Sandra C. Deshors,
Sandra Götz and
Samantha Laporte (eds.)

Rethinking Linguistic Creativity in Non-native Englishes

BENJAMINS CURRENT TOPICS

98

Rethinking Linguistic Creativity in Non-native Englishes

Benjamins Current Topics

ISSN 1874-0081

Special issues of established journals tend to circulate within the orbit of the subscribers of those journals. For the Benjamins Current Topics series a number of special issues of various journals have been selected containing salient topics of research with the aim of finding new audiences for topically interesting material, bringing such material to a wider readership in book format.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/bct

Volume 98

Rethinking Linguistic Creativity in Non-native Englishes
Edited by Sandra C. Deshors, Sandra Götz and Samantha Laporte

These materials were previously published in *International Journal of Learner Corpus Research* 2:2 (2016).

Rethinking Linguistic Creativity in Non-native Englishes

Edited by

Sandra C. Deshors
Michigan State University

Sandra Götz
Justus Liebig University Giessen

Samantha Laporte
Université catholique de Louvain

John Benjamins Publishing Company
Amsterdam / Philadelphia

™ The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/bct.98

Cataloging-in-Publication Data available from Library of Congress.
ISBN 978 90 272 0146 1 (HB)
ISBN 978 90 272 6352 0 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Introduction

Articles

Linguistic innovations in EFL and ESL

Rethinking the linguistic creativity of non-native English speakers

Sandra C. Deshors, Sandra Götz and Samantha Laporte
Michigan State University / Justus Liebig University Giessen /
Université catholique de Louvain

1. Introduction

The distinction between English as a native language (ENL), English as a second language (ESL) and English as a foreign language (EFL) has exerted an enormous influence on the modeling of Englishes worldwide (see Kachru 1982, 1985). In ENL and ESL contexts, English is used widely and 'naturally' for intranational purposes, while in EFL contexts English is taught and learned primarily as an international means of communication. In previous research, 'institutionalized' ESLs (such as Singaporean and Indian English; also referred to as New Englishes) and EFLs (such as French- and German-English interlanguages) have usually been treated as fundamentally different categories in different research paradigms. Despite an early call for a rapprochement between EFL and ESL to "bridge the paradigm gap" (Sridhar & Sridhar 1986) between the two research areas, it was not until 2008 that corpus linguists met for the first time to discuss possible ways to bridge the gap and to set an agenda for the development of more integrated approaches to EFL and ESL (Mukherjee & Hundt 2011). Since then, corpus-based research in both learner and second-language Englishes has undergone a significant shift and increasing efforts have been dedicated to bringing together research on EFL and ESL. Already, this has led a number of analysts to suggest that "the distinction between EFL and ESL should be viewed as a continuum" (Gilquin & Granger 2011: 56; see also Deshors 2014; Nesselhauf 2009).

However, approaching linguistic innovations from the perspective of this continuum raises interesting questions and challenges that invite us to explore how innovative non-native English speakers actually are when using their L2, to what extent the EFL and the ESL speaker populations can be investigated contrastively when it comes to assessing their linguistic creativity, and how this creativity can

https://doi.org/10.1075/bct.98.01des

be investigated using corpus data. While those questions certainly have a place in the wider discussion of how to bridge the paradigm gap, crucially, they are also opening up new directions for corpus-based research on learner Englishes as well as New Englishes. In an attempt to address those questions, we organized a pre-conference workshop on the occasion of the 36th ICAME conference at Trier University on May 27th 2015. Together, due to the variety of topics they cover, these papers portray innovations as being a multifaceted linguistic phenomenon. Ultimately, it is our hope that, collectively, those papers will provide an opportunity for scholars to pause and rethink what it means for language learners and second language users to be innovative in their L2. With this purpose in mind, this introduction aims to take stock of linguistic innovations in two ways. First, by defining the notions of errors and innovations and, second, by considering how those notions have so far been approached in EFL and ESL (Section 2.1). We will particularly keep in mind the dividing line between what stands as an error and what counts as an innovation, which will lead us to discuss the status of English learners as innovative L2 users (Section 2.2). In the remainder of the introduction, we will discuss corpus resources and the types of corpora that are best suited to capture innovations (Section 3), consider the emergence and the development of innovations and how they can be best explained (Section 4). Finally, we will summarize to what extent innovations have been shown to differ, if at all, across the EFL and ESL speaker populations. Concretely, we will show how so far state-of-the-art research on linguistic innovations has helped us capture their structural variation patterns and how innovations are generally perceived (Section 5).

2. Errors vs. innovations

2.1 Where should we draw the line?

Although "[t]he line is thin between errors and creative uses" (Gilquin & Granger 2011: 72), the distinction between the notions of error and innovation is essential to understand whether and how new varieties develop new conventions (Van Rooy 2011). However, despite the central aspect of this distinction in any discussion on linguistic creativity in L2, the dividing line between the two notions remains, to a large extent, very unclear. Throughout the literature, there is often an indeterminacy between what counts as an innovation and what is regarded as an error (Bamgbose 1998). As a result, it is somewhat difficult to assess, with precision, to what extent the deviation of a linguistic pattern from a native norm constitutes — or not — a characteristic feature of a particular type of non-native English (Hamid & Baldauf Jr 2013). Broadly, this lack of a clear-cut distinction between the two notions emerges from the fact that because they do not belong to the linguistic norm of the English language (Kachru 1982: 62), errors are generally considered unacceptable by native speakers. In addition, although innovations tend to be recognized as allowable deviations from the native English norm (Bennui 2013), there is, to date, no set criteria that objectively allow analysts to set errors and innovations apart. Further, in contrast with innovations that tend to result from a productive process and that, in that sense, are considered "systemic within a variety" (Kachru 1982: 62; see also Buschfeld 2013; Mollin 2006), errors tend to reflect gaps in a learner's knowledge.[1] Given this context, a main but yet unresolved issue that blurs a clear distinction between errors and innovations is how much deviation from the norm is acceptable (Kachru 1982: 61–62).

Traditionally, in order to draw the line between errors and innovations, scholars have relied on theoretical frameworks such as Kachru's (1982) Three Circles Model. Broadly, the three concentric circles, the Inner, Outer and Expanding Circles, represent patterns of acquisition, functional domains in which English is used across cultures and languages as well as types of spread (Kachru 1985). Concretely, the Inner Circle includes Englishes used as a mother tongue (e.g. British English, American English, Australian English) and the Outer Circle is composed of Englishes used in former British and American colonies and which are acquired in a relatively naturalistic environment. In contrast, the Expanding Circle includes EFLs primarily learnt as a Lingua Franca in classroom settings. Importantly, the model assumes that EFL and ESL differ in that EFLs are intrinsically norm-dependent and ESLs are norm-developing. In other words, ESLs "have a potential to develop their own norms and standards which are generally accepted as being characteristic features of a 'new' English variety" (Mukherjee 2010: 219). According to Kachru (2006: 91), this process is made possible by the fact that "[t]he substrate languages and the target language enhance each other's style potential and release creative energies of a language in a unique way". In contrast, EFLs are norm-dependent in the sense that "foreign learners are bound to orient themselves towards exonormative standards set by speakers outside their own speech community" (Mukherjee 2010: 238).

The general reliance on Kachru's model has had two important repercussions in the way linguistic creativity has so far been approached in non-native Englishes:

[1] While a discussion on the distinction between an error and a mistake is beyond the scope of this paper, we direct the reader to Corder (1967) for a summary of the features that characterize both phenomena. In a nutshell, Corder (1967) argues that errors result from a lack of L2 knowledge, that they are systematic, that learners are unaware of them and that they reflect deficits in a learner's competence. In contrast, mistakes tend to be slips of the tongue, temporary, often realized by learners who are able to fix them and they tend to merely reflect a performance phenomenon.

first, it has triggered a division of innovations and errors primarily based on the institutional status of the EFL or ESL in which they occur and second, resulting from this categorical division, it has encouraged a somewhat systematic labeling of potential linguistic innovations as deviations and thus errors in EFL, and as innovations in ESL. For instance, while Indian English has been shown to yield some of its most creative forms and structures on the lexico-grammatical level in speakers' innovative uses of prepositional verbs, ditransitive verbs and light-verb constructions (Mukherjee 2010; Mukherjee & Hoffmann 2006), within the paradigm of EFL research, linguistically very similar forms have mainly been associated with errors rather than innovations (see Mukherjee 2010). Thus, emerging from this distinction is the question whether (and if so to what extent) foreign language learners can (fully) receive any recognition for their linguistic creativity (Bamgbose 1998), given that their linguistic structures may coincide with those labeled as innovations in ESL (Edwards 2014a).

2.2 Towards a recognition of EFL users as innovative L2 speakers

As part of the ongoing collective effort to bridge the paradigm gap, a handful of recent (corpus) studies have already begun to challenge the above-described dichotomy between errors and innovations as well as the general view that the distinction between innovations and errors should solely rely on institutional status (Bruthiaux 2003; Deshors 2014; Edwards 2014a; Edwards & Laporte 2015; Gilquin 2011, 2015; Laporte 2012; Li & Mahboob 2012). Generally, this has been done in several different ways: empirically, methodologically and theoretically. Empirically, a number of scholars have started to draw parallels between EFL and ESL (Davydova 2012; Deshors 2014; Edwards 2014a; Gilquin 2011; Götz & Schilk 2011; Laporte 2012; Nesselhauf 2009). In the case of Gilquin (2011: 5), for instance, it emerges that "some innovations are […] shared by World Englishes, as for example the phrasal verb *cope up (with)*, which is identified by Platt (1989) as a typical feature of Singapore English, but actually occurs in other indigenized varieties of English as well as in learner Englishes". Similarly, Laporte (2012: 285) finds that "prepositional uses are very prone to innovation, and this, across a wide range of non-native populations, be they ESL or EFL". Methodologically, sophisticated approaches to corpus analysis such as multifactorial approaches as illustrated in Deshors (2014), have demonstrated how rewarding regression modeling is when used to study EFL and ESL in a unified way and how such approaches should be considered in order to investigate more closely than ever the notion of error vs. innovation in EFL. Finally, at a more theoretical level, studies such as Bruthiaux (2003), Li & Mahboob (2012) or Mukherjee & Hundt (2011) have questioned the suitability of theoretical frameworks based on historical and geographical legacy to accommodate discussions of language varieties. Importantly, the above body of research has already started to change the way we collectively approach (advanced) EFL learners by attributing to the learners more creative abilities than before (Gilquin & Granger 2011). Two main contributing factors can explain this important shift, namely the recognition that (i) both EFL and ESL share a number of innovations (increasing the credibility of EFL learners in terms of their own ability to be creative in their L2) and (ii) the fact that English is gradually playing an increasingly important role in identity construction and transcends its typical EFL functions. In this regard, Gilquin & Granger (2011: 75) present Tswana-English interlanguage as an interesting case of learner English that "shares features with both inner/outer circle varieties of English and […] varieties of the expanding circle" (see also Edwards (2014b) for an in-depth illustration of the case of English in the Netherlands). Crucially, with all the above-mentioned developments, scholars are now in a position to portray the creative potential of EFL learners with much sharper contours. Further, as a result of those developments, a range of 'new' research questions have started to emerge, such as: What do innovations look like in EFL and ESL? How do they compare and how are they perceived? How can we explain the emergence and development of innovations? How can corpora and corpus resources help us capture innovations? Ultimately, we consider these questions to be a valuable starting point to rethink the linguistic creativity of EFL and ESL users and we will address each of those questions in the remainder of this paper.

3. Exploiting corpus resources to capture innovations

Corpora represent a particularly rewarding data type for the study of innovations. Contrary to experimental data characteristic of the SLA paradigm, they offer access to contextualized and naturally produced language use that is representative of a particular population. This makes corpora an ideal resource to uncover (potential) innovations. However, they only provide an indirect means towards identifying innovations: one first needs to unearth phenomena of interest and later, whatever feature a corpus reveals, it is ultimately the analysts' call to label a structure an innovation, an error or a mistake.

To capture the new structures that corpora (may) host, we need to find ways to best exploit the corpora at our disposal. This has often been done in a top-down fashion by taking a specific lexical item as a starting point to look for innovations (e.g. Nesselhauf (2009) selects a number of specific prepositional verbs and phraseological chunks). Other studies have however relied on a more data-driven approach by capitalizing on annotated data and using automatic procedures that allow less expected innovations to surface. For example, Mukherjee & Hoffmann (2006) make use of a part-of-speech (POS) tagged corpus to identify and retrieve new ditransitive verbs. Resorting to parsed data, Schneider & Zipp (2013) and Schneider & Gilquin (this volume) automatically retrieve a wide range of new prepositional verbs (e.g. *join into* in ICE-Fiji, or *study about* in ICE-India), thereby (i) complementing the limited set of new prepositional verbs previously identified via lexical searches and (ii) offering a better appraisal of verb-preposition combinations in the data at hand.

The question that arises after extracting 'new' structures is whether these structures qualify as innovations, for which systematicity is often considered a prerequisite (see Section 2). As low-frequency phenomena that exist alongside standard forms (Mukherjee 2010), innovations represent a significant challenge for corpus linguists, namely that of establishing which linguistic forms yield traces of systematicity and are therefore likely to develop and ultimately qualify as full-fledged innovations. Just as some rare but conventional forms of British English appear only once (if at all) in the British component of the *International Corpus of English* (ICE-GB) (Greenbaum & Nelson 1996) or the *British National Corpus* (BNC 2007) (in morphology, for instance, a number of words ending with the suffix *-ness*, such as *overtness* or *effortlessness*, are hapax legomena in the BNC, but are recorded in the *Oxford English Dictionary* (OED 2015) and are thus conventional forms), so it remains to be determined whether rare instances of 'new' structures are used systematically in the speech community. This is compounded by the fact that most available corpora of EFL and ESL are (i) of limited size and (ii) synchronic in nature, which makes it difficult to trace the evolution of innovations (Gilquin 2015). However, despite these hurdles, a number of (new) corpora make it possible, at least in part, to overcome these difficulties.

One of these corpora is the recently developed *Corpus of Global Web-based English* (GloWbE; Davies 2013). As a mega-corpus of 1.9 billion words collected from the web and representing English as used in twenty different countries (traditionally Inner and Outer Circle countries), it is a goldmine for research into innovations. If only by its size, this database makes it, at least to some extent, possible to verify the systematicity of features captured in smaller corpora. In addition, with data produced in the 2000s and collected in 2012, GloWbE also makes it possible to trace the evolution of innovations uncovered on the basis of smaller corpora that represent data from the 1990s, such as the ESL subcorpora of the *International Corpus of English* (ICE; Nelson 1996).[2]

[2] See Davies & Fuchs (2015) for a discussion of the pros and cons of this database, and responses by Mair (2015), Mukherjee (2015), Nelson (2015) and Peters (2015).

However, beyond size, there are other important aspects of corpora that can help researchers capture innovations. As Mukherjee (2010) argues, newspaper corpora provide a different way of legitimizing a form as an innovation. Given the acrolectal, highly monitored, and even norm-providing nature of newspaper language, even low-frequency structures can be identified, with relatively high confidence, as accepted forms and thereby labeled innovations (e.g. the verbs *explain*, *inform* or *remind* as new ditransitive verbs which occur only a few times in the *Statesman Corpus* (Mukherjee & Hoffmann 2006), a 31-million-word newspaper corpus of Indian English). In a similar vein, Van Rooy & Kruger (this volume) use parallel corpora of edited and unedited versions of academic texts which, more than ever before, make it possible to trace the dynamic process of the emergence and acceptance of innovative linguistic structures.

Finally, corpus resources also play a crucial part in bridging the paradigm gap. One current challenge concerns data comparability between EFL and ESL. The price for high comparability is often a restriction to student writing due to the fact that most corpus data for EFL stem from the Learner Corpus Research framework. One notable exception is Edwards's (2014b) *Corpus of Dutch English* (CoDE), which is the first EFL corpus to follow the same design as the written component of the ICE and thereby covers a wide range of genres such as creative writing, written correspondence, and press reports and editorials. The development of CoDE is in line with the view that EFL speakers are users rather than merely learners (as is also core in the English as a Lingua Franca (ELF) framework). This view has led to the emergence of ELF corpora that represent a wider range of written and spoken registers (e.g. the *Corpus of English as a Lingua Franca in Academic settings* (ELFA 2008), *Vienna-Oxford International Corpus of English* (VOICE 2009), *Corpus of Academic Spoken English* (CASE, Diemer et al., 2018), the *Corpus of English in Finland* (Laitinen 2010), thus providing data that will make comparisons between ELF, EFL and ESL increasingly possible across a number of genres.

4. Explaining the emergence and development of innovations

A better understanding of innovations in non-native Englishes requires exploring the processes that lead to their emergence and later their adoption by a speech community. Croft (2000) proposes a usage-based theory of language change that offers an integrated explanation for these processes, irrespective of the status of a language in the speech community. In a nutshell, he argues that language change is the result of two distinct, but jointly required, mechanisms: (i) a mechanism for **innovation**, understood here as any "creation of novel forms in the language" (2000: 4), even if only ephemeral; and (ii) a mechanism for **propagation**, which is a selection mechanism that is largely driven by social forces and leads to the conventionalization of certain innovations. The following sections respectively focus on each of these mechanisms and consider how they relate to innovations in EFL and ESL.

4.1 The emergence of innovations

According to Croft (2000: 8), any innovation involves some sort of restructuring between language form (or structure) and language function (or meaning). This restructuring process is rarely random or accidental. Rather, it is likely to occur with a certain systematicity as a result of intra- and extra-lingual processes. More specifically, the mechanism for innovation seems driven by a combination of (at least) (i) cognitive processes that lead to certain types of restructuring (e.g. analogy); (ii) language-internal structures and irregularities (e.g. *talk about* sth. vs. *discuss* ø sth.) that facilitate the emergence of certain innovations; and (iii) language contact and transfer from another language. While Croft identifies these processes as driving language change in general, that is, also in native-speaker settings, this section attempts to explain how these mechanisms operate to lead to innovations outside of L1 settings in particular.

A number of specific **cognitive processes** have been argued to underlie the emergence of innovations found in EFL and ESL. For instance, drawing on cognitive mechanisms identified in Second Language Acquisition, Williams (1987), and more recently Schneider (2012), list a number of processes that are likely to be shared by EFL and ESL speakers and to give rise to new forms. These are processes such as regularization (e.g. the use of the plural *mouses* instead of *mice*), redundancy (e.g. redundant prepositions as in *enter into*), or simplification (omission of the noun plural marker -*s*).[3] Van Rooy (2011) argues that for such processes, EFL and ESL speakers are not qualitatively different from each other because in both settings, their cognitive representation is that of a second language. However, while there is certainly ground for shared cognitive processes, this common cognitive representation across EFL and ESL speakers might arguably be a relative rather than absolute one. For example, Szmrecsanyi & Kortmann (2011) test the hypothesis that EFL and ESL, due to similar cognitive processes of second language acquisition, are more analytical than ENL. They however find EFL to be significantly more analytical than ESL, which suggests some differences in terms of cognitive processes and leads the authors to even argue that EFL and ESL are "different animals" (2011: 175).

[3] As Schneider (2012) himself notes, these processes may well overlap: regularization can be construed as a special case of simplification, for example.

Interestingly, the above-mentioned cognitive processes are likely to interact with **language-internal configurations** that facilitate the emergence of new forms. That is, some irregularities in form and meaning intrinsic to (standard) English enhance the possibility for processes like regularization or analogy and thus favor particular kinds of innovations. The previously mentioned lexis-grammar interface (see Section 2) has been found to constitute a fertile breeding ground for innovations in non-native Englishes, exactly for this reason. One case in point is Mukherjee (2010) who shows how lexico-grammatical innovations such as new prepositional verbs, new light-verb constructions and new ditransitive verbs are cases of what Mukherjee & Hoffmann (2006: 166) have dubbed "semantico-structural analogy". The term itself highlights the fact that there is a re-mapping between form and function by drawing on existing formal and semantic templates, that is language-internal structures. For example, the new light-verb construction *have/take a glimpse* found in Indian English is based on the formal template of *catch a glimpse*, and the semantic template of *have/take a look*. Phenomena that arguably arise from the same process have also been identified in EFL, e.g. Nesselhauf (2005) finds *give a statement* in EFL data, which can be analysed as based on the formal template of *make a statement* and the semantic template of *give a speech*.

Finally, another important process that drives the emergence of innovations in non-native settings is that of **language transfer** or **substrate influence**. Non-native speakers, be they EFL or ESL, come with their L1-specific form-meaning structures that are likely to influence and interact with the above-mentioned forces, sometimes facilitating them, sometimes constraining them (Nesselhauf 2009). For example, Edwards & Laporte (2015: 21–22) show that there is an "intricate interplay between shared tendencies stemming from language internal (ir)regularities and L1 influence that accounts for pockets of idiosyncrasy in some varieties". Such observations warrant further research to uncover how exactly these interact.

4.2 From emergence to conventionalization

The emergence of an innovation does not *per se* lead to its adoption or conventionalization. Following Croft's (2000) account of language change, after a new form is created, that form undergoes a process of **propagation**. In a nutshell, this process involves social forces that determine whether innovations are ultimately adopted, that is, whether they become systematic and conventionalized in a language community.

As pointed out above, these social considerations have been at the core of most studies focusing on ESL. For example, Schneider's (2003, 2007) Dynamic Model of the Evolution of New Englishes highlights how adopting innovations goes hand in hand with the social process of identity-construction and an increasingly endonormative attitude of speech communities. Importantly, different social forces can pull the fate of innovations in different directions. This is, for instance, illustrated by Rosen (2014, this volume), who shows how innovations in Jersey English are developing in response to antagonist social forces such as local identity on the one hand, and pressures of globalization on the other. In addition to social factors, Schneider (2007: 110–112) adds that linguistic factors relating to the nature of innovations (such as markedness, transparency, regularity or salience of innovations) may also influence their propagation. For example, salient new features are more likely to spread and be adopted than non-salient ones.

When it comes to the conventionalization of innovations, such social factors have been argued to be at the root of the most important and long-lasting difference between EFL and ESL. According to Van Rooy (2011), (i) there is an identity dimension at play in ESL that is not present in EFL, (ii) there is greater opportunity for diffusion in ESL settings, and (iii) there is a more endonormative attitude in ESL, while in EFL settings, speakers' attitude is largely exonormative (see also Section 5). Ultimately, for Van Rooy, these crucial differences are what lead to the spread and conventionalization of innovations in ESL settings and not in EFL ones.

However, the dynamics of English worldwide appear to be gradually changing in response to new forces of globalization. Edwards (2014b), for example, shows how in the Netherlands, an Expanding Circle nation, English adopts increasingly intranational functions (e.g. in education, advertising, or business) and is a means of identity expression among young Dutch people. Similarly, Schneider (2014: 24) notes that "[we] can observe many innovative uses and sociolinguistic settings in which English is […] 'crossing' clear-cut distinctions and traditional taxonomies, defying standard norm-orientations, and transcending boundaries of language and nation as distinct entities". Although it is reasonable to expect that these new dynamics of English are likely to affect the propagation and status of linguistic innovations worldwide, at this point it is too soon to anticipate how exactly these developments will manifest linguistically.

From the above considerations, it seems clear that the emergence and development of innovations is a dynamic process in which linguistic and ever-changing social forces play an important part. The complexity of this dynamic process, in our view, calls for sound (corpus) studies that provide an empirical basis for the investigation of innovations, but also highlights the imperative to abstract away from these empirical studies in order to be able to explain theoretically the emergence and the development of innovations that reflect both linguistic and social factors.

5. How are innovations perceived in research on EFL and ESL and what do they look like?

Until a few decades ago, there was a very conservative ENL-centered view on how non-attested uses are perceived and evaluated in research on ESL, as summarized by Schneider (2003: 239):

> In many statements on global Englishes there is an inherent but hardly visible tendency to regard and portray Britain and other ENL countries as the 'centers', thus entitled to establish norms of correctness, and, conversely, New Englishes as peripheral, thus in some sense deviating from these norms and, consequently, evaluated negatively.

This view has, however, drastically changed as to how deviations are perceived within ESL. A large body of research on ESL gives thorough empirical descriptions of innovative features in ESL varieties, which have mainly been interpreted as being signs of a variety to have reached the phase of "nativization" in variety formation. This phase is "the most important, the most vibrant one, the central phase of both cultural *and linguistic* transformation" (Schneider 2003: 247; our emphasis). In the ESL paradigm, then, linguistic innovations are essential for the "identity construction" (ibid.) of the speakers of a new English variety. Consequently, 'New Englishes' emerge and gain acceptance only through the nativization of linguistic innovations in the respective variety. These innovations "for a time may occur or exist side by side with the corresponding traditional forms, and eventually may become established as traditional themselves" (Andersen 1989: 11). In fact, in ESL, an innovation might be the result of the conventionalized use of what was initially an error (in the sense of a deviant use of the norm prevailing in a given speech community) over a long period of time and across a wide range of speakers in a given speech community. Ultimately, it is through the generalized use of an error that innovations gain acceptance and are considered to characterize individual ESL varieties (see also Section 4). In contrast, within the EFL paradigm, all kinds of deviations from native norms have been perceived and categorically classified either as idiosyncratic or systematic errors (see Section 2.2). This is mainly due to two factors: first, in EFL speech communities, the native speaker model is put forward by language politics as the (only) target in English language teaching, and, second, there is a tendency of learners of English themselves to aim for those norms (see e.g. surveys by Mukherjee & Rohrbach 2006 and, more recently, by Krenz 2015). In stark contrast to this, in established ESL speech communities such as India, adhering to native target norms is not propagated by language politicians and would be highly unnatural to ESL speakers, as it would seem rather "'foreign' — unnatural and affected — if they imitated BRP [i.e. British Received Pronunciation; SCD, SG, SL]" (Nihalani et al. 2004: 203). However, despite this background, in research on EFL, corpus linguists have recently started to pay attention to the use of innovative structures by EFL learners as well and the number of studies devoted to the subject has been increasing fast and steadily (see Section 2).

As speaker communities, non-native English users are likely to develop innovations at various linguistic levels. As Kachru (2006: 89) points out, some of the most creative innovations can be found in grammar, vocabulary, discourse strategies, and genres and styles. However, this list can easily be extended to studies describing innovations at the phonological level (e.g. D'Arcy 2005), at the semantic level (e.g. Robbin 2013), at the pragmatic level (e.g. Isingoma 2013), at the lexico-grammatical level (e.g. Schilk et al. 2012), etc. In what follows, we illustrate this with some selected examples of how innovative features at different linguistic levels have been described and perceived in previous EFL and ESL studies.

