5

Between simplification and complexification: non-standard varieties of English around the world

BENEDIKT SZMRECSANYI AND BERND KORTMANN

1 Introduction

This contribution is an empirical study of morphosyntactic complexity variance in more than four dozen varieties of English, based on four different complexity notions and combining two different data sources.[1] Our point of departure is previous research (e.g. Szmrecsanyi and Kortmann forthcoming) according to which varieties of English – be they native L1 vernaculars, non-native L2 varieties, or English-based pidgins and creoles (P/Cs) – can be thought as varying along two underlying dimensions of morphosyntactic variance. Crucially, Szmrecsanyi and Kortmann (forthcoming) demonstrate that variety type (L1, L2, or P/C) and not, for example, geographical distance or proximity, is the best predictor of a given variety's location relative to these two dimensions.

The work reported here seeks to complement this line of research in three ways. First, we endeavour to analyse and interpret language-internal variation in English in terms of varying complexity and simplicity levels. More specifically, we will be concerned with measuring local morphological and syntactic complexities. The following notions of linguistic complexity will be subject to numerical quantification in the present study:

ornamental rule/feature complexity – the number of "ornamentally complex" features (cf. McWhorter 2001a) attested in a given variety's morphosyntactic inventory;

[1] We wish to thank Christian Mair (Freiburg) for giving us access to the Jamaican component of ICE (being compiled at Freiburg University) even though the component is not officially released yet, and Johanna Gerwin, our research assistant, who manually coded a large portion of our corpus database with utmost precision.

L2 acquisition difficulty, also known as "outsider complexity" (cf. Kusters 2003; Trudgill 2001) or "relative complexity" (cf. Miestamo 2008);

grammaticity and redundancy – the token frequency of grammatical markers, synthetic or analytic, in naturalistic discourse (cf. Greenberg 1960);

complexity deriving from *irregularities* – more specifically, the text frequency of irregular, lexically conditioned grammatical allomorphs in naturalistic discourse (cf. McWhorter 2001a; Trudgill 2004a).

Second, in addition to survey data (the classic data type in typological–dialectological research), we shall also tap naturalistic corpus data, a procedure which will yield a range of frequency-based complexity measures. And third, building on sociolinguistic work suggesting that there is a typological continuum of L1 varieties (cf. Trudgill forthcoming a), we expand our earlier threefold typological classification to a fourfold split: high-contact L1 vernaculars (e.g. Australian E),[2] v. low-contact L1 vernaculars (e.g. East Anglia E), v. English-based P/Cs (e.g. Tok Pisin), v. L2 varieties (e.g. Hong Kong E). In this connection we would like to note that L2 varieties are rather under-researched, especially from a dialectological/typological point of view, which is a gap in the literature that the present study will seek to remedy. Our overall research interest in this chapter will lie in the degree to which variety type correlates with complexity variance.

2 Data sources

2.1 *The World atlas of morphosyntactic variation in English*

The *World atlas* accompanies the *Handbook of varieties of English* (Kortmann et al. 2004). It is available – along with a phonological survey (which will not be subject to analysis in this chapter) – on CD-ROM and online (<http://www.mouton-online.com>). A catalogue of seventy-six features – essentially, the usual suspects in previous dialectological, variationist, and creolist research – was compiled and sent out to the authors of the chapters in the morphosyntax volume of the *Handbook*. For each of the seventy-six features, the contributors were asked to specify whether the feature in question is attested in the variety at hand. Kortmann and Szmrecsanyi (2004: 1142–5) discuss the survey procedure in considerable detail. Suffice it to say here that forty *Handbook* authors provided us with data on forty-six non-standard varieties of English. All seven Anglophone world regions (British Isles, America, Caribbean, Australia, Pacific, Asia, Africa), as well as a fair mix of traditional

[2] Since the many language-varieties discussed in this chapter are all varieties of or derived from English, the name "English" is abbreviated as "E" in order to enable readers to focus on the distinctive parts of the variety names.

(low-contact) L1 vernaculars (N = 8), high-contact L1 varieties (N = 12), L2 varieties (N = 11), and English-based P/Cs (N = 15), are represented in the survey.[3] Table 5.1 gives the breakdown by variety type.

The features in the survey are numbered from 1 to 76 (see the Appendix for the entire feature catalogue) and include all major phenomena discussed in previous survey articles on grammatical properties of (individual groups of) non-standard varieties of English. They cover eleven broad areas of morpho-syntax: pronouns, the noun phrase, tense and aspect, modal verbs, verb morphology, adverbs, negation, agreement, relativization, complementation, and discourse organization and word order.

