follows:

1. Formal grammatical *analyticity* comprises all those coding strategies where grammatical information is conveyed by free grammatical markers, which we in turn define in a fairly standard way as closed-class (or: function) word tokens that have no independent lexical meaning and thus belong to one (or more) of the following word classes: determiners (e.g. *who*), pronouns (e.g. *he*), prepositions (e.g. *in*), conjunctions (e.g. *and*), infinitive markers (e.g. *to*), so-called primary verbs (*be*, *have*, *do*), modal verbs (e.g. *can*), and negators (e.g. *not*).

2. Formal grammatical *syntheticity* comprises all those coding strategies where grammatical information is conveyed by bound grammatical markers, such as verbal, nominal, and adjectival inflectional affixes (e.g. past tense –ed, plural –s, *comparative* –er, and so on), the Saxon genitive (e.g. *Tom's house*) as a clitic, as well as allomorphies such as ablaut phenomena (e.g. past tense *sang*), i-mutation (e.g. plural *men*) and other non-regular yet clearly bound grammatical markers. Our model of morphological analysis is thus, at base, an item-and-process model (Hockett 1954: 396) where grammatically marked forms are derived from simple forms via some sort of process – in our diction, via adding some sort of overt grammatical, not necessarily segmentable

bound grammatical marker, be it a (regular) regular grammatical affix, a stem vowel change, and the like.

3. Derived from these two notions is a third one, which we choose to label *grammaticity* and which comprises the sum of the former two, i.e. all coding strategies where grammatical information is conveyed either by free or bound grammatical markers.

4.3 The frequency-based metrics

Inspired by Joseph Greenberg's (1960) paper, "A Quantitative Approach to the Morphological Typology of Language", we conducted a morphological/grammatical-functional analysis of random samples spanning 1,000 de-contextualized tokens (word forms) for every one of the geographic varieties sampled (see Szmrecsanyi to appear 2009 for more detail on the robustness of the method). For each token in the database, we established

- whether the token bears a bound grammatical marker (fusional or suffixing), as in *sing-s* or *sang*; and/or

- whether the token is a free grammatical marker, or a so-called function word, belonging to a closed grammatical class as defined in section 4.2.

On the basis of this analysis, we established three indices: *a syntheticity index* (measuring the text frequency of bound grammatical markers per 1,000 words of running text), an *analyticity index* (gauging the text frequency of free grammatical markers per 1,000 words of running text), and an overall *grammaticity index* (the sum of the former two indices). The syntheticity and analyticity indices have a lower bound of 0 and an upper bound of 1,000; the grammaticity index has a lower bound of 0 and an upper bound of 2,000. Thus a, syntheticity index value of 100 would indicate that a variety attests, on average, 100 bound grammatical markers per 1,000 words of running text; an analyticity index value of 500 would indicate that a variety attest, on average, 500 free grammatical markers (i.e. function words) per 1,000 words of running text; and a grammaticity index value of 600 would indicate that a variety attests, on average, 600 overt grammatical markers, bound or free, per 1,000 words of running text.

4.4 Results and discussion

The results of this exercise in index calculation can be summarized as follows. There is a strikingly consistent hierarchy that governs grammaticity levels: traditional L1-vernaculars > high-contact L1-vernaculars > L2-varieties. Hence, traditional L1-varieties exhibit the highest degree of grammaticity, L2-varieties the lowest degree, and high-contact L1-varieties cover the middle ground. This hierarchy dovetails nicely with claims, as pointed out in section 2.2 above, that a history of contact and adult language learning can eliminate certain types of redundancy, especially those found in grammatical marking (cf. for example McWhorter 2007, Siegel 2004 and 2008, Trudgill 2001 and, especially, Trudgill to appear 2009, Wunderlich 2008). L2-varieties in particular seem to follow this strategy most radically: our results suggest that L2-speakers do not generally opt for "simple" features instead of "complex" features, but rather appear to prefer zero marking over explicit marking, be it (presumably) L2-easy or complex.

The interplay between syntheticity and analyticity is visualized by way of a scatterplot in Figure 2. The dialects of the Southeast and East Anglia (in the top right corner) and Hong Kong English (in the bottom left corner) are the extreme cases in our dataset. The former are both highly analytic and above-average synthetic varieties, while Hong Kong English exhibits the lowest figures for either type of grammaticity. In general, the traditional L1 vernaculars are to be found in the upper right half of the diagram (i.e. they exhibit above-average grammaticity), whereas L2-varieties are located in the lower left half, being neither particularly analytic nor synthetic. High-contact L1-varieties cover the middle ground, along with the two standard varieties (colloquial American E and spoken British E) and, interestingly, the Southwest of England.