At the level of phonology, in EFL it has been noted that the interdental fricative /θ/ or /ð/ is often substituted either by /s, z/ or by /t, d/ (e.g. Yavaș 2009). The same phenomenon is described in various ESL varieties (e.g. Nihalani et al. (2004) for Indian English or Olajide & Olaniyi (2013) for Nigerian English). The difference between EFL and ESL does not lie in the formal realization of this feature, but in the perception and evaluation of its use: In EFL, this has been summarized as "interference" or (negative) "transfer" (Yavaș 2009: 177), whereas in ESL these substitutions are summarized as being "phonemic markers of identity" (Olajide & Olaniyi 2013: 284) that ESL speakers have in common "that supersedes L1 transfer" (Dako 2001: 26).

Lexical innovations have also been described in great detail in ESL. Typically, those innovations include borrowed and/or anglicized indigenous lexemes that refer to concepts for which no (British) English terms exist and thereby serve to "adapt to the socio-cultural reality in the country" (Dako 2001: 26). Studies that examine descriptions of nativized indigenous lexemes in ESL include Dako (2001) on nativized "Ghanaianisms" found in Ghanaian English, Meyler (2007) on nativized lexemes from Sinhala or Tamil in Sri Lankan English, Nihalani et al. (2004) on nativized indigenous Indian English lexemes, to name but a few. Recently, research has become less intuition-based, as Bernaisch (2015), for example, takes a corpus-based approach to identifying lexemes that are exclusively used in English spoken in the South Asian region and that are not used in British English, i.e. *gram* (referring to chick peas), *rupee* (the currency in Sri Lanka) and *sari/saree* (the traditional female dress worn in South Asia) (Bernaisch 2015: 106–107). Other lexical innovations in ESL concern the use of English terms in a semantically extended or slightly shifted fashion (see Dako 2001). In contrast to research on ESL, research on EFL shows that learners rarely borrow lexical items from their native language to use them innovatively in their foreign language. This may be due to the fact that English does not serve intranational purposes in the EFL community and there is simply no need to use genuinely borrowed lexemes. However, EFL and ESL show many parallels when it comes to the formation of new words and the coining of new lexemes, which happens with great systematicity. This will be demonstrated by Callies (this volume), by Horch (this volume) and by Schneider & Gilquin (this volume).

One further innovative linguistic feature worth mentioning in this context is code-switching and code-mixing. These have so far mainly been investigated and documented as successful communication strategies in research on bilingualism (e.g. Duran 1994; Grosjean 1989) and second language acquisition (e.g. Söderberg Arnfast & Jørgensen 2003), but we also find these in both ESL and EFL. Although the forms of code-switching and code-mixing are very similar in EFL and ESL, again, there is a difference in their interpretation and perception across the two non-native Englishes: When an EFL speaker resorts to their L1, this is typically treated as a communicative weakness or even a lexical error (e.g. in the EFL classroom; e.g. Berg 2013, S. Dose-Heidelmayer, personal communication, March 10, 2016); in ESL research, however, the functions and forms of code-switching and code-mixing are investigated intensively as contributing factors to the development of new dialects (e.g. "Hinglish, the code-switching between Hindi and English" (Sailaja 2011: 473)). Interestingly, despite this dichotomy in the way code-switching and code-mixing are approached across EFL and ESL, learner corpus research is nevertheless starting to witness a shift in scholars' perception of code-switching from communicative weakness to effective communicative strategy. This was recently documented in a study by Nacey & Graedler (2013) on Norwegian Learners of English and De Cock (2015) for French, Spanish, German and Italian Learners of English.

At the stylistic level, we find further illustrations of what can be classified — broadly speaking — as innovations in the sense of a restructuring in form-function mapping (see Croft 2000). Here, many ESL speakers "(continue to) use a stock of words which is either restricted to more formal contexts or considered to be rather archaic in the present-day usage of the erstwhile input variety, namely British English (BrE)" (Meyler 2007: xiv). For example, Bernaisch (2015) finds a significantly lower frequency of more formal lexemes in BrE compared to Indian and Sri Lankan English (i.e. the South-Asian speakers prefer more formal variants than the British English speakers, such as *commence*, *purchase* or *refrigerator*). Similarly, Bernaisch (2015) also documents a frequent use of archaic markers, such as the use of *madam* as an address term. Although EFL learners have not been studied in as much detail as ESL speakers with respect to register and/or formality, it has nonetheless been recognized that, similarly to ESL users, English learners lack a nativelike "text-type sensitivity" (Lorenz 1999: 64) or "register awareness"

(Gilquin & Paquot 2008). In other words, EFL yields many typically written features in their speech and typically spoken features in their writing.

At the level of pragmatics, particles and discourse markers provide an interesting case of shared innovations between ESL and EFL, particularly with regard to the creative use of discourse markers from the speakers' L1 when they speak English. Here, a very well-documented example is the use of *la/lah* in Singaporean English, an established ESL variety (Schneider 2007). It has been identified to be a solidarity marker between interlocutors as well as to be a pragmatically multifunctional marker (see Low & Deterding (2003) for a survey on previous studies). Similarly, EFL speakers have also been found to use discourse markers from their L1 when speaking English. For instance, French learners use *enfin*, *hein* and *allez*, and *ach* or *ja* has been documented in German EFL learners (e.g. Gilquin 2008). Again, the structural form of the innovative use is similar in EFL and ESL (i.e. the integration of discourse markers from the L1 when speaking English); however, their interpretation and perception are different. For instance, in EFL, "[i]f a non-native speaker uses discourse particles incorrectly […] this may lead to misunderstandings" (Aijmer 2002: 3), whereas in ESL the focus is on the description of the innovative forms and functions as a sign of nativization (e.g. Low & Deterding 2003).

The lexis-grammar interface has been claimed to be particularly prone to display innovative forms in ESL. This is mainly due to the fact that "certain words but not others of the same word class prefer specific grammatical rules or patterns" (Schneider 2007: 83). That is, even though neither the patterns nor the words are new, "what is novel is the habitual association between them in specific varieties" (Schneider 2007). One very well researched case in point at the lexico-grammatical level is the use of phrasal verbs which have been claimed to be one of the most "notoriously challenging aspects" (Gardner & Davies 2007: 339) in EFL. As such, they have been the subject of a variety of EFL and ESL studies (e.g. Deshors 2016; Gilquin 2015; Schneider 2014; Zipp & Bernaisch 2012). Because they are relatively frequent, phrasal verbs with *up* have attracted much attention and constructions such as *cope up with* have been shown to appear both in ESL and EFL (see Edwards & Laporte 2015; Gilquin 2015; Zipp & Bernaisch 2012). However, Gilquin (2015) reports on other innovative uses that EFL and ESL do not share and that are characteristic of individual variety types of English, e.g. *meddle up* (in Singaporean English), *fashion (your jeans) up* (in German learner English) or *spray up* (in British English) are not shared between the three Englishes.[4] Ultimately, however,

[4] We are aware, however, that some potentially shared constructions might simply not be found in corpus research and thus similarities might simply go unnoticed, often due to the differences in corpus designs of the corpora typically used for EFL-ESL comparisons (including different topics and compilation procedures).

those uses show that the underlying processes for using the particles are shared and can serve to "testify to the common creative potential of both types of varieties" (Gilquin 2015: 107).

Focusing on comparable linguistic innovations across ESL and EFL, it emerges that main differences between the two Englishes do not lie in the formal realization of innovations, as they seem to be quite similar in EFL and ESL. Rather, those differences emerge in both the interpretation and the perception of these linguistic innovations: The predominant terms used in EFL research being "deviation" "misuse", "errors" or "non-attested", as compared to being markers of "nativization", "identity construction" or simply giving neutral descriptions of innovative forms and functions in ESL studies. However, more and more, studies are starting to not only show clear structural parallels between ESL and EFL (e.g. Callies or Koch et al. this volume) but also propose theoretical explanations of innovations in EFL (e.g. Callies or Schneider & Gilquin this volume). For EFL research, this is no trivial development as it suggests that innovative features are starting to be accepted by ENL-editors (as illustrated by Van Rooy & Kruger (this volume) in the context of South African English). Crucially, this might eventually lead to a significant change in the way EFL speakers are perceived; that is, as creative language users instead of 'defective native speakers'.

While, within the World Englishes community, "it took a great deal of persistence to convince linguists and educationists that the post-colonial grammars, lexicons and phonologies were worthy of study and not some deviation to be scrubbed away" (Mesthrie & Bhatt 2008: 23), it is likely that more time will be needed until attitudes towards innovative EFL features begin to change in a similar vein. It is our hope that this volume will contribute a (baby) step in this direction.

Acknowledgements

We would like to thank Sebastian Hoffmann, Chair of the organizing committee of the ICAME 36 conference, for supporting our idea of a pre-conference workshop on the theme of *Corpus Linguistics and Linguistic Innovations in Non-native Englishes*. We are also very grateful to the participants and the audience of the workshop who all contributed to stimulating discussions. Thank you also to the editors of the IJLCR, Marcus Callies and Magali Paquot, who welcomed the idea of a special issue and who provided us with valuable feedback on this introductory chapter and to Kees Vaes who suggested this reprint in the *Benjamins Current Topics* series. Last but not least, we wish to thank all the reviewers and statistics advisor who generously contributed their time and expertise to the original papers in the special issue.

References

Aijmer, K. 2002. *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.10

Andersen, H. 1989. "Understanding linguistic innovations". In L.E. Breivik & E.H. Jahr (Eds.), *Language Change: Contributions to the Study of its Causes*. Berlin: Mouton de Gruyter, 5–27.

Bamgbose, A. 1998. "Torn between the norms: Innovations in World Englishes", *World Englishes* 17(1), 1–14. https://doi.org/10.1111/1467-971X.00078

Bennui, P. 2013. "Some syntactic innovations in new literatures in English", *International Journal of Linguistics* 5(5), 208–224. https://doi.org/10.5296/ijl.v5i5.3875

Berg, N. 2013. *Codeswitching in ESL Teaching*. Degree project. University of Stockholm. Available at: https://www.diva-portal.org/smash/get/diva2:634259/FULLTEXT01.pdf (accessed March 2016).

Bernaisch, T. 2015. *The Lexis and Lexicogrammar of Sri Lankan English*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g54

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Bruthiaux, P. 2003. "Squaring the circles: Issues in modeling English worldwide", *International Journal of Applied Linguistics* 13(2), 159–178. https://doi.org/10.1111/1473-4192.00042

Buschfeld, S. 2013. *English in Cyprus or Cyprus English*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g46

Corder, P. 1967. "The significance of learner's errors", *International Review of Applied Linguistics* 5, 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.

Dako, K. 2001. "Ghanaianisms. Towards a semantic and a formal classification", *English World-Wide* 21(2), 23–53. https://doi.org/10.1075/eww.22.1.03dak

D'Arcy, A. 2005. "The development of linguistic constraints: Phonological innovations in St. John's English", *Language Variation and Change* 17(3), 327–355.

Davies, M. 2013. Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries. Available at: http://corpus2.byu.edu/glowbe/ (accessed November 2015).

Davies, M. & Fuchs, R. 2015. "Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE)", *English World-Wide* 36(1), 1–28. https://doi.org/10.1075/eww.36.1.01dav

Davydova, J. 2012. "Englishes in the outer and expanding circles: A comparative study", *World Englishes* 31(3), 366–385. https://doi.org/10.1111/j.1467-971X.2012.01763.x

De Cock, S. 2015. "The use of foreign words in interviews with EFL learners: An effective communication strategy?" Paper presented at *Learner Corpus Research 2015*, Radboud University Nijmegen, Netherlands, 11-13 September 2015.

Deshors, S.C. 2014. "A case for a unified treatment of EFL and ESL: A multifactorial approach", *English World-Wide* 35(3), 279–307. https://doi.org/10.1075/eww.35.3.02des

Deshors, S.C. 2016. "Inside phrasal verbs constructions: A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations", *International Journal of Learner Corpus Research* 2(1), 1–30.

Diemer, S., Brunner, M.-L., Collet, C. & Schmidt, S. 2018. *Corpus of Academic Spoken English*. Birkenfeld: Trier University of Applied Sciences (coordination) / Saarbrücken: Saarland University / Sofia: St Kliment Ohridski University / Forlì: University of Bologna-Forlì /

Santiago: University of Santiago de Compostela / Helsinki: Helsinki University & Hanken School of Economics / Birmingham: Birmingham City University / Växjö: Linnaeus University / Lyon: Université Lumière Lyon 2 / Louvain-la-Neuve: Université catholique de Louvain / Boise: Boise State University. Available at http://umwelt-campus.de/case (accessed February 2018).

Duran, L. 1994. “Toward a better understanding of code switching and interlanguage in bilinguality: Implications for bilingual instruction”, *The Journal of Educational Issues of Language Minority Students* 14, 69–88.

Edwards, A. 2014a. “The progressive aspect in the Netherlands and the ESL/EFL continuum”, *World Englishes* 33(2), 173–194. https://doi.org/10.1111/weng.12080

Edwards, A. 2014b. *English in the Netherlands: Functions, Forms and Attitudes*. PhD dissertation, University of Cambridge.

Edwards, A. & Laporte, S. 2015. “Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels”, *English World-Wide* 36(2), 135–169. https://doi.org/10.1075/eww.36.2.01edw

ELFA. 2008. *The Corpus of English as a Lingua Franca in Academic Settings*. Director: Anna Mauranen. Available at: http://www.helsinki.fi/elfa/elfacorpus (accessed March 2016).

Gardner, D. & Davies, M. 2007. “Pointing out frequent phrasal verbs: A corpus-based analysis”, *TESOL Quarterly* 41(2), 339–359. https://doi.org/10.1002/j.1545-7249.2007.tb00062.x

Gilquin, G. 2008. “Hesitation markers across EFL learners: Pragmatic deficiency or difference?”. In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Berlin and New York: Mouton de Gruyter, 119–149.

Gilquin, G. 2011. “Corpus linguistics to bridge the gap between World Englishes and Learner Englishes”, *Communicación en el siglo XXI* Vol. II: 638–642. Available at: http://dial.uclouvain.be/downloader/downloader.php?pid=boreal%3A112509&datastream=PDF_01&disclaimer=5dded109ee97b89072e796cddd5219c599cbdbda547c241bb6bbe87d65203f8f (accessed October 2015).

Gilquin, G. 2015. “At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared”, *English World-Wide* 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Gilquin, G. & Granger, S. 2011. “From EFL to ESL: Evidence from the International Corpus of Learner English”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 55–78. https://doi.org/10.1075/scl.44.04gra

Gilquin, G. & Paquot, M. 2008. “Too chatty: Learner academic writing and register variation”, *English Text Construction* 1(1), 41–61. https://doi.org/10.1075/etc.1.1.05gil

Götz, S. & Schilk, M. 2011. “Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced German learners”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 79–100. https://doi.org/10.1075/scl.44.05sch

Greenbaum, S. & Nelson, G. 1996. “The International Corpus of English (ICE) project”, *World Englishes* 15(1), 3–15. https://doi.org/10.1111/j.1467-971X.1996.tb00088.x

Grosjean, F. 1989. “Neurolinguists, beware! The bilingual is not two monolinguals in one person”, *Brain and Language* 36(1), 3–15. https://doi.org/10.1016/0093-934X(89)90048-5

Hamid, M.O. & Baldauf Jr., R.B. 2013. “Second language errors and features of World Englishes”, *World Englishes* 32(4), 476–494. https://doi.org/10.1111/weng.12056

Isingoma, B. 2013. “Innovative pragmatic codes in Ugandan English: A relevance-theoretic account”, *Argumentum* 9, 19–31.

Kachru, B.B. 1982. “Models for non-native Englishes”. In B.B. Kachru (Ed.), *The Other Tongue: English across cultures*. Urbana and Chicago: University of Illinois Press, 48–74.

Kachru, B.B. 1985. “Standards, codification and sociolinguistic realism: The English language in the outer circle”. In R. Quirk & H.G. Widdowson (Eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.

Kachru, B.B. 2006. *World Englishes in Asian Contexts*. Aberdeen and Hong Kong: Hong Kong University Press.

Krenz, J. 2015. *Attitudes of German University Students towards Varieties of English: An Empirical Study*. Unpublished B.A. dissertation, University of Giessen.

Laitinen, M. 2010. “Describing ‘orderly differentiation’: Compiling the *Corpus of English in Finland*”, *English Today* 26(1), 26–33. https://doi.org/10.1017/S0266078409990459

Laporte, S. 2012. “Mind the gap! Bridge between world Englishes and learner Englishes in the making”, *English Text Construction* 5(2), 264–291. https://doi.org/10.1075/etc.5.2.05lap

Li, E. & Mahboob, A. 2012. *English Today: Forms, Functions, and Uses*. Hong Kong: Pearson Education.

Lorenz, G. 1999. *Adjective Intensification. Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Low, E.L. & Deterding, D. 2003. “A corpus-based description of particles in spoken Singapore English”. In D. Deterding, E.L. Low & A. Brown (Eds.), *English in Singapore: Research on Grammar*. Singapore: McGraw-Hill, 58–66.

Mair, C. 2015. “Response to Davies and Fuchs”, *English World-Wide* 36(1), 29–33. https://doi.org/10.1075/eww.36.1.02mai

Mesthrie, R. & Bhatt, R.M. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511791321

Meyler, M. 2007. *A Dictionary of Sri Lankan English*. Colombo: Mirisgala.

Mollin, S. 2006. *Euro-English: Assessing Variety Status*. Tübingen: Gunter Narr.

Mukherjee, J. 2010. “Corpus-based insights into verb-complementational innovations in Indian English: Cases of nativised semantico-structural analogy”. In A.N. Lenz & A. Plewnia (Eds.), *Grammar between Norm and Variation*. Frankfurt am Main: Peter Lang, 219–241.

Mukherjee, J. 2015. “Response to Davies and Fuchs”, *English World-Wide* 36(1), 34–37. https://doi.org/10.1075/eww.36.1.02muk

Mukherjee, J. & Hoffmann, S. 2006. “Describing verb-complementational profiles of New Englishes: A pilot study of Indian English”, *English World-Wide* 27(2), 147–173. https://doi.org/10.1075/eww.27.2.03muk

Mukherjee, J. & Hundt, M. (Eds.). 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.44

Mukherjee, J. & Rohrbach, J.-M. 2006. “Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research”. In B. Kettemann & G. Marko (Eds.), *Planing, Gluing and Painting Corpora: Inside the Applied Corpus Linguist’s Workshop*. Frankfurt/Main: Peter Lang, 205–232.

Nacey, S. & Graedler, A.-L. 2013. “Communication strategies used by Norwegian students of English”. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve: Presses Universitaires de Louvain, 345–356.

Nelson, G. 2015. “Response to Davies and Fuchs”, *English World-Wide* 36(1), 38–40. https://doi.org/10.1075/eww.36.1.02nel

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.14

Nesselhauf, N. 2009. “Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties”, *English World-Wide* 30(1), 1–26. https://doi.org/10.1075/eww.30.1.02nes

Nihalani, P., Tongue, R.K., Hosali, P. & Crowther, J. 2004. *Indian and British English: A Handbook of Usage and Pronunciation* (2nd ed.). New Dehli: Oxford University Press.

Oxford English Dictionary (OED). 2015. Online version. Oxford: Oxford University Press. Available at: http://www.oed.com (accessed January 2016).

Olajide, S.B. & Olaniyi, O.K. 2013. “Educated Nigerian English phonology as core of a regional ‘RP’”, *International Journal of Humanities and Social Science* 3(14), 277–286.

Peters, P. 2015. “Response to Davies and Fuchs”, *English World-Wide* 36(1), 41–44. https://doi.org/10.1075/eww.36.1.02pet

Platt, J. 1989. “The nature of indigenized Englishes: Interference – creativity – universals”, *Language Sciences* 11(4), 395–407. https://doi.org/10.1016/0388-0001(89)90028-4

Robin, A.A. 2013. “Old words, new meanings: A survey of semantic change amongst Yoruba-English bilingual undergraduates”, *Journal of Capital Development in Behavioural Sciences* 1, 55–79.

Rosen, A. 2014. *Grammatical Variation and Change in Jersey English*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g48

Sailaja, P. 2011. “Hinglish: Code-switching in Indian English”, *ELT Journal* 65(4), 473–480. https://doi.org/10.1093/elt/ccr047

Schilk, M., Bernaisch, T. & Mukherjee, J. 2012. “Mapping unity and diversity in South Asian English lexicogrammar: Verb-complementational preferences across varieties”. In M. Hundt & U. Gut (Eds.), *Mapping Unity and Diversity World-Wide: Corpus-Based Studies of New Englishes*. Amsterdam: John Benjamins, 137–165. https://doi.org/10.1075/veaw.g43.06sch

Schneider, E.W. 2003. “The dynamics of New Englishes: From identity construction to dialect birth”, *Language* 79(2), 233–281. https://doi.org/10.1353/lan.2003.0136

Schneider, E.W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618901

Schneider, E.W. 2012. “Exploring the interface between World Englishes and Second Language Acquisition – and implications for English as a Lingua Franca”, *Journal of English as a Lingua Franca* 1(1), 57–91. https://doi.org/10.1515/jelf-2012-0004

Schneider, E.W. 2014. “‘Transnational Attraction’: New reflections on the evolutionary dynamics of World Englishes”, *World Englishes* 33(1), 9–32. https://doi.org/10.1111/weng.12069

Schneider, G. & Zipp, L. 2013. “Discovering new verb-preposition combinations in New Englishes”, *Studies in Variation, Contacts and Change in English* 13. Available at: http://www.helsinki.fi/varieng/series/volumes/13/schneider_zipp.pdf (accessed November 2015).

Söderberg Arnfast, J. & Jørgensen, N. 2003. “Code-switching as a communication, learning, and social negotiation strategy in first-year learners of Danish”, *International Journal of Applied Linguistics* 13(1), 23–53. https://doi.org/10.1111/1473-4192.00036

Sridhar, K.K. & Sridhar, S.N. 1986. “Bridging the paradigm gap: Second language acquisition research and indigenized varieties of English”, *World Englishes* 5(1), 3–14. https://doi.org/10.1111/j.1467-971X.1986.tb00636.x

Szmrecsanyi, B. & Kortmann, B. 2011. "Typological profiling: Learner Englishes versus indigenized L2 varieties of English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 168–187. https://doi.org/10.1075/scl.44

Van Rooy, B. 2011. "A principled distinction between error and conventionalized innovation in African Englishes". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 189–208. https://doi.org/10.1075/scl.44.10roo

VOICE. 2009. The Vienna-Oxford International Corpus of English (version 1.0 online). Director: B. Seidlhofer; Researchers: A. Breiteneder, T. Klimpfinger, S. Majewski, M.-Luise Pitzl. Available at: http://voice.univie.ac.at (accessed March 2016).

Williams, J. 1987. "Non-native varieties of English: A special case of language acquisition", *English World-Wide* 8(2), 161–199. https://doi.org/10.1075/eww.8.2.02wil

Yavaş, M. 2009. *Applied English Phonology*. Malden: Blackwell.

Zipp, L. & Bernaisch, T. 2012. "Particle verbs across first and second language varieties of English". In M. Hundt & U. Gut (Eds.), *Mapping Unity and Diversity World-Wide: Corpus-based Studies of New Englishes*. Amsterdam: John Benjamins, 167–196. https://doi.org/10.1075/veaw.g43.07zip

“This hair-style called as ‘duck tail’”

The ‘intrusive *as*’-construction in South Asian varieties of English and Learner Englishes

Christopher Koch[1], Claudia Lange[2] and Sven Leuckert[2]
[1]Philipps Universität Marburg / [2]Dresden University of Technology

This paper focuses on the ‘intrusive *as*’-construction in complex-transitive verb complementation which was so far only attested for Indian English. Our data show that ‘intrusive *as*’ is a common feature in South Asian Englishes generally, albeit to different degrees. Comparing the South Asian data with data from Learner Englishes allows to test several hypotheses concerning the origin of ‘intrusive *as*’; the most robust correlation within the data points to redundancy as a motivating factor for both ESL and EFL contexts.

Keywords: ‘intrusive *as*’, innovation, verb complementation, complex transitives, South Asian Englishes, Learner Englishes

1. Introduction

This paper is concerned with a feature that has been a staple topic of the Indian English complaint tradition at least since the 1930s, when the first references appear in print: In his short tract “Some Notes on Indian English” published by the *Society for Pure English*, Goffin lists “[a] few other examples of ‘ignorant’ Indian English [that] are universal and flourishing enough to be worth special quotation: […] In conjunctions: […] ‘He **called** me ***as*** a nonsense.’” (1934: 26, emphasis ours).[1] In the same year, the first edition of Smith-Pearse’s booklet on *The English Errors of Indian Students* was published, remaining in print virtually unchanged up to the present day.[2] His chapter on ‘Common mistakes with conjunctions’ presents as ‘incorrect’ example 14: ‘He **called** me ***as*** a fool’, juxtaposing it with the ‘correct’

[1] We owe this first attestation to James Lambert, whose support we gratefully acknowledge.

[2] The first edition was published under the title *English Errors in Indian Schools*.

https://doi.org/10.1075/bct.98.02koc

version ‘He called me a fool’ (Smith-Pearse 1968: 23, emphasis ours). Nihalani et al. then do not only provide the term ‘intrusive *as*’ (2004: 178) for the usage of *as* “in places where it would be considered superfluous in BS [British Standard English]” (2004: 22), but also confirm that ‘*called* ***as***’ has persisted and “is often treated as a ‘common error’ by Indian teachers of English” (ibid.).

This paper investigates the extent of the use of ‘intrusive *as*’ with complex-transitive predicates best exemplified by *call* (*as*) (e.g. “I call it as tragic” (Nihalani et al. 2004: 22))[3] in South Asian and learner varieties of English. The ‘intrusive *as*’-construction lends itself particularly well to reconsidering the tension between ‘errors’ and ‘innovations’, two concepts that already played a major role when the field of World Englishes/Postcolonial Englishes was initially charted. As the quotes above indicate, there is as yet no agreement within the Indian English speech community whether to ‘promote’ the ‘intrusive *as*’-construction from an ‘error’ to an innovative feature of an endonormative variety of English.

This paper, then, attempts to contribute to these discussions by, firstly, providing empirical evidence for the distribution of ‘intrusive *as*’-constructions across South Asian Englishes and, secondly, by comparing significant patterns of use with data from learner corpora. The corpus-based results will then be embedded in more general considerations reflecting the main focus of this special issue, namely the emergence and stabilization of innovations in ESL/EFL speech communities.

