

# Typological profiling

## Learner Englishes versus indigenized L2 varieties of English\*

Benedikt Szmrecsanyi and Bernd Kortmann  
University of Freiburg

Drawing on naturalistic corpus data, this study is an exercise in establishing typological profiles of learner varieties (as sampled in the *International Corpus of Learner English*) vis-à-vis indigenized L2 varieties of English (as represented in the *International Corpus of English*), though we also include in our dataset, for benchmarking purposes, a number of European languages as well as three stylistic varieties drawn from the *British National Corpus*. Our research is informed by two typological parameters widely used in the crosslinguistic classification of languages: overt *grammatical analyticity*, which we operationalize as the text frequency of free grammatical markers, and overt *grammatical syntheticity*, which we define as the text frequency of bound grammatical markers. The principal insight afforded by this study is that learner Englishes and indigenized L2 varieties of English have strikingly different typological profiles, a finding which we trace back to a number of grammatical markers whose function and frequency differs between the two groups. We also present a methodology to explore if learner Englishes are sensitive to typological properties of their substrate languages, and find that this is not generally the case.

### 1. Introduction

In this paper, our primary research interest lies with the typological profiles of learner Englishes (as sampled in the *International Corpus of Learner English*), on the one hand, and of indigenized L2 varieties of English (as represented in the *International Corpus of English*), on the other hand. To this purpose, we take an

---

\* We thank the following colleagues and student assistants for coding a number of European languages for their analyticity and syntheticity profiles: Alice Blumenthal (French), Johanna Gerwin (German), Stefan Madeja (Italian), and Tatiana Perevozchikova (Bulgarian, Czech, and Russian). All interpretational flaws are, of course, ours.

interest in the coding of grammatical information and draw on terminology, concepts, and ideas developed in quantitative morphological typology. Specifically, we will be concerned with two time-honored and well-known typological parameters, *analyticity* and *syntheticity*, which go back at least to August Wilhelm von Schlegel (cf., for instance, 1818). While this is not the place to review the rich history of thought about these notions that has unfolded since the 19th century, we feel compelled to point out here that the terms “are used in widely different meanings by different linguists” (Anttila 1989: 315). Thus, to fix terminology right at the outset, we define *formal grammatical analyticity* as comprising all those coding strategies where grammatical information is conveyed by free grammatical markers, which we in turn define as function words that have no independent lexical meaning. Conversely, we take *formal grammatical syntheticity* to comprise all those coding strategies where grammatical information is signaled by bound grammatical markers.

We have shown elsewhere (Szmrecsanyi & Kortmann 2009a; Kortmann & Szmrecsanyi 2009, to appear; Szmrecsanyi 2009) that variability along the analyticity-syntheticity continuum is, in fact, endemic among synchronic and short-term diachronic varieties of English. Specifically, our research has highlighted the fact that first, there is a good deal of geographic variation (for instance, Southeast Asian varieties of English are comparatively economical as far as the overall extent of grammatical coding is concerned); second, that we find significant differences according to variety type (for instance, low-contact, traditional English dialects tend to be more synthetic than other variety types); third, that there is substantial register variability (as a rule, written varieties of English prefer syntheticity, spoken varieties of English go for analyticity); and, lastly, that written English appears to have become more synthetic and less analytic over the past four decades or so.

In the present contribution, then, we aim to add learner Englishes to our variety portfolio. The primary research question that will guide our empirical analysis is whether learner Englishes and indigenized L2 varieties share typological properties thanks to certain concomitants of second language acquisition (SLA). For instance, much research in the SLA vein has emphasized that learners – especially in early interlanguage stages – avoid synthetic structures and opt for analytic marking whenever possible (see, for instance, Seuren & Wekker 1986; Wekker 1996; Klein & Perdue 1997). One would thus hypothesize that both learner Englishes and indigenized L2 varieties of English will exhibit less syntheticity and more analyticity than, e.g., standard British English reference varieties, all other things being equal. A secondary research question that we shall be concerned with is whether and to what extent typological profiles of individual learner Englishes are

conditioned on learners' native language background. To investigate such substrate effects, we shall present a methodology that will also involve profiling a number of European languages in terms of their analyticity and syntheticity levels.

In this connection, a few comments on our general methodological orientation are in order. First, note that this study is an exercise in typological profiling rather than error analysis, which is why we shall also remain fairly agnostic about the distinction between nativeness and non-nativeness. Second, we maintain that the appropriateness of an integrated model for learner Englishes and indigenized L2 Englishes, and the appropriateness of labels such as *English as a Second Language* (ESL) and *English as a Foreign Language* (EFL), is not an a-priori issue but rather an actual empirical question, which the present study will attempt to shed light on. Third, we claim that the analysis of naturalistic corpus data is but one method to profile varieties of English, which can and – we believe – should be complemented by, e.g., survey-based evidence in the spirit of Kortmann & Szmrecsanyi (2004) and Szmrecsanyi & Kortmann (2009a,b,c).

This paper is structured as follows. In Section 2, we present our dataset. In Section 3, we detail our empirical method. Section 4 will present our results. Section 5 offers a discussion of our findings and some concluding remarks.

## 2. Data

This study investigates 25 data points: 11 learner Englishes, 5 indigenized L2 varieties of English, 3 standard British English benchmark registers, and 6 European mother-tongue languages.

### 2.1 Learner Englishes

To study learner Englishes, we tapped the *International Corpus of Learner English* (ICLE) Version 1.1 (Granger 1998; Granger et al. 2002), a resource providing a large number of essays by advanced learners of English with different mother tongue backgrounds. We selected 11 subcorpora which sample texts by learners with the following native and first languages at home: Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. With a view to targeting medium-advanced learners of English, we typically only included essays by learners who had studied 5–6 years of English at school, 2–3 years of English at university, and who spent a maximum of 2 months in an

English-speaking country.<sup>1</sup> The 11 subcorpora thus selected span a total of approximately 266,000 words of running text.