2.2 *Corpus data*

To supplement the survey data described above – which provide a clear "attested"/"not attested" signal, albeit at the price of being essentially dichotomous (and thus simplistic) in nature – we also sought to investigate naturalistic corpus data, a data type which can yield gradient frequency information. We accessed four major digitized speech corpora sampling

Table 5.1. Varieties sampled in the *World atlas*

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

[3] In this study, categorization of individual L1 varieties into the categories at hand (traditional L1 v. high-contact L1) was carried out somewhat impressionistically, taking into account factors such as the size of speech community, a prolonged history of adult L2 acquisition (as in the case of, e.g., Welsh E), and so on. Observe that our categorization also glosses over the notion of "shift varieties" (cf. Mesthrie 2004: 806).

Table 5.2. Speech corpora and varieties of English investigated

Corpus	Subcorpus	Variety/varieties	Variety type
Freiburg Corpus of English Dialects (FRED) (cf. Hernández 2006)	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
International Corpus of English (ICE) (cf. Greenbaum 1996)	ICE-NZ-S1A	New Zealand E (NZE)	high-contact L1
	ICE-HK-S1A	Hong Kong E (HKE)	L2
	ICE-JA-S1A	Jamaican E (JamE)	L2
	ICE-PHI-S1A	Philippines E (PhilE)	L2
	ICE-SIN-S1A	Singapore E (SgE)	L2
	ICE-IND-S1A	Indian E (IndE)	L2
	ICE-GB-S1A	colloquial British E (collBrE)	high-contact L1
Northern Ireland Transcribed Corpus of Speech (NITCS) (cf. Kirk 1992)		Northern Irish E (NIrE)	high-contact L1
Corpus of Spoken American English (CSAE) (Du Bois et al. 2000)		colloquial American E (collAmE)	high-contact L1

fifteen spoken varieties of English, including traditional (low-contact) vernaculars (N = 5), high-contact L1 varieties (N = 5), and L2 varieties of English (N = 5) (see Table 5.2 for an overview).[4]

All of the corpus material subject to analysis in the present study is spoken-conversational (ICE, CSAE) or drawn from rather informal interview situations (FRED, NITCS). Technically, we utilized an automated algorithm to extract 1,000 random, decontextualized tokens (i.e. orthographically transcribed

[4] Note though that we did not include corpus data on English-based pidgins and creoles here, the reason being that this step would have necessitated devising a set of tailor-made coding schemes, which would have gone beyond the scope of the present study.

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 morphological/grammatical analysis, on the basis of which we eventually computed a set of Greenberg-inspired indices (cf. Greenberg 1960). Sections 5 and 6 will provide more detail on the procedure.

3 Ornamental rule/feature complexity

We define *ornamental rule/feature complexity* as complexity deriving from the presence, in a given variety's morphosyntactic inventory, of features or rules that add contrasts, distinctions, or asymmetries (compared to a system that does not attest such features/rules) without providing a clearly identifiable communicative or functional bonus. In a nutshell, we mean to capture here "ornamental accretions" (McWhorter 2001c: 390) similar to human hair, which serves no real functional purpose and is thus "a matter of habit, doing no harm and thus carried along" (p. 389). A popular example for such complexity is grammatical gender (Trudgill 1999: 148).

Let us illustrate on the basis of our feature catalogue. Among the seventy-six features covered there are, indeed, many features that add contrasts, distinctions, or asymmetries – for instance, feature [26], *be* as perfect auxiliary (yielding additional selection criteria concerning verb type) and feature [3], special forms or phrases for the second person plural pronoun (adding an additional singular/plural contrast). Notice though that in our definition, only the former (*be* as perfect auxiliary) would qualify as complexifying. This is because special forms or phrases for the second person plural clearly add complexity, but they also yield a clearly identifiable functional/communicative advantage – the ability, that is, to distinguish one from two or more addressees. In the spirit of considerations like these, we classified the following items in our survey as "ornamentally complex":

[7] *she/her* used for inanimate referents (e.g. *She was burning good* [said of 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* – a construction with highly specific verb selection criteria, coexisting with the *-ing* progressive though not replacing it)

Table 5.3. Mean ornamental rule/feature complexity (number of ornamentally complex items) by variety type

Variety type	Mean no. of ornamentally complex features/rules attested[a]
Traditional L1	2.40
High-contact L1	1.17
L2	1.00
P/C	1.14

[a] marginally significant at $p = \cdot 06$ (ANOVA: $F = 2 \cdot 97$).