Previous research into the ‘intrusive *as*’-construction (Lange 2014, 2015) has already provided a corpus-based description of the phenomenon in South Asian varieties of English in addition to the largely intuition-based assessments before (e.g. Nihalani et al. 2004; Yadurajan 2001). Lange’s work demonstrated that the phenomenon is not restricted to the Indian English context but indeed represents a truly pan-South Asian feature shared by all varieties under scrutiny, even though Indian English and Sri Lankan English emerge as the two most progressive ones (as far as the present feature is concerned).

In the current study, we aimed at simultaneously widening and deepening our approach to find and describe this innovative pattern. The present study builds upon previous work and offers an extension of the approach, resulting in three major research questions:

1. Are the principal verbs of naming, labelling or depicting someone/-thing as already described in e.g. Nihalani et al. (2004) and Yadurajan (2001) the only verbs sanctioning the use of the ‘intrusive *as*’-construction, or are there further verbs that South Asian English speakers use in conjunction with this pattern?

[3] Our data show that *call* is the most frequently used verb in complex-transitive complementation across all varieties considered, cf. Section 5 below.

While Lange’s (2014, 2015) work was restricted to previously established lists of verb lemmas reported to employ the ‘intrusive *as*’-pattern in complex-transitive complementation, the present study employs a semi-automatic approach to find and retrieve potential instances of ‘intrusive *as*’ also for other verb lemmas.

2. Which grammatical context factors influence the selection of the ‘intrusive *as*’-pattern as opposed to ‘regular’ complex-transitive complementation without *as*, and are these different between the varieties and verb lemmas under scrutiny? In contrast to previous research, where only the general pattern preferences (‘intrusive *as*’-pattern or not) with regard to verb lemmas and varieties had been described, the dataset of the current study was expanded in terms of the amount of annotation, extending the coding to three further context factors which were assumed to potentially influence pattern selection: (a) voice of the (main) verb in the clause under scrutiny, (b) ‘distance’, i.e. number of words between main verb and complement (excluding ‘as’ to avoid introducing a bias), and (c) part of speech information on the head of the object’s complement.
3. Can similar frequencies of the ‘intrusive *as*’-construction be attested for the learner data; and, if so, do learners follow comparable selection patterns as the speakers evidenced in the South Asian data? By widening the scope of description from published ESL material to include both learner data from related ESL scenarios as well as EFL data from a variety of further backgrounds, the roots of the construction can be more firmly established. Related to this is the question whether ‘intrusive *as*’ should be regarded as a contact-induced feature that is only exhibited by those varieties of English in contact with local languages allowing similar constructions — or whether the pattern is motivated by more general acquisitional universals that have been shown to play a role in both ESL and EFL contexts. Several principles and processes related to Second Language Acquisition (SLA) and particularly relevant in the context of World Englishes have been proposed, frequently relying and extending Williams’s (1987) original classification. One taxonomy of such processes comes from Schneider (2012); his notion of ‘redundancy’ (“the unmotivated repeated (or double) marking of the same piece of information, characterizes language in general but appears particularly frequent in WEs” (2012: 65)) might be applicable if it can be shown that ‘intrusive *as*’ serves the purpose of creating redundancy, making the complex-transitive structure more explicit.[4]

[4] Another pillar study that dealt with this problem is Rohdenburg (1996), who called the phenomenon ‘explicitness’. We chose to stick to Schneider’s terminology, which acknowledges the fact that processes such as redundancy/explicitness are not restricted to the ESL/EFL context, but builds on a large body of specialized research within the field and (unlike Rohdenburg) integrates further processes that have been shown to play a role, both in terms of perception as well as production.

The structure of this paper is as follows: Section 2 contextualizes the issues relating to the dichotomies — or continua — marked by the terms native vs. non-native and error vs. innovation respectively. Section 3 outlines the grammatical properties of the feature under discussion, while Section 4 is concerned with the methodology in pursuit of our research questions. Section 5 gives an account of our data for both South Asian and Learner Englishes, with Section 6 evaluating and summing up our findings.

2. Old and new dichotomies: Native vs. non-native, error vs. innovation

The development of World Englishes as a legitimate field of linguistic enquiry is now, after more than 30 years, a matter for introductory textbooks, where key notions such as native/non-native, error/innovation as well as Inner Circle/Outer Circle etc. have been widely discussed. The categorial status and the overall usefulness of these notions were subject to considerable and frequently heated debates, such as the one between Randolph Quirk (1990) and Braj Kachru (1991) in the journal *English Today*.[5] Quirk's insistence on a categorial distinction between native and non-native (i.e. learner) Englishes entailed that only native Englishes were capable of being 'institutionalized' or standardized. In his reply, Kachru highlighted the term "non-native institutionalized variety of English" to underline his criticism of the then prevailing concept of 'nativeness' predicated upon monolingual speech communities. More specifically, Kachru rejected Quirk's implication that all divergences from institutionalized Standard English are by definition 'errors',[6] and more than two decades of research later, it seems as if Kachru has won the day: "recent realities seem to be rendering the ENL — ESL distinction increasingly obsolete" (Schneider 2007: 13), since

> […] there are no structural features, at any level of grammatical description, that characterize all "non-native" varieties of English to the exclusion of all "native" varieties. Given that most linguists who have made serious efforts to find such features acknowledge/concede that there aren't any […], we are fully justified in concluding that the dichotomy native variety/non-native variety cannot be structurally or grammatically sustained. And if it indeed cannot be sustained, speakers of at least **the varieties that can be shown to have their own norms**, such as Indian English and Singapore English, must be classified as native speakers of English […]. (Singh 2007: 39–40, emphasis ours)

[5] Quirk's paper originally appeared 1989 in the *JALT Journal* (http://jalt-publications.org/jj/issues/1989-05_11.1) and has since been anthologized several times.

[6] Quirk only recognized British and American English as institutionalized varieties of English.

This quote effectively sums up the state of the art in the field of Postcolonial Englishes: the binary — and normative — distinction between native and non-native (i.e. L2) Englishes has largely been discarded, and most scholars today acknowledge Edgar Schneider’s model of a varietal cline, where socio-historical factors pertaining to the speech community in question are the main driving force towards endonormative stabilization (cf. Schneider 2007). That is, applying the term ‘error’ to PCEs would in effect set the clock back to 1990, when Quirk insisted that all non-native varieties including PCEs

> are inherently unstable, ranged along a qualitative cline, with each speaker seeking to move to a point where the varietal characteristics reach vanishing point, and where thus, ironically, each variety is best manifest in those who by commonsense measures speak it worst (Quirk 1990: 5–6).

However, even if this stance and with it the distinction between ENL and ESL has lost its legitimacy, the one between ESL and EFL remains subject to debate. Singh concludes his pronouncement above by adding that “[t]he only thing to remember is that we are talking about speakers and NOT learners” (Singh 2007: 43). In this, he is seconded by Schneider (2014), who considers EFL or Expanding Circle Englishes to be beyond the scope of his Dynamic Model. Even though there are some passing similarities between ESL and EFL, the EFL varieties he considers crucially lack “Phase 4 components” (ibid.: 27), that is indicators of a shared norm among the speech community. What is implied here is that EFL speakers are unlikely to develop the shared norms which constitute a speech community and which effectively demarcate plain errors from potential innovations.[7]

Recent studies, however, took a critical stance with regard to what is frequently considered a paradigm gap, i.e. the separate treatment of ESL and EFL in the fields of Second Language Acquisition and World Englishes (cf. Buschfeld 2014; Deshors 2014; Edwards & Laporte 2015; Gilquin 2015), respectively. Findings in Buschfeld (2011, 2013) and Edwards (2014) suggest “that the psycholinguistic processes underlying the development of learner language and second-language varieties seem to be fundamentally similar” (Buschfeld 2014: 183), ultimately calling for a widespread re-evaluation of the relation between ESL and EFL and a more differentiated treatment of linguistic developments in Learner Englishes.

Due to the fact that neither traditional models such as Kachru’s (1985) Three Circles model nor more recent models such as Schneider’s (2007) Dynamic Model are particularly well suited for the analysis and explanation of developments in Learner Englishes, new models have been proposed by Schneider (2014) and

[7] Cf. Schneider (2012) for references to divergent positions which challenge the dividing line between ESL and EFL.

Buschfeld & Kautzsch (2016). Both Schneider's Transnational Attraction and Buschfeld & Kautzsch's model of Extra- and Intra-Territorial Forces (EIF) take into account more general processes of globalization and seek to provide a means for a unified treatment of ESL and EFL or, more precisely, of Englishes with and without colonial background.

Since 'intrusive *as*'-constructions are listed in dictionaries of common learner errors (e.g. Heaton & Turton 1997: 60), a rationale for a comparison of usage patterns in South Asian Englishes and Learner Englishes is provided. An in-depth corpus analysis of 'intrusive *as*' in both ESL (represented by the SAVE corpus, cf. Section 4.1) and EFL (represented by ICLE and ICNALE, cf. Section 4.1) establishes the basis for a systematic comparison and, hence, a discussion of whether the feature can be considered an innovation shared between ESL and EFL or whether the feature should indeed be treated as an error in Learner Englishes.

We will return to the issue of error vs. innovation in Sections 5.3 and 6 below; the next sections will firstly elucidate the construction under scrutiny, namely the complex-transitive complementation pattern with and without *as*, and then proceed to a discussion of our methodology and our actual findings.

3. Complex-transitive predication

According to the *Cambridge Grammar of the English Language* (CGEL; Huddleston & Pullum 2002), complex-transitive constructions involve additional predication about the direct object by means of what Huddleston (2002: 217) calls a "predicative complement (PC)". Rather than referring to an entity or a person, a PC serves to denote "a property that is predicated of the person [or entity]" referenced by the object. Syntactically speaking, this means that PCs are complements, but semantically, they serve a predicative function. Examples (1) and (2), taken from the CGEL (ibid.), illustrate the phenomenon:

(1) Ed seemed quite competent. [complex-intransitive: S-P-PCS]

(2) She considered Ed quite competent. [complex-transitive: S-P-O-PCO]

In complex-intransitives, the predicative complement provides additional information about the subject, whereas in complex-transitives, the information in the complement refers to the direct object. Since our study is concerned with the latter type, we will exclusively discuss object complements from here onwards.

Complements in complex-transitives can be realized by noun phrases, adjective phrases or nonfinite clauses, although constructions with nonfinite clauses are rare. Complex-transitives can take depictive and resultative complements, a "primarily semantic distinction [that] is not always easy to draw" (Huddleston

2002: 265). As pointed out in Lange (2015), *name* stands out as a particularly difficult example in this regard, although other verbs such as *designate* and *call* prove a similar challenge. The following examples found in the SAVE corpus represent clear cases of *name* being used in a resultative sense (3) and, in innovative fashion, in a depictive sense (4):[8]

(3) Additional Secretary Badiur Rahman, **named *as*** the acting power secretary on Monday night, will move to the Implementation Monitoring and Evaluation Division (IMED). (SAVE-BD_DS_2006–11__pt1.txt)

(4) […] we come across a teacher of literature **named *as*** Mr. Keating, acted by Robin Williams. (SAVE-SL_DN_2002-05-07.txt)

Verbs in complex-transitives may occur with or without *as* (or, albeit very rarely, with *for*) before the object complement (see Quirk et al. 1985: 1196). For some verbs such as *describe*, the occurrence with *as* is felicitous in any variety. However, in Indian English and other Asian varieties of English, we find the so-called ‘intrusive *as*’-construction, i.e. complex-transitives with *as* where no preposition or particle would be expected in the historical input variety.[9]

4. Methodology

4.1 Corpus data and data extraction

Previously, corpus-based analyses of the ‘intrusive *as*’-phenomenon had to rely on intuition-based assessments of the main verbs sanctioning the ‘intrusive *as*’-construction, such as provided in Nihalani et al. (2004) or Yadurajan (2001). For the current study, it was decided that this list of verb lemmas should be based on a completely corpus-based approach building on an exhaustive search for all verbs used in conjunction with *as* in South Asian Englishes. The data were derived from the *South Asian Varieties of English Corpus* (SAVE; Bernaisch et al. 2011), a newspaper corpus of 18 million words covering data from two major newspapers for each of the six contexts, i.e. Bangladeshi, Indian, Maldivian, Nepali, Pakistani and Sri Lankan English, to the extent of three million words per variety.

[8] The examples come from the Sri Lankan and the Bangladeshi subcorpus of the *South Asian Varieties of English Corpus (SAVE)* , which will be introduced in Section 4 below.

[9] *As* in complex-transitive constructions is a preposition (Quirk et al. 1985: 1200, Huddleston & Pullum 2002: 255); in the following, we use the functionally neutral term ‘particle’ to encompass both the possible source and the target constructions, i.e. quotative particles (cf. example (5) in Section 5.1) and prepositions.

For this purpose, a semi-automatic approach was employed using an *R* script (R Development Core Team 2015) building on the CLAWS C7-annotated version of the SAVE corpus, which allowed us to search for any verb used in conjunction with *as* without relying on a list of verb lemmas. However, while CLAWS proved useful for the detection of potential cases of 'intrusive *as*', it does not recognize the pattern for what it is, marking *as* up as a conjunction in most cases. Still, the syntactic slots to be filled in a complex-transitive construction made it possible to restrict the context around *as* to those that could be potential cases of 'intrusive *as*'. So, while the CLAWS annotation could not be relied upon for the positive detection of 'intrusive *as*' cases, it could still be employed for the definition of an exclusion pattern that, if it matched, would describe a non-'intrusive *as*' case for the occurrence currently processed. The overall approach can thus be summarized as follows:

1. Search sentence-internally for 'VERB + X + *as*', where X…
   a. … may be any number of words, excluding other verbs,
   b. … may include quotations, but no other punctuation.[10]
2. Exclude a potential finding from step (1) in case *as* only represents the beginning of another clause (operationalized by excluding several irrelevant uses of *as*, e.g. conjunctions (followed by NP + V)).[11]

After applying this script to all components of the SAVE corpus, we arrived at a list of approximately 6,000 potential uses of 'intrusive *as*' per component, which were then annotated manually for whether they instantiate true cases of 'intrusive *as*' (i.e. verb lemmas that do not usually allow complex-transitive complementation with *as*). Even in the historical input variety there is a wide variety of verb lemmas allowing or requiring the use of *as* in complex-transitive predication (cf. Section 3), which is why a reference list was produced following the same computing strategy as above, building on the periodicals section of the *British National Corpus* (c. 8.8m words). This list was consulted, in addition to the *Oxford English Dictionary* and *Merriam Webster Dictionary*, to decide whether a particular verb lemma also sanctions the use of *as* in complex-transitive contexts in British or American English, i.e. the main exonormative standards in South Asian varieties of English.

[10] Even though these two restrictions certainly entail the risk of ignoring potentially relevant (nested) clausal structures, they are required to make the results more reliable. If they were not in place, a second verb with 'intrusive *as*'-complementation at the end of a sentence could cause the first main verb in the sentence to be recognized as exhibiting the pattern.

[11] In addition to this exclusion pattern based on typical parts of speech sequences after a conjunction, the CLAWS annotation for 'ditto' tags could be used to recognize common multi-word sequences with *as* (e.g. *as well as*) and exclude these instances.

As a result of the extensive search procedure described above, a list of around 60 verb lemmas allowing 'intrusive *as*' in South Asian varieties of English was obtained. However, the vast majority of these verb lemmas only showed relatively low frequencies of occurrence with the 'intrusive *as*'-pattern (many were indeed hapax legomena), and frequently only occurred in a subset of the varieties in question. Since their inclusion in the same quantitative paradigm together with relatively high-frequency items would not allow drawing reliable conclusions, their discussion will not be a main focus in the present study. These low-frequency 'intrusive *as*' verb lemmas will instead be covered in the form of a brief excursus in Section 5.1, while the focus of the current study will be on an in-depth analysis of six verb lemmas found to display relatively high frequencies of 'intrusive *as*'.

In order to estimate whether 'intrusive *as*' is best described as a South Asian phenomenon or whether it might be a more universal feature of second/foreign-language acquisition, the analysis of these six verb lemmas was then extended to corpora of learner English. Suitable datasets were found in the *International Corpus of Learner English* (ICLE; Granger et al. 2009) as well as the *International Corpus Network of Asian Learners of English* (ICNALE; Ishikawa 2011). While the latter incorporates data from learner essays in ten Asian countries or areas with a scope of c. 1.2 million words,[12] the former corpus covers speakers from an even wider range of English-learning scenarios, including 3.7 million words from 16 mother tongue backgrounds, most of these European. It must be noted, though, that there is only very little overlap between the varieties covered in SAVE and those in ICNALE: in fact, it is only Pakistani speakers that are included in both datasets. On the other hand, ICNALE and ICLE both include data from Japan as well as from China. With these two corpora in place, the question whether 'intrusive *as*' is restricted to the Asian context, and thus likely emerged there, or may indeed be seen as a general language-learning strategy can be addressed.[13]

[12] The ICNALE also includes a category ENS, which incorporates essays from English native speakers from the USA, UK, Canada, Australia and New Zealand for reference and contrast. Since the ENS category did not display any instances of 'intrusive *as*' at all, it will not be further relevant for the purposes of this study.

[13] Of course, it must be noted that the design of the corpora is very different, since both the ICLE and ICNALE build upon essay data, while SAVE represents published material. The extent to which exonormative editing of the texts may have taken place is difficult to ascertain, however, it appears plausible to assume that the post-editing process within a publishing house will be more rigorous on average. Following this line of reasoning, the high number of findings for the 'intrusive *as*'-pattern in South Asian Englishes (see Section 5) seems even more striking.

4.2 Data preparation and coding

Only six verb lemmas with moderate to high frequency of ‘intrusive *as*’ remained after the search procedure described above: CALL, DECLARE, DEEM, DUB, NAME, and TERM. New concordances were then computed for each of these lemmas in order to (a) arrive at a list of all instances of complex-transitive constructions with these verb lemmas, both with and without ‘intrusive *as*’, and (b) also include those ‘intrusive *as*’ cases which were excluded by the semi-automatic approach due to the restrictions that were set in place (see Section 4.1). This final list of moderate- to high-frequency verbs allowing the ‘intrusive *as*’-construction was then annotated for several grammatical context features that were assumed to potentially influence the selection of the ‘intrusive *as*’-construction over its ‘regular’ counterpart without *as*. Table 1 summarizes the variables that were coded for (in addition to the dependent one PATTERN), as well as their potential values.

Table 1. Variables and variable levels used for annotation of verb lemmas exhibiting ‘intrusive *as*’-complementation

Variable	Description	Values
PATTERN	Complex-transitive predication realized with ‘intrusive *as*’ or without (i.e. regular pattern)	INTR-AS, REG
VAR	SAVE corpus component	BD, IN, LK, MV, NP, PK
VERB	Verb lemma of the main verb in the current clause	CALL, DECLARE, DEEM, DUB, NAME, TERM
VOICE	Voice of the main verb	ACTIVE, PASSIVE[a]
DIST	‘Distance’ between main verb and complement (excl. *as* if present)	0, 1, 2, 3, 4+
COMPL	Complement type, i.e. head of the complement (noun, adjective, nonfinite clause, other)	N, A, I, O[b]

[a] In some cases, the voice of the main verb was ambiguous, e.g. ‘… *it is call* [?] *as…*’, or the relevant verb appeared in a nonfinite clause modifying a constituent which was anonymized, e.g. ‘*[…], call* [?] *as […]*. In cases like these, the choice for coding of active vs. passive voice was based on the form of the full verb so as not to ‘correct’ the linguistic material; that is in both instances quoted, *call* was coded as active voice

[b] The variable level O was reserved for those cases in which the form of the complement was highly ambiguous or impossible to discern (e.g. if the complement had been anonymized in the data).

‘Voice’ and ‘distance’ were selected as coding categories in the hope of shedding some light on the origin of the ‘intrusive *as*’-construction in South Asian Englishes. One obvious explanatory parameter in the context of Postcolonial Englishes is always language contact, and the ‘intrusive *as*’-construction has been linked to the quotative construction available in many South Asian languages (cf. Sridhar 1992: 142–43), which may also be used “to name or label persons or things” (Hock 1982: 42), as in the following example from Marathi, an Indo-Aryan language:

(5) britiš yeṇyāpūrvī bāmbelā mumbaī mhaṇūn olakhat asat
British coming before Bombay Mumbai q. known was
‘Bombay was called Mumbai before the British came.’ (quoted after Kachru 1979: 69)

Since the putative source construction and/or its translation equivalent in Marathi and other South Asian languages employs a passive construction, we hypothesized that a higher incidence of ‘intrusive *as*’-constructions with verbs in the passive might indicate that the construction is indeed contact-induced. Another hypothesis to be tested required the coding of ‘distance’; if syntactic contexts where the verb and its complement were far removed from each other, as illustrated in example (6), favour ‘intrusive *as*’, then redundancy could be considered as a likely explanation. Redundancy “contributes to not only the securing of information (encoded twice) but also the ease of processing” (Schneider 2012: 65) and appears as a motivating factor in both ESL and EFL contexts. In examples such as (6), ‘intrusive *as*’ functions as a (technically redundant) means to signal the upcoming object complement and helps to process the structure of the sentence. Redundancy should be taken into consideration particularly with regard to longer sentences, then, since constructions with a distance of three or more words between the object and its complement are arguably more difficult to process than a construction where the complement immediately follows the object.

(6) Local BJP leaders **termed** Dr Yonzone, former principal of Kalimpong College and former chairman of School Service Commission, Hills, ***as*** an “outsider”. (SAVE-IN_SM_2004-04-22.txt)

Subsequent annotation of the learner data was performed in the same format as detailed above (cf. Table 1), with an obvious difference in the values of the VAR variable, which was coded by adding a shorthand for the country or area of the speakers (based on the corpus filenames, e.g. ICLE_BG or ICNALE_CHN) to the abbreviated name of the corpus (Table 2):

Table 2. Learner corpora components and abbreviations used (based on filenames)

Corpus	Components
ICLE	BG (Bulgarian), CN (Chinese), CZ (Czech), D (Dutch), FI (Finnish), FR (French), GE (German), IT (Italian), JP (Japanese), NO (Norwegian), PO (Polish), RU (Russian), SP (Spanish), SW (Swedish), TR (Turkish), TS (Tswana)
ICNALE	CHN (China), HKG (Hong Kong), IDN (Indonesia), JPN (Japan), KOR (Korea), PAK (Pakistan), PHL (Philippines), SIN (Singapore), THA (Thailand), TWN (Taiwan)

4.3 Statistical analysis

In order to establish whether speakers followed an interpretable pattern of choice for either complementation type based on the context factors as laid out in Table 1,[14] we turned to an analysis building on conditional inference trees (Hothorn et al. 2006). This method applies recursive partitioning algorithms to the data in order to determine along which independent variables (and combinations thereof) the dataset should best be separated into binary groups in order to predict best the outcomes of the dependent variable. The process is recursive in the sense that consecutive binary splits are calculated until further divisions no longer increase the predictive accuracy (previous applications of this technique can be found in Bernaisch et al. 2014, Lohmann 2013 as well as Tagliamonte & Baayen 2012), The analysis was performed using *R*'s (R Development Core Team 2015) `partykit` package, which offers an overhaul of the frequently used implementation within the `party` package, as well as the additional advantage of being coded entirely in R and thus offering advanced options for customization of the visual output. Within the current study, this approach had the dual benefit of producing readily interpretable tree structures capturing the most likely layers of choice that speakers may follow, as well as delivering significantly better results than a regular binary logistic regression using *R*'s `glm` function in terms of the amount of variation within the dataset explained by the model (cf. Section 5.1).

5. Results

5.1 'Intrusive *as*' in South Asian Englishes

Within the South Asian data, the six remaining main verb lemmas as identified in Section 4.2 covered 6,156 instances of either regular complex-transitive complementation or cases of 'intrusive *as*', albeit the latter group constitutes the clear minority, with only 753 (approximately 12.2%) complex-transitive cases exhibiting the 'intrusive *as*'-pattern. Figure 1 provides the overall findings and distributions across varieties and verb lemmas for both cases of 'intrusive *as*'-complementation (upper panel) as well as for regular cases for reference (lower panel). It is immediately recognizable that there are wide discrepancies between (a) the relative frequencies of occurrence of 'intrusive *as*' for the different verb lemmas in contrast to their 'regular' counterparts (i.e. differing widths of the cells/columns), as well as 'intrusive *as*' preferences between the six verb lemmas and the six components of

[14] This excludes PATTERN, which does not constitute a context factor but the dependent variable.

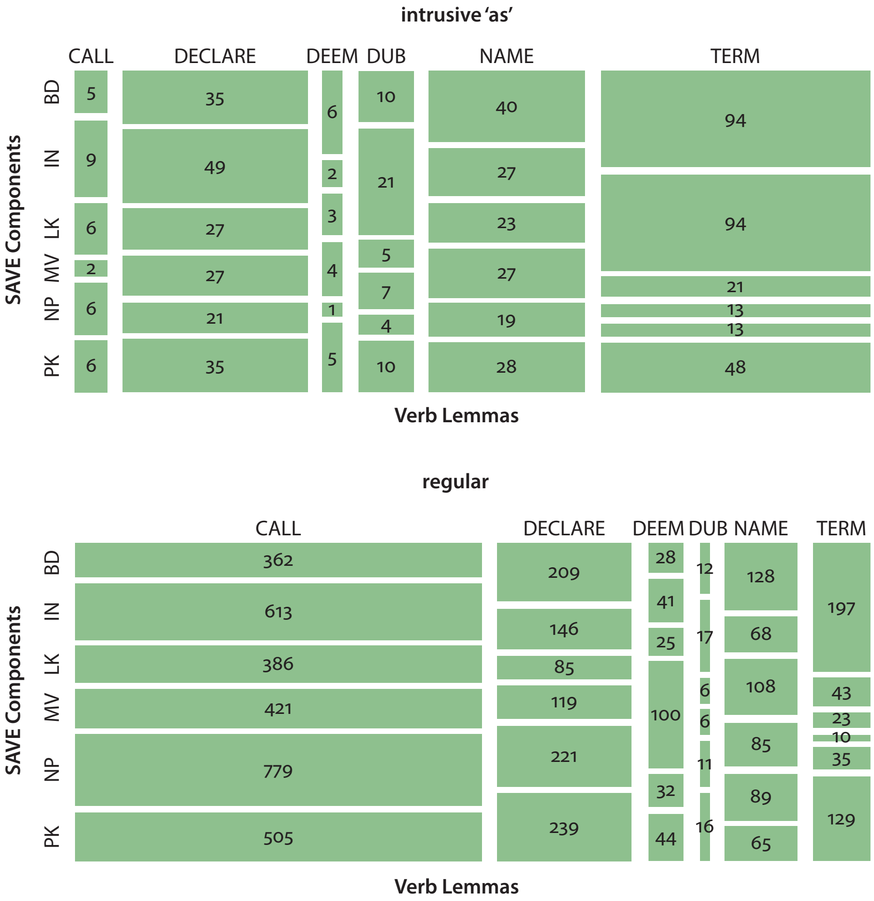


Figure 1. Absolute frequencies of ‘intrusive *as*’ (upper plot) and regular (lower plot) complementation cases across SAVE components and verb lemmas

the SAVE corpus (i.e. the heights of the cells/rows). The differences are particularly striking for the verb lemmas CALL and TERM: While no verb lemma truly prefers the ‘intrusive *as*’-pattern over ‘regular’ complementation, this difference is smallest for TERM, which is the third most frequent lemma in our South Asian data but advances to the most prominent one for ‘intrusive *as*’-complementation, displaying a relative frequency of 39.3% of the pattern. On the other hand, the roles are reversed in the case of the most frequent lemma, CALL, which is certainly intriguing, particularly given the fact that CALL AS is the most stereotypical verb appearing with ‘intrusive *as*’ and much more frequent in spoken IndE, where CALL in the complex-transitive pattern occurs with *as* in 18.3% of all cases (cf. Lange

2015).[15] TERM AS, on the other hand, occurs in 88.9% of all complex-transitive constructions featuring TERM in ICE-India spoken. Together with the relative prevalence of the 'intrusive *as*' option in the current data the pattern appears to have crossed the threshold from the spoken to the written language at least for this verb lemma and, since we are dealing with data derived from acrolectal newspaper English, has apparently become entrenched in standard(izing) Indian and other South Asian Englishes.