## 2.2 Indigenized L2 varieties

To obtain data on indigenized L2 varieties of English, we turned to the *International Corpus of English* (ICE) (cf. Greenbaum 1996). Specifically, we were interested in the student essay and exam script sections (genre code w1a), thus matching as far as possible the text type sampled in ICLE. We accessed the following ICE components: ICE-East Africa, ICE-Hong Kong, ICE-India, ICE-Philippines, and ICE-Singapore, which left us with five data points on indigenized L2 varieties of English based on a total of 281,000 words of running text. The five L2 varieties thus included in this study are classic ‘New Englishes’ (Platt et al. 1984) or, in Kachru’s parlance, ‘outer circle’ varieties (e.g. Kachru 1985), even though they represent very different developmental stages.

## 2.3 Standard British English benchmark varieties

Building on a previously published dataset (cf. Szmrecsanyi 2009) drawn from the *British National Corpus* (BNC) (cf. Aston & Burnard 1998), we also included in our analysis three standard British English registers for benchmarking purposes: school essays (genre classification code W\_essay\_school; 147,000 words of running text), university essays (genre classification code W\_essay\_univ; 65,000 words of running text), and – as the lone spoken register subject to analysis in the present study – spontaneous face-to-face conversation (genre classification code S\_conv; approx. 4.3m words of running text).

## 2.4 European mother-tongue languages

To investigate the issue of substrate effects, we also profiled six European languages: Bulgarian, Czech, French, German, Italian, and Russian. For each of these languages, we drew on comparatively small corpora (approx. 10,000 words of running text each) sampling quality newspaper prose.

---

1. To obtain sufficiently large subcorpora, slight adaptations of these criteria were necessary for Finnish learner essays (adaptation: 4–7 years of English at school, 1–3 years at university, max. 4 months abroad) and Swedish learner essays (adaptation: 4–7 years of English at school, 2–4 years at university, max. 4 months abroad).

### 3. Method

Methodically, the present study is going to explore part-of-speech (henceforth: POS) frequencies to gauge typological profiles, utilizing an aprioristic but theory-informed categorization of POS categories to derive two frequency-based indices: an *analyticity index* and a *syntheticity index*. Our method is inspired by Joseph Greenberg's (1960) seminal paper entitled 'A Quantitative Approach to the Morphological Typology of Language'. Greenberg (1960) demonstrated that seemingly abstract typological notions can be sufficiently precisely measured by calculating a number of indices, on the empirical basis of naturalistic texts. Succinctly put, the present study will apply Greenberg's index method to the corpus material described in the previous section.

#### 3.1 Coding varieties of English

In the case of ICLE and ICE (which unlike the BNC are not POS-annotated in the first place), an algorithm selected 1,000 random tokens (i.e. orthographical words) per variety studied. This yielded a dataset of 16,000 word tokens (11 ICLE varieties plus 5 ICE varieties, multiplied by 1,000). Subsequently, all of these word tokens were annotated manually for their POS class using the BNC's CLAWS5 tag set (cf. Aston & Burnard 1998) with a minor extension.<sup>2</sup> The technicalities are discussed in Szmercsanyi (2009), a paper that also reports measures of interrater reliability and of the robustness of results deriving from 1,000-token samples. In the case of the BNC (a corpus that is POS-annotated in the first place, such that manual coding efforts are not a limiting factor), our results will be based on POS-frequencies not in random token samples but in the respective BNC texts as a whole.

Given our definition of analyticity and syntheticity offered at the beginning, POS-tags (or rather the tokens annotated with POS-tags) were placed into two relevant categories:

- *Analytic tokens*: conjunctions, subjunctions, and prepositions (tags CJ\*, PRF, PRP); determiners, articles, and *wh*-words (D\*, AT0, AVQ, PNQ); existential *there* (EX0); pronouns (PNI, PNP, PNX); the tokens *more* and *most*; the infinitive marker *to* (TO0); modals (VM0); the negator *not* (XX0), auxiliary BE ((A)VB\* + V\*, (A)VB\* + \* + V\*, (A)VB\* + XX0), auxiliary DO ((A)VD\* + V\*,

---

2. As in the CLAWS8 tagset, the primary verbs *be*, *do*, and *have* were explicitly annotated for whether they occurred in auxiliary function by prefixing the character 'A' to the CLAWS5 tag; note that in the analysis of the BNC itself, primary verbs were automatically disambiguated contextually for auxiliary or main verb usage.

- (A)VD\* + \* + V\*, (A)VD\* + XX0), and auxiliary HAVE ((A)VH\* + V\*, (A)VH\* + \* + V\*, (A)VH\* + XX0).
- *Synthetic tokens*: the *s*-genitive (POS); comparative and superlative adjectives (AJC, AJS); plural nouns (NN2); plural reflexive pronouns (PNX + word token ending in \*ves); inflected verbs ((A)V\*D, (A)V\*G, (A)V + N, (A)V\*Z).

Perl retrieval scripts were subsequently run on the dataset and established the text frequencies of the relevant POS-tags (or POS-tag categories), utilizing the above categorization to generate two Greenberg-inspired index values per data point as output:

- The *analyticity index*: the ratio of the number of free grammatical markers in a sample (F) to the total number of words in the sample (W), normalized to a sample size of 1,000 tokens. Hence:  $\text{ANALYTICITY INDEX} = F/W \times 1,000$ .
- The *syntheticity index*: the ratio of the number of words in a sample that bear a bound grammatical marker (B) to the total number of words in the sample (W), normalized to a sample size of 1,000 tokens. Hence:  $\text{SYNTHETICITY INDEX} = B/W \times 1,000$ .

Both indices have a lower bound of 0, and an upper bound of 1,000 index points.

### 3.2 Coding European mother-tongue languages

The method used to code the non-English data points was overall very similar to the method utilized to code varieties of English. An algorithm selected 1,000 random word tokens from each of the six 10,000 word corpora sampling written Bulgarian, Czech, French, German, Italian, and Russian. These randomly selected word tokens – in all, 6,000 – were then coded, typically by native speakers, (i) for whether or not they are function words (defined as conjunctions, subordinations, prepositions, determiners, articles, pronouns, infinitive markers, modals, negators, or auxiliary verbs), and (ii) for the number of bound grammatical markers borne by each token (note here that unlike English, Russian, for example, can affix several inflections to a single word token, in which case every one of those inflections loads on the syntheticity index). Retrieval scripts were then run on the coded random word token samples to calculate the indices.

