World Englishes between simplification and complexification

Bernd Kortmann and Benedikt Szmrecsanyi
Freiburg Institute for Advanced Studies

This paper offers a broad empirical morphosyntactic study contributing to three debates in linguistics, one of long standing (the so-called equi-complexity axiom), the other two rather more recent, namely McWhorter's claim (2001a,b, 2007) that (pidgins and) creoles have the simplest grammars, and Trudgill's claim (2009) that high-contact varieties of English are characterized by structural simplification processes while low-contact varieties are the result of complexification processes. We will present the results of comparative analyses covering three notionally different morphosyntactic complexity metrics applied to two different data types for about 50 largely non-standard varieties of English (low-contact L1s, high-contact L1s, L2s, pidgins and creoles). Ultimately, we believe to be in a firm position to reject the axiom that the morphosyntax of all languages (and varieties of a language) is equally complex and to support both the claims by McWhorter and Trudgill.

Keywords: morphosyntax, complexity, analyticity, syntheticity, grammaticity

1. Introduction

This paper is to be seen against a triple background: two of immediate relevance to the World Englishes community and all those interested in language contact, and one of much wider and more fundamental concern, reaching far beyond the study of English — and these three backgrounds are closely related.

Let us turn to the latter, more far-reaching issue first, namely to the notions of *complexity / complexification* vs. *simplicity / simplification*. These are currently widely debated notions in the study of language typology and language variation, as can be seen from recent book publications such as Kusters (2003), Dahl (2004), Miestamo, Sinnemäki and Karlsson (2008) or Sampson, Gil and Trudgill (2009). One of the hotly debated issues is, for example, the so-called "equi-complexity axiom" or "equi-complexity dogma", which goes back at least to the American Structuralism of the 1950s

(but can ultimately be found in various guises since Wilhelm von Humboldt's days), according to which all languages, more exactly the grammars of all languages, are equally complex. Here is a relevant quotation from Charles Hockett (1958: 180–1):

> … it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically. Fox, with a more complex morphology than English, thus ought to have a somewhat simpler syntax; and this is the case.

The authors of the present paper disagree with this axiom, like others before them. Take, for example, such a prominent representative of typological linguistics as Robert M.W. Dixon (1997: 75):

> While it is the case that all languages are roughly equal (that is, no language is six times as complex as any other, and there are no primitive languages), it is by no means the case that they are exactly equal … There is no doubt that one language may have greater overall grammatical complexity.

Based on a fairly extensive set of pilot studies on varieties of English around the world, we will provide evidence which, in support of Dixon, leads us to challenge the equi-complexity axiom and to stating the following: (1) There **are** languages and, in our case, even varieties of a language which have a simple(r) or less complex grammar and others with a (considerably more) complex grammar, and (2) there is not necessarily a trade-off between syntheticity and analyticity.

The second debate against which this paper needs to be seen involves recent (controversial) statements concerning simplicity / simplification and complexity / complexification relevant for the World Englishes community and, it needs to be added, to all those interested in pidgins and creoles and, more generally, in language contact. The one who started this debate and has remained at its centre until today is John McWhorter (cf. especially his 2001a article "The world's simplest grammars are creole grammars" and his 2007 monograph *Language Interrupted*). The bottom line of these and other relevant publications by McWhorter is that (1) there is no such thing as equi-complexity among the grammars of the world's languages, and more exactly that (2) the grammars of pidgins and creoles tend to be among the simplest, or located at the pole of higher simplicity, among the world's languages. Just compare some relevant quotations from McWhorter (2001a, b):

> Our claim is indeed 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. (2001a: 162)
> … in the final analysis, there would be a healthy band of languages beginning at the "simplicity" pole which would all be creoles. (2001a: 162)

> My claim is that all of the world's least complex languages will be creoles, not that all creoles are simpler than older languages. (2001b: 392)
> I can confidently state that while creoles do vary typologically, they consistently demonstrate significant reduction of what I have called "ornament" in comparison to older languages. (2001b: 410)

There was even a special issue of the journal *Linguistic Typology* (vol. 5–2/3, 2001a,b) devoted to a (partly heated) debate of these and related claims by McWhorter. One of the contributors to this volume was Peter Trudgill. In his article "Contact and simplification" (2001: 371–4), he essentially sides with McWhorter ("I entirely agree that creole grammars are the simplest grammars, …"; 2001b: 372), identifying as the major motivation for the (high degree of) structural simplicity of creoles the strategies that adults universally tend to resort to in coping with the problem of learning another language or dialect:

> My thinking was, and is, that "linguistic complexity", although this, as McWhorter says, is very hard to define or quantify, equates with "difficulty of learning for adults". (371)
> Just as complexity increases through time, and survives as the result of the amazing language-learning abilities of the human child, so complexity disappears as a result of the lousy language-learning abilities of the human adult. Adult language contact means adult language learning; and adult language learning means simplification, most obviously manifested in a loss of redundancy and irregularity and an increase in transparency. This can indeed be seen at its most extreme in pidgins and hence in creoles … But it is not confined to these types of language. (372)

It is the last sentence of the latter quotation which leads on right away to the third debate in the context of which this paper needs to be seen. Trudgill's statement that loss of redundancy, loss of irregularity, and increase in transparency can be observed in types of language other than pidgins and creoles, too, plays a central role in the so-called "vernacular universals debate" launched by Jack Chambers (2001, 2004) and standing at the centre of a volume edited by Filppula, Klemola and Paulasto (2009). By *vernacular universals* (or: *vernacular roots*) Chambers understands "a small number of phonological and grammatical processes [which] recur in vernaculars wherever they are spoken" (2004: 128), such as levelling of irregular verb forms, multiple negation, or copula deletion. The major divide that Chambers thus (rather unsurprisingly) draws among varieties of English is the one 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 and Paulasto (2009). According to Trudgill, not enough vernacular universals have been found to make the concept fruitful (for a more moderate and differentiated evaluation of Chambers' notion cf.