The overall differences between verb lemmas and SAVE components in terms of the frequencies of the 'intrusive *as*'-pattern were highly significant, if only with a weak correlation ($X\text{-}squared = 64.3$, $df = 20$, p-value < 0.001, Cramer's $V = 0.148$).[16] However, this initial overview cannot of course account for all variables within the study and their potentially complex interactions, which is why we turned to conditional inference trees as a method of analysis (cf. Section 4.3). The tree in Figure 2 explains 89.8% of variation in the dataset, which also represents a significant improvement over a baseline model that always predicts the most common pattern (i.e. 'regular' complementation, which accounts for 87.8% of the data; $p_{\text{binomial test}} < 0.001$) without considering any of the independent variables. All predictors are included in the tree and thus offer significant influences on the selection of one complementational pattern over the other. The tree structure should be interpreted as follows (cf. also Bernaisch et al. 2014): Starting at the top node (node 1), one moves along the edges (branches) of the tree towards either the left or right subsequent node, thus restricting the further inspection of the dataset to (a) the independent variable indicated within the node as well as (b) the variable level to the one given on the respective edge of the graph. This process is repeated until a terminal node is reached, which then provides (a) information on the absolute number of points in the dataset with the combination of variables and their levels selected while moving along the edges, and (b) a bar plot of the relative distribution of the two complementation patterns.[17] For example, the leftmost terminal node reveals that there are 1,380 instances of complex-transitive predication in the active voice (edge between nodes 1 and 2) and with either the verb lemmas CALL or DEEM (edge between nodes 2 and 3), which is later further restricted to only the lemma CALL (after node 4) as well as with 0–3 words between object and

[15] IndE is the only South Asian variety of English for which spoken data (from ICE-India) are available.

[16] The lemma DEEM had to be excluded from this test, since expected frequencies were too low.

[17] The standard settings for `ctree` not only include numerical identifiers for the inner nodes but also for the terminal nodes. These were not displayed in Figure 2 in order to improve readability, but this explains why node numbers are not consecutive in the figure.

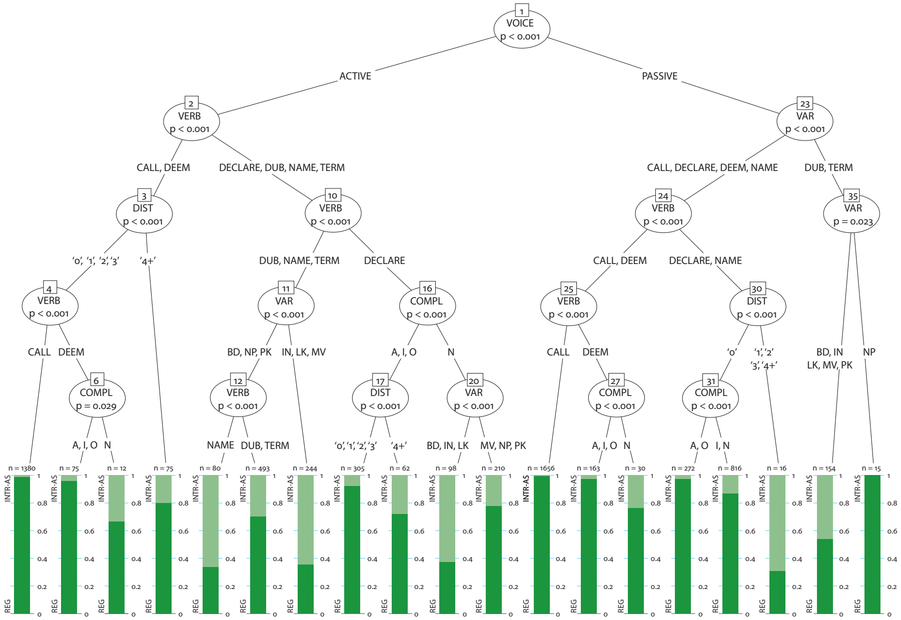


Figure 2. Conditional inference tree for the South Asian (SAVE) dataset. This plot has been restricted to five layers (out of six significant splits) to fit it on the page.

complement (node 3; excluding *as* from the count if present). Of these 1,380 cases, only a small fraction (actually c. 1.1%) are constituted by instances of ‘intrusive *as*’.

Inspecting all terminal nodes, one quickly comes to the conclusion that, with the exception of four combinations of variable levels, ‘intrusive *as*’ is never the preferred pattern in any of the SAVE components (i.e. accounts for >50% of occurrences). These four exceptions are the following:

1. Active NAME in BD, NP, or PK,
2. Active DUB, NAME or TERM in IN, LK, or MV,
3. Active DECLARE with a nominal complement in BD, IN or LK,
4. Passive DECLARE or NAME with a non-consecutive object and complement.

In addition to those few cases of clear preferences for ‘intrusive *as*’, however, the tree also captures some intriguing recurrent shifts in frequency distributions with regard to single predictors as well as sub-structures that can be repeatedly observed. VOICE emerges as the most important distinction within the dataset. Given the fact that languages having a quotative particle use it particularly frequently in passive voice, this initial split may not be very surprising. In fact, however, it is only in 28.2% of all ‘intrusive *as*’ cases that the main verb is passive (as opposed to 53.9% of all cases of ‘regular’ complementation), and only 6.8% of all passive cases display ‘intrusive *as*’ (against 17.8% for active cases). A tentative explanation for this unexpected distribution may be seen in the fact that in passive cases, the complement is usually separated from the constituent it refers to by the main verb, while in active cases, object and complement are usually in a consecutive sequence, making the 'intrusive *as*'-construction more relevant in the latter situation.

The second-level binary splits on each of the two main branches of the tree follow a less readily understandable pattern. Following the various branches of the tree structure that emerge from splits based on verb lemmas, it can be seen that, both in active and passive cases, there appear to broadly be three main verb groups: CALL and DEEM appear to share the greatest degree of similarities, remaining on the same tree branches until the fourth layers/splits both in active and passive. DUB and TERM also remain on the same branch for most of the splits (see nodes 12 for active and 35 for passive cases), while also showing the greatest dissimilarities to the first group (indicated by distances within the tree). NAME and DECLARE, finally, follow a less clear pattern, being clustered together (with CALL and DEEM) in passive cases while showing differing preferences for 'intrusive *as*' in active cases (where NAME follows a similar pattern with DUB and TERM).

As for the third independent variable, distance, a clearer pattern can both be found in active and passive cases, in that the higher distances seem to make 'intrusive *as*' increasingly more likely. For active cases, this only significantly affects pattern selection if the number of words between the main verb and the complement (or *as*) exceeds three (nodes 3 and 17). In passive cases, only a distance of zero (i.e. complement directly after main verb) significantly prefers the 'regular' pattern, while higher distances show a clear preference for realization with 'intrusive *as*' (see node 30). This can be understood as a measure of employing redundancy in cases where distance is uncharacteristically high (i.e. higher than the typical zero in passive cases; as well as for long objects or other uncharacteristic features such as postponed adverbs, appositions, etc. in active cases).

The above interpretation also matches the distribution of complementation choices according to the final criterion, complement type. In numerous cases, significant distinctions are made between longer/heavier (primarily NP) and shorter constituents (particularly AP), see nodes 6, 16, 27, and 31. The case of nonfinite clauses as complements is less clear, with them sometimes behaving similarly to NP complements, while in most cases being more similar to AP complements. It should be noted, however, that both nonfinite clauses and 'other' complements represent a clear minority of data points, making their results potentially far less informative than in the case of the two main types of complements observed in the data.

Summing up the tree model for the SAVE data, the overall situation appears to be that, while the main distinguishing criterion is the voice of the main verb, cognitive processing demands of the immediate linguistic contexts appears to be a major influence on pattern selection, with higher distances between main verb and complement as well as heavier constituents in complement position generally favouring the selection of 'intrusive *as*' as opposed to 'regular' complementation.

As far as the rarer types we found in our South Asian data are concerned, we believe that they lend support to a hypothesis that views semantico-structural analogy as the main reason for the usage of 'intrusive *as*'. Semantico-structural analogy is defined by Mukherjee as "a process by means of which nonnative speakers of English as a second language introduce new forms and structures into the English language on grounds of semantic and formal templates that already exist in the English language system" (2007: 175–6). The following sentence includes an example beyond the range of the principal verbs of labelling, naming etc. listed in Nihalani et al. (2004):

(7) Mr. Hardingham went to school with Maldivians in UK and has been to the Maldives, which he **finds *as*** a second home. (SAVE-MV_DO_082.txt)

As mentioned before, our analysis yielded 60 different verb lemmas in the 'intrusive *as*'-pattern, only few of which occurred more than once or a few times. However, the high number of different verbs and examples such as (7) suggest the possibility of more verbs occurring with 'intrusive *as*' in the future (cf. Koch & Bernaisch 2013 for a related case involving 'new ditransitives' in South Asian Englishes). As Bybee (1995) and Bybee & Thompson point out, "the type frequency of a pattern determines its degree of productivity" (2007: 275), suggesting that an increased overall usage of 'intrusive *as*' might eventually lead not only to an even higher frequency with the verbs that already occur rather frequently, but to an extension to other verbs as well. However, speaking of the 'intrusive *as*'-construction as a productive one should be avoided, since the notion that frequency and productivity cannot be equated (Bauer 2001: 48), originally discussed with regard to morphological productivity, applies here, too. Rather, thinking of the 'intrusive *as*'-construction as being (potentially) influenced by frequency effects in the long run seems sensible. If the construction occurs with even more lemmas in the future, it is likely that creative usage and increasing frequencies with certain verbs played a role, although this is a hypothetical assumption requiring long-term studies.

5.2 'Intrusive *as*' in Learner Englishes

Given the findings discussed above, it seems plausible to describe 'intrusive *as*' as a potentially substrate-derived feature that is preferred in those cases where, for contextual reasons, additional marking of the complex-transitive situation is required but not offered by Standard English grammar. As the question of contact-induced change is difficult to answer conclusively with the current results based only on South Asian data, we turned to the ICLE and ICNALE corpora of learner English (cf. Section 4) in order to estimate whether the 'intrusive *as*'-construction constitutes a specific South Asian phenomenon or indeed a more general feature of language acquisition.

Again, analysis of the dataset was carried out by means of a conditional inference tree. Before further analysis, the dataset was also reduced to only include those learner varieties that displayed 'intrusive *as*' at least once, excluding five components from both ICLE (FR, GE, IT, NO, TS) and ICNALE (CHN, HKG, KOR, PHL, TWN). Additionally, from the list of verb lemmas, DUB needed to be excluded since it could only be attested with 'regular' complementation, and that only once. However, the number of 'intrusive *as*' findings is still very low in comparison to the SAVE data, with only 5.3% of all complex-transitive complementations following this pattern (the differences in complementation pattern-frequencies being significantly different; $X\text{-}squared = 44.6$, $df = 1$, $p\text{-value} < 0.001$). While a number of significant distinctions arise within the dataset (cf. Figure 4), the overall tree model is not significantly better at predicting pattern choice than a baseline model (which always predicts 'regular' complementation, i.e. 94.7% of all instances in this dataset). Figure 3 summarizes the distribution of 'intrusive *as*' and 'regular' complementation cases across the remaining five verb lemmas and the ICLE and ICNALE components, while Table 3 additionally quantifies the absolute and relative frequencies of both patterns ($X\text{-}squared = 67.1$, $df = 15$, $p\text{-value} < 0.001$).[18] Again, CALL represents the most prominent verb lemma but shows a much lower than expected relative frequency of 'intrusive *as*'-predication. On the other hand, TERM provides relatively more 'intrusive *as*' cases than 'regular' ones (even if only by a slight margin). The differences in Figure 3 are significant at $p < 0.001$ ($X\text{-}squared = 116.92$, $df = 4$).[19]

Within the remaining varieties in the dataset, the number of 'intrusive *as*' cases differed considerably, from individual occurrences only (BG, PO) to over 25% of all instances of the relevant verb lemmas. Interestingly, it is the Asian varieties which consistently display the highest relative frequencies of the pattern (but remember that these are based on some very low absolute frequencies and that some Asian varieties were amongst those excluded due to zero occurrences of 'intrusive *as*') — with the exception of Turkish learners of English, where the highest relative frequency of 'intrusive *as*' could be attested, both in terms of absolute and

[18] Since about a third of all expected frequencies of this test were smaller than five, and a Fisher exact test failed to compute, a simulation-based chi-squared test was run, which returned the same significance result.

[19] Four out of ten expected frequencies were smaller than five, which is why a Fisher exact test was run which also yielded highly significant results.

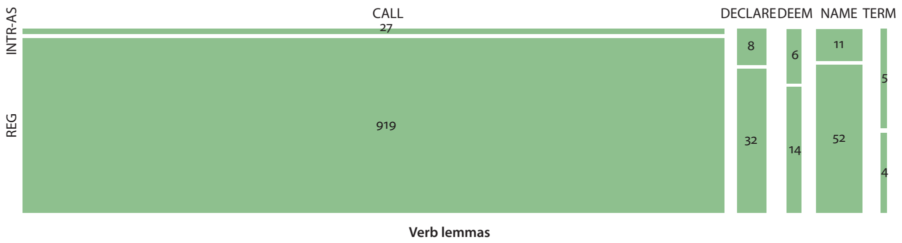


Figure 3. Absolute frequencies of 'intrusive *as*' and 'regular' complementation cases across the remaining verb lemmas and varieties in the learner dataset (ICLE & ICNALE)

Table 3. Absolute and relative frequencies of 'intrusive *as*'-complementation (and 'regular' for comparison) in the remaining learner corpus components (ICLE & ICNALE)

Corpus	**Variety**	**'intrusive *as*' vs. 'regular' (abs. freq.)**		**'intrusive *as*' (rel. freq.)**
ICLE	BG	1	70	1,4%
	CN	7	57	10,9%
	CZ	3	92	3,2%
	D	3	108	2,7%
	FI	2	99	2.0%
	JP	3	74	3,9%
	PO	1	63	1,6%
	RU	4	101	3,8%
	SP	4	109	3.5%
	SW	3	87	3,3%
	TR	14	41	25.5%
ICNALE	IDN	3	37	7.5%
	JPN	2	37	5.1%
	PAK	3	20	13.0%
	SIN	2	9	18.1%
	THA	2	18	10.0%

relative numbers. A possible explanation for this might be the Turkish quotative verb *demek* and its derived particle *diye* (cf. Kornfilt 1997), both of which could, in theory, be seen as models for the 'intrusive *as*'-construction. Despite the presence of a quotative in Turkish, however, there are two factors potentially mitigating its explanatory value. Most importantly, the Turkish quotative is not used in contexts similar to those we identified in the 'intrusive *as*'-pattern; rather, it functions as a grammaticalized *verbum dicendi* to indicate direct or reported speech.[20] Secondly, the mother tongues of other learners featured in the ICLE — most notably the Bulgarian learners — feature quotatives as well, which is not suggested by their corresponding 'intrusive *as*' frequencies. For the present discussion, the assumption is reasonable that the presence of a quotative in Turkish favours 'intrusive *as*'-usage, but an exhaustive discussion of the problem remains work for future analyses.

As could already be inferred from Figure 3, CALL follows very different patterns than the other verbs in the dataset, and this is also represented in the conditional inference tree in Figure 4. The tree model never predicts 'intrusive *as*', and only frequency differences can be observed, but these appear to again favour (if only relatively slightly) distance as a recurrent predictor. For both groups of verbs (CALL vs. all other lemmas), increasing distance actually leads to higher frequencies of 'intrusive *as*', but for the lack of data for all lemmas except CALL, the latter group's findings cannot be significantly differentiated internally, and thus this is not represented in the tree. However, averaging across all verbs and (remaining) varieties (and thus ignoring the inherent differences) for the moment, we can indeed find significant distinctions between relatively little 'intrusive *as*'-usage for distances 0 (3.9%) and 1 (7.6%) and markedly higher relative frequencies for the longer distances (16.4%) if we combine all distances ≥2 due to low frequencies). Neither the complement type (only 7 instances of 'intrusive *as*' without a nominal head) nor the voice of the main verb (quite in contrast to the SAVE data), with an almost even distribution between active and passive cases, provide significant splits in the data even at this very general level, and are likewise excluded as predictors by the tree model. While no statistically significant distinctions can be observed within the Turkish speakers' data, in the rest of the data there appear to be two major groups of varieties (Asian varieties with the exception of ICNALE_SIN vs. the rest), and while the latter group basically never uses 'intrusive *as*' for distances 0, 1, and 4+ (0.6%), they employ the pattern in 10.8% of cases with distances 2 and 3.

[20] This usage is in line with the traditional definition of quotatives; cf. Chapter 2 in Buchstaller (2014).

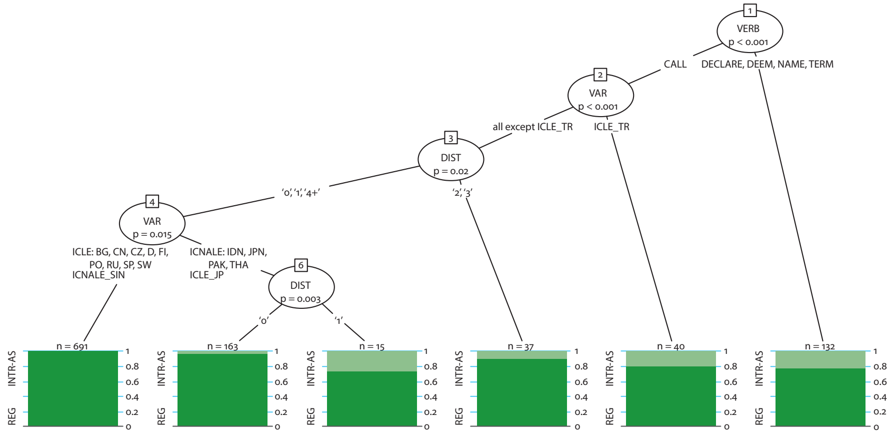


Figure 4. Conditional inference tree for the learner (ICLE & ICNALE) dataset

The former group does not behave significantly differently for these two distance values, but appear to differentiate between very little 'intrusive *as*'-usage for distance 0 (4.3%) and markedly higher preference for distance 1 (26.7%). Overall, these numbers lend some support to the idea that 'intrusive *as*' is more likely to occur in contexts with longer distance between the verb and the complement.

5.3 Discussion of findings

First of all, our data clearly show that 'intrusive *as*' is not a recurrent feature of Learner Englishes: five varieties both out of 16 represented in the ICLE corpus and 15 in the ICNALE corpus are devoid of tokens, and generally the frequency of 'intrusive *as*'-constructions is much lower than in the SAVE data. Further, the frequencies are not evenly distributed within the datasets: with the exception of Turkish, it is the Asian Learner varieties which display the highest frequencies. These straightforward descriptive facts should be kept in mind before entering into a more detailed scrutiny of the correlations emerging from the conditional inference tree analysis. This difference in frequency and distribution already indicates that at least with regard to the feature under discussion, ESL and EFL varieties pattern quite differently. Further differences become apparent with the analysis based on conditional inference trees.

Some of the significant patterns in the data analysis presented above lend tentative support to our initial hypotheses. A contact explanation for the occurrence of 'intrusive *as*' becomes more likely when we complement our ESL data with EFL data from those speakers whose linguistic repertoires comprise languages which have a quotative construction: 'intrusive *as*'-constructions show the highest frequencies in Asian and Turkish Learner Englishes. However, this correlation, interesting as it is, appears too weak to be taken as the sole witness for a case of contact-induced language change. Much more detailed analyses of the source constructions in the relevant background languages would be required in order to identify what multilingual speakers identify as the 'pivot' of the construction (cf. Matras & Sakel 2007), which they then map onto the target language English. Additionally, it must be noted that there was also variation in the South Asian speakers' preferences for 'intrusive *as*' both in terms of varieties as well as verb lemmas which could not be sufficiently explained with the current annotation scheme.

Only one similarity between South Asian Englishes and Learner Englishes emerging from our data analysis can be attributed to redundancy as a motivating factor. Heavy NPs and/or a considerable distance between the verb and its complement definitely favour the occurrence of 'intrusive *as*'. In that respect, ESL and EFL speakers resort to similar strategies for enhancing ease of processing both in production and reception; and if we follow Weinreich et al. (1968) or Matras & Sakel (2007) in distinguishing actuation and propagation of an innovative linguistic feature, then the issue of demarcating errors from innovations fades into the background. Multilingual speakers' creativity drives actuation on an individual, spontaneous level and is motivated by

> the need to perform effectively in communicative interaction while adhering, on the one hand, to the rules about the selection of clearly-identifiable phonological substance (matter) from the language that is appropriate in the particular context, while at the same time exploiting constructions that are available to the speaker in his/her entire repertoire of linguistic-communicative structures. It is this underlying motivation, and the similarities among the creative processes that arise from it in different types of situations, which in our opinion justify examining cases of diachronic change alongside cases of learners' and bilinguals' spontaneous performance. (Matras & Sakel 2007: 854)

That is, the distinction between error and innovation only plays a role at the level of propagation, impinging upon the likelihood that an original feature stabilizes within the repertoire of a speech community. The same creative processes and cognitive mechanisms may have radically different outcomes in the long run, solely depending on sociolinguistic factors.

6. Conclusion

In this paper, we analysed the distribution and frequency of the 'intrusive *as*'-construction in South Asian Englishes represented by the SAVE corpus, and Learner Englishes as represented by the ICLE and ICNALE corpora. Although we found tokens of ‘intrusive *as*’ both in South Asian Englishes and some Learner Englishes, the feature occurred much more frequently and consistently in the SAVE corpus.

Despite the differences in terms of frequency, we found redundancy to be a motivating factor for the use of ‘intrusive *as*’ in both ESL and EFL, suggesting similar processing strategies across different varietal types. Ultimately, this finding further justifies comparing individual (learner) varieties to prototypical cases of second-language varieties and Learner Englishes (as well as ENL) on a continuum (as suggested in, amongst others, Biewer 2011: 28 and Buschfeld 2011: 219), where some of the Learner Englishes we analyzed appear closer to ESL and others fall closer to the EFL pole of the continuum. It should be noted, however, that the complete absence or relatively low frequency of the ‘intrusive *as*’-construction in some of the ICLE and ICNALE subcorpora turns the feature under investigation into a bit of a critical case, because the differences in distribution across ESL and EFL are potentially too drastic to promote an integrated approach (at least with regard to ‘intrusive *as*’). Again, though, the idea of distinguishing between actuation and propagation might be helpful in this case, if, as Edwards claims, “the distinction between ESL and EFL is largely sociolinguistic in nature” (2014: 24). In addition, future work on the ‘intrusive *as*’-construction needs to take a closer look at the typological structure of the involved contact languages in order to substantiate a contact-based argumentation. If the usage of ‘intrusive *as*’ in both ESL and EFL is indeed systematically influenced by constructions found in the speakers’ L1s, there would be yet another reason to dismiss a sharp distinction between variety types. Moreover, the datasets could be expanded both in terms of varieties as well as predictor variables in order to be able to confirm or re-evaluate the initial findings presented in this paper.

Acknowledgements

We would like to thank Tobias Bernaisch for his invaluable input on the statistical evaluation of the data, as well as Hannes Föst for his help in annotating the many thousands of data points. Of course, any shortcomings in the text remain our responsibility alone.

References

Bauer, L. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486210

Bernaisch, T., Gries, S. Th. & Mukherjee, J. 2014. “The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes”, *English World-Wide* 35(1), 7–31. https://doi.org/10.1075/eww.35.1.02ber

Bernaisch, T., Koch, C., Schilk, M. & Mukherjee, J. 2011. *Manual to the South Asian Varieties of English (SAVE) Corpus*. Giessen: Justus Liebig University, Department of English.

Biewer, C. 2011. "Modal auxiliaries in second language varieties of English: A learner's perspective". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 7–33. https://doi.org/10.1075/scl.44.02bie

Buchstaller, I. 2014. *Quotatives. New Trends and Sociolinguistic Implications*. Malden: Wiley.

Buschfeld, S. 2011. *The English Language in Cyprus: An Empirical Investigation of Variety Status*. PhD dissertation, University of Cologne.

Buschfeld, S. 2013. *English in Cyprus or Cyprus English. An Empirical Investigation of Variety Status*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g46

Buschfeld, S. 2014. "English in Cyprus and Namibia: A critical approach to taxonomies and models of World Englishes and Second Language Acquisition research". In S. Buschfeld, T. Hoffmann, M. Huber & A. Kautzsch (Eds.), *The Evolution of Englishes. The Dynamic Model and beyond*. Amsterdam: John Benjamins, 181–202. https://doi.org/10.1075/veaw.g49

Buschfeld, S. & Kautzsch, A. 2016. "Towards an integrated approach to postcolonial and non-postcolonial Englishes", *World Englishes*. https://doi.org/10.1111/weng.12203

Bybee, J. 1995. *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins.

Bybee, J. & Thompson, S. 2007. "Three frequency effects in syntax". In J. Bybee (Ed.), *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press, 269–278. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

Deshors, S.C. 2014. "A case for a unified treatment of EFL and ESL: A multifactorial approach", *English World-Wide* 35(3), 277–305. https://doi.org/10.1075/eww.35.3.02des

Edwards, A. 2014. *English in the Netherlands. Functions, Forms and Attitudes*. PhD dissertation, University of Cambridge.