[41] *a*-prefixing on *ing*-forms (e.g. *They wasn't a-doin' nothin' wrong*)
[60] Northern Subject Rule (e.g. *I sing* [v. **I sings*], *Birds sings, I sing and dances*)

So, which of the forty-six varieties in our survey attest most of these features? Table 5.3 cross-tabulates ornamental rule/feature complexity with variety type. The overall thrust of the effect is clear: the typical traditional L1 vernacular attests between two and three ornamentally complex features/rules, while high-contact L1 varieties, L2 varieties, and English-based P/Cs typically attest only about one ornamentally complex feature.[5] Thus, ornamental complexity is clearly a function of the degree of contact (and, possibly, of a history of L2 acquisition among adults), which is a result that ties in well with the literature (Trudgill 2001; 2004a; forthcoming).

4 L2 acquisition difficulty

We will now consider a measure of *outsider complexity* (cf. Kusters 2003; Trudgill 2001) or *relative complexity*. Following the terminology in Miestamo (2008), we shall refer to this type of complexity as *difficulty* (and, correspondingly, to "relative" simplicity as *ease*). The particular reference point that we will use is an adult L2 learner as an outsider whose difficulty or ease of acquiring a language or language variety is theoretically highly relevant, especially in a sociolinguistic perspective. Against this backdrop, we operationalize *L2 acquisition difficulty* as the degree to which a given variety does *not* attest phenomena that L2 acquisition research has shown to recur in interlanguage varieties. The following interlanguage universals (or near-universals), then, may be extrapolated from the literature:

[5] The difference between high-contact L1 vernaculars, L2 varieties, and English-based pidgins and creoles is not statistically significant (ANOVA: $F = \cdot 19$, $p = \cdot 83$).

- avoidance of inflectional marking (– INFLECTION), preference for analyticity (+ ANALYTICITY) (Klein and Perdue 1997: 311; Seuren and Wekker 1986; Wekker 1996)
- pronoun systems are minimal (– PRONOUN) (Klein and Perdue 1997: 312)
- preference for semantic transparency (+ TRANSPARENCY) (Seuren and Wekker 1986)
- tendency to overgeneralize, as in *he goed* (+ GENERALIZATION) (Towell and Hawkins 1994: 227)
- typically, one particle for negation (Klein and Perdue 1997: 312) which is preverbal, especially in early stages of L2 acquisition (Hawkins 2001: 84; Littlewood 2006: 510) (+ PREVERBAL NEG)
- avoidance of agreement by morphological means, for instance, third person singular *-s* (– AGREEMENT) (Dulay and Burt 1973; 1974; Klein and Perdue 1997: 311)
- widespread copula absence (– COPULA) (Klein and Perdue 1997: 320)
- resumptive pronouns are frequent (+ RESUMPTIVE) (Hyltenstam 1984)
- overt syntactic subordination is dispreferred (– SUBORDINATION) (Klein and Perdue 1997: 332)
- inversion as a relatively late development (– INVERSION) (Littlewood 2006: 510)

Given this body of research, we classified the following twenty-four items in our seventy-six-feature catalogue as diagnostics for ease of L2 acquisition:

[6] lack of number distinction in reflexives (– INFLECTION)
[8] generic *he/his* for all genders (– PRONOUN)
[14] absence of plural marking after measure nouns (– INFLECTION)
[27] *do* as a tense and aspect marker (+ ANALYTICITY)
[28] completive/perfect *done* (+ ANALYTICITY)
[29] past tense/anterior marker *been* (+ ANALYTICITY)
[31] *would* in *if*-clauses (+ TRANSPARENCY)
[36] regularization of irregular verb paradigms (+ GENERALIZATION)
[37] unmarked verb forms (– INFLECTION)
[40] zero past tense forms of regular verbs (– INFLECTION)
[45–47] *ain't* (– INFLECTION)
[48] invariant *don't* in the present tense (– INFLECTION)
[50] *no* as preverbal negator (+ PREVERBAL NEG)
[52] invariant non-concord tags (– INFLECTION)
[53] invariant present tense forms: no marking for third person singular (– AGREEMENT)
[55] existential/presentational *there's* etc. with plural subjects (– AGREEMENT)

Table 5.4. Mean L2-ease (number of L2-easy items) by variety type

Variety type	Mean no. of L2-easy features[a]
Traditional L1	6.14
High-contact L1	6.23
L2	6.00
P/C	12.73

[a] highly significant at $p < \cdot 01$ (ANOVA: $F = 16.63$).