Szmrecsanyi/Kortmann 2009a in the same volume) and, more fundamentally, the "true typological split" among varieties of English lies elsewhere, not between vernacular and non-vernacular varieties. The major split rather lies between high-contact and low-contact varieties of English. The former include (1) koineised non-standard urban varieties in the British Isles and colonial varieties of North America, Australasia and South Africa; (2) non-native, indigenized L2 varieties like Indian English or Nigerian English, (3) shift varieties like Irish English and Welsh English, (4) English-based pidgins and creoles (as extreme cases resulting from language contact), and, notably, (5) 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. The major point now for the purposes of the present paper is this: According to Trudgill, and following the line of reasoning given in the two quotations from Trudgill (2001) above, what characterizes the grammars of high-contact varieties are processes of simplification as opposed to processes of complexification to be observed in the grammars of low-contact varieties.

It is against this triple background of (1) the equi-complexity axiom, (2) the "(pidgins and) creoles have the simplest grammars" debate, and (3) the typological split between high- and low-contact varieties of English that the main purpose of the present paper — which heavily draws on data and ideas first discussed in Szmrecsanyi and Kortmann (2009b) — needs to be seen. Underlying all three debates are two foundational questions: (1) "What exactly is meant by complexity?", and once this question has been answered, (2) "How can complexity be quantified?". For the purposes of the present paper, the first question can immediately be answered: We will exclusively be concerned with structural, more exactly morphosyntactic, complexity. It is the second question which is at the heart, and constitutes the bulk, of this paper. We will introduce a set of metrics which are capable of quantifying structural, or "surfacy", morphosyntactic complexity. More exactly, what will be presented here are the results of some first large-scale empirical, comparative analyses covering three notionally different complexity metrics (Sec. 2.1) applied to a number of different types of varieties of English (traditional or low-contact, L1 varieties, high-contact L1 varieties, L2 varieties, pidgins and creoles). Two types of data will be used (cf. Section 2.2): survey data on 46 varieties of English based on the *Handbook of Varieties of English* (Kortmann *et al*. 2004; Kortmann and Szmrecanyi 2004) and naturalistic corpus data largely taken from individual dialect corpora and the *International Corpus of English* on 15 L1 and L2 varieties of English. Different facets of structural complexity will be investigated for each of the two data types in turn (cf. Sec. 3 for the survey data and Sec. 4 for the corpus data). As a result of our empirical study, we believe to be in a firm position to back up our rejection of the equi-complexity axiom and to evaluate both the claims by McWhorter and Trudgill concerning different degrees of simplicity and complexity inherent in the morphosyntax of (largely non-standard) varieties of English around the world.

2. Types of complexity and data

2.1 Complexity

In our endeavour to measure local morphological and, to a more limited extent, syntactic complexities, we are operating with the following facets of structural complexity:

1. **Quantitative complexity**, also known as "more is more complex"- complexity (cf. Arends 2001: 180): This involves, on the one hand, the size of the marker and rule inventory of a language or variety. What we are interested in here are, in particular, the number of "ornamental" markers and rules, i.e. those involving more form/code and/or more rules without added communicative bonus (cf. McWhorter 2001a, b, also 2007: 21–32 on overspecification and structural elaboration). This will be henceforth referred to as **ornamental rule complexity** (for examples of, in our view, ornamental morphosyntactic features see Sec. 3 below). On the other hand, we will take quantitative complexity to involve the text frequency of grammatical markers, i.e. their token frequency in naturalistic discourse. This is what henceforth will be referred to as **grammaticity** (a coinage we prefer over Dahl's 2004 term *verbosity*). Grammaticity can be further subdivided into **grammatical analyticity** (the text frequency of free grammatical markers in naturalistic discourse) and **grammatical syntheticity** (the text frequency of bound grammatical markers in naturalistic discourse).
2. **L2-acquisition complexity**, also known as "outsider complexity" (cf. Kusters 2003; Trudgill 2001) or "relative complexity" (Miestamo et al. 2008), i.e. the difficulty individual morphosyntactic features pose to adults in acquiring a second language (cf. Sec. 3 for examples). Interesting for the present purposes is the number of morphosyntactic features in a given variety which (in research on language contact and, especially, Second Language Acquisition) are known to recur in interlanguage varieties. In other words, we are primarily interested in the number of L2-simple features (i.e. features known to facilitate second language acquisition by adults) in the inventory of a given variety of English.
3. **Complexity deriving from irregularities and low transparency** (cf. McWhorter 2001a, b, 2007; Trudgill 2004): Here we are interested in the text frequency of irregular, lexically conditioned grammatical morphemes (or rather: allomorphs) in naturalistic discourse.

Three central questions that will be explored with regard to these types of complexity are the following: To what extent are complexity levels sensitive to variety type? Are there trade-offs between complexity types? In terms of grammaticity specifically, are there trade-offs between syntheticity and analyticity?

2.2 Data

We will draw on two types of data sources: survey data, the classic data type in typological and dialectological research, and naturalistic corpus data. The latter data type is quintessential for all (text) frequency-based complexity measures.

The source for our survey data 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 *et al*. 2004) and subjected to a first close examination in Kortmann and Szmrecsanyi (2004). For this survey, material has been 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, including a discussion of problems and potential drawbacks, cf. Kortmann and Szmrecsanyi 2004: 1142–5). The features in the survey are numbered from 1 to 76 and cover eleven broad areas of morphosyntax: pronouns, the noun phrase, tense and aspect, modal verbs, verb morphology, adverbs, negation, agreement, relativization, complementation, discourse organization and word order. This, for example, is the complete set of negation features in the survey (including the feature numbering):

[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*?)

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):

Table 1. Varieties in the *World Atlas of Morphosyntactic Variation in English*

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

For the importance of sorting the 46 varieties according to variety type, and not geography (i.e. Anglophone world region) — an issue that is also of towering importance in this paper — see Szmrecsanyi and Kortmann (2009a, fc.).