The dotted trend line merits particular attention in Figure 2: as a rather robust ($R2 = 0.40$) statistical generalization, this line indicates that on the inter-variety level, there is *no* trade-off between analyticity and syntheticity. Instead, analyticity and syntheticity correlate positively such that a variety that is comparatively analytic will also be comparatively synthetic, and vice versa. Once again, in terms of L2-varieties this is another way of saying that these tend to opt for a coding strategy of less overt marking, rather than trading off synthetic marking for analytic marking, which is purportedly L2-easy:

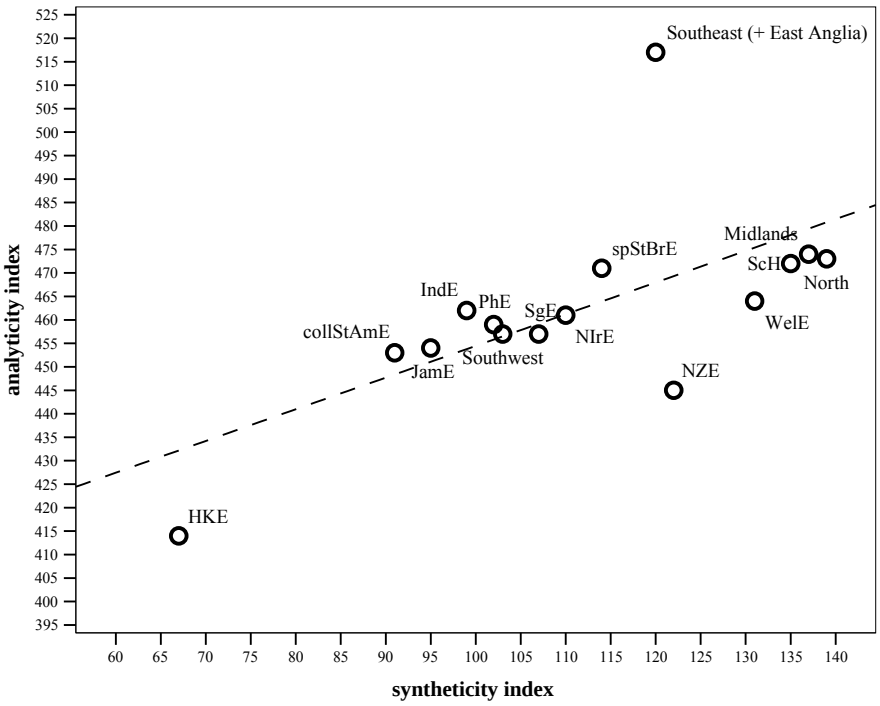

Figure 2: Analyticity by simplicity. Dotted trend line represents linear estimate of the relationship (cf. Szmrecsanyi and Kortmann to appear 2009a).

5. Strategy-based approaches III: analyticity, syntheticity, and the typological relevance of the distinction between spoken and written language

On the basis of the general method outlined in section 4, Szmrecsanyi (to appear 2009) surveys genre (or: register) variation in the *British National Corpus* with regard to analyticity and syntheticity and finds instructive differences between spoken and written genres. It turns out that, for one thing, spoken texts are significantly more analytic than written texts: the typical spoken text exhibits in excess of 50 more analytic markers per 1,000 words of running text than the typical written text. Secondly, written texts are significantly more synthetic than spoken texts, in that the former

exhibit, on average, in excess of approximately 30 more synthetic markers per 1,000 words of running text than the latter. Furthermore, spoken texts exhibit significantly more grammaticity than written texts, and as far as the scope of variability is concerned, variability among written texts is more sizable than among spoken texts. In sum, variability in the analyticity-syntheticity dimension is endemic in stylistic varieties of English.

The crucial link of this sort of register variation study in English to language typology is the following. We have conducted preliminary investigations – utilizing the methodology detailed above – into written and spoken registers of four European languages (English, Italian, German, and Russian), and the results (cf. Figure 3) show that spoken-written differences can be substantial, and indeed more substantial than in English. While the overall pattern of variability seems to be the same throughout (i.e. spoken registers are less synthetic and more analytic than written registers), Figure 3 makes amply clear that, intriguingly, intra-lingual register differences can be more pronounced than inter-lingual differences. For example, spoken Italian is more similar to spoken English than it is to written Italian.