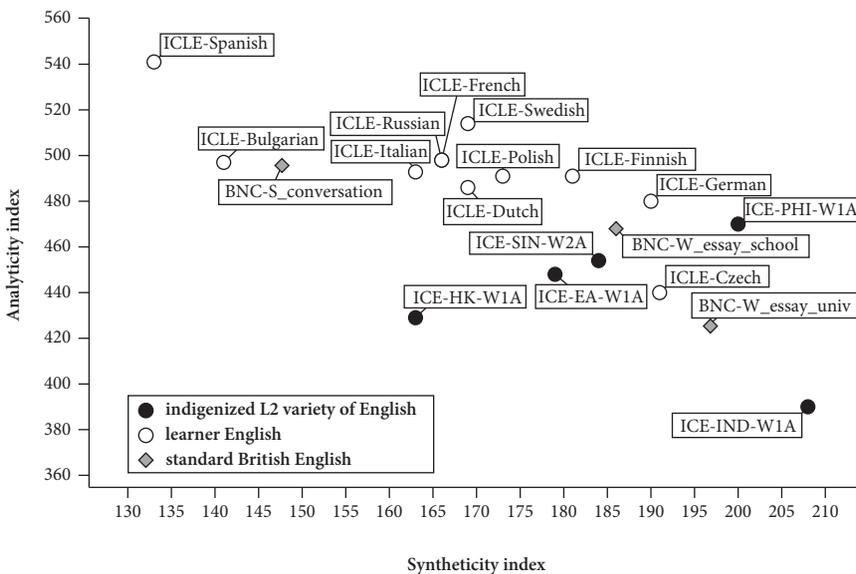
## 4. Results

We move on to a discussion of our results. Section 4.1. canvasses the big picture by projecting index scores to typological space. Section 4.2. deconstructs the index

scores to isolate grammatical markers that are implicated in the overall variability. Section 4.3. investigates the issue of substrate effects between learner Englishes and their respective substrate languages.

#### 4.1 The big picture

Figure 1 is a two-dimensional plane visualizing overall analyticity-syntheticity variability in typological space. The vertical axis plots analyticity index scores while the horizontal axis indicates syntheticity index scores. Thus, ICLE-Spanish, in the top left corner of the diagram, turns out to be the most analytic and least synthetic variety in our sample: the data point is associated with an analyticity index score of 541 (meaning that in 1,000 words, 541 words are function words) and a syntheticity index score of 133 (hence, of 1,000 words, 133 bear a bound grammatical marker). At the other end of the spectrum, in the bottom right corner of Figure 1, we find ICE-India as the most synthetic and least analytic variety in the sample (analyticity index score: 390, syntheticity index score: 208). Among the ICLE data points sampling learner Englishes, it is ICLE-Czech that stands out as the most synthetic (syntheticity index score: 191) and least analytic (analyticity index score: 440) learner variety. As for the ICE components sampling indigenized L2 varieties of English, we find that ICE-Hong Kong is the least synthetic data

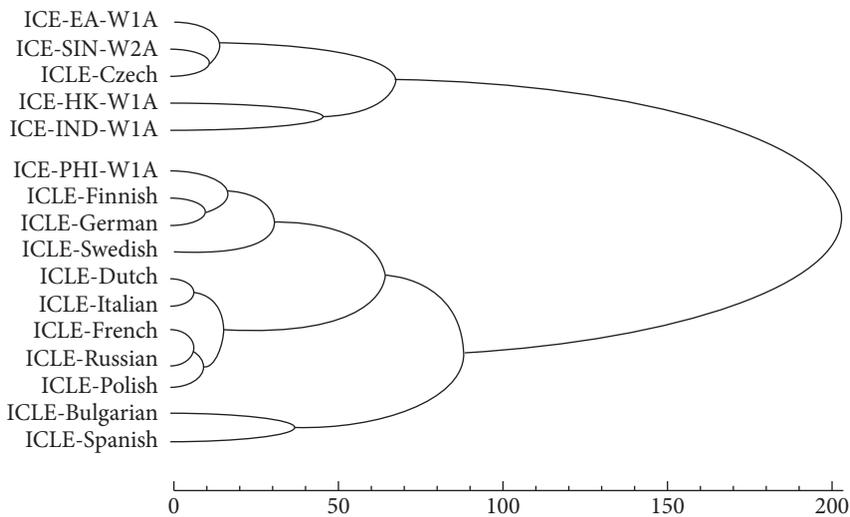


**Figure 1.** Analyticity by syntheticity (in index points, per thousand words): indigenized L2 varieties of English as sampled in ICE (black dots), learner Englishes as sampled in ICLE (white dots), and standard British English registers as sampled in the BNC (grey diamonds)

point (syntheticity index score: 163) while it is ICE-Philippines that yields the most analytic essay material (analyticity index score: 470). As for the three Standard British English registers (all drawn from the BNC) included in our inquiry, observe that these form a neat continuum from more analytic and less synthetic to more synthetic and less analytic: face-to-face conversation (BNC-S\_conversation) tends towards the analytic pole (analyticity index score: 496, syntheticity index score: 148), university essays (BNC-W\_essay\_univ) are situated close to the synthetic pole (analyticity index score: 427, syntheticity index score: 197), and school essays (BNC-W\_essay\_school) cover the middle ground (although they are a lot closer to university essays than to face-to-face conversation).