The bulk of our naturalistic corpus data stems from two major digitized speech corpora: the *Freiburg Corpus of English Dialects* (Anderwald and Wagner 2007; FRED; cf. Hernandez 2006; Kortmann and Wagner 2005) and the *International Corpus of English* (ICE; cf. Greenbaum 1996). On the whole, we sampled 12 spoken varieties (two high-contact varieties, five low-contact varieties, five L2 varieties), from these two corpora plus, as another high-contact L1 variety, the *Northern Ireland Transcribed Corpus of Speech* (NITCS; cf. Kirk 1992). Purely for benchmarking purposes, we also included spoken data from two high-contact standard varieties of British English (from the ICE-GB) and American English (from the *Corpus of Spoken American English* [CSAE]; cf. Du Bois *et al*. 2000). Table 2 provides the total picture of the 15 samples drawn from digitized speech corpora.

Table 2. Speech corpora and varieties of English investigated

corpus	subcorpus	variety/varieties	variety type
FRED	FRED-SE	English Southeast + East Anglia (SE+EA)	traditional L1
	FRED-SW	English Southwest (SW)	traditional L1
	FRED-MID	English Midlands (Mid)	traditional L1
	FRED-N	English North (N)	traditional L1
	FRED-SCH	Scottish Highlands (SCH)	traditional L1
	FRED-WAL	Welsh English (WelE)	high-contact L1
ICE	ICE-NZ-S1A	New Zealand English (NZE)	high-contact L1
	ICE-HK-S1A	Hong Kong English (HKE)	L2
	ICE-JA-S1A	Jamaican English (JamE)	L2
	ICE-PHIL-S1A	Philippines English (PhilE)	L2
	ICE-SING-S1A	Singapore English (SingE)	L2
	ICE-IND-S1A	Indian English (IndE)	L2
	ICE-GB-S1A	colloquial British English (collBrE)	high-contact L1
NITCS		Northern Irish English (NIrE)	high-contact L1
CSAE		colloquial American English (collAmE)	high-contact L1

Technically, we utilized an automated algorithm to extract 1 000 random, decontextualized tokens (i.e. orthographically transcribed words) per variety and (sub)corpus, yielding in all a dataset of 15 000 tokens (15 varieties × 1 000 tokens). This dataset was then subjected to morphosyntactic analysis, on the basis of which we eventually computed a set of Greenberg-inspired indices (cf. Greenberg 1960; for further details cf. Sec. (4).

3. Complexity in survey data

For measuring complexity in our survey data, the features in the survey were classified, where applicable, into three groups, which are detailed below. Note here that some features in the survey are "neutral" such that they cannot meaningfully be classified into any particular group (for instance feature [2]: *me* instead of possessive *my*). Also notice that a given feature may have been classed into more than one group at once — for instance feature [37]: levelling of preterite and past participle verb forms; unmarked forms — is categorized as simplifying as well as L2-simple.

1. **"ornamentally complex" features**, i.e. those that complicate the system *vis-à-vis* the standard system, without clearly yielding an added communicative bonus (cf. also McWhorter 2001a, b, 2007). Of course we, in principle, agree with Siegel (2004: 150) "…that there are no principled means of determining which

grammatical features are fundamental and which are ornamental". But then again, if carefully applied, the criterion of additional morphosyntactic rules and/or markers (whether bound or free) without carrying an additional semantic or pragmatic meaning can serve as a useful guideline for separating "ornament" from "essence". Some examples from the 76 features catalogue used in the survey may help. In all, we consider the following seven features as meeting the criterion of "ornamental complexity":

[7] *she*/*her* used for inanimate referents (e.g. *She was burning good* [about a house])

[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*)

[26] *be* as perfect auxiliary (e.g. *They're not left school yet*)

[32] *was sat* / *stood* with progressive meaning (e.g. *when you're stood* 'are standing' *there you can see the flames*)

[41] *a*-prefixing on *ing*-forms (e.g. *They wasn't a-doin' nothin' wrong*)

[60] Northern Subject Rule (e.g. *I sing* [*vs*. *I sings], *Birds sings*, *I sing and dances*)

[26], for instance, implies a grammar with two ways of forming the perfect (*be*- and *have*-perfect). Thus additional selection criteria are necessary in order to be able to determine which perfect goes with which verb type. Clearly, this is a complicating factor without yielding a special semantic or pragmatic bonus. Similarly, the Northern Subject Rule [60] in the North of England and Ireland complexifies the agreement system of the relevant grammars by being sensitive both to type-of-subject constraint (pronoun *vs*. common or proper noun) and position-of-subject constraint (immediately adjacent to finite verb or not), again however without a "communicative bonus".

2. **simplifying features**, i.e. those that simplify the system, *vis-à-vis* the standard system. On the whole, 41 out of the 76 features qualify as simplifying in this way. Here are just three examples:

[4] regularized reflexives-paradigm (e.g. *hisself*, *theirselves* / *theirself*), i.e. regularization of an irregular pattern in standard English (both personal pronoun + *-self*/*selves* and possessive pronoun + *-self*/*selves*)

[30] loosening of sequence of tenses rule (e.g. *I noticed the van I came in*), i.e. the elimination of selection restrictions (past perfect *vs*. simple past in subordinate clause)

[38] levelling of preterite and past participle verb forms: past form replacing the participle (e.g. *He had went)*, i.e. reducing a three-way contrast (*go*, *went*, *gone*) to a two-way contrast (*go*, *went*, *went*)

3. **L2-simple features**, i.e. those that are known to recur in interlanguage varieties, such as the following (for more extensive lists cf. Seuren and Wekker 1986; Wekker 1996; and especially Klein and Purdue 1997):

– avoidance of inflectional marking, preference for analyticity
– preference for semantic transparency
– tendency to overgeneralize, as in *he goed*
– typically, one particle for negation which is preverbal, especially in early stages of L2-acquisition
– avoidance of agreement by morphological means, for instance, no 3rd person singular *–s*
– widespread copula absence
– resumptive pronouns are frequent