Edwards, A. & Laporte, S. 2015. "Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels", *English World-Wide* 36(2), 135–169. https://doi.org/10.1075/eww.36.2.01edw

Gilquin, G. 2015. "At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared", *English World-Wide* 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Goffin, R.C. 1934. "Some notes on Indian English". In A.A. Goffin & R.C. Goffin (Eds.), *S.P.E. Tract No. XLI*. Oxford: Clarendon Press, 20–32.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (Eds.). 2009. *International Corpus of Learner English*. Version 2 (Handbook + CD-ROM). Louvain-la-Neuve: Presses universitaires de Louvain.

Heaton, J.B. & Turton, N.D. 1997. *Longman Dictionary of Common Errors*. Harlow: Longman.

Hock, H.H. 1982. "The Sanskrit quotative: A historical and comparative study", *Studies in the Linguistic Sciences* 12(2), 39–85.

Hothorn, T., Hornik, K. & Zeileis, A. 2006. "Unbiased recursive partitioning: A conditional inference framework", *Journal of Computational and Graphical Statistics* 15(3), 651–674. https://doi.org/10.1198/106186006X133933

Huddleston, R. 2002. "The verb". In R. Huddleston & G.K. Pullum (Eds.), *The Cambridge Grammar of the English Language (CGEL)*. Cambridge: Cambridge University Press, 213–322.

Huddleston, R. & Pullum, G.K. (Eds.). 2002. *The Cambridge Grammar of the English Language (CGEL)*. Cambridge: Cambridge University Press.

Ishikawa, S. 2011. "A new horizon in learner corpus studies: The aim of the ICNALE project". In G. Weir, S. Ishikawa & K. Poonpon (Eds.), *Corpora and Language Technologies in Teaching, Learning and Research*. Glasgow: University of Strathclyde Publishing, 3–11.

Kachru, B.B. 1985. "Standards, codification and sociolinguistic realism: The English language in the outer circle". In R. Quirk & H.G. Widdowson (Eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press for The British Council, 11–30.

Kachru, B.B. 1991. "Liberation linguistics and the Quirk concern", *English Today* 25(1), 3–13. https://doi.org/10.1017/S026607840000523X

Kachru, Y. 1979. "The quotative in South Asian Languages", *South Asian Languages Analysis* 1, 63–77.

Koch, C. & Bernaisch, T. 2013. "Verb complementation in South Asian English(es): The range and frequency of 'new' ditransitives". In G. Andersen & K. Bech (Eds.), *English Corpus Linguistics: Variation in Time, Space and Genre*. Amsterdam: Rodopi, 69–89.

Kornfilt, J. 1997. *Turkish*. Oxon and New York: Routledge.

Lange, C. 2014. People call it as city of garden – tracing the 'intrusive *as*' construction in South Asian varieties of English. Paper presented at the *30th South Asian Languages Analysis Roundtable* (SALA 30), University of Hyderabad/India, 6-8 February 2014.

Lange, C. 2015. "The 'intrusive *as*'-construction in South Asian varieties of English", *World Englishes* 35(1), 133–146. https://doi.org/10.1111/weng.12173.

Lohmann, A. 2013. "Is tree hugging the way to go? Classification trees and random forests in linguistic study", *Views (Vienna English Working Papers)* 22, 1–17. Available at: https://anglistik.univie.ac.at/fileadmin/user_upload/i_anglistik/Department/Views/Uploads/VIEWS_22_2013_Lohmann.pdf (accessed March 2018).

Matras, Y. & Sakel, J. 2007. "Investigating the mechanisms of pattern replication in language convergence", *Studies in Language* 31(4), 829–865. https://doi.org/10.1075/sl.31.4.05mat

Mukherjee, J. 2007. "Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium", *Journal of English Linguistics* 35(2), 157–187. https://doi.org/10.1177/0075424207301888

Nihalani, P., Tongue, R.K., Hosali, P. & Crowther, J. 2004. *Indian and British English: A Handbook of Usage and Pronunciation*. Delhi: Oxford University Press.

Quirk, R. 1990. "Language varieties and standard language", *English Today* 21, 3–10. https://doi.org/10.1017/S0266078400004454

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

R Development Core Team. 2015. *The R Project for Statistical Computing*. Version 3.2.2. Available at: http://www.r-project.org/ (accessed March 2018).

Rohdenburg, G. 1996. "Cognitive complexity and increased grammatical explicitness in English", *Cognitive Linguistics* 7(2), 149–182. https://doi.org/10.1515/cogl.1996.7.2.149

Schneider, E.W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618901

Schneider, E.W. 2012. "Exploring the interface between World Englishes and Second Language Acquisition – and implications for English as a Lingua Franca", *Journal of English as a Lingua Franca* 1(1), 57–91. https://doi.org/10.1515/jelf-2012-0004

Schneider, E.W. 2014. “New reflections on the evolutionary dynamics of World Englishes”, *World Englishes* 33(1), 9–32. https://doi.org/10.1111/weng.12069

Singh, R. 2007. “The nature, structure, and status of Indian English”. In R. Singh (Ed.), *Annual Review of South Asian Languages and Linguistics 2007*. Berlin and New York: Mouton de Gruyter, 33–46.

Smith-Pearse, T.L.H. 1968. *The English Errors of Indian Students*. Delhi: Oxford University Press.

Sridhar, S.N. 1992. “The ecology of bilingual competence: Language interaction in indigenized varieties of English”, *World Englishes* 11(2–3), 141–150. https://doi.org/10.1111/j.1467-971X.1992.tb00058.x

Tagliamonte, S. & Baayen, H. 2012. “Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice”, *Language Variation and Change* 24, 135–178. https://doi.org/10.1017/S0954394512000129

Weinreich, U., Labov, W. & Herzog, M.I. 1968. “Empirical foundations for a theory of language change”. In W.P. Lehmann & Y. Malkiel (Eds.), *Directions for Historical Linguistics: A Symposium*. Austin and London: University of Texas Press, 97–195.

Williams, J. 1987. “Non-native varieties of English: A special case of language acquisition”, *English World-Wide* 8(2), 161–199. https://doi.org/10.1075/eww.8.2.02wil

Yadurajan, K.S. 2001. *Current English. A Guide for the User of English in India*. New Delhi: Oxford University Press.

Detecting innovations in a parsed corpus of learner English

Gerold Schneider and Gaëtanelle Gilquin
University of Konstanz & University of Zurich / University of Louvain

In research on L2 English, recent corpus-based studies indicate that some non-standard forms are shared by indigenized (ESL) and foreign (EFL) varieties of English, which challenges the idea of a clear dichotomy between innovation and error. We present a data-driven large-scale method to detect innovations, test it on verb + preposition structures (including phrasal verbs) and adjective + preposition structures, and describe similarities and differences between EFL and ESL. We use a dependency-parsed version of the International Corpus of Learner English to automatically extract potential innovations, defined as patterns of overuse compared to the British National Corpus as reference corpus. We measure overuse by means of collocation measures like O/E or T-score, and compare our results with similar results for ESL. In both quantitative and qualitative analyses, we detect similarities between the two varieties (e.g. *discuss about*) and dissimilarities (e.g. *accuse for*, only distinctive for EFL). We report more verb/adjective + preposition combinations than previous studies and discuss the roles of analogy and transfer.

Keywords: Learner English, English as a Foreign Language (EFL), English as a Second Language (ESL), data-driven approach, corpus linguistics, verb-preposition constructions, Cognitive Linguistics, Error Analysis, collocations, linguistic innovations

1. Introduction

Since the era of Error Analysis, much focus in interlanguage studies has been on non-native-like features. Initially restricted to cases of misuse, the advent of learner corpus research has made it possible to identify cases of under- and/or overuse, which equally contribute to the non-nativeness of learner production (e.g. Granger 2009; Nesselhauf 2005; Salazar 2014: 180).

https://doi.org/10.1075/bct.98.03sch

Recent theoretical and technological developments, however, have changed the way non-native features are considered and investigated. From a theoretical perspective, attempts have been made to bridge the paradigm gap that has long existed between research on learner language and on indigenized second-language varieties (see Gilquin 2015a; Gut et al. 2015; Mukherjee & Hundt 2011). Adopting an empirical approach, these studies have shown that learner English, or English as a Foreign Language (EFL), and indigenized varieties of English, or English as a Second Language (ESL), share certain non-standard features, be it in the domain of syntax (e.g. Edwards 2014), lexis/lexico-grammar (e.g. Edwards & Laporte 2015; Gilquin 2011; Gilquin & Granger 2011; Götz & Schilk 2011; Laporte 2012; Nesselhauf 2009) or phonology (e.g. Fuchs & Wunder 2015; Götz 2015). It has therefore become impossible to simply disregard any differences between EFL and ENL (English as a Native Language) as errors that should be eliminated, especially when they coincide with the "innovations" that are found in ESL.

From a technological perspective, it has become increasingly common to enrich learner corpora with different kinds of annotation (cf. van Rooy 2015), including syntactic annotation (e.g. Dickinson & Ragheb 2009; Rosén & Smedt 2010). This, in turn, has allowed for more sophisticated types of automatic data extraction, including extraction of L2 patterns (e.g. Díaz-Negrillo et al. 2013; Ng et al. 2014; Schneider & Hundt 2009).

In this paper, we take advantage of these theoretical and technological developments to examine non-native-like combinations of verb + prepositional phrase (PP) and adjective + PP. Starting from the assumption that not all non-native-like combinations are necessarily errors, we set out to identify potential instances of innovations in a corpus of learner English. The first steps of this identification are fully automatic, thanks to the syntactic parsing of the learner corpus and the comparison with a large parsed corpus of native English by means of collocation statistics. Such a corpus-driven approach is what distinguishes our study from most other studies that have sought to bridge the gap between EFL and ESL research (see above). It greatly facilitates the retrieval of phenomena that may be relatively rare in the data and would normally require large amounts of manual work (cf. Schneider & Zipp 2013).

In particular, we address the following research questions. First, can the patterns of overuse which we observe using collocation statistics deliver combinations that are specific to EFL and/or to ESL (RQ1)? Second, does the same method also allow us to detect which patterns of verb + PP and adjective + PP are more typical for EFL and which for ESL (RQ2)? Third, does the method give us the tools to find more patterns than have been previously described (RQ3)? Fourth, does the method give us the tools to distinguish between error and innovation (RQ4)?

The paper is structured as follows. In Section 2 we show that verb/adjective + PP constructions are an important characteristic of L2 and that using parsed data can lead to insightful observations. In Section 3, we present our data and our method using collocation statistics. In Section 4 we give quantitative results, comparing the triangular relationship between EFL, ESL and ENL, while in Section 5 we provide a qualitative analysis. Section 6 addresses the question whether our method allows us to make a distinction between error and innovation, before the conclusion in Section 7.

2. Motivation

2.1 Verb + preposition and adjective + preposition combinations

In order to investigate differences between EFL, ESL and ENL use, one can in principle search for differences in linguistic patterns at any linguistic level: phonological, lexical, morphological, syntactic. The first of these is not available, given our selection of corpus data. According to Schneider (2004: 229), crucial differences between varieties occur at the level of lexico-grammar. It is the interaction between lexis and grammar that is open to variation, and it typically involves collocational preferences and verb complementation.

Collocational preferences can be captured by collocation measures, which we introduce in Section 3.1. Concerning complementation, we investigate combinations of verbs/adjectives and prepositions or verbal particles. Verb + PP combinations constitute an important and frequent (Cornell 1985) subgroup of verb complementation, and exhibit a high rate of innovation, both in ESL and EFL. In ESL, Indian English, for instance, presents a high degree of innovation in its use of prepositional verbs (Mukherjee & Hoffmann 2006); in EFL phrasal verbs represent "one of the most notoriously challenging aspects of English language instruction" (Gardner & Davies 2007: 339; see also Gilquin 2015b or Deshors 2016). New verb + PP combinations are a promising research object, as demonstrated by Nesselhauf (2009), who describes instances of combinations (e.g. *discuss about*, *enter into*, *request for*) which she found both in ESL and EFL varieties. The comparison between ESL and EFL also highlights the paradox that some of the "innovations" identified in ESL varieties coincide with those held up as common "errors" in EFL (cf. Gilquin 2017).

Prepositions have been shown to be difficult to acquire for non-native speakers of English, leading to avoidance, non-standard uses, etc. (see Gilquin & Granger 2011: 59–60). Investigations of selected prepositions and verbal particles, for example the preposition *into* (Gilquin & Granger 2011) or the particle *up* (Gilquin

2011), revealed interesting correlations between EFL, ESL and ENL use. Gilquin (2011) shows that both EFL and ESL speakers tend to overuse phrasal verbs in writing, while at the same time underusing them in speech, which indicates lacking ability to adapt to register conventions, although the degree differs, with ESL speakers being more sensitive to register variation. In order to address the question of how other prepositions and verbal particles pattern, the manual annotation work would be enormous. Fortunately, we can use automatic annotation, as shown in the following.

2.2 Syntactically parsed data

Corpus-based descriptions of ESL varieties (see Sand 2004; Schneider 2004 or Sedlatschek 2009) and EFL varieties (e.g. Nesselhauf 2005) have typically been conducted on orthographic, i.e. not annotated, corpora. Automatic annotation has risks and benefits. It has the risk of errors adding to the noise of corpus imbalances, which is why we propose to use a semi-automatic approach, in which type-based candidates are presented to the user. For her investigation, Gilquin (2011) had to manually differentiate between *up* as a verbal particle and other uses, while we can now rely on automatically annotated data. Automatic annotation also offers the advantage that unrestricted amounts of data can be processed, which in comparison to Gilquin (2011, 2015a) allowed us to include the whole of the International Corpus of Learner English (ICLE), but also most components of the International Corpus of English (ICE) and the written part of the British National Corpus (BNC) (see Section 3.2 for a presentation of the corpora), and in addition made it possible to step up from selected particles/prepositions to all particles/prepositions, and all combinations they may have with their head verb, be they adjacent or not.

2.3 Innovation vs. error

Errors are traditionally associated with EFL, and innovations with ESL. However, the partial overlap between EFL and ESL non-standard features (see Section 1) means that the distinction between errors and innovations may have to be reconsidered. We start from the assumption that both errors and innovations may be found in either variety, and we seek to operationalize the distinction by objective means.

Van Rooy (2011: 189) points out that "[a] distinction between error and conventionalized innovation is essential to understanding if and how New Varieties of English develop new conventions". He suggests that the two key criteria for distinguishing innovations and errors are systematicity and acceptability.

Systematicity, which is required "to show that these variants are not mere random errors, but have found a place in emergent linguistic systems" (ibid.), is easily operationalized in our approach by means of collocation measures, and by discarding infrequent combinations (we discard hapax legomena). Acceptability is more difficult to operationalize. ICLE is not error-tagged, and there is no corpus-based way to find out if an innovative expression used by one EFL learner would be acceptable to other EFL learners. As far as the sparse data allows, we do check, however, if an expression is used by several learners, and if it is used by learners with different L1 backgrounds. The former may point to acceptability by a part of the community; the latter may point to a psycholinguistic base for an innovation, or to typologically related L1 backgrounds. Absence of the latter may indicate L1 transfer errors.

According to usage-based linguistics, acceptability typically follows from frequency, with a certain time lag. Frequency of co-occurrence is not only an effect of entrenchment, it is also often described as a contributor, as functional and cognitive linguists increasingly point out, e.g. Bybee (2007: 337). In practical terms, this entails that after a new combination (which is initially seen as an error) has occurred frequently enough and attains collocational status for some speakers, it has increasingly better chances to become accepted as an innovation.

Gut (2011: 120) notes that "[t]he labeling of a structure as an error (…) has an attitudinal and political rather than a linguistic basis". If this is the case, the systematicity-based continuum of chance co-occurrence to strong collocation, which can be directly measured by collocation statistics, may suffice as a first operationalization. We do not distinguish between innovation and error in Section 4, although the fact that we remove hapax legomena means that two very obvious types of error, typos and single production errors, are excluded. We attempt to distinguish between innovation and error again in Sections 5 and 6 and give partial answers.

Gradient continua and attitudinal preferences can be captured by collocation statistics, which we introduce now.

3. Methods and data

3.1 Method: Collocations and overuse

Schneider (2004: 229) mentions collocational differences as a feature of indigenized varieties of English. Collocations signify conventionalized use of linguistic expressions. Criteria include non-compositionality, non-substitutability, limited modifiability, non-literal translations and statistical co-occurrence. While only the last of these can be measured trivially in corpora, it has proven to be a surprisingly appropriate measure, both in terms of measuring collocation strength (e.g. Wulff 2008) and in approximating the psycholinguistic entrenchment which is behind collocations: Gries & Wulff (2005) and Gries & Wulff (2009) find strong correlations between collocation strengths and experimentally obtained sentence completions from advanced EFL learners, which means that collocation measures lend themselves as a model of listener expectations.

A wide array of frequency-based collocation statistics has been suggested, see e.g. Evert (2008) and Pecina (2009). We restrict our investigation to O/E and T-score. O/E (which literally means Observed divided by Expected) and its variant MI (Mutual Information) are information-theoretic measures (Shannon 1951) of the extent to which two words appear more often together (O=Observed) than expected (E) if all words were randomly distributed in the corpus (or inside the frame of a construction). O/E is defined as

$$\frac{O}{E} = \frac{p(x,y)}{p(x)\cdot p(y)} = \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N}\cdot\frac{f(y)}{N}} = \frac{f(x,y)\cdot N\cdot N}{f(x)\cdot f(y)\cdot N} = \frac{f(x,y)\cdot N}{f(x)\cdot f(y)}$$

where x is the first word, y is the second word, $p(x)$ is the independent probability of x, $f(x)$ is the frequency of x in the corpus, $p(x,y)$ is the joint probability of x and y occurring together, and N is the size of the corpus. If co-occurrence of x and y is due to chance, i.e. if there is no collocational force, then the independent probability of seeing both (*Expected*) and the joint probability of seeing the combination (*Observed*) are roughly equal.

In order to describe innovations in ESL and EFL, we need to find verb/adjective + PP combinations which (i) are conventionalized, i.e. frequent enough to reach collocation status, (ii) are collocations in the non-native corpora, and (iii) are not collocations, or much less so, in the native corpora. If we apply traditional collocation measures we fail to see point (iii). A successful measure for (iii) is the collocation ratio (Schneider & Zipp 2013): if $c_{\mathrm{L1}}(x,y)$ is a collocation measure c for L1 of words x and y, then

$$\text{Collocation ratio} = c_{\mathrm{L2}}(x,y) \;/\; c_{\mathrm{L1}}(x,y)$$

The collocation ratio is a measure of overuse, of "overcollocability". Our suggested overuse statistics is an information-theoretic measure of surprise at seeing learner data when actually expecting native speaker data. For the collocation measure O/E, with c_{L2} as ICLE and c_{L1} as BNC, the ratio is defined as

$$\text{O/E ratio} = \frac{\text{O/E(ICLE)}}{\text{O/E(BNC)}} = \frac{\frac{O(ICLE)}{E(ICLE)}}{\frac{O(BNC)}{E(BNC)}} = \frac{\frac{O_{ICLE}(R,w_1,w_2)\cdot N_{ICLE}}{O_{ICLE}(R,w_1)\cdot O_{ICLE}(R,w_2)}}{\frac{O_{BNC}(R,w_1,w_2)\cdot N_{BNC}}{O_{BNC}(R,w_1)\cdot O_{BNC}(R,w_2)}}$$

where w_1 = verb or adjective, w_2 = preposition or verbal particle, R = syntactic relation expressing prepositional phrase attached to a verb, N = corpus size in words.

The O/E-ratio is itself an O/E measure, in which O = O/E(ICLE) and E = O/E(BNC), or in words: the observed value is the O/E measure as found in the application corpus ICLE, while we expected the O/E measure from the native speaker reference corpus BNC. O/E is an information theoretic measure of surprise: the interpretation of O/E-ratio is equally straightforward, it is also a measure of surprise.

The O/E measure has the tendency to over-represent rare events. The opposite characteristic has been attributed to the T-score measure. There are several answers to these two opposing characteristics. One is that as they are complementary, and if we thus apply both, we maximize recall. For the T-score collocation measure a formulation in terms of O and E (Evert 2008) is

$$T = \frac{O - E}{\sqrt{(O)}} \rightarrow T\,ratio = \frac{T(ICLE)}{T(BNC)} = \frac{\frac{O(ICLE) - E(ICLE)}{\sqrt{O(ICLE)}}}{\frac{O(BNC) - E(BNC)}{\sqrt{O(BNC)}}}$$

We also test and apply the T-ratio, but its statistical interpretation is more involved.[1]

3.2 Data: Parsed EFL, ESL and ENL corpora

For the comparison of EFL, ESL and ENL, we use the following corpora. For EFL, we use the *International Corpus of Learner English* (ICLE; Granger et al. 2009). It is a corpus of learner English from university students with 16 different mother tongue backgrounds. It contains 3.7 million words from essays of higher intermediate to advanced learners of English.

For ESL, we use selected components of the *International Corpus of English* (ICE; Nelson et al. 2002). Each ICE component contains 1 million words of spoken and written text and has the same genre distribution. Among the 11 currently publicly available complete ICE components, 4 are from native language variants (GB, Canada, Ireland, New Zealand), while 7 contain ESL data, in which we are

[1] An approach which is complementary to our collocation-based method is presented in Graën and Schneider (2007), who exploit multiparallel corpora.

interested. We have excluded two components: ICE-East Africa, as it is made up of several subcomponents, and ICE Nigeria, as its spoken part contains no punctuation. We have kept all other ESL data, i.e. the following 5 components: ICE-Singapore, ICE-Philippines, ICE-Jamaica, ICE-India and ICE-Hong Kong.

For ENL, we use the written part of the *British National Corpus* (BNC; Aston & Burnard 1998). It contains 90 million words of written texts from a wide range of registers. We use it as a reference corpus of native British English.

We are aware that these corpora are not an ideal base for comparison: the mix of genres and the level of formality are different between the corpora: unedited student essays make up the entire ICLE but only small subsets of the ICE, and have no counterpart in the BNC; they are also less formal than the written BNC, which consists largely of published material. This feature of the BNC, on the other hand, makes it suitable as a reference corpus of formal, high-level usage of British English. The ICE components which we use as an ESL reference have a much higher contingent of spoken language, which includes spontaneous, unedited usage. This characteristic is not only a disadvantage, but also an advantage when comparing the learner language represented in the ICLE, which contains similarly spontaneous forms, many of which were not edited in the written essays, as the learners may not have been aware that they are infelicitous or incorrect. For these reasons, the ICE components are still a good alternative to the much larger GloWbE corpus (Davies & Fuchs 2015).

We use richly annotated corpus material: the corpora are annotated syntactically using the automatic dependency grammar parser Pro3Gres (Lehmann & Schneider 2012; Schneider 2008). An evaluation of the performance of the parser on ESL varieties is given and our approach is tested on selected phenomena in Schneider & Hundt (2009) and Schneider & Zipp (2013).

We do not distinguish between verbal particle and preposition, because often confusion between the two categories is at the core of the difference between the ENL use and the EFL or ESL use (e.g. *result in* vs. *result into*). For the same reason, we also include verb + PP combinations in which the PP is attached as an adjunct according to the automatic parser. We also include adjective + PP combinations, as they, too, have collocational status. For example, Benson et al. (2009) recognize adjective + preposition as an independent category in addition to verb + preposition (and noun + preposition, e.g. in nominalizations, which we have not included). Adjective + preposition combinations are often similarly difficult to acquire for learners of English.

4. Data-driven detection of verb/adjective + PP innovations/errors in EFL

In this section, we present and interpret our quantitative results. In the ranked lists that we show, we only give the top 10 to 30 entries, for space reasons. Our first operationalization of systematicity of innovations, which our algorithms (see Section 3.1) return and which we discuss, validate and interpret in the following, allows us to introduce a limited step of acceptability judgment by the authors, and a base for the qualitative analysis in Section 5.

4.1 Collocation ratio with O/E

We first apply the O/E-ratio introduced in Section 3.1, using ICLE as application corpus and BNC as reference corpus. The top 30 candidates for EFL innovations/errors are given in Table 1, sorted by decreasing O/E-ratio (first column). The second column contains the verb or adjective lemma, which is modified by the preposition or verbal particle given in column 3. Column 4 lists the frequency of the construction in ICLE. We have only excluded hapax legomena. Columns 5 and 6 give the collocation measure O/E for the application and reference corpora.

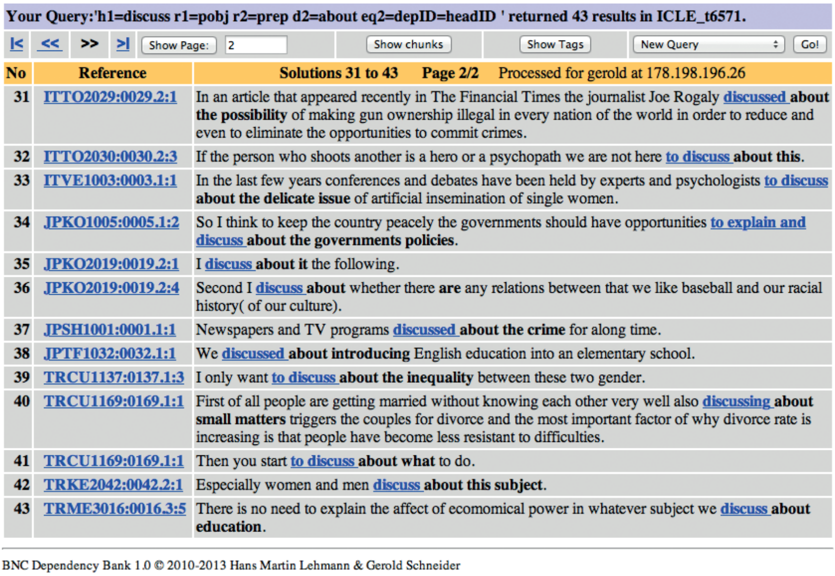
The last column is not output of our algorithm, but shows our comments and interpretation based on our inspection of the hits (see Figure 1, which lists the hits of line 24), in particular whether the type in this line is a learner innovation/error or not (for example because it is particularly frequent due to the essay topics, or a consistent parsing error). In uncertain cases we consulted dictionaries such as Benson et al. (2009). If our comment starts with "instead of" the hit is a true positive, i.e. the line represents a usage which is specific to learner English. The comment "CORPUS essay topic" means that this verb/adjective + preposition pair is overrepresented in ICLE because it appears very frequently due to the essay topics that are used in ICLE. *Handicap after*, for example, is overrepresented due to the essay topic "Discuss the pros and cons of abortion", where many students write that abortion should be allowed if a child would be *handicapped after birth*.