**Table 1.** Mean syntheticity and analyticity indices by variety type: indigenized L2 varieties of English (ICE) versus learner Englishes (ICLE) (significance of inter-group differences according to an independent samples *t*-test)

	indigenized L2 varieties of English	learner Englishes	significance of group difference
analyticity index	438	494	$t = -3.92 (p = .002)$
syntheticity index	187	167	$t = 2.01 (p = .064)$



**Figure 2.** Hierarchical agglomerative cluster analysis: indigenized L2 varieties of English as sampled in ICE versus learner Englishes as sampled in ICLE (cluster algorithm: Ward)

All this is another way of saying that there is, in our dataset, a good deal of variability along the analyticity and syntheticity dimensions. This variability notwithstanding, we must not miss an important generalization, which is that as a rule, learner Englishes are significantly more analytic than indigenized L2 varieties of English. There is also a tendency for indigenized L2 varieties to be more synthetic than learner Englishes. Table 1 details that the average ICLE variety has 56 more function words per 1,000 words than the average ICE variety. Conversely, the average ICE variety exhibits 20 more bound grammatical markers per 1,000 words than the average ICLE variety. Mean index values aside, Figure 1 moreover makes clear that the dispersion around the mean index scores displayed in Table 1 is sufficiently small to warrant treating ICLE varieties and ICE varieties as two fairly discrete and internally coherent variety groups. For the sake of backing up this impressionist assessment in a statistically more robust way, a supplementary cluster analysis – see Figure 2 for the resulting dendrogram – run on an Euclidean distance matrix derived from the  $16 \times 2$  index matrix confirms that on the whole, ICLE varieties and ICE varieties indeed split up quite nicely into two different clusters. There are, in fact, only two outliers. For one, ICLE-Czech is grouped with the indigenized L2 varieties; second, ICE-Philippines ends up in the ICLE cluster. The bottom line of all this is that the ICLE varieties and ICE varieties, as groups, are clearly two different animals with regard to their degrees of analyticity and syntheticity – they have different typological profiles with only minimal overlap.

How do the Standard British English registers fit into the picture? A glance at Figure 1 tells a simple story: Standard British English university essays are located right in the center of gravity of the ICE cluster, Standard British English face-to-face conversation is situated in the ICLE cluster, and school essays are to be found in the no man's land between the ICE and ICLE cluster. While we concede that conflating variety *and* register variation in one graph is not unproblematic, we still believe that this sort of linguistic geography allows for two interesting interpretations. It might be the case that ICE essay writers are simply better at targeting Standard British English norms (if indigenized L2 variety users actually are in the business of targeting 'standard' norms, an issue whose discussion is beyond the scope of the present study) than ICLE essay writers, who underuse synthetic marking and overuse analytic marking at the expense of proximity to target norms. In this interpretation, then, ICLE essays conform to the well-known anti-syntheticity SLA universal (Seuren & Wekker 1986; Wekker 1996; Klein & Perdue 1997) whereas ICE essays do not, and it is a mere accident – one that is due to the fact that spoken texts always tend to be more analytic and less synthetic than written texts (see Szmrecsanyi to appear) – that Standard British English face-to-face conversation is close to the ICLE cluster. An alternative interpretation of the facts at hand is that we are dealing here with a phenomenon known as *register interference*

(Aijmer 2002: 55): both ICE varieties and ICLE varieties are close, in their own ways, to Standard British English. It is just that ICLE writers (or the teachers that instruct them, or the text books used in the classroom) may not be fully aware of certain stylistic implications of analytic and synthetic modes of grammatical marking, which leads them to adopt inappropriately oral and conversational norms. Our data do not put us in a position to settle for good the question which interpretation is the correct one, but both probably have merit.

By way of an interim summary, we have seen in this section that learner Englishes as sampled in ICLE and indigenized L2 varieties of English as sampled in ICE have demonstrably dissimilar typological profiles. In short, ICLE varieties are more analytic and less synthetic than ICE varieties.

#### 4.2 Sources of variability

The task before us now is to identify those grammatical markers and/or marker categories which are responsible for the bulk of variability in index scores. To this end, we will deconstruct the indices considered in the previous section, exploring which of the component categories loading on the two indices discriminate between learner Englishes, on the one hand, and indigenized L2 varieties of English, on the other hand.

In exactly this spirit, Table 2 lists those five grammatical markers where we see significant or marginally insignificant frequency contrasts between the ICLE dataset and the ICE dataset. The most marked discrepancies concern the text frequency of pronouns, an analytic marker category. In indigenized L2 varieties of English, pronouns have a mean text frequency of 28 *ptw*; in learner Englishes, the frequency is more than twice this figure. It seems to us that in many cases, ICLE writers use finite subordinate clauses, as in (1) (*is that you must not wait* instead of *is not*

**Table 2.** Mean frequencies (in frequency per thousand words) of grammatical markers by variety type: indigenized L2 varieties of English (ICE) versus learner Englishes (ICLE); significant or marginally insignificant differentials only (significance of inter-group differences according to an independent samples *t*-test)

	indigenized L2 varieties of English	learner Englishes	significance of group difference
pronouns	28	58	$t = -6.77 (p < .001)$
negator <i>not</i> , <i>n't</i>	5	10	$t = -2.82 (p = .014)$
auxiliary <i>do</i>	1	4	$t = -2.31 (p = .037)$
auxiliary <i>have</i>	4	7	$t = -1.85 (p = .086)$
inflected verbs	120	107	$t = 1.82 (p = .090)$

*to wait*) while ICE writers tend to opt for the nonfinite – and pronoun-less – construction, as in (2): *is to fragment* rather than *is that he fragments*.

- (1) The most important thing here is that *you* must not wait until the child is “big enough” to learn <ICLE-Swedish FIAB1001>
- (2) His idea, in a nutshell, *is to fragment* an input text according to key words. <ICE-Singapore W1A-006>

It is particularly interesting to note that even those ICLE writers whose native language is pro-drop (Italian and Spanish) overuse pronouns vis-à-vis ICE writers: in ICLE-Italian, pronouns have a text frequency of 46 *ptw*, in ICLE-Spanish it is 53 *ptw*.

Moving down in Table 2, the negator *not* (including its contracted variant *n't*), likewise an analytic category, is twice as frequent in learner Englishes (mean frequency: 10 *ptw*) than in indigenized L2 varieties (mean frequency: 5 *ptw*). Browsing through the concordance lines, it appears that in many cases, learners' overuse of negators is a strategy to deal with certain limitations of their lexicon – in other words, we are likely to be dealing here with a lexically motivated structural difference. In (3), for instance, the ICLE writer uses the phrase *treatments we still do not have* (rather than its lexical alternative, *treatments we still lack*); the ICE writer in (4), however, does use the verb *lack* instead of the paraphrase *they do not have school fees*.