Out of the 76 features in the survey, 24 qualify as L2-simple in our view. Here is a selection (see Szmrecsanyi and Kortmann 2009b for the complete list):

[14]	absence of plural marking after measure nouns (e.g. *five pound*)	LOSS OF INFLECTION
[27]	*do* as a tense and aspect marker (e.g. *This man what do own this*)	GAIN IN ANALYTICITY AND TRANSPARENCY
[31]	*would* in *if*-clauses (e.g. *If I'd be you,…*)	GAIN IN TRANSPARENCY
[36]	regularization of irregular verb paradigms (e.g. *catch-catched-catched*)	GAIN IN GENERALIZATION
[40]	zero past tense forms of regular verbs (e.g. *I walk* for *I walked*)	LOSS OF INFLECTION
[48]	invariant *don't* in the present tense	LOSS OF INFLECTION
[50]	*no* as preverbal negator (e.g. *me no iit brekfus*)	RULE SIMPLIFICATION AND GAIN IN PROCESSING EASE
[53]	invariant present tense forms: no marking for the 3rd person singular (e.g. *So he show and say, What's up?*)	LOSS OF AGREEMENT
[57]	deletion of *be*	LOSS OF COPULA
[67]	resumptive / shadow pronouns (e.g. *This is the house which I painted* ***it*** *yesterday*)	GAIN IN PROCESSING EASE DUE TO RESUMPTIVE ELEMENT (see Hawkins 2004)

Table 3. Mean ornamental rule / feature complexity, rule simplicity, and L2-simplicity (based on number of relevant items) by variety type

variety type	mean number of three types of features/rules attested		
	ornamentally complex (total number: (7)	**simplifying** (total number: 41)	**L2-simple** (total number: 24)
traditional L1	2.40	14.86	6.14
high-contact L1	1.17	14.00	6.23
L2	1.00	8.55	6.00
P/C	1.14	16.20	12.73

For each of the four variety types we then calculated the mean number for each of the three types of features. The results are given in Table 3.

The major observations to be made in Table 3 can be summarized as follows:

1. **ornamental complexity**: Traditional L1 varieties are most ornamental, exhibiting between two and three ornamental morphosyntactic features, while high-contact L1 varieties, L2 varieties, and English-based pidgins and creoles attest only about one ornamentally complex feature, with L2 varieties at the bottom end (1.00). Thus, ornamental complexity is clearly an inverse function of the degree of contact, possibly also of a history of L2-acquisitions among adults, which in turn confirms McWhorter's (2001a, 2007) and especially Trudgill's position (as, for example, in Trudgill 2001, 2004, 2009).
2. **feature / rule simplicity**: English-based pidgins and creoles attest the most simplifying features (16.2), while L2 varieties attest by far the fewest simplifying features (8.55). There are no major differences between low- and high-contact L1 varieties in this respect (14.86 *vs.* 14.00). Most importantly, these relatively high figures clearly reflect that L1 varieties, too, even low-contact L1s, can not simply be characterized as exhibiting complexity in the domain of morphosyntax.
3. **L2-acquisition simplicity**: Again (and again as one would have expected) English-based pidgins and creoles attest most, this time by far the most, L2-simple features (12.73), namely more than twice as many as any of the other three variety types, all of which hover around a mean number of six L2-simple features.

In sum, the results for traditional dialects, on the one hand, and for English-based pidgins and creoles, on the other hand, are consonant with what one would expect from reading the recent literature. What is most astonishing about the figures in Table 3 are the low values for both simplifying and L2-simple features / rules in L2 varieties. This clearly runs counter to what one may assume given what has been voiced on the (rather low) degree of morphosyntactic complexity in previous L2-research.

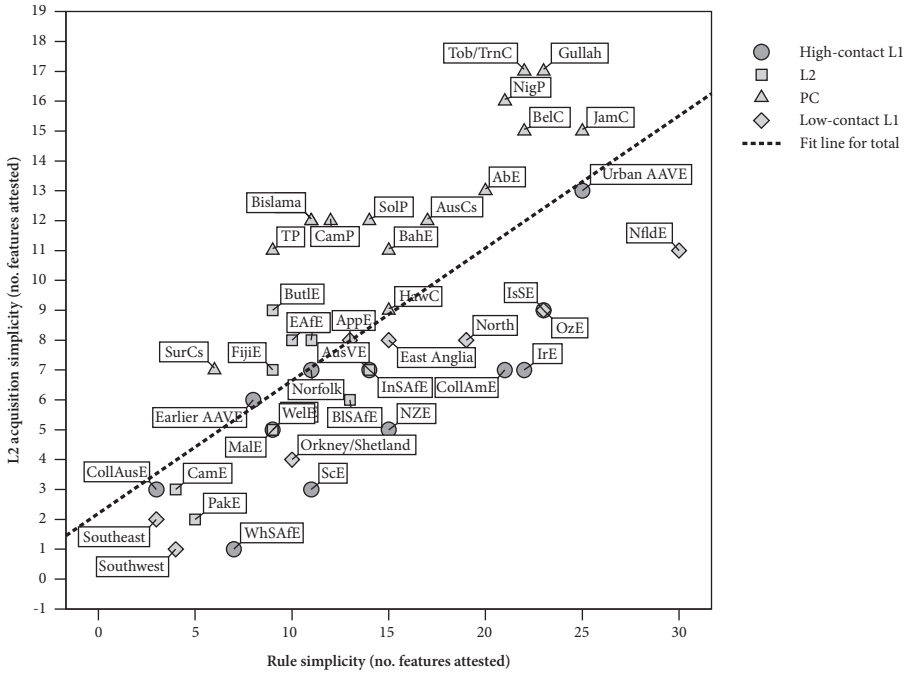

Diagram 1. L2-simplicity by rule simplicity. The dotted trend line represents linear estimate of the relationship

For the survey-based complexity types introduced in this section, let us finally consider the question whether there are trade-offs between the different complexity types. Diagram 1 plots the two simplicity indices against each other, with rule simplicity on the horizontal axis, and L2-acquisition simplicity on the vertical axis. It is interesting how the three main variety types (L1s, L2s, pidgins and creoles) cluster together: The L1 varieties (indicated by circles and diamonds for high- and low-contact L1s respectively) are all on or below the trend line; pidgins and creoles (indicated by triangles) all cluster above the trend line, largely in the top right corner of the diagram; the L2 varieties (indicated by squares) cluster in the lower left corner, attesting neither many simplifying nor many L2-simple features.