The last column of Table 1 thus shows that 12 of the top 30 candidates were indeed validated as EFL innovations/errors. In terms of the evaluation measure precision (e.g. Jurafsky & Martin 2009: 489),[2] this corresponds to 40% precision, which on the one hand may seem low, but on the other hand is sufficiently high, because manual filtering based on inspecting the hits is quite simple. We can easily increase precision by setting a filter on O/E(BNC) corresponding to the criterion that innovations/errors should not have high collocational status in the native

[2] In words, precision measures how many hits are true positives; recall measures how many of all the true positives are found by the automatic system.

Table 1. Verb/adjective + preposition overuse in ICLE, sorted by decreasing O/E-ratio

O/E ratio	VERB/ADJ	PREP	F	O/E(ICLE)	O/E(BNC)	Comment
414.02	straight	out	2	1599.65	3.86	CORPUS essay topic
256.95	handicap	after	30	2211.46	8.61	CORPUS essay topic
201.30	responsible	of	19	23.31	0.12	instead of *responsible for*
150.95	worth	for	7	81.81	0.54	instead of *worth something*
144.47	view	upon	3	268.71	1.86	instead of *viewed on* (old-fashioned)
111.27	toss	about	2	505.05	4.54	
111.03	balance	from	2	47.87	0.43	
100.77	boil	by	2	45.97	0.46	
83.77	base	amongst	2	300.08	3.58	
77.10	attack	against	2	125.61	1.63	instead of *attack somebody*
72.87	alarm	of	2	92.95	1.28	
69.04	diverse	by	2	91.95	1.33	instead of *different according to*
65.18	exist	out	4	18.01	0.28	
53.54	design	before	2	304.28	5.68	
53.22	cool	down	4	6657.67	125.11	
50.78	bath	without	2	640.14	12.61	
50.31	sleep	around	13	420.93	8.37	
49.99	synonymous	to	2	26.10	0.52	instead of *synonymous with*
48.51	select	among	3	751.98	15.50	instead of *select from*
42.36	credit	for	2	233.73	5.52	
41.44	benefit	out	2	24.74	0.60	instead of *benefit from*
39.91	lower	than	4	198.58	4.98	
39.11	basic	for	2	58.43	1.49	
35.81	discuss	about	43	65.68	1.83	instead of *discuss something*
35.42	separate	between	4	189.54	5.35	instead of *distinguish between*
32.67	pour	onto	3	9928.44	303.87	
32.64	dependent	from	2	5.26	0.16	instead of *dependent on*
32.45	comment	by	2	22.19	0.68	
32.06	helpless	for	4	66.78	2.08	
31.47	stretch	beyond	4	6360.12	202.11	
30.22	understand	towards	2	54.88	1.82	instead of *understand sth.*



Your Query:'h1=discuss r1=pobj r2=prep d2=about eq2=depID=headID ' returned 43 results in ICLE_t6571.

No	Reference	Solutions 31 to 43 Page 2/2 Processed for gerold at 178.198.196.26
31	ITTO2029:0029.2:1	In an article that appeared recently in The Financial Times the journalist Joe Rogaly discussed **about the possibility** of making gun ownership illegal in every nation of the world in order to reduce and even to eliminate the opportunities to commit crimes.
32	ITTO2030:0030.2:3	If the person who shoots another is a hero or a psychopath we are not here to discuss **about this**.
33	ITVE1003:0003.1:1	In the last few years conferences and debates have been held by experts and psychologists to discuss **about the delicate issue** of artificial insemination of single women.
34	JPKO1005:0005.1:2	So I think to keep the country peacely the governments should have opportunities to explain and discuss **about the governments policies**.
35	JPKO2019:0019.2:1	I discuss **about it** the following.
36	JPKO2019:0019.2:4	Second I discuss **about** whether there **are** any relations between that we like baseball and our racial history(of our culture).
37	JPSH1001:0001.1:1	Newspapers and TV programs discussed **about the crime** for along time.
38	JPTF1032:0032.1:1	We discussed **about introducing** English education into an elementary school.
39	TRCU1137:0137.1:3	I only want to discuss **about the inequality** between these two gender.
40	TRCU1169:0169.1:1	First of all people are getting married without knowing each other very well also discussing **about small matters** triggers the couples for divorce and the most important factor of why divorce rate is increasing is that people have become less resistant to difficulties.
41	TRCU1169:0169.1:1	Then you start to discuss **about what** to do.
42	TRKE2042:0042.2:1	Especially women and men discuss **about this subject**.
43	TRME3016:0016.3:5	There is no need to explain the affect of ecomomical power in whatever subject we discuss **about education**.

BNC Dependency Bank 1.0 © 2010-2013 Hans Martin Lehmann & Gerold Schneider

Figure 1. Hits for *discuss about* from ICLE, shown in Dependency Bank (Lehmann & Schneider 2012)

variant. If we set a filter of O/E(BNC) < 5, precision rises to above 50%, but at the trade-off of a cost in recall: for example, *select among* and *separate between* would not be returned. Equally, only including results which the automatic parser annotates as PP-arguments would increase precision, but lead to a large loss in recall. For example, *discuss about* and *attack against* are parsed as PP-adjuncts.

In Table 2, we use such a filter of O/E(BNC) < 5, and in addition we take into consideration the fact that verb/adjective + preposition combinations which were not seen in the BNC may never appear there because they are unacceptable in native British English. We thus added a smoothing count of 0.5 (new fifth column) to types unseen in the BNC. We used a frequency threshold of f(ICLE) > 3. As one can see in the last, again manually added column, 17 of the top 30 candidates (corresponding to 57% precision) are innovations/errors.

In Table 2, it is striking to see that the majority of true positives (12 out of 17) can be analyzed as involving the use of a semantic, compositional preposition instead of a functional, idiomatic preposition, namely *critical towards*, *critical against*, *indulge into*, *destructive for*, *discuss about*, *conscious about*, *belong into*, *aware about*, *aspire for*, *guilty for*, *accuse for*, *deal about*.

Table 2. Verb/adjective + preposition overuse in ICLE, sorted by decreasing O/E ratio, with filter O/E(BNC) < 5 and smoothing for events unseen in BNC

O/E ratio	VERB/ADJ.	PREP	F(ICLE)	F(BNC)	O/E(ICLE)	O/E(BNC)	Comment
488.81	critical	towards	7	0.5	1511.26	3.09	instead of *critical to*
201.30	responsible	of	19	2	23.31	0.12	instead of *responsible for*
189.01	critical	against	4	0.5	370.22	1.96	instead of *critical to*
150.95	worth	for	7	1	81.81	0.54	instead of *worth something*
145.67	superior	than	22	0.5	434.65	2.98	instead of *superior to*
138.75	indulge	into	6	0.5	61.11	0.44	instead of *indulge in*
110.11	overcrowd	at	32	0.5	485.00	4.40	CORPUS essay topic
69.11	destructive	for	5	1	166.95	2.42	instead of *destructive to*
65.18	exist	out	4	2	18.01	0.28	
39.91	lower	than	4	2	198.58	4.98	
35.81	discuss	about	43	7	65.68	1.83	instead of *discuss something*
34.27	conscious	about	10	2	124.19	3.62	instead of *conscious of*
32.06	helpless	for	4	1	66.78	2.08	
31.55	possible	out	4	5	30.37	0.96	
30.60	recur	to	4	7	125.26	4.09	
29.94	dependent	of	8	4	19.34	0.65	instead of *dependent on*
24.63	belong	into	4	2	6.63	0.27	instead of *belong to*
23.59	renounce	to	9	3	108.40	4.60	
23.07	decide	over	7	13	102.14	4.43	CORPUS essay topic
21.96	inherent	to	9	13	78.29	3.56	
20.46	relate	with	49	76	32.98	1.61	instead of *relate to*

Table 2. (*continued*)

O/E ratio	VERB/ADJ.	PREP	F(ICLE)	F(BNC)	O/E(ICLE)	O/E(BNC)	Comment
19.80	aware	about	4	1	5.94	0.30	instead of *aware of*
19.67	aspire	for	4	3	51.94	2.64	instead of *aspire to*
18.21	guilty	for	22	28	59.11	3.25	instead of *guilty of*
17.72	little	by	11	36	70.80	4.00	
17.67	produce	out	4	30	44.85	2.54	
17.19	accuse	for	8	19	18.33	1.07	instead of *accuse of*
15.39	interest	to	7	0.5	11.54	0.75	
15.01	specialize	on	4	4	40.24	2.68	
15.01	deal	about	4	2	3.91	0.26	instead of *deal with*

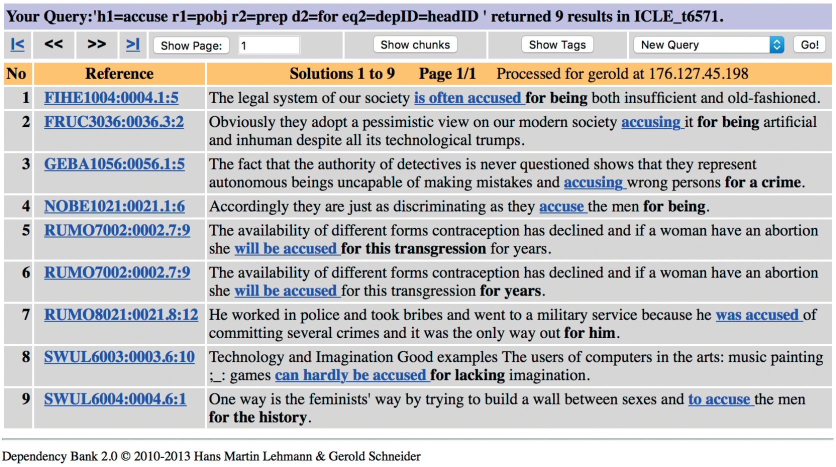
4.2 Collocation ratio with T-score

We next apply T-ratio from Section 3.1, using ICLE as application corpus and BNC as reference corpus. The top 10 candidates for EFL innovations/errors are given in Table 3, sorted by decreasing T ratio (first column); all other columns are analogous. Figure 2 shows the hits of line 3.

Table 3. Verb/adjective + preposition overuse in ICLE, sorted by decreasing T-ratio

T ratio	VERB/ADJ	PREP	F	T(ICLE)	T(BNC)	Comment
5.9820	impose	to	10	5336.86	892.15	instead of *impose on*
3.5860	replace	to	3	1168.35	325.81	instead of *replaced by* (partly)
2.1133	accuse	for	8	5143.81	2433.98	instead of *accuse of*
2.0275	addict	on	4	3431.99	1692.68	instead of *addict to*
1.4296	better	than	87	17920.70	12535.47	
1.3929	alarm	of	2	2691.03	1932.01	instead of *alarm about*
1.3322	handicap	after	30	10530.89	7905.03	CORPUS essay topic
1.2812	better	for	59	14564.98	11367.88	
1.2074	diverse	by	2	2690.71	2228.48	instead of *different according to*
1.1541	discuss	about	43	12421.43	10762.54	instead of *discuss sth.*

In terms of precision, 15 of the top 30 candidates with T-ratio are innovations/errors, which corresponds to 50% precision. We could increase precision by setting a filter on T(BNC) corresponding to the criterion that innovations/errors should not have high collocational status in the native variant. With a filter of, e.g. T(BNC) < 5,000, precision rises to above 50%, but at the trade-off of a cost in recall (*discuss about* and *relate with*, for example, would not be returned).



Your Query:'h1=accuse r1=pobj r2=prep d2=for eq2=depID=headID ' returned 9 results in ICLE_t6571.

No	Reference	Solutions 1 to 9 Page 1/1 Processed for gerold at 176.127.45.198
1	FIHE1004:0004.1:5	The legal system of our society is often accused **for being** both insufficient and old-fashioned.
2	FRUC3036:0036.3:2	Obviously they adopt a pessimistic view on our modern society accusing it **for being** artificial and inhuman despite all its technological trumps.
3	GEBA1056:0056.1:5	The fact that the authority of detectives is never questioned shows that they represent autonomous beings uncapable of making mistakes and accusing wrong persons **for a crime**.
4	NOBE1021:0021.1:6	Accordingly they are just as discriminating as they accuse the men **for being**.
5	RUMO7002:0002.7:9	The availability of different forms contraception has declined and if a woman have an abortion she will be accused **for this transgression** for years.
6	RUMO7002:0002.7:9	The availability of different forms contraception has declined and if a woman have an abortion she will be accused for this transgression **for years**.
7	RUMO8021:0021.8:12	He worked in police and took bribes and went to a military service because he was accused of committing several crimes and it was the only way out **for him**.
8	SWUL6003:0003.6:10	Technology and Imagination Good examples The users of computers in the arts: music painting ;_: games can hardly be accused **for lacking** imagination.
9	SWUL6004:0004.6:1	One way is the feminists' way by trying to build a wall between sexes and to accuse the men **for the history**.

Dependency Bank 2.0 © 2010-2013 Hans Martin Lehmann & Gerold Schneider

Figure 2. Hits for *accuse for* from ICLE, shown in Dependency Bank

4.3 Quantitative analysis of verb/adjective + PP combinations in ESL

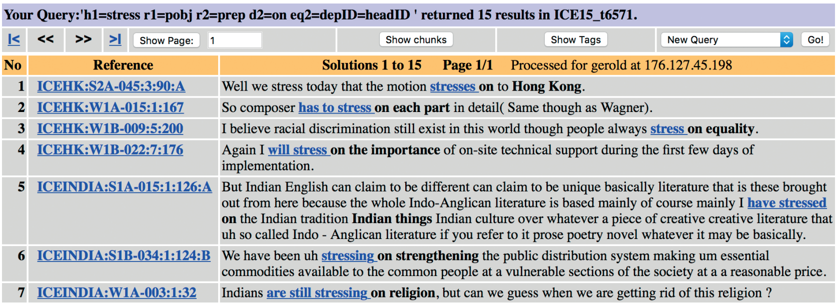
We have so far detected EFL innovations/errors by our collocation-based approach. We can apply the same approach to ESL varieties; Schneider & Zipp (2013) have done so to describe innovations in ICE-Fiji and ICE-India.

Any individual ESL variety could be analyzed in the same fashion. Here, we use a collection of ESL varieties, the 5 ICE components described in Section 3.2, henceforth ICE-5 ESL. Using a collection of ESL corpora has the advantages that sparse data issues are reduced and that psycholinguistically based innovations, i.e. innovations not due to L1 transfer but due to general cognitive processes like analogy, are boosted. Analogy is seen as a key ability for language acquisition (Tomasello 2003) and generally in usage-based approaches to language (Bybee 2007). This mirrors our EFL approach of using all linguistic backgrounds in ICLE, but also has the disadvantage that innovations specific to one variety are likely to be overlooked.

Table 4 shows the top 22 candidates, again sorted by decreasing O/E-ratio. Some of the non-native-like combinations that we have seen in EFL also appear in ESL, for example *discuss about*. 6 of the top 22, for example *study about*, are innovations. As in ICLE, corpus buildup mismatches between application and reference corpus are responsible for some overused expressions (cf. "CORPUS"). The tendency to transfer nominal subcategorization patterns to the corresponding verb as in *emphasize* or *stress* (see Figure 3) may be a universal psycholinguistic mechanism (cf. Section 5).

Table 4. Verb/adjective + preposition overuse in ICE-5 ESL, sorted by decreasing O/E-ratio

O/E ratio	VERB/ADJ.	PREP	F	O/E(ICE-5 ESL)	O/E(BNC)	Comment
128.08	lower	than	12	637.31	4.98	
110.85	immerse	into	6	213.99	1.93	
55.27	canvass	before	31	2743.07	49.63	CORPUS: all from ICE-IND, legal term
54.04	preside	by	6	65.45	1.21	CORPUS: most from ICE-IND
50.31	play	inside	8	171.08	3.40	CORPUS: sports news
45.57	discuss	about	35	83.59	1.83	instead of *discuss sth.* / noun
35.95	understand	between	12	348.19	9.69	tagging error
28.70	elect	into	6	47.15	1.64	
26.90	emphasise	on	8	116.84	4.34	instead of noun
22.57	switch	over	12	292.55	12.96	instead of *switch to*
20.14	print	over	6	159.57	7.93	CORPUS: all from 1 ICE-JAM article
19.88	run	toward	11	515.50	25.94	CORPUS: sports news
19.76	study	about	14	26.05	1.32	instead of *study sth.* / noun
19.19	branch	into	6	505.80	26.36	
18.74	awaken	as	7	87.52	4.67	CORPUS: most from 1 ICE-IND article
18.15	coat	on	8	97.37	5.37	
16.73	better	than	80	1862.09	111.31	
16.67	sort	of	17	218.04	13.08	tagging error
14.92	accuse	before	6	123.02	8.24	
14.91	dress	on	8	39.61	2.66	
14.49	emphasize	on	9	69.18	4.78	instead of *emphasize sth.* / noun
13.35	stress	on	14	83.46	6.25	instead of *stress sth.* /noun



Your Query:'h1=stress r1=pobj r2=prep d2=on eq2=depID=headID ' returned 15 results in ICE15_t6571.

No	Reference	Solutions 1 to 15 Page 1/1 Processed for gerold at 176.127.45.198
1	ICEHK:S2A-045:3:90:A	Well we stress today that the motion stresses **on** to **Hong Kong**.
2	ICEHK:W1A-015:1:167	So composer has to stress **on each part** in detail(Same though as Wagner).
3	ICEHK:W1B-009:5:200	I believe racial discrimination still exist in this world though people always stress **on equality**.
4	ICEHK:W1B-022:7:176	Again I will stress **on the importance** of on-site technical support during the first few days of implementation.
5	ICEINDIA:S1A-015:1:126:A	But Indian English can claim to be different can claim to be unique basically literature that is these brought out from here because the whole Indo-Anglican literature is based mainly of course mainly I have stressed **on** the Indian tradition **Indian things** Indian culture over whatever a piece of creative creative literature that uh so called Indo - Anglican literature if you refer to it prose poetry novel whatever it may be basically.
6	ICEINDIA:S1B-034:1:124:B	We have been uh stressing **on strengthening** the public distribution system making um essential commodities available to the common people at a vulnerable sections of the society at a a reasonable price.
7	ICEINDIA:W1A-003:1:32	Indians are still stressing **on religion**, but can we guess when we are getting rid of this religion ?

Figure 3. Hits for *stress on* in ICE-5 ESL, shown in Dependency Bank

The generally smaller O/E-ratio in ESL as compared to EFL (Section 4.1) shows that ESL (represented by ICE-5 ESL) is closer to the BNC reference than is EFL (represented by ICLE). This is probably due to the following reasons. First, there are fewer innovations/errors in ESL than in EFL. Particularly errors, i.e. those choices which are not accepted by more experienced speakers of the same community, are less frequent in ESL. Second, the semantic similarities of the texts are probably less strong between individual ICE documents than between individual ICLE documents, which often have the same essay title. The fact that collecting many L1 backgrounds glosses over many of the characteristics of an individual variety equally applies to ICE-5 ESL and to ICLE, and is therefore probably not a major reason for the large differences in O/E-ratio.

4.4 Quantitative analysis of verb/adjective + PP combinations in EFL vs. ESL

Until now we have compared EFL and ESL to a native British English reference. We can also compare EFL to an ESL reference corpus, indicating which innovations are more EFL-like. When using the same parameters as in Section 4.1 (Table 1), precision is quite low (6/30), indicating that EFL is closer to ESL than to the native reference corpus.

To boost precision, we ran a version of the innovation extraction method seen in Table 2, with particularly strict O/E(ICE 5 ESL) < 2, counting unseen instances again as 0.5, aiming at a core set of typical verb/adjective + preposition innovations which only EFL speakers but not ESL speakers use. The top results are given in Table 5. 12 of the 21 top hits are true positives.

Looking at Table 5 reveals that noun-analogies (noun complementation patterns which are taken over to the verb) are very rare (only one, *assist to*) compared to ESL (Section 4.3, Table 4), and that the preposition *to* seems to be used too generically: 7 out of the 13 true positives involve *to*. There might be a trend to use *to* as a generic marker for indirect objects.

Table 5. Verb/adjective + preposition overuse in ICLE, sorted by decreasing O/E-ratio, using ICE-5 ESL as a reference corpus, with threshold O/E(ICE 5 ESL) < 2, and smoothing for events unseen in ICE-5 ESL

O/E ratio	VERB/ADJ	PREP	F(ICLE)	F(ICE-5 ESL)	O/E(ICLE)	O/E(ICE-5 ESL)	Comment
35.97	equivalent	in	5	0.5	35.34	0.98	
34.19	assist	to	6	1	27.63	0.81	instead of *assist sth.*
25.68	accuse	for	8	0.5	18.33	0.71	instead of *accuse of*
22.29	wrong	at	6	0.5	24.38	1.09	
21.61	explain	from	8	0.5	16.03	0.74	
21.28	stay	like	5	0.5	13.53	0.64	
15.45	participate	to	8	1	8.46	0.55	instead of *participate in*
14.10	arise	by	6	0.5	12.14	0.86	instead of *due to/from*
12.60	employ	of	5	0.5	18.19	1.44	parsing error
11.35	benefit	to	13	1	10.49	0.92	instead of *be of benefit to*
9.10	impose	to	10	1	8.15	0.90	instead of *impose on*
8.06	oppose	in	6	0.5	5.05	0.63	
5.63	equal	for	9	0.5	4.22	0.75	instead of *equal to*
5.51	discuss	of	5	0.5	4.22	0.77	
5.40	remain	to	5	2	4.33	0.80	
5.34	necessary	with	6	0.5	6.70	1.25	instead of *necessary for*
5.08	keep	into	5	1	4.22	0.83	instead of *keep at*
5.05	reflect	to	5	1	5.12	1.01	instead of *reflect sth.*
4.95	confront	to	6	0.5	7.17	1.45	instead of *confront with*
4.93	discuss	for	13	2	6.13	1.24	
4.72	popular	to	6	0.5	4.84	1.03	instead of *popular for*

5. Qualitative analysis

We now examine the non-native-like verb/adjective + PP combinations found in EFL, using the method described above, and also briefly compare the results with those found for ESL. Our approach is more qualitative here, seeking to identify the processes that may have led to these combinations.

When compared to the native reference corpus BNC, some verb/adjective + PP combinations overused by learners are reported close to the top of the ranked lists by both O/E- and T-ratio, for example *basic for*, *discuss about*, *helpless for* or *relate with*. However, it also turns out that each measure brings up its own combinations. Interestingly, this includes the use of different prepositions with one and the same verb or adjective, for instance *independent from* (with the O/E-ratio) and *independent on* (with the T-ratio). This shows that neither of the two measures is sufficient in itself and that they should be combined with each other. Consequently, no distinction will be made between the two measures in the following qualitative analysis.

If we exclude typos, we can distinguish several types of major combinations. Some involve the use of a prepositional complement instead of a transitive use of the verb. Thus, instead of *discuss sth* some learners use the combination *discuss about* (1); instead of *consider sth* they use *consider about sth* (2); and instead of *phone sb* they use *phone to sb* (3).

(1) First of all let's **discuss about** the goodness of having PC cafes. (ICLE:CNHK1224)

(2) In this essay I am going to **consider about** the advantages and disadvantages of banning smoking in restaurants. (ICLE:CNHK1371)

(3) He the boss **phoned to** his friends from Mafia and asked to get rid of his friends with whom he was bore to death. (ICLE:RUMO7025)

In other cases, it is the verb or the adjective that is inadequate. In example (4) the learner has used *insensible* instead of *insensitive*, and in example (5) *helpless* instead of *useless*.

(4) One does not have to be a Marxist to understand what he meant:_that religion was an escape from the hard everyday life making people ignorant and **insensible to** the wrongs that existed at that time. (ICLE:SWUL6001)

(5) To conclude not all of qualifications people can get from universities are useful some of subjects are **helpless for** their future jobs (…). (ICLE:CNUK2015)

There are also many cases where a non-standard preposition is used, for example *concentrate to* (instead of *concentrate on*) and *intolerant to* (instead of *intolerant of*):

(6) When the demand for these machines is big enough the production **concentrates to** certain areas and to certain people and the first step towards industrialism is then taken. (ICLE:FIJO3011)

(7) As people usually get married at the young age they can be quite **intolerant to** any kind of disturbance in their new home. (ICLE:TRCU1169)

Very often, these non-native-like combinations have not been coined by chance, but seem to be the result of analogy, or more precisely "nativised semantico-structural analogy" (Mukherjee 2005, cited in Mukherjee & Hoffmann 2006). The basis of this analogy can be a word of the same family but with a different part-of-speech. The use of *about* with the verb *discuss* (see example (1) above) could be related to the preposition that is used with the noun *discussion*. The same is true of *attack_V against* (cf. *attack_N against*), *be credited for* (cf. *credit_N for*) or *relate with* (cf. *relation with*), e.g. (8). In the case of *independent on*, cf. (9), it is the preposition of the positive form of the adjective, *dependent*, which is borrowed by the learners.

(8) For example in the Gulf War the USA **attacked against** the Iraqis in order to prevent the price of petrol from going up. (ICLE:FIJO2003)

(9) The first reason why childhood does not end when you become economically **independent on** your parents is that maturity is a mental condition which has nothing to do with money (…) (ICLE:ITVE2003)

In other combinations, the analogy is based on a synonym. Thus, the use of the preposition *between* after the verb *separate* (10) could be due to the use of the same preposition with the synonymous verb *distinguish*. *To be viewed upon as* (11) could be formed by analogy with *to be looked upon as*, *to arrive to* by analogy with *to get to*, and *afraid about* by analogy with *scared about*.

(10) It looks like it can be hard to **separate between** what is reality and what is TV-entertainment. (ICLE:NOHO1037)

(11) Women have always been **viewed upon as** the weaker part of the population that had to be led and helped by men. (ICLE:CZPR3005)

Finally, some combinations seem to be due to transfer from similar combinations in the learners' mother tongue, for example the use of *inherent to* by French-speaking learners (12), who have the combination *inhérent à* in their L1, where the preposition *à* corresponds to English *to*.

(12) By reading ancient stories we realize that suffering is **inherent to** the human condition and we feel taken in a timeless feeling. (ICLE:FRUL1010)

When we compare these non-native-like combinations with those found for ESL, a number of similarities emerge. The combination *discuss about*, for example, is mentioned in the literature on Indian English (Mukherjee 2007), and also in a study that specifically compares EFL and ESL (Nesselhauf 2009). The results of Schneider & Zipp's (2013) study on ESL also partly overlap with our results, with combinations like *discuss about*, *benefit out of* and *aware about* being found in the two studies; compare (13) and (14). Similar phenomena are also attested in both analyses, like the use of a redundant particle, illustrated by *viewed upon as* (instead of *viewed as*) in (11) above or *listed down* in (15).