[57] deletion of *be* (− COPULA)
[65] use of analytic *that his* etc. instead of *whose* (+ TRANSPARENCY)
[67] resumptive/shadow pronouns (+ RESUMPTIVE)
[72] serial verbs (− SUBORDINATION)
[73] lack of inversion/auxiliaries in *wh*-questions (− INVERSION/− COPULA)
[74] lack of inversion in main clause *yes/no* questions (− INVERSION)

Table 5.4 illuminates how the number of L2-easy features in a variety's inventory tallies with variety type. English-based P/Cs clearly stand out in that they attest, typically, between twelve and thirteen L2-easy features, while other varieties of English only attest about six L2-easy features on average.[6] This is another way of saying that English-based P/Cs are substantially more L2-easy than any other variety type in our survey. In itself, it is not actually surprising that P/Cs are particularly L2-easy, given the nature of creole genesis and the great deal of adult L2 acquisition that accompanies it (cf. Seuren and Wekker 1986). Yet it is noteworthy that L2 varieties of English do not attest significantly more L2-easy features than L1 varieties of English, as one might have expected. This is a puzzle that the subsequent sections will attempt to shed light on.

5 Grammaticity and redundancy

Let us go on to a discussion of some frequency-based, corpus-derived complexity metrics, all of which are "absolute" (cf. Miestamo 2008) because they do not draw on an extragrammatical reference point. We operationally define a given variety's morphosyntactic *grammaticity* as the text frequency with which that variety attests grammatical markers in naturalistic, spontaneous spoken discourse. We take more grammaticity to be indicative of higher complexity, and draw a further distinction between (i) *overall grammaticity*, (ii) *synthetic*

[6] The difference between low-contact L1 vernaculars, high-contact L1 varieties, and L2 varieties is not statistically significant (ANOVA: $F = \cdot 02$, $p = \cdot 98$).

Table 5.5. 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 $p = \cdot02$ (ANOVA: $F = 5{\cdot}80$).
[b] marginally significant at $p = \cdot07$ (ANOVA: $F = 3{\cdot}35$).
[c] significant at $p = \cdot02$ (ANOVA: $F = 6{\cdot}04$).

grammaticity (i.e. the incidence of bound grammatical morphemes), and (iii) *analytic grammaticity* (i.e. the incidence of free grammatical morphemes). We suggest that *grammaticity*, thus defined, can be roughly equated with (grammatical) *redundancy*, in the sense of *repetition of information* (cf. e.g. Trudgill forthcoming a).

Our particular method here is broadly modelled on Joseph Greenberg's (1960) paper, "A quantitative approach to the morphological typology of language". This means that we conducted a morphological/grammatical–functional analysis of our corpus database spanning 15,000 tokens (recall here from section 2 that we compiled fifteen sets – one for each variety of English investigated – of 1,000 orthographically transcribed, randomly selected words). For each token in the database, we established:

- whether the token contains a bound grammatical morpheme (fusional or suffixing), as in *sing-s* or *sang*;
- and/or 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).[7]

On the basis of this analysis, we established three indices: a *syntheticity index* (the percentage of bound grammatical morphemes per 1,000 tokens), an *analyticity index* (the percentage of free grammatical morphemes per 1,000 tokens), and an *overall grammaticity index* (the sum of the former two indices). Table 5.5 cross-tabulates these indices with variety type.

The figures in Table 5.5 may be interpreted as follows: in traditional L1 vernaculars, 13 per cent of all orthographically transcribed words (tokens) carry a bound grammatical morpheme, 48 per cent of all tokens are function words,

[7] In the case of inflected auxiliaries (e.g., *he* is *singing*), the token was counted as being both inflected and belonging to a closed class.

and approximately 61 per cent 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 most grammaticity, L2 varieties exhibit least grammaticity, and high-contact L1 varieties occupy the middle ground. Assuming, as we do, that grammaticity is just another name for redundancy, this hierarchy dovetails nicely with claims in the literature that a history of contact and adult language learning can eliminate certain types of redundancy (cf. Trudgill 1999; 2001; and esp. forthcoming a). We also note along these lines that the fact that L2 varieties generally exhibit the smallest amount of grammatical marking resolves our earlier puzzle (cf. section 4) as to why L2 varieties do not exhibit particularly many L2-easy features. It just seems as though L2 speakers do not generally opt for "simple" features in preference to "complex" features. As a matter of fact, L2 speakers appear to prefer zero marking – and thus, possibly, hidden pragmatic complexity (cf. Walter Bisang's Chapter 3 above) – over explicit marking, be it L2-easy or complex.