4. Complexity in naturalistic corpus data

In this section, we will introduce a total of four frequency-based complexity metrics calculated on the basis of the 15 samples of naturalistic corpus data described in Section 2.2 (see tab. (2)). Note that pidgins and creoles are not represented in these samples, thus the complexity metrics explored here only provide us with comparative in-

dices for high-contact L1 varieties (three samples: NZE, NIrE, WelE), low-contact L1 varieties (five samples, all representing dialects in the British Isles: SE and EA, SW, Mid, N, SCH), and L2 varieties (five samples: HKE, SingE, PhilE, IndE, JamE). For benchmarking purposes, we also calculated the relevant frequency-based indices for two spoken standard varieties (for BrE and AmE respectively).

The following are the frequency indices we calculated for each sample (and thus, per variety): (1) the grammaticity index, i.e. the total frequency of overt grammatical markers per sample; (2) the analyticity index, i.e. the total frequency of **free** grammatical morphemes (or: function words) per sample; (3) the syntheticity index, i.e. the total frequency of **bound** grammatical morphemes per sample; and (4) the transparency index, i.e. the percentage of **bound** grammatical morphemes which are **regular** per sample. In terms of the three complexity types distinguished in Section 2.1, the first three frequency-based indices are all measures of quantitative complexity, whereas the fourth one clearly relates to complexity deriving from irregularities and low transparency.

The method we have used for calculating these four frequency-based complexity metrics by way of a pilot study (cf. Szmrecsanyi and Kortmann 2009b) was inspired by Joseph Greenberg and his 1960 paper "A Quantitative Approach to the Morphological Typology of Language", which in turn was inspired by Sapir's work on morphological typology (cf. Greenberg 1960: 185). For each sample we extracted a random set of 1 000 orthographically transcribed words. For each token in the relevant database, we subsequently established (1) whether the token bears a bound grammatical morpheme (fusional or agglutinative), as in *sing-s* or *sang*; and/or (2) whether the token is a free grammatical morpheme, or a so-called function word, belonging to a closed grammatical class (essentially, determiners, pronouns, *wh*-words, conjunctions, auxiliaries, prepositions, negators). On the basis of this analysis, we established for each sample the syntheticity index (how many bound grammatical morphemes per 1 000 tokens?), the analyticity index (how many free grammatical morphemes per 1 000 tokens?), and the overall grammaticity index (simply the sum of the former two indices). The relevant count for the fourth index, the transparency index, was the share of bound inflectional morphemes per sample whose meaning was transparent, i.e. the percentage of regular (as opposed to lexically conditioned) allomorphs.

How reliable are findings deriving from a sample size of 1 000 tokens? To address this issue, simulations on the basis of the oral history interviews in the *British National Corpus* (BNC) were conducted such that for a number of different sample sizes, 10 000 random samples each were obtained to assess the samples' dispersion around the means. It turns out that a 1 000-tokens random sample has a satisfactorily precise 95% confidence interval for the mean of ±.0003 points for the analyticity index, ±.0002 points for the syntheticity index and ±.0003 points for the grammaticity index (in the present study, the three indices have a lower ceiling of 0 and an upper ceiling of 1.

Table 4. Syntheticity, analyticity and overall grammaticity indices by variety type

variety type	mean syntheticity index[a]	mean analyticity index[b]	mean overall grammaticity index[c]
traditional L1	.13	.48	.61
high-contact L1	.11	.46	.57
L2	.09	.45	.54

[a] significant at $p = .02$ (one-way ANOVA: $df = 2$, $F = 5.80$)
[b] marginally insignificant at $p = .07$ (one-way ANOVA: $df = 2$, $F = 3.35$)
[c] significant at $p = .02$ (one-way ANOVA: $df = 2$, $F = 6.04$)

Table 4 cross-tabulates the first three of these indices with variety type. In traditional L1-vernaculars, 13% of all orthographically transcribed words (tokens) carry a bound grammatical morpheme, 48% of all tokens are function words, and approximately 61% of all tokens bear grammatical information. 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 in the literature that a history of contact and adult language learning can eliminate certain types of redundancy, especially those found in grammatical marking (cf. for example Trudgill 2001 and, especially, Trudgill 2009). The fact that L2-varieties generally exhibit the smallest amount of grammatical marking also resolves our earlier puzzle (cf. Sec. (3) as to why L2-varieties do not exhibit particularly many L2-easy features. It just seems as if L2-speakers do not generally opt for "simple" features instead of "complex" features. As a matter of fact, L2-speakers appear to prefer zero marking over explicit marking, be it (presumably) L2-easy or complex.

The interplay between syntheticity and analyticity is visualized in the scatterplot in Diagram 3. The dialects of the SE and EA (in the top right corner) and HKE (in the bottom left corner) are the extreme cases in our dataset. The former are both highly analytic **and** above-average synthetic varieties, while HKE 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 (collAmE and collBrE) and, interestingly, the SW of England (a variety that we classed as a traditional, low-contact vernacular, even though there is a well-known history of adult L2-acquisition and Celtic-Anglo-Saxon contact; see, for instance, Wakelin 1975). The dotted trend line merits particular attention in Diagram 3: As a rather robust ($R^2 = 0.40$) statistical generalization, this line indicates that on the inter-variety level, there is **no** trade-off between analyticity and syntheticity. Needless to say, such a trade-off is often claimed to be one that has governed the

history of English, involving the growing importance of functional word classes and the rise of many periphrastic constructions compensating for the loss of inflectional endings. (Note, though, that we have nothing to say about rigidity of word order as a factor to be reckoned with in the development of English from a synthetic language with a free, pragmatic word order to an analytic language with a fairly fixed, grammaticalized word order.) Instead, according to this diagram, 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 less overt marking, rather than trading off synthetic marking for analytic marking, which is purportedly L2-easy (cf. for instance Seuren and Wekker 1986).