- (3) There are plenty of treatments we still *do not* have, for example, for HIV and cancer. <ICLE-Polish POLU1009>
- (4) Some time they *lack* school fees and the educational supply materials for learning. <ICE-East Africa Essays-T>

Two other analytic markers are substantially more frequent in the ICLE texts than in the ICE texts: auxiliary *do*, as in (3) above, and auxiliary *have*. The significantly higher frequency of auxiliary *do* in learner Englishes is probably related to the aforementioned tendency by learners to use negative paraphrases instead of their lexical alternatives (this is the issue of *lack* versus *do not have* discussed in the foregoing paragraph). The marginally insignificant overuse of auxiliary *have* in ICLE texts, on the other hand, is due to a preference for the perfect construction, which we regularly find in contexts that would be coded with the simple past in many ICE L2 Englishes. Compare example (5), where the ICLE writer employs the present perfect (*as I have said before*), to example (6), where the ICE writer uses the simple past (*as I said before*).

- (5) We can say that the man destroys not only the nature which surrounds him but also other people, as *I have said* before. <ICLE-Czech CZPU1006>

- (6) As *I said* before lingua francas arise when there is a need of communication between/among groups with different languages.

<ICE- East Africa Essays-K>

The last item in Table 2 concerns overtly inflected verbs, a category that loads on the syntheticity index. Inflected verbs have a text frequency of 120 *ptw* in ICE essays and 107 *ptw* in ICLE essays (this differential is marginally insignificant, but still warrants, we believe, the analyst's attention). In other words, learners comparatively often use the base form of lexical verbs instead of inflected forms, as in (7): *don't* and *benefit* (overt past tense marker absent) instead of *didn't* and *benefited* (overt past tense marker present).

- (7) There are several technological elements that one hundred years ago, *don't* exist, such as television, radio, video, cars, and all kind of machines in the job, in house, in everyplaces which have done us better the life. These improvements have *benefit* us in part, for example to make your life more comfortable ...

<ICLE-Spanish SPM04034>

At this point, it will be instructive to additionally investigate exactly which grammatical markers are most implicated in setting apart the two extreme varieties in our dataset, ICLE-Spanish (as the most analytic and least synthetic variety in our dataset) and ICE-India (as the most synthetic and least analytic variety). Let us begin by comparing ICLE-Spanish to the other ICLE varieties in our dataset. A series of chi-square tests of independence reveals that Spanish learners of English use significantly ( $p = .001$ ) fewer inflected verb forms than learners with other mother tongue backgrounds – example (7) above nicely illustrates this phenomenon. But compared to other ICLE writers, Spanish learners also significantly ( $p = .01$ ) overuse determiners. This phenomenon is exemplified in (8), where we find two noun phrases (*the society*, *the money*) where other writers would not necessarily employ a determiner.

- (8) From the beggining of *the society* there are the problem of *the money*.

<ICLE-Spanish SPM04048>

Turning to Indian English, a statistical analysis of the text frequency of grammatical markers in ICE-India reveals that writers in this corpus use explicit synthetic plural marking on nouns, as in (9) and (10), significantly ( $p = .01$ ) more frequently than other ICE writers.

- (9) Impacts Of Science On Human Life <ICE- India W1A-002>  
 (10) Business without planning would not bring desired effects or would not achieve the objectives of a business in fairly manner. The main objectives of a business (every) is to survive in the market. <ICE- India W1A-016>

Note in this connection that the use of plural *-s* with non-count nouns is a well-known characteristic of Indian English (cf., for instance, Kachru 1983; Bhatt 2008). We would also like to mention here that relative to other ICE essays, ICE-India essays exhibit significantly ( $p = .002$ ) fewer conjunctions, especially fewer subordinating conjunctions; a detailed discussion of this phenomenon is reserved for another occasion.

In sum, the results discussed in this section suggest that learner Englishes and indigenized L2 varieties of English differ primarily thanks to five grammatical categories with different usage patterns and frequencies in the two variety types: pronouns, the negator *not*, auxiliary *do*, auxiliary *have*, and verbal inflections. In a similar vein, we have probed those grammatical markers that make ICLE-Spanish and ICE-India special: determiners/articles and verbal inflections in the case of the former, and explicit plural marking and (subordinating) conjunctions in the case of the latter.

#### 4.3 Substrate effects?

We will now turn to the issue of whether we can predict a given learner variety's typological profile by considering the typological profile of the learners' mother tongue language – in other words, the question is if we can demonstrate substrate effects. To this end, we are going to relate the typological profiles of a convenience sample of six European languages (Bulgarian, Czech, French, German, Italian, and Russian) to the profiles of a matching subset of six ICLE varieties (ICLE-Bulgarian, ICLE-Czech, ICLE-French, ICLE-German, ICLE-Italian, and ICLE-Russian).

The relevant information can be gleaned from Table 3. Thus, for instance, ICLE-Bulgarian has an analyticity index of 497 and a syntheticity index of 141; the corresponding index values for (written) Bulgarian are 372 and 395, respectively. The problem is that these values cannot be compared directly, as variability between languages is substantially more pronounced than variability between ICLE varieties, as a cursory glance at the standard deviations reported in Table 3 makes clear. So, what we need is a way to normalize index values, and we will draw on *z*-score transformation to achieve this normalization.<sup>3</sup> Consider, again, ICLE-Bulgarian: the mean analyticity index in the ICLE dataset is 485, and the standard deviation in this dataset is 23; ICLE-Bulgarian's analyticity index score of 497 therefore translates into a *z*-score of .5 (which is another way of saying that ICLE-Bulgarian's analyticity index score is .5 standard deviations *above* the mean of all ICLE varieties under study). In a similar fashion, the analyticity index score of