The interplay between the different types of grammaticity is visualized as a scatterplot in Fig. 5.1. The Southeast/East Anglia and Hong Kong E are the extreme cases in our dataset. The former are highly analytic *and* synthetic

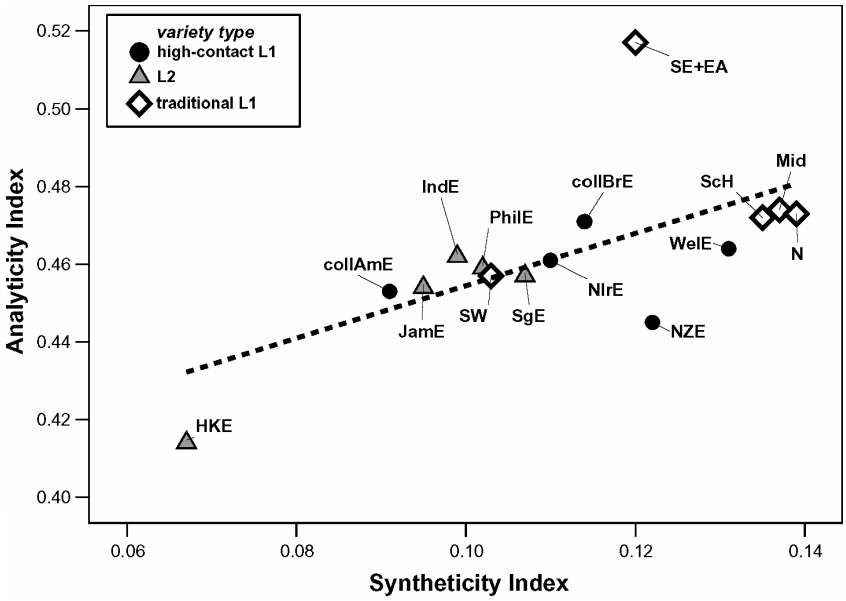

Fig. 5.1. Analyticity by syntheticity. Dotted line represents linear estimate of the relationship ($R^2 = 0{\cdot}40$).

varieties, while HKE is none of these things. In sum, the traditional L1 vernaculars are to be found in the upper right half of the figure (exhibiting above-average grammaticity), whereas L2 varieties are located in the lower left half, being neither particularly analytic nor synthetic. High-contact L1 varieties occupy the middle ground, along with the two standard varieties (collAmE and collBrE) and the Southwest of England. What merits particular attention in Fig. 5.1, we believe, is the slope of the dotted trend line: 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. Needless to say, such a trade-off is often claimed to be one that has governed the history of English. In reality, according to Fig. 5.1, analyticity and syntheticity correlate positively, so 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. e.g. Seuren and Wekker 1986).

6 Irregularity and transparency

This section will look more closely at the text frequency of bound grammatical morphemes, adding a twist to our findings on syntheticity in the previous section by distinguishing, additionally, between regular and irregular grammatical allomorphs. We propose that a given variety A is more complex than another variety B if, in naturalistic spoken discourse, variety A exhibits a greater percentage of bound grammatical allomorphs which are irregular and lexically conditioned (in turn, variety A would be *less complex* and *more transparent* if it exhibited a higher share of regular-suffixing allomorphs). The rationale is a theme in the complexity literature (e.g. McWhorter 2001a: 138) that while inflectional marking and syntheticity is not *per se* complex, it typically adds to linguistic complexity thanks to collateral "nuisance factors" such as allomorphy and morphophonemic processes (cf. Braunmüller 1990: 627).

We investigated all bound grammatical morphemes in our 15 × 1,000 token corpus database, classifying them into either (i) regular-suffixing, phonologically conditioned allomorphs (e.g. *he walk-ed*) or (ii) irregular, lexically conditioned allomorphs (e.g. *he sang*), and establishing corresponding *transparency indices*, which yield the share of regular allomorphs as a percentage of all bound grammatical morphemes (cf. Table 5.6). As can be seen, in typical L2 discourse, 82 per cent of all bound grammatical allomorphs are regular; the figure decreases to 71 per cent in the case of high-contact L1 varieties and to 65

Table 5.6. Mean transparency indices (percentage of regular-suffixing bound grammatical morphemes) by variety type

Variety type	Mean transparency index[a]
Traditional L1	.65
High-contact L1	.71
L2	.82

[a] highly significant at $p < \cdot 01$ (ANOVA: $F = 11{\cdot}31$).

per cent in traditional L1 vernaculars. The upshot, then, is that 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, occupy 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 2004a: 307).