The fourth complexity metric we now investigate on the basis of naturalistic corpus data is the transparency index, which determines the degree of regularity for the synthetic markers of grammaticity. The idea behind this is that of two varieties A and B that one is structurally more complex which exhibits the higher degree (i.e. text frequency) of irregular, lexically conditioned bound grammatical allomorphs (as opposed to the text frequency of regular-agglutinative bound grammatical allomorphs). Methodologically, we investigated all bound grammatical morphemes in our 15 × 1 000 token corpus database, classifying them into either (1) regular-suffixing,

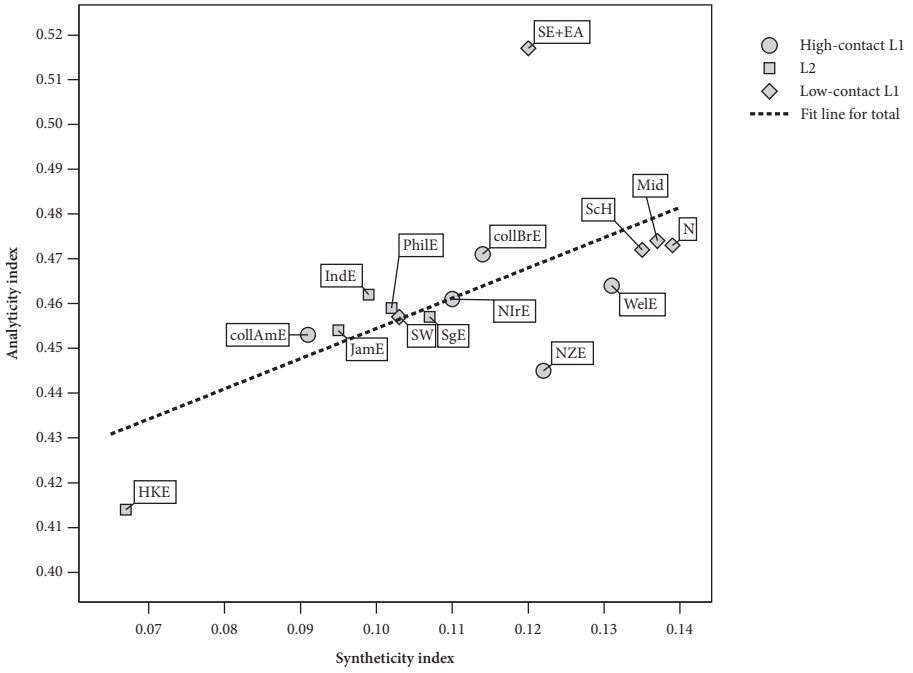

Diagram 2. Analyticity by syntheticity. Dotted trend line represents linear estimate of the relationship

phonologically conditioned allomorphs (e.g. *he walk-ed*) or (2) irregular, lexically conditioned allomorphs (e.g. *he sang*) and establishing corresponding transparency indices. These yield the share of regular allomorphs as a percentage of all bound grammatical morphemes. It emerges that, in typical L2-discourse, 82% of all bound grammatical allomorphs are regular; the corresponding figures for high-contact and traditional L1-varieties decrease to 71% and 65% respectively. In other words, as one would have expected, in terms of irregularity L2-varieties are least complex and most transparent while traditional, low-contact L1-vernaculars are most complex and least transparent (high-contact L1-varieties, once again, cover the middle ground). These results can be taken to suggest that higher degrees of contact and, in particular, adult language acquisition both appear to level irregularities. The likely reason is that "[i]mperfect learning … leads to the removal of irregular and non-transparent forms which naturally cause problems of memory load for adult learners, and to loss of redundant features" (Trudgill 2004: 307).

In Diagram 3 the interplay between transparency and grammaticity is visualized once again by a scatterplot. And again, similar to Diagram 2, HKE is to be found at one end of the diagram (the top left corner reflecting the lowest degree of grammaticity and a fairly high degree of transparency) whereas the dialects of the SE and EA are

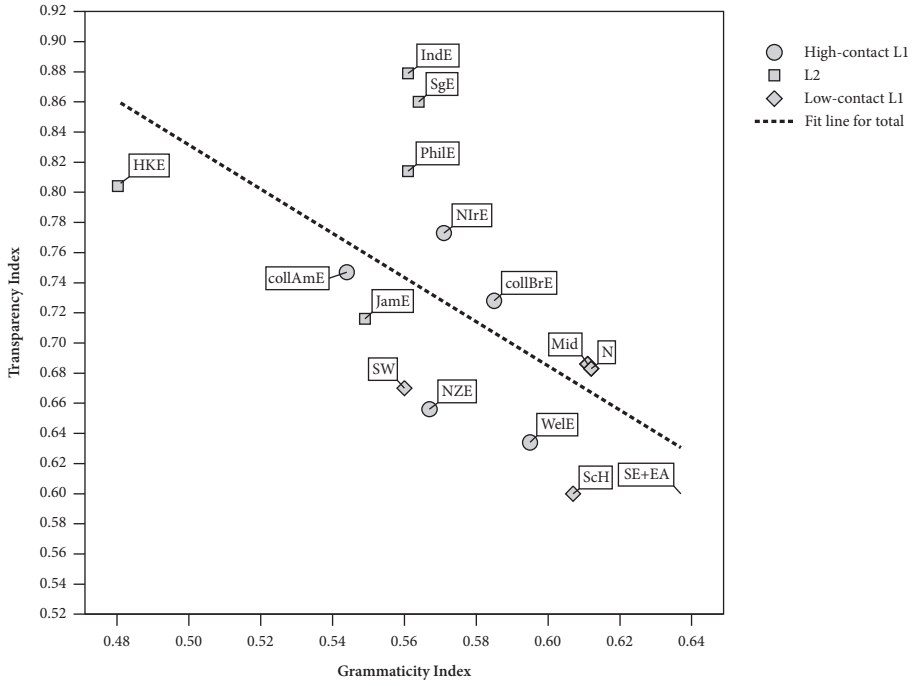

Diagram 3. Transparency by grammaticity. Dotted trend line represents linear estimate of the relationship