(13) To sum it up I belive that the E. C will be a paradise for the middle-classes since they mostly have white-collar jobs often provided with a certain position they'll **benefit** most **out of** tax-deregulations etc. (ICLE:SWUL9013)

(14) So they'll **benefit out of** the faculty teaching (ICE IND:S1A-064)

(15) Adi Asenaca said an Asian Development Bank poverty participation survey **listed down** forms of poverty in the country and her ministry was following up on the recommendations. (ICE FJ:W2C 013)

At the same time, we also observe differences between EFL and ESL. If we consider the types of combinations found in the two varieties (cf. Section 4), it is striking that 7 of the 9 true positives in the ESL data (Table 4, Section 4.3) involve a preposition where Standard English would use an object-complement (e.g. *study sth* instead of *study about sth*). 5 of these (*discuss about*, *emphasise on*, *study about*, *emphasize on*, *stress on*) could be based on an analogy to noun usage. Our data suggests that analogy to the complementation patterns of nouns is particularly frequent among ESL speakers. In comparison, only 3 of the 12 true positive types showed a noun-analogy in EFL (Table 1, Section 4.1). On the other hand, innovations/errors involving the use of a semantic, compositional, often directional preposition instead of a functional, idiomatic preposition are slightly more common in EFL (12 out of 17, see Table 2) than in ESL (4 out of 9: *discuss about*, *study about*, *mention about*, *call as*). This indicates that ESL may prefer grammatical analogies, while EFL may overuse spatial and directional analogies.

6. Discussion

We now return to the discussion of error versus innovation. So far, the concept of innovation has mainly been limited to the description of native English and ESL in the literature. Yet, the presence of similar non-native-like combinations in EFL and ESL makes it difficult to maintain a sharp distinction between the two and treat these combinations as errors in the case of EFL and as innovations in the case of ESL. The results of our automatic detection of non-native-like combinations include instances that probably no one would want to consider linguistic innovations, for example misspellings. Others, however, should perhaps be treated on a par with ESL innovations.

In Section 4.3 (Table 4) we have learnt that analogy to the complementation patterns of nouns seems particularly frequent among ESL speakers. In Section 4.4 (Table 5) we have used an automatic method to detect those verb/adjective + preposition combinations that are considerably different in EFL and ESL. We can use the same method to detect those which are similar. For this purpose, we use the settings from Section 4.1 (Table 2), i.e. O/E(BNC) < 5, f(ICLE) > 3, smoothing for events unseen in BNC, but we only report those verb/adjective + preposition combinations whose O/E from ICLE is not very different from the one in ICE-5 ESL. As a threshold we set that O/E(ICLE) is maximally 3 times larger than O/E(ICE-5 ESL) or vice versa. Results are given in Table 6. These are verb/adjective + preposition combinations which, according to our data, are shared between EFL and ESL. As they exist independently in both, with similar O/E-ratios, we hypothesize that they are more likely to be based on psycholinguistic trends than on L1 transfer or acquisition processes.

Among the combinations which can be treated on a par, it is important to distinguish between combinations that seem to be the result of L1 influence and those that seem to be the result of cognitive operations such as analogy. The latter, which we have called psycholinguistically based innovations, are probably more likely to be recognized as innovations than the former (L1 transfer innovations). Table 6 includes particularly many types that can be due to analogy, as we show in Table 7, which filters by those types that are found in speakers from several L1 backgrounds, (penultimate column). In the last column, we suggest a possible analogy. For example (as discussed in Table 2: a semantic preposition replaces a functional one), in *indulge into* the preposition iconically reduplicates a directionality instigated by the verb; in *aspire for* the subcategorization frame is derived from the corresponding nominalization. In future research, we want to test if ESL (and ENL) speakers are more willing to accept unusual patterns based on analogy than other patterns.

Table 6. Results from Table 2, filtered for similar O/E in ICLE and ICE-5 ESL

O/E ratio	VERB/ADJ.	PREP	F (ICLE)	O/E (ICLE)	O/E(ICE-5 ESL)	O/E (BNC)	Comment
145.67	superior	than	22	434.65	565.61	2.98	instead of *superior to*
138.75	indulge	into	6	61.11	28.10	0.44	instead of *indulge in*
35.81	discuss	about	43	65.68	83.59	1.83	instead of *discuss sth.*
34.27	conscious	about	10	124.19	78.30	3.62	instead of *conscious of*
19.67	aspire	for	4	51.94	31.93	2.64	instead of *aspire to*
17.72	little	by	11	70.80	38.50	4.00	
15.39	interest	to	7	11.54	6.08	0.75	
14.29	point	by	6	13.23	5.57	0.93	
13.49	commensurate	to	4	22.37	49.29	1.66	
13.24	interest	for	26	63.97	41.70	4.83	
12.94	speak	over	5	33.16	13.06	2.56	
10.65	own	to	8	23.20	8.80	2.18	instead of *owing to* (partly)
10.28	watch	than	4	17.52	18.76	1.70	
9.75	capable	in	5	2.83	2.97	0.29	instead of *capable of/to*
9.10	deprive	from	10	18.64	12.64	2.05	
8.84	study	about	8	11.66	26.05	1.32	instead of *study sth.*
8.62	charge	of	4	30.98	11.88	3.59	instead of *change sth*/noun
7.86	shut	to	7	36.53	27.73	4.65	
7.28	face	to	35	19.64	7.86	2.70	instead of *face sth.*
7.24	state	about	4	25.04	11.77	3.46	
6.81	invest	to	5	5.44	2.93	0.80	instead of *invest in*
6.66	speed	in	5	33.13	27.33	4.98	
6.65	waste	for	8	24.28	18.73	3.65	
6.52	reward	to	6	18.07	24.65	2.77	
6.37	associate	to	4	3.89	3.29	0.61	instead of *associate with*
6.36	strike	to	6	16.48	6.16	2.59	
6.02	know	over	4	16.60	9.30	2.76	
5.95	afford	with	4	18.63	33.91	3.13	
5.89	steal	to	6	9.39	3.21	1.59	instead of *steal from* (partly)

Table 6. (*continued*)

O/E ratio	VERB/ADJ.	PREP	F (ICLE)	O/E (ICLE)	O/E(ICE-5 ESL)	O/E (BNC)	Comment
5.88	sum	in	4	22.32	30.50	3.80	
5.51	influence	on	15	15.21	6.40	2.76	instead of noun (partly)
5.30	depend	from	9	4.84	1.76	0.91	instead of *depend on*
5.19	search	from	5	15.06	7.52	2.90	instead of *search on*

Table 7. Possible analogy interpretations of innovations which are common to EFL and ESL

O/E ratio	VERB/ADJ.	PREP	Comment	# L1 BACKG.	Possible analogy
145.67	superior	than	instead of *superior to*	8	better than
138.75	indulge	into	instead of *indulge in*	4	iconic
35.81	discuss	about	instead of *discuss sth.*	7	discussion (noun)
34.27	conscious	about	instead of *conscious of*	6	
19.67	aspire	for	instead of *aspire to*	3	aspiration (noun)
10.65	own	to	instead of *owing to* (partly)	7	
9.75	capable	in	instead of *capable of/to*	4	diligent in
8.84	study	about	instead of *study sth.*	4	
8.62	charge	of	instead of noun	3	noun
7.28	face	to	instead of *face sth.*	>9	face *up to* w/o *up*
6.81	invest	to	instead of *invest in*	2	devote to
6.37	associate	to	instead of *associate with*	4	relate to
5.89	steal	to	instead of *steal from* (partly)	6	
5.51	influence	on	instead of noun(partly)	4	noun
5.30	depend	from	instead of *depend on*	3	iconic
5.19	search	from	instead of *search on*	2	

The purpose of one's approach should also be taken into account when trying to identify innovations. For descriptive purposes, one might be more inclined to recognize the learner's right to be creative, and hence the existence of linguistic innovations, whereas for pedagogical purposes teachers will teach native-like combinations and reject most non-native-like combinations — and perhaps rightly so. Finally, the setting is important too. In an EFL setting, which focuses on competence, non-native-like combinations are less likely to be accepted as innovations than in an English as a Lingua Franca setting, where communication takes precedence over competence.

7. Conclusion

We have described innovations in verb/adjective + preposition combinations (including phrasal verbs) in learner English, using ICLE as application corpus and BNC as reference corpus. We have applied overuse statistics like O/E and T-score, known from collocation analysis to detect and describe errors and innovations in learner English. Overuse statistics are an information-theoretic measure of surprise at seeing learner data when actually expecting native speaker data. We have given a first evaluation of the precision of our method, which allows us to answer the first part of RQ1 (can the patterns of overuse which we observe using collocation statistics deliver combinations that are specific to EFL and/or to ESL?) positively: the patterns of overuse which we observe using collocation statistics deliver combinations that are specific to EFL language. Our method, which we call collocation ratio, is corpus-driven and as far as we are aware reports more combinations than have previously been described (it should be borne in mind that for space reasons we could only show the top entries of considerably longer lists). Using more and larger EFL and ESL corpora would likely deliver further patterns. We can thus also answer RQ3 (does the method give us the tools to find more patterns than have been previously described?) positively.

We have applied the same method to ESL varieties using selected components from ICE and have provided a first evaluation, which allows us to answer the second part of RQ1 positively: the patterns of overuse which we observe using collocation statistics also deliver combinations that are specific to ESL language. In order to assess differences between EFL and ESL, we have compared EFL data against ESL data as a reference. This delivers combinations which are likely to be seen as unacceptable by ESL speakers, and are thus candidates for errors. Concerning RQ2 (does the same method also allow us to detect which patterns of verb + PP and adjective + PP are more typical for EFL and which for ESL?), we thus give a tentatively positive answer. Our data suggests that analogy to the complementation patterns of nouns is particularly frequent among ESL speakers, while EFL speakers tend to overuse the preposition *to*. The use of compositional, semantic prepositions instead of idiomatic, functional ones (e.g. *indulge into*, *discuss about*) seems to be a shared pattern.

In order to assess similarities between EFL and ESL, we have further performed a qualitative analysis, and we have also reported which verb/adjective + preposition innovations in ESL attain similar O/E ratios in EFL. The approach comparing O/E ratios delivers combinations which are likely to be seen as acceptable by ESL speakers, and are thus candidates for innovations. The instances found in this way all occur in a variety of L1 backgrounds, which increases the probability that they are not caused by L1 transfer, but are based on more psycholinguistic mechanisms such as processes of analogy (e.g. the subcategorization frame is derived from the corresponding nominalization) or iconicity (e.g. the preposition iconically reduplicates a directionality instigated by the verb). In the qualitative step of our analysis, we have discussed relevant examples and performed a manual classification of the combinations. We infer that neither O/E nor T-score are sufficient on their own, as each brings up results that the other misses, and that they thus need to be combined to increase recall.

Concerning RQ4 (does the method give us the tools to distinguish between error and innovation?), we have partly narrowed down the candidates by excluding hapax legomena, by restricting innovations to combinations that are found in both EFL and ESL, and that are used by speakers of several L1 backgrounds. We have also singled out cases that can be explained by analogy. However, the results obtained cannot be evaluated, unlike in the other RQs. On the one hand this means that we can only give a speculative answer to RQ4, on the other hand it means that we are treading on new scientific ground by presenting lists of shared verb/adjective + PP combinations to the research community.

Our method thus offers a powerful means of automatically extracting from corpora a large number of patterns distinctive for EFL and/or ESL, and gives some clues as to the status of these patterns (errors or innovations). It therefore contributes to the recent efforts to bridge the paradigm gap between EFL and ESL, by providing new techniques that facilitate the analysis and should make it possible to collect further evidence for the link between the two varieties.

References

Aston, G. & Burnard, L. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Benson, M., Benson, E. & Ilson, R. 2009. *The BBI Combinatory Dictionary of English* (3rd ed.). Amsterdam: John Benjamins.

Bybee, J. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

Cornell, A. 1985. "Realistic goals in teaching and learning phrasal verbs", *International Review of Applied Linguistics in Language Teaching (IRAL)* 23(4), 269–280.

Davies, M. & Fuchs, R. 2015. "Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE)", *English World-Wide* 36(1), 1–28. https://doi.org/10.1075/eww.36.1.01dav

Deshors, S.C. 2016. "Inside phrasal verb constructions: A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations", *International Journal of Learner Corpus Research* 2(1), 1–30.

Díaz-Negrillo, A., Ballier, N. & Thompson, P. (Eds.). 2013. *Automatic Treatment and Analysis of Learner Corpus Data*. Studies in Corpus Linguistics 59. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.59

Dickinson, M. & Ragheb, M. 2009. "Dependency annotation for learner corpora". In Proceedings of the *Eighth Workshop on Treebanks and Linguistic Theories (TLT)*. Milan, Italy.

Edwards, A. 2014. "The EFL-ESL continuum and the case of the Netherlands: A comparative analysis of the progressive aspect", *World Englishes* 33, 173–194. https://doi.org/10.1111/weng.12080

Edwards, A. & Laporte, S. 2015. "Outer and expanding circle Englishes. The competing roles of norm orientation and proficiency levels", *English World-Wide* 36(2), 135–169. https://doi.org/10.1075/eww.36.2.01edw

Evert, S. 2008. "Corpora and collocations". In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: de Gruyter, 1212–1248.

Fuchs, R. & Wunder, E.-M. 2015. "A sonority-based account of speech rhythm in Chinese learners of English". In U. Gut, R. Fuchs & E.-M. Wunder (Eds.), *Universal or Diverse Paths to English Phonology? Bridging the Gap between Research on Phonological Acquisition of English as a Second, Third or Foreign Language*. Berlin: de Gruyter, 165–184.

Gardner, D. & Davies, M. 2007. "Pointing out frequent phrasal verbs: A corpus-based analysis", *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect* 41(2), 339–359.

Gilquin, G. 2011. "Corpus linguistics to bridge the gap between World Englishes and Learner Englishes". In L. Ruiz Miyares & M.R. Álvarez Silva (Eds.), *Comunicación Social en el Siglo XXI*, Vol. II. Santiago de Cuba: Centro de Lingüística Aplicada, 638–642.

Gilquin, G. 2015a. "At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared", *English World-Wide* 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Gilquin, G. 2015b. "The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach", *Corpus Linguistics and Linguistic Theory* 11(1), 51–88. https://doi.org/10.1515/cllt-2014-0005

Gilquin, G. 2017. "Applied cognitive linguistics and second/foreign language varieties: Towards an explanatory account". In J. Evers-Vermeul & E. Tribushinina (Eds.), *Usage-based Approaches to Language Acquisition and Language Teaching*. Berlin: de Gruyter, 47–71.

Gilquin, G. & Granger, S. 2011. "From EFL to ESL: Evidence from the International Corpus of Learner English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 55–78. https://doi.org/10.1075/scl.44.04gra

Götz, S. 2015. "Fluency in ENL, ESL and EFL: A corpus-based pilot study". In Proceedings of Disfluency in Spontaneous Speech, DISS 2015. Glasgow, UK. Available at: http://disfluency.org/DiSS_2015/Programme_files/Goetz-DISS2015.pdf (accessed April 2016).

Götz, S. & Schilk, M. 2011. "Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced German learners". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 79–100. https://doi.org/10.1075/scl.44.05sch

Granger, S. 2009. "Prefabricated patterns in advanced EFL writing: Collocations and formulae". In A.P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, 185–204.

Graën, J. & Schneider, G. 2017. "Crossing the Border Twice: Reimporting Prepositions to Alleviate L1-Specific Transfer Errors". In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition* at NoDaLiDa, Gothenburg, 22nd May 2017, 18–26.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English*. Version 2 (Handbook + CD-ROM). Louvain-la-Neuve: Presses universitaires de Louvain.

Gries, S.T. & Wulff, S. 2005. "Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora", *Annual Review of Cognitive Linguistics* 3, 182–200. https://doi.org/10.1075/arcl.3.10gri

Gries, S.T. & Wulff, S. 2009. "Psycholinguistic and corpus linguistic evidence for L2 constructions", *Annual Review of Cognitive Linguistics* 7, 163–186. https://doi.org/10.1075/arcl.7.07gri

Gut, U. 2011. "Studying structural innovations in New English varieties". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 100–124. https://doi.org/10.1075/scl.44

Gut, U., Fuchs, R. & Wunder, E.-M. (Eds.). 2015. *Universal or Diverse Paths to English Phonology*. Berlin: de Gruyter. https://doi.org/10.1515/9783110346084

Jurafsky, D. & Martin, J.H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Laporte, S. 2012. "Mind the gap! bridge between world Englishes and learner Englishes in the making", *English Text Construction* 5(2), 265–292. https://doi.org/10.1075/etc.5.2.05lap

Lehmann, H.M. & Schneider, G. 2011. "A large-scale investigation of verb-attached prepositional phrases". In S. Hoffmann, P. Rayson & G. Leech (Eds.), *Studies in Variation, Contacts and Change in English, Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*. Varieng, Helsinki. Available at: http://www.helsinki.fi/varieng/series/volumes/06/lehmann_schneider/ (accessed April 2016).

Lehmann, H.M. & Schneider, G. 2012. "Dependency Bank". In Proceedings of *LREC 2012 Workshop on Challenges in the Management of Large Corpora*, 23–28.

Mukherjee, J. 2005. "All mine, mine alone…". Emerging local norms in Indian English lexicogrammar. Paper presented at the University of Zurich.

Mukherjee, J. 2007. "Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium", *Journal of English Linguistics* 35(2), 157–187. https://doi.org/10.1177/0075424207301888

Mukherjee, J. & Hoffmann, S. 2006. "Describing verb-complementational profiles of New Englishes: A pilot study of Indian English", *English World-Wide* 27(2), 147–173. https://doi.org/10.1075/eww.27.2.03muk

Mukherjee, J. & Hundt, M. 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.44

Nelson, G., Wallis, S. & Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Varieties of English Around the World: G29. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g29

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.14

Nesselhauf, N. 2009. "Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties", *English World-Wide* 30(1), 1–25. https://doi.org/10.1075/eww.30.1.02nes

Ng, H.T., Wu, S.M., Briscoe, T., Hadiwinoto, C., Susanto, R.H. & Bryant, C. (Eds.). 2014. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, Baltimore, Maryland, June. https://doi.org/10.3115/v1/W14-17

Pecina, P. 2009. *Lexical Association Measures: Collocation Extraction. Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Charles University in Prague.

Rosén, V. & Smedt, K.D. 2010. "Syntactic annotation of learner corpora". In H. Johansen, A. Golden, J.E. Hagen & A.-K. Helland (Eds.), *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag* [Systematic, Varied, but not Arbitrary. Anthology about Norwegian as a Second Language on the Occasion of Kari Tenfjord's 60th Birthday]. Oslo: Novus forlag, 120–132.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. 2001. *Multi-word expressions: A pain in the neck for NLP*. Technical Report LinGO Working Paper No. 2001-03, Stanford University, CA.

Salazar, D. 2014. *Lexical Bundles in Native and Non-native Scientific Writing*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.65

Sand, A. 2004. "Shared morpho-syntactic features in contact varieties of English: Article use", *World Englishes* 23(2), 281–98. https://doi.org/10.1111/j.0883-2919.2004.00352.x

Schneider, E.W. 2004. "How to trace structural nativization: Particle verbs in world Englishes", *World Englishes* 23(2), 227–249. https://doi.org/10.1111/j.0883-2919.2004.00348.x

Schneider, G. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. PhD Thesis. Institute of Computational Linguistics, University of Zurich.

Schneider, G. & Hundt, M. 2009. "Using a parser as a heuristic tool for the description of New Englishes". In Proceedings of *Corpus Linguistics 2009*, Liverpool.

Schneider, G. & Zipp, L. 2013. "Discovering new verb-preposition combinations in New Englishes", *Studies in Variation, Contacts and Change in English* 13. Available at: http://www.helsinki.fi/varieng/series/volumes/13/schneider_zipp (accessed April 2016).

Sedlatschek, A. 2009. *Contemporary Indian English: Variation and Change*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g38

Shannon, C. 1951. "Prediction and entropy of printed English", *The Bell System Technical Journal* 30, 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Tomasello, M. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Van Rooy, B. 2011. "A principled distinction between error and conventionalized innovation in African Englishes". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging the Paradigm Gap*. Amsterdam : John Benjamins, 189–207.

Van Rooy, B. 2015. "Annotating learner corpora". In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 79–105. https://doi.org/10.1017/CBO9781139649414.005

Wulff, S. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. London: Continuum.

The innovative progressive aspect of Black South African English

The role of language proficiency and normative processes

Bertus van Rooy and Haidee Kruger
North-West University / Macquarie University & North-West University

Conflicting findings are reported for New Englishes and Learner Englishes: similarities are identified mainly on psycholinguistic grounds and differences on sociolinguistic grounds. This chapter offers an analysis of the progressive form in Black South African English, in which the interaction between gradual increases in proficiency and normative interventions by explicit feedback and editing of published texts is examined to establish the route towards conventionalisation of innovative features. The results indicate that one innovative feature, the extension of the progressive to longer time spans, becomes established as a feature of the variety, but other potential innovations gradually disappear under normative influence and with increased proficiency. Innovations are likely to be accepted if they are insufficiently salient to be targeted for normative correction and sufficiently present in the written and spoken input to become entrenched in the grammatical representations of learners as they turn into advanced users of the New English.

Keywords: innovation, conventionalisation, progressive, Black South African English, normative processes, editing, feedback

1. Introduction

Empirical research aimed at bridging the gap between New Englishes (NEs) and Learner Englishes (LEs)[1] has, broadly speaking, arrived at one of two conclusions:

[1] New Englishes are also referred to as Outer Circle Englishes or English as a second language (ESL), while Learner Englishes are also called Expanding Circle Englishes or English as a foreign language (EFL). Theoretically, it is possible for a Learner English to develop into a New English, but as we argue in this chapter, it is not a straightforward process.

https://doi.org/10.1075/bct.98.04van

that the NEs and LEs are much more similar than recognised by scholars of World Englishes (e.g. Edwards & Laporte 2015; Gilquin 2015; Gilquin & Granger 2011; Laporte 2012; Nesselhauf 2009), or that the NEs are more different from native varieties of English, while the LEs are closer to the native varieties (Götz & Schilk 2011; Gries & Deshors 2015; Van Rooy 2006).

The reason for the differences in conclusions can be traced in part to differences in research focus. Researchers who find similarities between the forms, or functional extensions of forms, in NEs and LEs identify shared psycholinguistic processes of second language acquisition as the principal reason for the similarity (see Laporte 2012; Schneider 2012: 77). Formal and functional differences between NEs and other types of non-native Englishes are traced to sociolinguistic factors. The orientation towards an endogenous norm in NEs and a native-speaker norm in LEs is very salient in current research (Gilquin 2015: 96–97), alongside an awareness that the input variety of NEs is usually the local form, whereas a more concerted attempt is made to provide native speaker input in LE environments (Laporte 2012: 266). However, within the NE and LE settings, a fair degree of variability exists (Edwards 2014; Gilquin & Granger 2011; Schneider 2007: 155–161).

We accept as a given that similarities exist and that the psycholinguistic process of second language acquisition is one of the most important reasons for these similarities. However, closer investigation is required of the differences that are also attested in the literature, particularly with the aim of formulating explanatory hypotheses to account for these differences. Norm orientation and proficiency are important factors in explicating these differences, frequently referred to in existing research (Edwards & Laporte 2015); however, such research has seldom investigated the interaction between second-language proficiency and normative processes, which is the focus of this chapter.

The impasse in the current debate is that some linguistic features, or constructions, appear to show broad approximation to native-speaker constructions with increased proficiency in NEs and LEs alike (e.g. Edwards & Laporte 2015), whereas other features seem to indicate divergence between NEs and other varieties (e.g. Hundt & Vogel 2011). To take the debate further and propose a solution to the impasse, it is necessary to examine the effects of proficiency and normative processes in a single study, with data that control for the effects of both factors.

In this chapter, we examine the development and stabilisation of the innovative use of the progressive form[2] in Black South African English (henceforth BSAfE). We draw on corpora of timed student writing at different levels of proficiency

[2] Like Paulasto (2014), we use the term "progressive form" to denote the BE verb+-*ing* structure, and discuss the semantic and functional associations of the construction independently from progressive aspectuality as such.

alongside more advanced writing in order to control for the effects of proficiency. For published writing, we compare unedited and edited versions of the same texts to control for normative processes. In combination, the data allow us to examine the process of acceptance of the innovation by two key agents in the normative process: the teachers who give feedback on student writing, and professional editors. While the degree to which innovative forms are regarded as acceptable by the publishing industry is frequently raised as a measure of endonormativity in NEs (see Bamgbose 1998), and some scholars propose that the lower frequency of some innovative features in published written language may be the consequence of editorial intervention (see Götz & Schilk 2011; Mair 2006: 191–192; Mesthrie & Bhatt 2008), research in this area is limited. Our thesis is that candidate NE constructions that are in essence learner errors will decline in frequency in the data from increasingly proficient users partly due to teacher feedback, but conventionalised innovative constructions will remain in the data from the most proficient users, escaping censure from editors and teachers.

The influence of the context on conventionalisation is not exhaustively analysed with reference to normative processes — the frequency of occurrence of the progressive form and its different uses in the input data also play an important role. To capture the role of frequency effects, our analysis is couched in an emergentist, usage-based approach to language, and draws on concepts from Construction Grammar (Bybee 2010; Goldberg 1995). If a construction is defined as a conventionalised form-function pairing, our aim is to examine the process by which such conventionalisation takes place in new varieties. Following Croft (2000), we assume that stability is the norm in language change, and many possible innovations arise in language but disappear unnoticed most of the time, while only a small number of innovations eventually attain the status of being conventionalised. A very important factor in this process is frequency — both the frequency with which a user uses a construction, and the frequency of exposure to a variant in the overall language experience of a user (Bybee 2010). Awareness of normative acceptability mediates the associations that particular constructions have with register and formality. Over and above this more conscious association of constructions with different degrees of normative acceptability, norm orientations also overtly affect the published variety of a language that users are exposed to, thereby influencing frequency effects in the input that are usually below the level of conscious awareness of users. Norm orientation is also enforced by the education system and specifically by feedback that learners receive on their writing, which may further reinforce particular constructions and prevent the entrenchment of others. Where frequency effects can be predicted to be largely gradual and linear, normative interventions through feedback or editorial correction may introduce categorical or non-linear changes in the use of features to the extent that these interventions succeed in changing learners' constructional representations, aligning them with normative usage.

In the next section, we review current knowledge about the progressive in NEs and LEs, with specific attention to BSAfE. This is followed by a discussion of the research method, and the results, before the implications of the findings are discussed and conclusions offered.