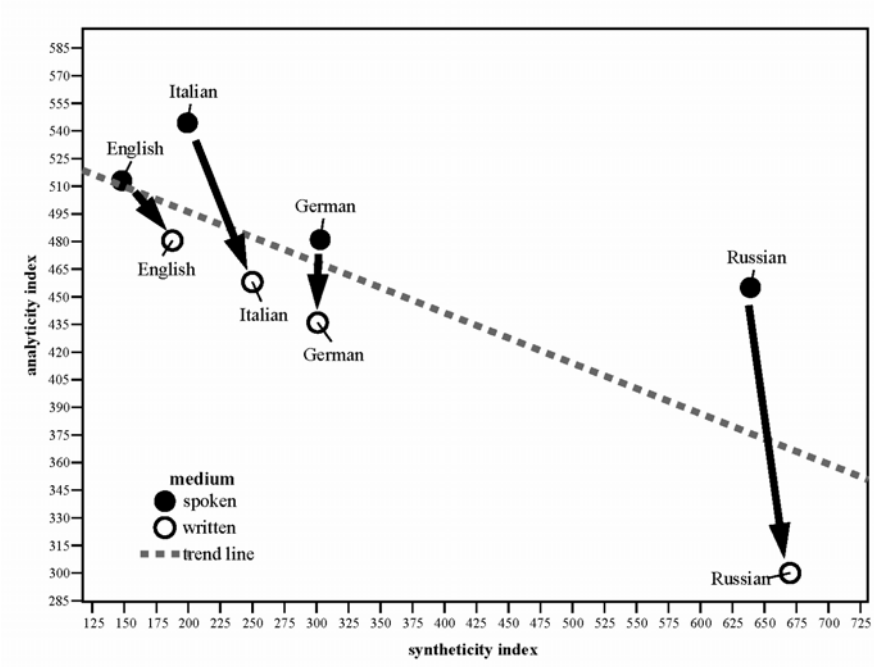

Figure 3: Analyticity vs. syntheticity in written and spoken standard varieties of four European languages

Against the backdrop of the overall argumentation in this paper, we take the findings of our small set of pilot studies represented in Figure 3 as support for our view that instead of looking for (typically rather more than less) specific individual features as candidates for vernacular universals, a strategy focussing on universal coding strategies dependent on variety type (in this case: spontaneous spoken vs. written varieties) holds considerably more promise.

6. Conclusions

In this paper we have sought to present a more balanced picture of Chambers' notion of vernacular universals. Limitations of his original, admittedly both rather crude and sweeping understanding of vernacular universalhood there are many, but ultimately, we believe, there are plenty of prospects as well. Chambers' original notion may have been found

unconvincing in numerous respects. Yet at the same time, it has undoubtedly proved to be a useful concept in triggering a fascinating controversy and promising line of research in dialectology, contact linguistics, and the study of morphosyntactic variation in English, which will undoubtedly be continued, possibly even finding its way into language typology. We hope that we have shown along which lines the notion of vernacular universals needs to be downsized, as it were, adjusted and refined in order for it to be turned into an even more fruitful concept that, ultimately, will help us (dialectologists, creolists, variationists, second language acquisition specialists, language historians, and typologists) to truly see the wood for the trees – which is exactly, in our view, what Chambers had in mind. Ways of making Chambers' original notion a less controversial and more fruitful concept include the following: first, taking seriously the constraints we formulated at the end of section 2.1; second, operationalizing it in ways as presented in sections 3-5; third, admitting standard ways of formulating universals in language typology to the search and formulation of vernacular universals, notably by looking not just for absolute, but also statistical (non-implicational as well as implicational) universals; fourth, keeping in mind the broad picture, i.e. looking for universal coding strategies which – lastly – are distinctive of individual variety types, regardless in which language we encounter them.

In Freiburg, we will continue to pursue this line of research by significantly broadening the feature catalogue to be investigated, widening the range of varieties (starting with naturalistic corpus data for pidgins and creoles, but also (adult) learner varieties of English and other languages) and the range of languages such that spontaneous spoken and written varieties can be systematically compared with regard to overall coding strategies in grammar. This research agenda is likely to contribute to a better understanding of those basic principles which underlie the structural variation we can observe within and across languages, adding a crucial new and truly integrative facet to existing research into micro- and macroparametric variation.