found in the diametrically opposed corner of the diagram (the bottom right corner reflecting the highest degree of grammaticity and lowest degree of transparency). As for the correlation with variety type, traditional L1 varieties clearly exhibit most grammaticity and least transparency while L2 varieties exhibit least grammaticity and most transparency (note again, though, that for this study no naturalistic corpus data for pidgins and creoles were investigated).

In general, then, the corpus material scrutinized here shows that, in a cross-variety perspective, there is no trade-off between syntheticity and analyticity. There is a trade-off, however, in that transparency correlates negatively with grammaticity, i.e. the more grammatical information a given variety explicitly codes *via* (bound or free) grammatical markers, the lower the number of regular, transparent grammatical markers tends to be and *vice versa*.

5. Conclusions and outlook

We hope to have demonstrated that, even if there is — as always — still room for refinement, the diverse sets of metrics introduced here allow us to numerically quantify (degrees of) complexity and simplicity in the morphosyntax of varieties of English in an objective way, at least in a much more objective way than ever before. As a first major result, given the differences we were able to identify, our study strengthens the position of all those who reject the equi-complexity axiom. At least when restricting the idea of an overall balance between the different structural levels of a language or dialect to morphosyntax, the idea of equi-complexity should be put at rest. We do not wish to rule out that simplicity on the morphosyntactic level may be compensated for by, for example, a higher degree of complexity in intonation or on the level of pragmatics. We simply have nothing to say on this.

We also succeeded in identifying a pattern underlying the differences in complexity between World Englishes. We believe to have shown that variety type (low- *vs.* high-contact L1 varieties, L2 varieties, pidgins and creoles) predicts observable complexity levels rather well — much along the lines of McWhorter (2001a, 2007) and Trudgill (2001, 2009). Thus, at least based on our English data, language contact appears to result very systematically in a lower degree of complexity due to the strategies preferred by adults in second language acquisition (cf. also Siegel 2004). At the same time, L2 varieties have been found to exhibit a strikingly different complexity profile from English-based pidgins and creoles. There is converging evidence from our survey and our corpus data that for L2 varieties, in particular, the alternative to L2-difficult syntheticity seems to be no grammatical marking at all, rather than analytic marking or "overtly simple" marking. This shows indeed how beneficial studies "…comparing simplicity in a creole to that found in L2 varieties of its lexifier" can be (Siegel 2004: 158). It is more of exactly these kinds of studies that Jeff Siegel called for at the end of his most insightful 2004 article on "Morphological simplicity in pidgins and creoles".

From a more general, methodological point of view, our two sets of metrics — survey-based and corpus-based — offer the following advantages: They serve as absolute holistic complexity measures (cf. Siegel 2004) which allow us to make (1) comparisons across varieties and variety types and to test (2) the trade-off between syntheticity and analyticity. What we have presented here are first results of what we consider to be no more than a set of pilot studies. Our next steps will be to extend both the feature catalogue for the survey data (bringing it up from currently 76 to about 200 morphosyntactic features) and the range of varieties investigated both on the basis of this feature set and naturalistic corpus data. But we are also going in the direction sketched below, which, as other sets of pilot studies have shown, promises important new insights into genre variation, language change, and language typology.

Our way of exploring large-scale study of complexity in language-internal varieties is, in principle, possible for any type of variety (e.g. also for learner varieties, stylistic — especially written *vs*. spontaneous spoken — varieties, or historical varieties) and for any language (especially such languages with an equally "rich" colonial history as English and thus spread around the world, such as Spanish, French, or Portuguese). Moreover, we believe that large-scale language-internal variation is a very useful testing ground for developing and calibrating complexity metrics which can be used for quantifying complexity variation across languages, too. This takes us back full circle, as it were, to language typology and Greenberg's (1960) quantitative approach to morphological typology since it was this approach which provided an important source of inspiration for our approach to the quantitative study of notions like simplicity / complexity or analyticity / syntheticity. Our paper is thus another nice example, we believe, of the potential held in stall by the partnership between the study of dialectology and variationist studies, on the one hand, and the study of cross-linguistic variation in language typology, on the other hand (in the spirit of Kortmann 2004).

In concluding, we would like to make a final remark on what we have presented here in relation to the title of this paper. The title was inspired by Trudgill (2009) where he suggests the typological split between high- and low-contact varieties of English which, in turn, is the result of simplification and complexification processes respectively. The main thrust of the present paper was not, however, to talk about these two processes. The approach taken here was actually purely synchronic and static, presenting metrics for both the absolute and relative quantification of overt morphosyntactic complexity and simplicity. In terms of language change, it is at most the **outcome** of simplification processes that we can make confident statements about, namely in connection with L2 varieties and pidgins and creoles — so, yes, extensive language or dialect contact does indeed seem to foster the growth of morphosyntactic simplicity. There is nothing at all, though, that we have said about the (complexification?) processes yielding the higher degree of morphosyntactic complexity we identified in the low-contact L1 varieties of English. In our view, far too little is known as yet to make firm statements on the two kinds of processes when looking at the four variety types investigated here. But three points seem worth pointing out: (1) The kind of metrics