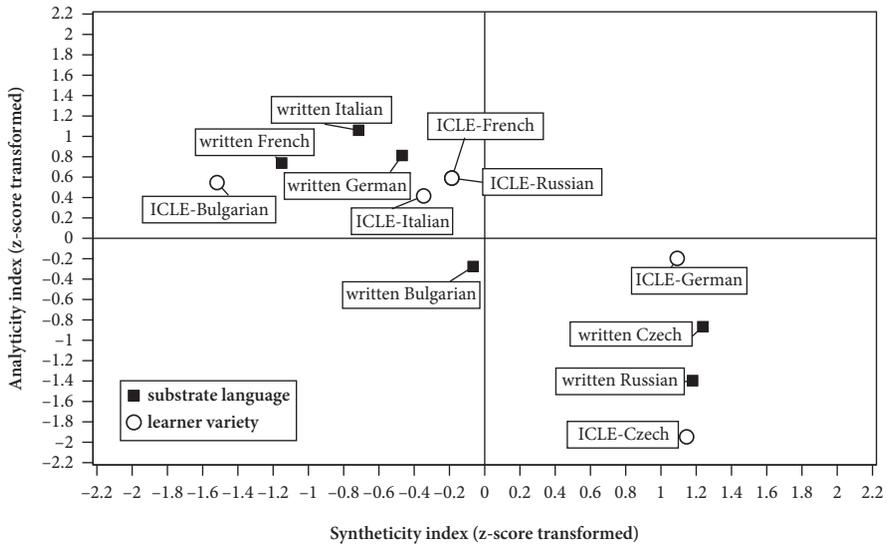
3. Notice that this sort of normalization is similar to speaker normalization of formant frequencies customary in acoustic phonetics.

**Table 3.** Mean index values and z-scores by ICLE variety and the respective substrate language. Z-scores were calculated on the basis of intra-group dispersion (ICLE varieties or substrate languages)

	analyticity index		syntheticity index	
	value	z-score	value	z-score
ICLE -Bulgarian	497	.5	141	-1.5
ICLE -Czech	440	-1.9	191	1.1
ICLE -French	498	.6	166	-.2
ICLE -German	480	-.2	190	1.1
ICLE -Italian	494	.4	163	-.3
ICLE -Russian	498	.6	166	-.2
<i>mean</i>		485		170
<i>standard deviation</i>		23		19
(written) Bulgarian	372	-.3	394	-.1
(written) Czech	334	-.9	683	1.2
(written) French	439	.8	153	-1.2
(written) German	436	.7	301	-.5
(written) Italian	458	1.1	250	-.7
(written) Russian	300	-1.4	670	1.2
<i>mean</i>		390		409
<i>standard deviation</i>		64		222

372 associated with (written) Bulgarian translates into a z-score of  $-.3$ : the mean analyticity index score in the European languages subset is 390, the corresponding standard deviation is 64 – and Bulgarian’s analyticity index score of 372 is .3 standard deviations *below* the mean value of 390. Interpretationally, we will consider matching signs of z-score-transformed index score pairings as a necessary – though not necessarily sufficient – condition for the assumption of substrate effects. Hence, we would argue that the typological profile of a substrate language X can be taken to have an effect on the typological profile of learner variety Y if *both* are more or less analytic than the respective group means, and likewise if *both* are more or less synthetic than the respective group means.

In Figure 3 we find a visualization of the data in Table 3. The diagram divides typological space in four quadrants (starting with the top left quadrant and moving clockwise): (I) above-average analytic/below-average synthetic, (II) above-average analytic/above-average synthetic, (III) below-average analytic/below-average synthetic, and (IV) below-average analytic/above-average synthetic (consider



**Figure 3.** Analyticity by syntheticity (in z-score transformed index scores): a subset of learner Englishes as sampled in ICLE (white dots) versus their respective substrate languages (black squares)

written Bulgarian: it is located in the below-average analytic/below-average synthetic quadrant since both of its index z-scores in Table 3 have a negative sign).<sup>4</sup>

Following our earlier definition of substrate effects, Figure 3 indicates substrate effects to the extent that learner varieties and their respective substrate languages end up in the same quadrants, thanks to matching signs of index z-score pairings. This is the case for exactly three learner variety/substrate language pairings:

- ICLE-French is above-average analytic and below-average synthetic, and so is written French;
- ICLE-Italian is likewise above-average analytic and below-average synthetic, as is written Italian;
- ICLE-Czech is below-average analytic and above-average synthetic, just as written Czech is.

ICLE-Bulgarian, however, is above-average analytic and below-average synthetic, while – as we have seen – written Bulgarian is above-average analytic and above-average synthetic. ICLE-Russian is similarly both above-average analytic and

4. As an aside relating to Figure 3, it is quite interesting to point out in this context that the above-average analytic/above-average synthetic and below-average analytic/below-average synthetic quadrants (II and III) are almost empty – this may be taken to suggest that the old idea of a trade-off between synthetic and analytic grammatical marking does seem to hold in our dataset.

below-average synthetic, but its substrate language is below-average analytic and above-average synthetic. Conversely, ICLE-German is below-average analytic and above-average synthetic whilst written German is above-average analytic and below-average synthetic.

We conclude that one might argue for substrate effects in Czech, French, and Italian learner English, but there is no empirical basis for assuming such effects in the case of Bulgarian, German, or Russian learner English. It is fair to say that this is a fairly mixed picture, which makes it hard to talk about *systematic* substrate effects in terms of grammatical analyticity and syntheticity across all learner varieties.

## 5. Discussion and conclusion

Typological profiling was the name of the game in this study, and has yielded four very clear results for the parameter *grammatical analyticity vs. syntheticity*. Three of these relate to the two primary research questions which informed this study: do learner Englishes and indigenized L2 varieties share typological properties thanks to certain concomitants of SLA, and to what extent are typological profiles of individual learner Englishes conditioned by the learners' native language background?

Concerning the first question, Figure 1 has made amply clear that, as hypothesized, the majority of learner varieties and even indigenized L2 varieties exhibit, on average, less syntheticity and more analyticity than Standard British English reference varieties. This tendency is far more pronounced for learner varieties, however, than for indigenized L2 varieties. The vast majority of learner varieties are both significantly less synthetic and more analytic than the two *written* Standard British English reference registers, school essays and university essays.