Notes

[1] For example, there is no mention of non-standard varieties in the special issue "Whither Linguistic Typology – *an und für sich* and in relation to other types of linguistic pursuits?" of the journal *Linguistic Typology* 11.1 (2007: 1-306), which celebrates the tenth anniversary of the journal.

[2] Compare Kortmann and Szmrecsanyi (2004) for an extended empirical discussion of unrestricted angloversals, i.e. the most frequent features in varieties of English across the board (1153–1160), unrestricted areoversals (1160–1184), and unrestricted varioversals (1184–1194).

[3] Note however that these categories are not mutually exclusive. For instance, New Zealand English is a transplanted L1 variety, yet it also serves as a standard variety.

References

Anderwald, Lieselotte and Susanne Wagner (2007). The Freiburg English Dialect Corpus (FRED): Applying corpus-linguistic research tools to the analysis of dialect data. In: Joan Beal, Karen Corrigan and Hermann Moisl (eds.), *Using Unconventional Digital Language Corpora.* Vol. I: *Synchronic Corpora*, 35–53. Basingstoke: Palgrave MacMillan.

Anttila, Raimo (1989). *Historical and Comparative Linguistics.* Amsterdam/Philadelphia: Benjamins.

Auer, Peter (2004) Non-standard evidence in syntactic typology – Methodological remarks on the use of dialect data vs. spoken language data. In: Bernd Kortmann (ed.), *Dialectology meets Typology,* 69–92. Berlin/New York: Mouton de Gruyter.

Barbiers, Sjef, Leonie Cornips and Susanne van der Kleij (eds.) (2002). *Syntactic Microvariation*. Amsterdam: SAND. (`http://www. meertens.nl/books/synmic/`)

Bisang, Walter (2004). Dialectology and typology: An integrative perspective. In: Bernd Kortmann (ed.), *Dialectology meets Typology,* 11–45. Berlin/New York: Mouton de Gruyter.

Chambers, J.K. (2000). Universal sources of the vernacular. In: Ulrich Ammon, Klaus J. Mattheier and Peter H. Nelde (eds.), *The Future of European Sociolinguistics*, 11–15. (Special issue of *Sociolinguistica: International Yearbook of European Sociolinguistics* 14.) Tübingen: Niemeyer.

Chambers, J.K. (2001). Vernacular universals. In: Josep M. Fontana, Louise McNally, M. Teresa Turell and Enric Vallduví (eds.), *ICLaVE 1*: *Proceedings of the First International Conference on Language Variation in Europe*, 52–60. Barcelona: Universitat Pompeu Fabra.

Chambers, J.K. (2003²). *Sociolinguistic Theory: Linguistic Variation and Its Social Implications*. Oxford, UK/Malden, US: Blackwell.

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

Danchev, Andrei (1992). The evidence for analytic and synthetic developments in English. In: Matti Rissanen, Ossi Ihalainen, Terttu Nevalainen and Irma Taavitsainen (eds.), *History of Englishes: New Methods and Interpretations in Historical Linguistics,* 25-41. Berlin/New York: Mouton de Gruyter.

De Vogelaer, Gunther, Annemie Neuckermans and Guido Vanden Wyngaerd (2002). Complementizer agreement in the Flemish dialects. In: Sjef Barbiers, Leonie Cornips and Susanne van der Kleij (eds.), *Syntactic Microvariation*. Amsterdam. 97–115. Available online at http://www.meertens.nl/ books/synmic/.

Filppula, Markku, Juhani Klemola and Heli Paulasto (eds.) (to appear 2009). *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*. London/New York: Routledge.

Gil, David, Peter Trudgill and Geoffrey Sampson (eds.) (to appear 2009). *Language Complexity as a Variable Concept.* Oxford: Oxford University Press.

Greenbaum, Sidney (1996). *Comparing English Worldwide: The International Corpus of English.* Oxford/New York: Clarendon Press/Oxford University Press.

Greenberg, Joseph H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26**:** 178–194.

Haser, Verena and Bernd Kortmann (2008). Agreement in English dialects. In: Andreas Dufter, Jürg Fleischer and Guido Seiler (eds.), *Describing and Modeling Variation in Grammar*. Berlin/New York: Mouton de Gruyter.

Haspelmath, Martin, Matthew S. Dryer, David Gil and Bernard Comrie (eds.) (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press. http://wals.info/index

Hernández, Nuria (2006). User's guide to FRED. Available online at `http://www.freidok.uni-freiburg.de/volltexte/2489`. Freiburg: English Dialects Research Group.