To visualize the variance at hand here, the scatterplot in Fig. 5.2 plots transparency indices against grammaticity indices, i.e. the total text frequency of grammatical markers (see the preceding section for the technicalities). Traditional L1 vernaculars, displaying much grammatical marking but not being transparent, cluster in the lower right half of the diagram, whilst L2 varieties are to be found in the upper left half of the diagram, where one finds varieties characterized by low grammaticity but high transparency. As for the extreme cases, the Southeast and East Anglia are the single most verbose (in the sense of high grammaticity) and least transparent varieties in our sample; conversely, Hong Kong E exhibits, as we have seen already, the lowest levels of grammaticity overall, and Indian E turns out to be the most transparent, least irregular variety considered in our study. The Southwest of England, once again, rather patterns with the high-contact L1 varieties, and the two standard vernaculars (Standard American and British E) maintain a low profile – one that is akin to high-contact non-standard L1s – in every respect, a finding which is fully consonant with claims (cf. Trudgill forthcoming a) that these standard dialects constitute just another type of high-contact variety. The overall generalization emerging from Fig. 5.2, as suggested by the dotted trend line, is that transparency trades off against grammaticity – thus, morphosyntactic grammaticity implicates irregularity, and vice versa.

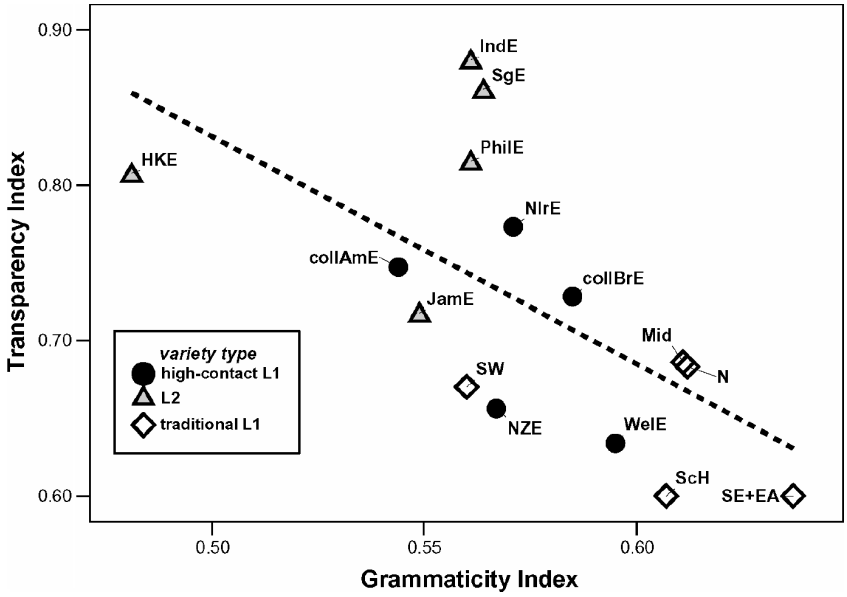

Fig. 5.2. Transparency by grammaticity. Dotted line represents linear estimate of the relationship ($R^2 = 0.38$).

7 Summary and conclusion

We have sought to present quantitative evidence in this study that variety type is a powerful predictor of complexity variance in varieties of English around the world.

In particular, we detailed how low-contact, traditional L1 vernaculars are on almost every count (ornamental complexity, grammaticity, irregularity) more complex than high-contact L1 varieties of English. Therefore, contact clearly is a crucial factor (cf. Trudgill 1999; forthcoming a). At the same time, "young", native varieties of English (i.e., English-based pidgins and creoles) turned out not to be any less ornamental than high-contact L1 varieties of English (partially contradicting e.g. McWhorter 2001a), although English-based pidgins and creoles are, according to our metric, more benign in terms of L2 acquisition difficulty than any other variety type in our sample.