This leads to our second, far more important finding. Learner varieties and indigenized L2 varieties clearly exhibit different typological profiles: the former are significantly more analytic and also exhibit a tendency to be less synthetic than the latter. In exhibiting these characteristics, the two variety groups are both fairly discrete and internally coherent. Our study therefore provides another piece of empirical evidence, based on a large-scale comparison of synthetic vs. analytic coding strategies in grammar, which confirms the need for drawing a distinction between *English as a Foreign Language* (EFL) and *English as a Second Language* (ESL) varieties on purely structural grounds.

We have further suggested that the learner varieties as represented in ICLE can be argued to exhibit a phenomenon which has justly been labeled *register interference* (Aijmer 2002: 55). More exactly, our study replicated findings deriving from a range of ICLE-based studies on (largely) morphosyntax which have been

conducted since the late 1990s, all of which point to “the speech-like nature of learner writing” (Granger & Rayson 1998: 129) such “that aspects of the language in the corpus are more speech-like than comparable native English writing” (Aijmer 2002: 73; cf. also, e.g., Biber & Reppen 1998, Meunier 2000). Recall that according to our evidence, ICLE varieties are, on average, substantially closer to British English conversation than to school or university essays. It is, then, interesting that “the learners’ stylistic immaturity” (Granger & Rayson 1998: 130), which is the almost expected outcome of natural developmental factors (texts written by young or inexperienced native speakers likewise exhibit a strongly oral style) that are reinforced by educational factors (namely the dominant communicative approach to teaching English as a foreign language), can be observed not only at the concrete level of individual morphosyntactic categories, such as the use of articles, personal pronouns, special tensed and non-tensed verb forms, modals, passive, adverbials, subordinators, finite vs. non-finite adverbial and complement clauses. Instead, we appear to be able to gauge this immaturity also by exploring rather abstract and much more coarse-grained typologically inspired parameters, such as analyticity and syntheticity. We speculate that the reason why ICE writers are better at approximating to Standard (here: British) English essay writing conventions than ICLE writers is that ICE writers are subject to the long-term effect of being trained in English (essay) writing on a wide range of topics in many different school subjects in an English-medium education system.

Finally, we failed to detect systematic substrate effects such that the degree of analyticity and syntheticity exhibited by the learner’s L1 influenced the degree of analyticity and syntheticity of the relevant learner variety of English. For some ICLE varieties such effects can be postulated, for others not. Future research will have to show, however, to what extent substrate effects along the parameters investigated here may come to the fore when zooming in on individual (bundles of) morphosyntactic categories. Take, for instance, the NP and the effect which the notorious underuse of articles by Russian (or, more generally, East Slavic) learners of English will have on the analyticity index of the relevant learner variety.

In conclusion, let us take a step back and sketch some methodological implications and desiderata of this study for future research. This relates to SLA research on non-native (learner) varieties and native (indigenized) L2 varieties of English, on the one hand, and to our own research focus, on the other hand, i.e. typology-driven morphosyntactic profiling of varieties and variety types of English in terms of recurrent bundles or ‘conspiracies’ of morphosyntactic features, grammatical surface complexity, and analytic vs. synthetic strategies for coding grammatical information. In the latter context, we have endeavored here to add learner’s varieties of English as a sixth variety type to our portfolio of variety types consisting of low-contact L1 varieties, high-contact L1 varieties, indigenized L2 varieties,

English-based Pidgins, and English-based Creoles. In a range of previous studies we have demonstrated that each of these variety types exhibits distinct morpho-syntactic, complexity and analyticity/syntheticity profiles (see, e.g., Szmrecsanyi & Kortmann 2009a, Kortmann & Szmrecsanyi 2009). Since all of our previous research on World Englishes and English-based Pidgins and Creoles has so far largely been based on comparisons of spoken data, it will be most interesting to see how *spoken* learner varieties fit into the picture. Excitingly, this will be possible once the LINDSEI corpus (*The Louvain International Database of Spoken English Interlanguage*) is released, since from that point onwards comparisons with the spoken components of the ICE corpora for indigenized L2 varieties can be conducted. In the context of the present study and previous ICLE-based research, one will want to probe if in the spoken medium, too, learner Englishes and indigenized L2 varieties exhibit different typological profiles. The expectation is that the typological profiles should approximate each other considerably: since genre interference can no longer be relevant for the spoken learner varieties, at least on a structural level, both learner and indigenized L2 varieties should exhibit features characteristic of spoken Standard English. As a result of the expected approximation of the typological profiles of LINDSEI varieties and spoken ICE-L2 varieties, we might also expect the (structural) distance between these two variety types to be considerably smaller than between either of them and any of the other variety types we have investigated so far (e.g. low-contact L1 varieties, Pidgins, or Creoles).

But even LINDSEI will not be able to remedy one general drawback which many currently (or soon) available corpora sampling learner English suffer from – namely, the very fact that both LINDSEI and ICLE represent fairly advanced learner Englishes. It is part and parcel of the ICLE design that the data stem from about 20-year-old non-native “university undergraduates in English Language and Literature in their third or fourth year” (Granger 1998: 10), and for the sake of comparability, this is also the profile of the LINDSEI informants. This is not a problem, of course, for SLA research interested in advanced learner varieties. It is a big problem, however, if learner varieties are tackled from a completely different angle, as done by the present authors. Our prime interest in learner varieties was, and still is, to take them as a means to an end in order to learn more about the genesis and evolution of indigenized L2 varieties of English, on the one hand, and English-based Pidgins and Creoles, on the other hand. In previous research of ours (cf. Kortmann & Szmrecsanyi 2009, to appear, Szmrecsanyi & Kortmann 2009 a,b,c), we found confirmed claims by Trudgill (2001, 2009) and McWhorter (2001, 2007) to the effect that the grammars of Pidgins and Creoles are characterized by simplification processes, which in turn is taken to be a result of intensive language contact and rapid adult language acquisition (of English). From that point of view – investigating learner varieties of English to learn more about typical

strategies and grammatical patterns used by adult foreign language learners – what is desperately needed is collections of (primarily spoken) data for early adult learners of English, produced ideally in a natural, non-instructional environment. Longitudinal data of this kind were collected and investigated, for example, as part of a major SLA project in the 1980s under the auspices of the European Science Foundation. In five European countries, the SLA process of adult immigrants was documented for a period of 30 months for five L2s acquired by speakers of six L1s (for the methodology cf. e.g. Klein & Purdue 1997: 308–310 and Trévisé & Porquier 1986). For English as L2, the project provides data from two adult immigrants with Italian and Punjabi as their native languages. This is a most valuable starting-point, but similar data need to be collected for several groups of adult learners of English with different L1 backgrounds.