we presented here provide a means of objectifying judgements in terms of simplification and complexification, especially once applying these metrics to older stages of standard English and especially older stages of non-standard L1 varieties. (2) Once including historical material, we should not be surprised to see that simplification processes have also taken place in low-contact L1 varieties over time, just as complexification processes have taken place in the other three types of varieties; the main difference lies in which type of processes, simplification or complexification, outweighs the other in the relevant variety or variety type. (3) This, finally, leads us to the crucial question of the standard of comparison in passing judgment on processes of simplification and complexification: simpler or more complex than what? For L1 varieties this will be the "standard", or dominant variety, of the relevant period, for L2 varieties and pidgins and creoles that variety (or those varieties) of English which served as superstrate in the relevant part of the world during the crucial formation periods of the New Englishes, like for example Irish English or the dialect of Southwest England. In determining the nature and degree of, for instance, simplification processes for a given New English we must therefore take the relevant (typically non-standard) founder L1 variety as the standard of comparison. Thus we may find, for example, that a given English-based pidgin or creole didn't have to go that long down the road of simplification as we may at first glance assume since the non-standard founder variety itself already exhibited morphosyntactic features resulting from simplification compared with the "standard" in the British Isles at the time. *Vice versa*, where a low-contact L1 variety served as founder variety, exhibiting, as argued in a rather global way by Trudgill and quantitatively confirmed in the present paper, a large (or above-average) number of complexifying morphosyntactic properties, the relevant contact variety, given its present degree of low complexity, must have undergone even more radical simplification processes than we would normally have assumed when taking "standard English" as the standard of comparison. Whichever scenario we are going to witness for the individual contact situation, our set of metrics (or if not these, then at least metrics of that kind), carefully applied to the appropriate sets of material, will allow us in the future to help objectifying step by step simplification and complexification processes in the historical morphology and syntax of individual varieties and variety types of English around the world.

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 C. Beal, Karen B. Corrigan and Hermann L. Moisl eds. *Creating and Digitizing Language Corpora*. Vol. 1: *Synchronic Databases*. Basingstoke: Palgrave MacMillan, 35–53.

Arends, Jacques. 2001. "Simple grammars, complex languages". *Linguistic Typology* 5: 180–2.

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*. Barcelona: Universitat Pompeu Fabra, 52–60.

_____. 2004. "Dynamic typology and vernacular universals". In Kortmann, ed.: 127–45.

Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity.* Amsterdam: John Benjamins.

Dixon, Robert M.W. 1997. *The Rise and Fall of Languages.* Cambridge: Cambridge University Press.

Du Bois, John W., Wallace L. Chafe, Charles Meyer and Sandra A. Thompson. 2000. *Santa Barbara Corpus of Spoken American English, Part 1*. Philadelphia PA: Linguistic Data Consortium.

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

Greenbaum, Sidney, ed. 1996. *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon Press.

Greenberg, Joseph H. 1960. "A quantitative approach to the morphological typology of language". *International Journal of American Linguistics* 26: 178–94.

Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hernández, Nuria. 2006. "User's Guide to FRED: Freiburg Corpus of English Dialects". Freiburg: English Dialects Research Group. <http://www.freidok.uni-freiburg.de/volltexte/2489>.

Hockett Charles Francis. 1958. *A Course in Modern Linguistics.* New York NY: Macmillan.

Kirk, John. 1992. "The Northern Ireland Transcribed Corpus of Speech". In Gerhard Leitner, ed. *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter, 65–73.

Klein, Wolfgang and Clive Perdue. 1997. "The basic variety (or: Couldn't natural languages be much simpler?)". *Second Language Research* 13: 301–47.

Kortmann, Bernd and Benedikt Szmrecsanyi. 2004. "Global synopsis: Morphological and syntactic variation in English". In Kortmann *et al*., eds.: 1142–202.

_____. 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*. Berlin: Mouton de Gruyter, 1–20.

_____. ed. 2004. *Dialectology Meets Typology*. Berlin: Mouton de Gruyter.

_____. Kate Burridge, Rajend Mesthrie, Edgar W. Schneider and Clive Upton, eds. 2004. *A Handbook of Varieties of English.* Vol. 2: *Morphology and Syntax*. Berlin: Mouton de Gruyter.

Kusters, Christiaan Wouter. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.

McWhorter, John H. 2001a. "The world's simplest grammars are creole grammars". *Linguistic Typology* 5: 125–66.

_____. 2001b. "What people ask David Gil and why: Rejoinder to the replies". *Linguistic Typology* 5: 388–413.

_____. 2007. *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars*. Oxford: Oxford University Press.

Matti Miestamo, Kaius Sinnemäki and Fred Karlsson, eds. 2008. *Language Complexity: Typology, Contact, Change.* Amsterdam: John Benjamins, 23–41.

Sampson, Geoffrey, David Gil and Peter Trudgill, eds. 2009. *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press.

Seuren, Pieter A.M. and Herman Wekker. 1986. "Semantic transparency as a factor in creole genesis". In Pieter Muysken and Norval Smith, eds. *Substrata versus Universals in Creole Genesis.* Amsterdam: John Benjamins, 57–70.

Siegel, Jeff. 2004. "Morphological simplicity in pidgins and creoles". *Journal of Pidgin and Creole Languages* 19: 139–62.

_____. 2008.*The Emergence of Pidgin and Creole Languages.* Oxford: Oxford University Press.

Szmrecsanyi, Benedikt and Bernd Kortmann. 2009a. "Vernacular universals and angloversals in a typological perspective". In Filppula, Klemola and Paulasto, eds.: 33–53.

_____. and —. 2009b. "Between simplification and complexification: Non-standard varieties of English around the world". In Sampson, Gil, and Trudgill, eds.: 64–79.

_____. and —. fc. "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: 371–4.

_____. 2004. "Linguistic and social typology: The Austronesian migrations and phoneme inventories". *Linguistic Typology* 8: 305–20.

_____. 2009. "Vernacular universals and the sociolinguistic typology of English dialects". In Filppula, Klemola and Paulasto, eds.: 304–322.

Wakelin, Martyn F. 1975. *Language and History in Cornwall*. Leicester: Leicester University Press.

Wekker, Herman. 1996. *Creole Languages and Language Acquisition.* Berlin: Mouton de Gruyter.