Hockett, Charles F. (1954). Two models of grammatical description. *Word* 10: 210–231.

Kortmann, Bernd (ed.) (2004). *Dialectology Meets Typology*. Berlin/New York: Mouton de Gruyter.

Kortmann, Bernd, Tanja Herrmann, Lukas Pietsch and Susanne Wagner (2005). *A Comparative Grammar of British English Dialects: Agreement, Gender, Relative Clauses.* Berlin/New York: Mouton de Gruyter.
Kortmann, Bernd, Edgar Schneider, Kate Burridge, Rajend Mesthrie and Clive Upton (eds.) (2004). *A Handbook of Varieties of English*, vol. 2: *Morphosyntax.* Berlin/New York: Mouton de Gruyter.

Kortmann, Bernd and Benedikt Szmrecsanyi (2004). Global synopsis: morphological and syntactic variation in English. In: Bernd Kortmann et al. (eds.), *A Handbook of Varieties of English*, vol. 2: *Morphosyntax*, 1142–1202. Berlin/New York: Mouton de Gruyter.

Kortmann, Bernd and Benedikt Szmrecsanyi to appear 2009 World Englishes between simplification and complexification. In: Lucia Siebers, and Thomas Hoffman (eds.), *World Englishes: Problems - Properties – Prospects*. Amsterdam/Philadelphia: Benjamins.

Kortmann, Bernd and Susanne Wagner (2005). The Freiburg English Dialect project and corpus. In: Bernd Kortmann, Tanja Herrmann, Lukas Pietsch and Susanne Wagner, *A Comparative Grammar of British English Dialects: Agreement, Gender, Relative Clauses*, 1-20. Berlin/New York: Mouton de Gruyter.

McWhorter, John (2007). *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars*. Oxford: Oxford University Press.

Mesthrie, Rajend (2006). World Englishes and the multilingual history of English. *World Englishes* 25(3/4): 381–390.

Platt, John Talbot, Heidi Weber and Mian Lian Ho (1984). *The New Englishes*. London: Routledge & Kegan Paul.

Pride, John B. (ed.) (1982). *New Englishes.* Rowley, Mass.: Newbury House.

Schwegler, Armin (1990). *Analyticity and Syntheticity: A Diachronic Perspective with Special Reference to Romance Languages*. Berlin/New York: Mouton de Gruyter.

Siegel, Jeff (2004). Morphological simplicity in Pidgins and Creoles. *Journal of Pidgin and Creole Languages* 19: 139–162.

Siegel, Jeff (2008). *The Emergence of Pidgin and Creole Languages.* Oxford: Oxford University Press.

Siemund, Peter (to appear 2009). Linguistic universals and vernacular data. In: Markku Filppula, Juhani Klemola and Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*. London/New York: Routledge.

Szmrecsanyi, Benedikt (to appear 2009). Typological parameters of intra-lingual variability: Grammatical analyticity vs. syntheticity in varieties of English. *Language Variation and Change*.

Szmrecsanyi, Benedikt and Bernd Kortmann (to appear 2009a). Between simplification and complexification: Non-standard varieties of English around the world. In: David Gil, Peter Trudgill and Geoffrey Sampson (eds.), *Language Complexity as a Variable Concept.* Oxford: Oxford University Press.

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

Szmrecsanyi, Benedikt and Bernd Kortmann (to appear 2009c). The morphosyntax of varieties of English worldwide: A quantitative perspective. Special issue of *Lingua.*

Trudgill, Peter (2001). Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* 5: 371–374.

Trudgill, Peter (to appear 2009). Vernacular universals and the sociolinguistic typology of English dialects. In: Markku Filppula, Juhani Klemola and Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*. London/New York: Routledge.

Wagner, Susanne (2004). 'Gendered' pronouns in English dialects – a typological perspective. In: Bernd Kortmann (ed.), *Dialectology meets*

Typology: Dialect Grammar from a Cross-Linguistic Perspective, 479–496. Berlin/New York: Mouton de Gruyter.

Winford, Donald (to appear 2009). The interplay of 'universals' and contact-induced change in the emergence of new Englishes. In: Markku Filppula, Juhani Klemola and Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*. London/New York: Routledge.

Wunderlich, Dieter (2008). Spekulationen zum Anfang der Sprache. *Zeitschrift für Sprachwissenschaft* 27: 229–265.