This takes us to L2 varieties of English. Our implicit working hypothesis was that these should be objectively "simpler" than native varieties of English in terms of almost any measure. This hypothesis does not mesh with the facts, though. L2 varieties are not any less ornamental than, say, high-contact L1 varieties, and intriguingly, they also do not appear to be any L2-easier than L1

varieties. Instead, our corpus-based analysis of text frequencies of grammatical markers showed that while traditional L1 vernaculars are most redundant, in the sense that they attest high text frequencies of grammatical markers (be they analytic or synthetic), L2 varieties tend towards the lower, non-redundant end of the grammaticity spectrum. Specifically, L2 varieties do *not* trade off purportedly L2-difficult inflectional marking (cf. e.g. Klein and Perdue 1997; Seuren and Wekker 1986) for analytical marking. Rather, in spontaneous discourse L2 speakers appear to avoid grammatical marking of any kind, rather than opting for analytic marking or overtly L2-easy marking. Be that as it may, our analysis did indicate that L2 varieties are significantly more transparent than L1 varieties in that they exhibit the highest share of regular bound grammatical allomorphs. Hence, L2 varieties trade off grammaticity against transparency.

In terms of methodology, we should like to argue that language-internal variation is an ideal research site for developing, testing, and calibrating different complexity metrics. Varieties of English, in particular, constitute for a host of reasons (rich variation in sociohistorical settings, broad availability of data, and so on) an ideal opportunity to understand the comparatively simple (i.e. language-internal complexity variance) before probing the comparatively complicated (i.e. cross-linguistic complexity variance).

Appendix: the feature catalogue

For a version of the feature catalogue illustrated with linguistic examples, see Kortmann and Szmrecsanyi (2004: 1146–8).

Pronouns, pronoun exchange, and pronominal gender

1 *them* instead of demonstrative *those*
2 *me* instead of possessive *my*
3 special forms or phrases for the second person plural pronoun
4 regularized reflexives paradigm
5 object pronoun forms serving as base for reflexives
6 lack of number distinction in reflexives
7 *she/her* used for inanimate referents
8 generic *he/his* for all genders
9 *myself/meself* in a non-reflexive function
10 *me* instead of *I* in co-ordinate subjects
11 non-standard use of *us*
12 non-coordinated subject pronoun forms in object function
13 non-coordinated object pronoun forms in subject function

Noun phrase

14 absence of plural marking after measure nouns
15 group plurals

16 group genitives
17 irregular use of articles
18 postnominal *for*-phrases to express possession
19 double comparatives and superlatives
20 regularized comparison strategies

Verb phrase: tense and aspect

21 *wider range of uses of the Progressive*
22 habitual *be*
23 habitual *do*
24 non-standard habitual markers other than *do*
25 levelling of difference between Present Perfect and Simple Past
26 *be* as perfect auxiliary
27 *do* as a tense and aspect marker
28 completive/perfect *done*
29 past tense/anterior marker *been*
30 loosening of sequence-of-tense rule
31 *would* in *if*-clauses
32 *was sat/stood* with progressive meaning
33 *after*-Perfect

Verb phrase: modal verbs

34 double modals
35 epistemic *mustn't*

Verb phrase: verb morphology

36 levelling of preterite and past participle verb forms: regularization of irregular verb paradigms
37 levelling of preterite and past participle verb forms: unmarked forms
38 levelling of preterite and past participle verb forms: past form replacing the participle
39 levelling of preterite and past participle verb forms: participle replacing the past form
40 zero past tense forms of regular verbs
41 *a*-prefixing on *ing*-forms

Adverbs

42 adverbs (other than degree modifiers) have same form as adjectives
43 degree modifier adverbs lack *-ly*

Negation

44 multiple negation/negative concord
45 *ain't* as the negated form of *be*

46 *ain't* as the negated form of *have*
47 *ain't* as generic negator before a main verb
48 invariant *don't* for all persons in the present tense
49 *never* as preverbal past tense negator
50 *no* as preverbal negator
51 *was/weren't* split
52 invariant non-concord tags

Agreement

53 invariant present tense forms due to zero marking for the third person singular
54 invariant present tense forms due to generalization of third person *-s* to all persons
55 existential/presentational *there's, there is, there was* with plural subjects
56 variant forms of dummy subjects in existential clauses
57 deletion of *be*
58 deletion of auxiliary *have*
59 *was/were* generalization
60 Northern Subject Rule

Relativization

61 relative particle *what*
62 relative particle *that* or *what* in non-restrictive contexts
63 relative particle *as*
64 relative particle *at*
65 use of analytic *that his/that's, what his/what's, at's, as'* instead of *whose*
66 gapping or zero-relativization in subject position
67 resumptive/shadow pronouns

Complementation

68 *say*-based complementizers
69 inverted word order in indirect questions
70 unsplit *for to* in infinitival purpose clauses
71 *as what/than what* in comparative clauses
72 serial verbs

Discourse organization and word order

73 lack of inversion/lack of auxiliaries in *wh*-questions
74 lack of inversion in main clause *yes/no* questions
75 *like* as a focusing device
76 *like* as a quotative particle

References

Bisang, W. (this volume). On the evolution of complexity -- sometimes less is more in East and mainland Southeast Asia.