And there is yet another set of data for which corpora need to be compiled, since these data offer a second, equally important window on the genesis and evolution of (ultimately) indigenized L2 Englishes, namely data for their early stages. This means data from the first few generations of L2 speakers or, to use a different measure, from (different sub-stages of) the second and third stages of Schneider's (2007) well-known evolutionary cycle of postcolonial Englishes, i.e. the stages of exonormative stabilization (stage 2) and, especially, nativization (stage 3). The sooner the SLA and the World Englishes communities embark on these challenging corpus compilation projects in order to explore early SLA in general, and early SLA in (nativized as well as non-nativized) L2 Englishes in particular, the better.

## References

- Aijmer, K. 2002. Modality in advanced Swedish learners' written interlanguage. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* [Language Learning & Language Teaching 6], S. Granger, J. Hung & S. Petch-Tyson (eds), 55–76. Amsterdam: John Benjamins.
- Anttila, R. 1989. *Historical and Comparative Linguistics* [Current Issues in Linguistic Theory 6] Amsterdam: John Benjamins.
- Aston, G. & Burnard, L. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: EUP.
- Bhatt, R. 2008. Indian English: Syntax. *Varieties of English*, 4: *Africa, South and Southeast Asia*, R. Mesthrie (ed.), 546–562. Berlin: Mouton de Gruyter.
- Biber, D. & Reppen, R. 1998. Comparing native and learner perspectives on English grammar: A study of complement clauses. *Learner English on Computer*, S. Granger (ed.), 145–158. London: Longman.
- Granger, S. 1998. The computer learner corpus: A versatile new source of data for SLA research. *Learner English on computer*, S. Granger (ed.), 3–18. London: Longman.

- Granger, S. & Rayson, P. 1998. Automatic profiling of learner texts. *Learner English on computer*, S. Granger (ed.), 119–131. London: Longman.
- Granger, S., Dagneaux, E. & Meunier, F. 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Greenbaum, S. (ed.). 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon.
- Greenberg, J.H. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26(3): 178–194.
- Kachru, B.B. 1983. *The Indianization of English: The English language in India*. New Delhi: OUP.
- Kachru, B.B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In *English in the World: Teaching and Learning the Language and Literatures*, R. Quirk & H.G. Widdowson (eds), 11–30. Cambridge: CUP.
- Klein, W. & Perdue, C. 1997. The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research* 13: 301–347.
- Kortmann, B. & Szmrecsanyi, B. 2004. Global synopsis: Morphological and syntactic variation in English. In *A Handbook of Varieties of English*, B. Kortmann, E. Schneider, K. Burrige, R. Mesthrie & C. Upton (eds), 1142–1202. Berlin: Mouton de Gruyter.
- Kortmann, B. & Szmrecsanyi, B. 2009. World Englishes between simplification and complexification. In *World Englishes: Problems – Properties – Prospects* [Varieties of English around the World 40], L. Siebers & T. Hoffmann (eds), 263–286. Amsterdam: John Benjamins.
- Kortmann, B. & Szmrecsanyi, B. To appear. Parameters of morphosyntactic variation in World Englishes: Prospects and limitations of searching for universals. In *Linguistic Universals and Language Variation*, P. Siemund (ed.). Berlin: Mouton de Gruyter.
- McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 6: 125–166.
- McWhorter, J. 2007. *Language Interrupted: Signs of Non-native Acquisition in Standard Language Grammars*. Oxford: OUP.
- Meunier, F. 2000. A Computer Corpus Linguistics Approach to Interlanguage Grammar: Noun Phrase Complexity in Advanced Learner Writing. PhD dissertation, Centre for English Corpus Linguistics. Université catholique de Louvain, Louvain-la-Neuve.
- Platt, J.T., Weber, H. & Ho, M.L. 1984. *The New Englishes*. London: Routledge & Kegan Paul.
- Schlegel, A.W.v. 1818. *Observations sur la Langue et la Littérature provençales*. Paris: Librairie grecque-latine-allemande.
- Schneider, Edgar W. 2007. *Postcolonial English. Varieties of English Around the World*. Cambridge: CUP.
- Seuren, P. & Wekker, H. 1986. Semantic transparency as a factor in creole genesis. In *Substrata versus Universals in Creole Genesis* [Creole Language Library 1], P. Muysken & N. Smith (eds), 57–70. Amsterdam: John Benjamins.
- Szmrecsanyi, B. 2009. Typological parameters of intralingual variability: Grammatical analyticity vs. syntheticity in varieties of English. *Language Variation and Change* 21(3).
- Szmrecsanyi, B. & Kortmann, B. 2009a. Between simplification and complexification: Non-standard varieties of English around the world. In *Language Complexity as a Variable Concept*, G. Sampson, D. Gil & P. Trudgill (eds), 64–79. Oxford: OUP.
- Szmrecsanyi, B. & Kortmann, B. 2009b. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua* 119(11): 1643–1663.

- Szmrecsanyi, B. & Kortmann, B. 2009c. Vernacular universals and angloversals in a typological perspective. In *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, M. Filppula, J. Klemola & H. Paulasto (eds), 33–53. London: Routledge.
- Trévise, A. & Porquier, R. 1986. Second language acquisition by adult immigrants: exemplified methodology. *Studies in Second Language Acquisition* 8: 265–275.
- Trudgill, P. 2001. Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* 5: 371–374.
- Trudgill, P. 2009. Vernacular universals and the sociolinguistic typology of English dialects. In *Vernacular Universals and Language Contacts: Evidence from varieties of English and beyond*. M. Filppula, J. Klemola & H. Paulasto (eds), 304–322. London: Routledge.
- Wekker, H. 1996. *Creole Languages and Language Acquisition*. Berlin: Mouton de Gruyter.