Braunmüller, K. (1990). Komplexe Flexionssysteme – (k)ein Problem für die Natürlichkeitstheorie? *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* **43:** 625-35.

Du Bois, J. W., W. L. Chafe, C. Meyer and S. A. Thompson (2000). *Santa Barbara corpus of spoken American English, Part 1.* Philadelphia: Linguistic Data Consortium.

Dulay, H. and M. Burt (1973). Should we teach children syntax? *Language Learning* **23:** 245-58.

Dulay, H. and M. Burt (1974). Natural sequences in child second language acquisition. *Language Learning* **24:** 37-53.

Greenbaum, S. (1996). *Comparing English worldwide: the International Corpus of English.* Oxford/New York: Clarendon Press/Oxford University Press.

Greenberg, J. H. (1960). A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics* **26:** 178-94.

Hawkins, R. (2001). *Second Language Syntax: A Generative Introduction.* Oxford: Blackwell.

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

Hyltenstam, K. (1984). The use of typological markedness conditions as predictors in second language acquisition: The case of pronominal copies in relative clauses. In Andersen, R. (ed.) *Second Languages.* Rowley, MA: Newbury. 39-58.

Kirk, J. (1992). The Northern Ireland Transcribed Corpus of Speech. In Leitner, G. (ed.) *New Directions in English Language Corpora.* Berlin/New York: Mouton de Gruyter. 65-73.

Klein, W. and C. Perdue (1997). The basic variety (or: Couldn"t natural languages be much simpler?). *Second Language Research* **13:** 301-47.

Kortmann, B., E. Schneider, K. Burridge, R. Mesthrie and C. Upton (eds.) (2004). *A Handbook of Varieties of English.* Berlin/New York: Mouton de Gruyter.

Kortmann, B. and B. Szmrecsanyi (2004). Global synopsis: morphological and syntactic variation in English. In Kortmann, B., E. Schneider, K. Burridge, R. Mesthrie and C. Upton (eds.) *A Handbook of Varieties of English.* Berlin/New York: Mouton de Gruyter. 1142-202.

Kusters, W. (2003). *Linguistic Complexity: The Influence of Social Change on Verbal Inflection.* Utrecht: LOT.

Littlewood, W. (2006). Second Language Learning. In Davies, A. and C. Elder (eds.) *The Handbook of Applied Linguistics.* Malden, MA: Blackwell. 501-24.

McWhorter, J. (2001a). The world's simplest grammars are creole grammars. *Linguistic Typology* **6:** 125-66.

McWhorter, J. (2001b). What people ask David Gil and why: Rejoinder to the replies. *Linguistic Typology* **6:** 388-413.

Mesthrie, R. (2004). Introduction – Varieties of English in Africa and South and Southeast Asia. In Kortmann, B., E. Schneider, K. Burridge, R. Mesthrie and C. Upton (eds.) *A Handbook of Varieties of English.* Berlin/New York: Mouton de Gruyter. 805-12.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In Miestamo, M., K. Sinnemäki and F. Karlsson (eds.) *Language Complexity: Typology, Contact, Change.* Amsterdam/Philadelphia: Benjamins.

Seuren, P. and H. Wekker (1986). Semantic transparency as a factor in creole genesis. In Muysken, P. and N. Smith (eds.) *Substrata versus Universals in Creole Genesis.* Amsterdam, Philadelphia: Benjamins. 57-70.

Szmrecsanyi, B. and B. Kortmann (forthcoming). Vernacular universals and angloversals in a typological perspective. In Filppula, M. and J. Klemola (eds.) *Vernacular Universals v. Contact-Induced Change.*

Towell, R. and R. Hawkins (1994). *Approaches to second language acquisition.* Clevedon, Philadelphia: Multilingual Matters.

Trudgill, P. (1999). Language contact and the function of linguistic gender. *Poznan Studies in Contemporary Linguistics* **35**.

Trudgill, P. (2001). Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* **5:** 371-4.

Trudgill, P. (2004). Linguistic and Social Typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* **8:** 305-20.

Trudgill, P. (forthcoming). Vernacular universals and the sociolinguistic typology of English dialects. In Filppula, M. and J. Klemola (eds.) *Vernacular Universals v. Contact-Induced Change.*

Wekker, H. (1996). *Creole languages and language acquisition.* Berlin, New York: Mouton de Gruyter.