2. The progressive in NEs, LEs and BSAfE

A relatively small number of studies have compared the use of the progressive in NEs and LEs, but until recently, the consensus (Hundt & Vogel 2011; Van Rooy 2006) has been that the LEs use the progressive with similar frequencies and in ways more similar to native varieties, specifically with a similar semantic range (Axelsson & Hahn 2001; Van Rooy 2006). Ranta (2006) examines the ELFA corpus and concludes that the vast majority of English Lingua Franca (ELF) uses are in agreement with standard English descriptions, but a residue of non-standard uses occur, although some of these uses are also attested, if with lower frequency, in the native-speaker control corpus. In contrast, many NEs use the progressive in an extended range of contexts (Hundt & Vogel 2011; Lunkenheimer 2012) although the degree to which extension takes place varies (Sharma 2009; Van Rooy 2014). Edwards (2014) challenges this view when she reports that in a number of ways, the progressive in Dutch English, traditionally classified as LE, does not conform to the predictions for an LE. These usages put Dutch English in a position on a continuum between Singaporean English, which is closer to native varieties, and Indian English, which is the furthest from native varieties. The most extensive previous comparative study, Hundt and Vogel (2011), makes use of student writing as data source, a limitation that also applies to Van Rooy (2006) and Axelsson and Hahn (2001). By contrast, Edwards (2014) compiles a corpus of advanced users of Dutch English that corresponds to the written part of an ICE-corpus, and her findings emphasise the importance of controlling for proficiency levels in comparing varieties (Edwards 2014: 189).

A remaining limitation in Edwards (2014) is the depth of the semantic analysis she undertakes. This is a limitation that Paulasto (2014) raises in a related context: the attestations of some similarities in extension or similarity in the frequency of non-standard usage may hide differences that lie below this level of analysis. Furthermore, the processes by which Dutch English arrived at this point cannot be inferred from the available data. Thus, while not rejecting Edwards's (2014) conclusions, it is necessary to examine constructional semantics and data that control for the gradual development and entrenchment of innovative uses across different proficiency levels more carefully to understand the significance and implications for the relationship between NEs and LEs.

Variation in the form and use of the progressive is captured by three of the features in the *Mouton World Atlas of Varieties of English* (Kortmann & Lunkenheimer 2012): the extension of the progressive to stative verbs, the extension of the progressive to habitual contexts, and the omission of the auxiliary BE before the progressive. As far as BSAfE is concerned, Mesthrie (2012: 497–498) notes that both semantic extensions, to stative and habitual contexts, are pervasive, while the omission of the auxiliary is neither rare nor pervasive. Previous research on BSAfE provides qualified support for these judgements, but also raises unanswered questions regarding the role of proficiency and normative processes.

Van Rooy (2014) reports that the progressive form is used with much higher frequency in BSAfE than in any other (native or NE) variety of English for which data are available. Register variation is extensive, with higher frequencies in speech than in writing, while student writing displays a much higher frequency than the published written registers — those registers that have been subject to the editorial process. This raises the question of whether the lower frequency in these registers is the consequence of the higher proficiency and increased register awareness of writers producing these texts, and/or of editorial intervention. The possibility of a proficiency effect receives support from research on spoken corpora, which indicates that more proficient users of BSAfE make less frequent use of the progressive than less proficient users (Minow 2010: 144; Siebers 2013: 145). Meierkord (2007: 335) finds that among student participants who attended a multi-racial school with native-speaker fellow pupils and native-speaker teachers, non-standard usages are very infrequent, whereas non-standard uses occur with high proportions among students who attended a school with black fellow pupils and teachers.

The use of the progressive to express states with longer duration is widely reported by researchers who have conducted semantic analyses (Meierkord 2007; Minow 2010; Siebers 2013; Van Rooy 2006, 2014). They find that the progressive is not merely extended to stative verbs and the dynamic/stative contrast is "overridden", as argued by Mesthrie and Bhatt (2008: 67), but rather that the meaning of a state or activity with longer duration than the conventional temporary duration of native varieties is denoted, as in examples (5) and (8) below. Habitual uses of the progressive aspect, illustrated by example (3) below, have received less attention in previous research. Siebers (2013: 159–161) identifies examples of the habitual use, especially in the past tense and in relative clauses, in the speech of some of her speakers, but notes that this is particularly characteristic of basilectal speakers with limited proficiency. Minow (2010: 198) reports the more widespread presence of habitual meanings in her data. Research on the omission of the auxiliary

BE, exemplified by (1) below, is limited to Minow (2010: 146–147), who reports that the deletion rate in her spoken BSAfE corpus is 9%. The feature occurs most often in the speech of the least proficient speaker groups.

Previous research on BSAfE confirms the extension of the progressive form to stative and habitual contexts. The association with proficiency is a consistent trend, with lower frequencies and fewer non-standard usages of the progressive in the speech of more proficient users. The omission of the auxiliary is attested with low frequency, and declines further with higher levels of proficiency. These findings may therefore be interpreted as evidence that non-standard uses of the progressive form are simply an effect of proficiency: as speakers become more proficient, they approximate the native-speaker target more closely. If this is the case, then these non-standard constructions should be classified as potential innovations that remained learner errors that did not attain conventionalised status.

However, two pieces of evidence are in conflict with this conclusion. The data on the role of proficiency in the frequency of the progressive form do not allow the construction of a continuous scale, but contain a discontinuity based on the educational background of an elite minority and the majority of speakers (Meierkord 2007; Minow 2010). Evidence for the disappearance of the non-standard uses (Meierkord 2012) is not conclusive and is problematised by methodological differences in classifying observations in terms of judgements of "standardness" or an analysis of constructional semantics. Resolving these two issues necessitates a new corpus investigation with closer controls of the relevant variables related to proficiency and normative processes.

3. Method

Five main corpora, representing different proficiency levels, were used in the analysis. To control for register variation, only argumentative writing, arguably the register that is most clearly subject to normative control, is used. At the level of school learners and undergraduate students, the corpora consist of argumentative essays, and at advanced levels, post-graduate dissertations and academic articles. Academic writing in English is a very typical activity that learners and mature users of BSAfE perform, and therefore the data have high ecological validity.

The composition of the corpora used are summarised in Table 1. At the lowest proficiency level, the corpus comprises argumentative essays written in the classroom by learners in the Eastern Cape Province, collected in 2003–2004.[3] The

[3] The data were collected by Mr. Madoda Nkani, for his PhD, which sadly never came to conclusion because of his untimely death.

learners were native speakers of isiXhosa, and were in Grades 10 and 11 at the time of data collection. The next proficiency level is represented by essays written by BSAfE speakers with a range of home languages[4] for the national matriculation examination (the final school examination, taken at the end of Grade 12) in 2003. The original sample was selected to represent a balance of achievement levels and from across the entire country, but only from public schools.

Undergraduate university writing is represented by timed essays written in the classroom and collected in 2003–2004 from two university campuses, one in the Eastern Cape Province, and one in Gauteng province. The essays were mostly produced by first-year students in academic support courses. In the former case, the vast majority of students are native speakers of isiXhosa. For the Gauteng university, all nine of the official African languages are represented in the sample.

Advanced writing is represented by two samples. One sample is extracted from a corpus of university dissertations at a university in the North-West province, where the majority of students were speakers of Setswana. A 1000-word sample was extracted from a random sample of 25 dissertations, obtained through the university database covering the first decade of the 21st century, and representing a range of disciplines. No dissertations that contained evidence of professional editing was included in the sample.

The second sample of advanced writing is a corpus of 21 complete academic texts, including postgraduate dissertations and scholarly articles written for domestic scholarly journals by BSAfE speakers, which were edited in full by professional editors. Both the original unedited and the professionally edited versions were included in this study. The texts span a range of disciplines across the humanities, and social and economic sciences, and were produced by postgraduate students and academics from at least four different institutions in Gauteng (not all metadata are known). The texts reflect the writing of at least 18 individual

Table 1. Composition of the corpora

Corpus	Texts	Word count
Grade 10–11 school essays	51	15,214
Matric examination essays	98	26,510
Undergraduate essays	141	54,856
Dissertation extracts	25	25,762
Published academic writing: unedited	21	254,093
Published academic writing: edited	21	256,479

[4] The Bantu languages in South Africa correspond very closely in their aspectual systems, and are not expected to lead to different types of native-language transfer (Piotrowska 2015: 57–71).

authors, of varying language backgrounds. The professional editors were either native South African speakers of English, or English/Afrikaans bilinguals. Each text was edited by one professional editor.

Data were extracted in WordSmith 6 (Scott 2015) by extracting all character strings ending in *-ing*, after which all non-progressive forms were removed manually. All forms of BE *going to* + VERB and the perfect progressive were excluded from the analysis. However, forms where the auxiliary verb BE was omitted but the clause was otherwise identified as a finite clause were retained, in order to determine the occurrence of auxiliary omission, as exemplified by (1):

(1) Poverty is the cause of HIV/AIDS in Africa becouse in this Country the is so mony people who Ø safaring in the long time (School essays, XRE047)[5]

The presence or absence of the auxiliary was coded as a separate variable, and the percentage omission was calculated per corpus.

For the analysis of overall frequency of the progressive, as well as the constructional semantics, values were normalised to a relative frequency per 1,000 words, reported to two decimal places. Where appropriate, differences between corpora are evaluated for statistical significance using a log-likelihood calculation, where the following levels of statistical significance are observed: if $\lambda > 3.84$, then $p < 0.05$, and if $\lambda > 10.83$, then $p < 0.001$.

The analysis of the uses of the construction is in the first place an analysis of constructional semantics. Based on previous attempts at undertaking similar classifications (Minow 2010; Paulasto 2014; Piotrowska 2015; Sharma 2009; Siebers 2013; Van Rooy 2006, 2014), each instance was assigned to a particular semantic class, e.g. temporary activity or state, stance, habitual, or ongoing activity. Thereafter, as a measure of the degree of acceptability or standardness of the semantic classes, each class was put in a subset, based on the extent to which the grammatical descriptions of native varieties of English (especially Biber et al. 1999: 470–475; Croft 2012: 152–155; Huddleston 2002: 162–172) afford central or prototype status to the construction, or treat particular extended usages as standard, borderline acceptable or non-standard (the latter often by omission, i.e. not discussing a particular usage as subtype of the progressive at all). Standardness is defined as centrality in the grammatical descriptions of native varieties of English. This classification method is used instead of a judgement of "acceptability" by native speakers. Native speakers do not always converge in their judgements (Kirsner 2014), and Axelsson & Hahn (2001) found that such a classification is especially

[5] Extracts from the corpora are unedited, so all original learner phenomena, including spelling mistakes, are included. Extracts are identified in terms of the corpus from which they are taken, followed by the filename for the particular text.

difficult to make consistently for the use of the progressive construction in non-native English usage. Therefore, acceptability is operationalised as degree of deviation from prototypical usages that are, by implication, presented as the standard forms in grammatical descriptions of native varieties of English.

Constructional meanings are grouped into four classes: (1) standard/prototypical native English usage, (2) acceptable extensions of the prototype, (3) borderline extensions that some native speakers might accept some of the time, and (4) clearly non-standard usages that are unlikely to be accepted by native speakers of English. A complete exposition of the constructional semantics falls outside the scope of the present chapter, and the discussion that follows focuses only on the classification in methodological terms (for more detail see Kruger & Van Rooy 2017; Piotrowska 2015; Van Rooy 2014).

The most prototypical standard usage has the meaning of a temporary activity or state that is ongoing at the time of reference, as illustrated in example (2). Two low-frequency usages were also included in the standard/prototypical category: the use of the progressive to present a time-frame within which another activity takes place, and simultaneity of two ongoing activities.

(2) the influx of WOMAN from different countries come to South Africa to look for the job but they didn't find it. They were hoping to to get job in South Africa… (Undergraduate essays, WZE041)

Various extensions of the progressive into new semantic territories are regarded as acceptable in standard English. These include the use of the progressive to refer to future events (with or without a modal auxiliary) or the stance use, as well as a number of related aspectual meanings — inceptive, iterative, habitual (ongoing habit for a delimited period of time following Sharma 2009) and the coercion of a punctual verb into a durative verb. A final subcategory in the extended standard group is the pluractional — where an event occurs repeatedly, but is not a habit (Piotrowska 2015: 53–55).

The contrast between the habitual and pluractional categories is illustrated by examples (3) and (4), where the agentivity and the regularity of the soccer players going to practice are attributed to the players as a habit in (3), but children losing their parents due to AIDS in (4) is not a habit of the children, yet it happens often.

(3) They [soccer players] must paid more because they loose their problems to concentrate with soccer Every day and everytime they must going to the field to practise the soccer ball. (School essays, XRE041)

(4) the most disastrous desease which has caused misery and frustration in our country each and everyday children are loosing their parents… (Undergraduate essays, XUE006)

Extensions on the border between standard and non-standard usage typically involve longer time-frames and do not have the attribute of temporariness. The characterising use, illustrated by (5), is restricted by definition to relative clauses that refer to a state associated with the antecedent of the relative clause, but without meeting the requirements for a habitual.

(5) The government supplies South African citizens with free condoms to ensure every individual's safety but since all the people that are suffering from poverty have got no time to listen they then give them to kids to play with. (Undergraduate essays, WZE013)

The contextual-frame use relates to the time-frame use, but there is no indication of temporariness, only incompleteness, as illustrated by (6):

(6) There is no free water, no electricity in their houses. They do not have jobs and also free education. What is happening they locked their water taps, they say he must pay rent althou peopple do not have jobs. (Matric exam essays, 051)

The ongoing activity or state interpretation represents an imperfective that is not presented as habitual or temporary, and with no immediate prospect of terminating, as exemplified by (7).

(7) Every town of South Africa has got prostitutes because people are triying to make money for food and other things (Undergraduate essays, XUE036)

Meanings that clearly fall beyond the typical usage described in standard grammars are unlimited states or, for dynamic verbs, temporal profiles that are interpreted as perfective, as in example (8), where no internal portion of the event is profiled, or perfect, as in example (9), where the prior activity has persistent relevance.

(8) Section 189 of the LRA with regard to strikers participating in a protected strike does not appear to be justifiable, because it is only protecting the employers and again it is possible to have the impasse resolved in some way… (Dissertations, MN)

(9) before going to the world cup they have mood to play because they know that after they are playing very hard they get more wages (School essays, XRE041)

To analyse the effect of editorial intervention in comparing the unedited and edited versions of the published writing, each instance of the use of the progressive in these two corpora was classified as a case where the progressive was accepted (present in identical form in both the unedited and edited versions), removed (present in the unedited text, but not in the edited), or added (present in the edited text, but not in the unedited) (see Kruger and Van Rooy 2017 for more detail).

4. Results and discussion

In this section, we report the findings of the study by focusing on three dimensions of the progressive construction: omission of the auxiliary, overall frequency of the progressive, and innovative semantic uses. The results indicate that the use of the progressive form changes with increased proficiency levels, as expected from the discussion above, but that the three dimensions of the progressive construction investigated change at differential rates, and have different outcomes. The differences provide new insight into the interaction between proficiency and normative processes that play a role in the conventionalisation of an innovative feature, or its demise as learner error.

4.1 Omission of the auxiliary BE

The omission of the auxiliary BE with the progressive form is the clearest instance of a learner feature, which occurs frequently in the writing of the least proficient writers, and disappears rapidly with increased proficiency, as shown in Figure 1. The omission of auxiliary BE in argumentative writing, illustrated by (1) above,

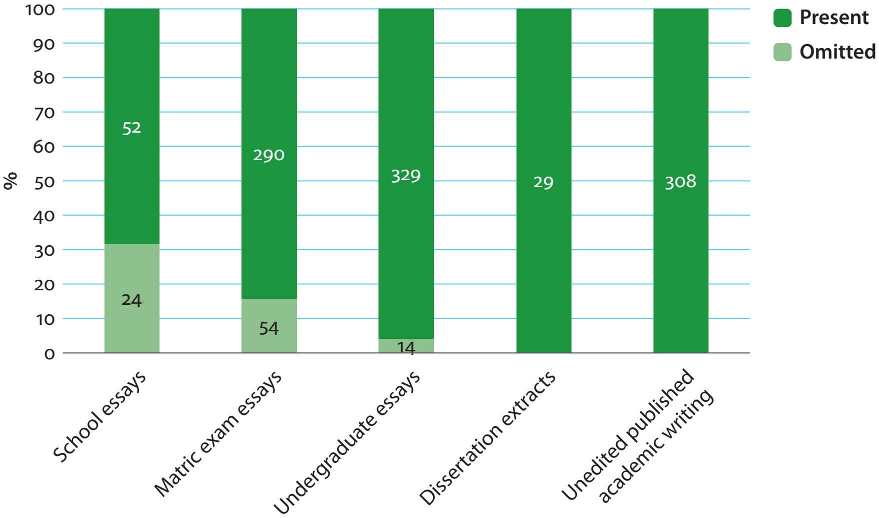


Figure 1. Percentage omission of auxiliary BE by corpus, with raw frequencies indicated in columns

follows a similar pattern of proficiency-related decrease as reported for spoken BSAfE by Minow (2010).

Two factors likely play a role in the early disappearance of the feature. Omission of the auxiliary BE is saliently marked as a grammatical error, and in all likelihood attracts censure from teachers, who, regardless of their own language background, will have a salient representation of this feature as an error. In addition, omission of auxiliary BE will not occur in the written-language input that learners receive in school environments. While there may be limited exposure in spoken language in school environments, the stigmatisation of this feature is likely to filter through to spoken language. In this way, the possibility for the error to become conventionalised is short-circuited by censure of the feature as an error (which becomes part of learners' cognitive representation of the construction, leading to avoidance of use) and the lack of input, which combine to prevent the construction from becoming entrenched.

4.2 Overall frequency of the progressive form

The overall frequency of the progressive form, as shown in raw frequencies in Table 2 and in normalised values in Figure 2, reveals a pattern that is not as neatly linear as the pattern for auxiliary omission. Rather, there is a binomial distribution with a relatively higher range of values for learners from school to undergraduate

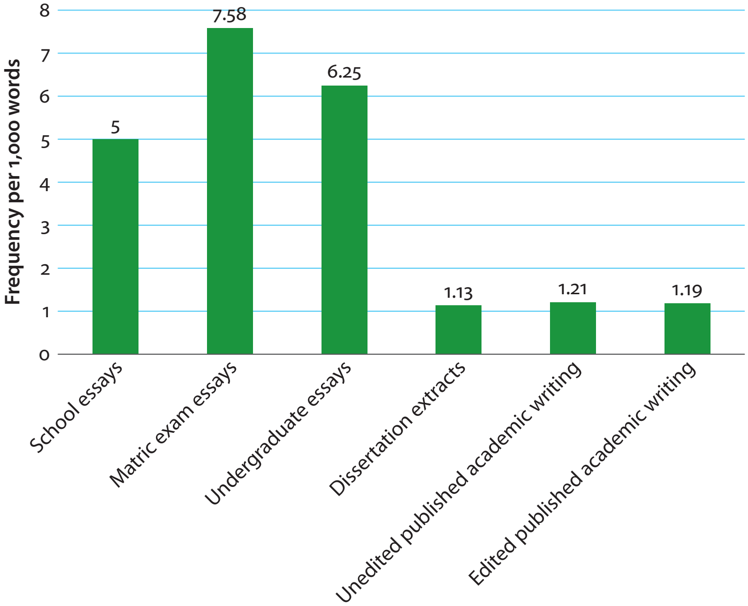


Figure 2. Normalised frequency of the progressive (per 1,000 words)

Table 2. Raw frequencies of progressive forms in the corpora

School essays	Matric exam essays	Undergraduate essays	Dissertation extracts	Unedited published academic writing	Edited published academic writing
76	201	343	29	308	306

level and much lower values for proficient users. The difference between dissertations and published writing (whether edited or unedited) is statistically not significant, whereas the values in dissertations and published writing are all significantly lower than in the school, matric and undergraduate corpora. For all relevant comparisons, $\lambda > 100$ ($p < 0.001$). The values for the three learner corpora exceed most previously reported values for student or learner writing (e.g. Axelsson & Hahn 2001; Hundt & Vogel 2011), and almost reach a similar frequency as reported for spoken BSAfE (Minow 2010; Siebers 2013; Van Rooy 2014). The low frequency of the progressive form in published academic writing is not due to extensive editorial intervention, since editors remove and add progressives in approximately equal measure (see Table 4 and related discussion).

The frequency increase of slightly more than 50% from the first to the second learner corpus is an unexpected finding ($\lambda = 10.17$, $p < 0.001$), since it runs counter to the expected gradual decline in frequency with increased proficiency. It appears that the learners who contributed to the school essay corpus have not mastered some senses of the progressive, as we will propose after reviewing the constructional semantics in Section 4.3. A slight decline in frequency is observed, in line with expectations, from the high school learners to the university students ($\lambda = 4.62$, $p < 0.05$), which corresponds in magnitude with the distance between two adjacent proficiency groups in the studies of Minow (2010) and Siebers (2013). However, the frequency change from undergraduate students to postgraduate students and published academic writing is qualitatively different ($\lambda = 126.61$, $p < 0.001$). The numbers for undergraduate students and academic writing are very similar to the data reported by Van Rooy (2014: 164), although he used two different corpora to represent student writing and academic writing respectively, which means that the sudden drop in frequency cannot be attributed to sampling variance.

In interpreting these findings, both the differences in input and explicit normative feedback across the school and higher-education contexts should be considered. The high frequency of the progressive in the writing of school learners and first-year student writing may in part be the consequence of teacher input combined with peer input. The teachers of these learners are mainly BSAfE speakers whose own speech contains high frequencies of the progressive form: Van Rooy's (2014) analysis of BSAfE classroom lectures demonstrates a frequency of approximately 10 instances per 1,000 words. Learners up to early university level get input reflecting the overuse of the progressive in spoken language not only from their peers, but also from teachers, the norm-setting authority in the school context, who are less likely to censure the overuse of the progressive since it is part of their own usage. Furthermore, although the material school learners read is unlikely to reflect the overuse of the progressive, it nevertheless contains many instances of the progressive form, in contrast to auxiliary BE omission, which is absent from written input. Frequency of input in particularly the spoken context, reinforced by written input, and combined with limited overt censure on use therefore allow the feature to become entrenched.

The notable decline from first-year writing to postgraduate writing may be accounted for by two factors. First years are distinguished from postgraduates in terms of the amount of reading they do, as well as in terms of the types of texts that they read. First years have just completed high school, which is characterised in the case of the majority of black South Africans by limited reading and lack of access to books (Pretorius & Mampuru 2007). Postgraduate students' exposure to the discourse norms of standard English in formal written contexts is both quantitatively and qualitatively different from first years' exposure. The exposure to texts produced in accordance with these norms most likely decreases students' propensity to use the progressive, particularly because there is a strong incentive to conform to these discourse norms which may incline students to privilege this input in the development of their own writing. Thus, students acquire the conventions of the genre of academic discourse, which in general uses the progressive less often than other written genres (see Van Rooy 2014: 164 for comparative numbers from BSAfE). Such acquisition comes from reading, but also from writing texts that approximate the genre conventions more and more as time passes. The second factor, related to more extensive opportunity to write academic texts, is that, over the course of undergraduate and postgraduate study, the effects of writing feedback on constructional representation may accrue. Importantly, this feedback comes from lecturers who are typically more proficient, and also less likely to be mesolectal speakers of BSAfE — white academic staff are in the majority in higher education (Department of Higher Education and Training 2014: 17). In this context, lecturers as norm authorities are more likely to comment on the "overuse" of the progressive in their feedback and also provide spoken input with lower frequencies of the progressive than school teachers.

The definite and substantial reduction in the frequency of use of the progressive only at the very advanced level of postgraduate academic writing demonstrates that the increased frequency of the progressive is an innovative feature that remains in circulation far longer than the omission of the auxiliary BE. However, it does not become fully conventionalised, and at high levels of proficiency the frequency of the progressive is vastly reduced, approximating the native norm.

4.3 Innovative semantic uses

The semantic range of the progressive, compared across the six corpora, shows an initial extension from mainly non-standard uses to a larger proportion of standard uses with advancing proficiency, but a subset of the non-standard uses remains extremely prominent even among the most proficient users, as shown in Table 3 and Figure 3.

A number of important findings emerge from the results in Table 3 and Figure 3. The aspectual meaning of temporariness, illustrated by (2) above, which is the most prototypical in standard English, occurs in high proportion in the data, with the exception of the school essays at the lowest end of the proficiency scale. The other meaning that persists throughout the data set, and is indeed even more frequent than temporariness in every single corpus, is the meaning classified as ongoing. This meaning, illustrated by (7) above, is at the outer limits of the semantic range of the native English progressive form, but is clearly the prototypical meaning of the progressive in BSAfE, as Van Rooy (2014) has already proposed. Apart from its high frequency across the corpora in this study, the fact that the temporariness meaning is underrepresented in the school essay corpus means that the ongoing meaning is entrenched first in the acquisition of English by BSAfE speakers.

The ongoing meaning is related to the claims about the extension of the progressive to stative verbs, but involves a semantic difference in the constructional schema that extends beyond simply overriding the stative/dynamic contrast. Van Rooy (2006) relates this extension to the persistitive aspect in the South African Bantu languages, encoded by the verbal prefix *-sa* in all nine indigenous African

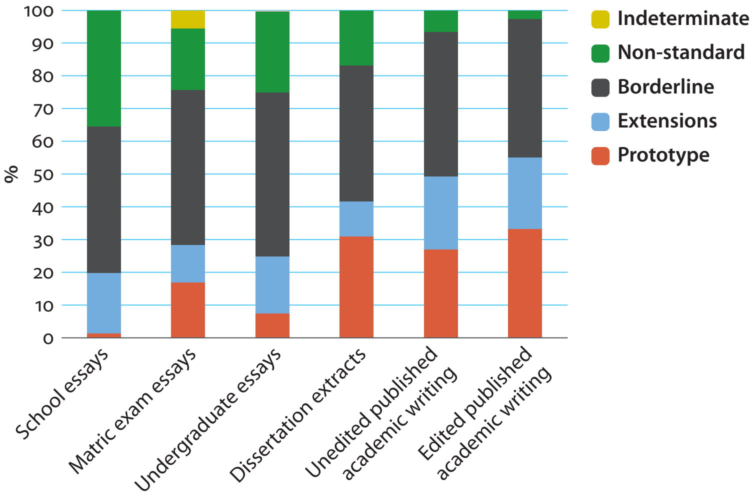


Figure 3. Distribution of uses of the progressive, proportionally by corpus

Table 3. Distribution of uses of the progressive (normalised per 1,000 words), with percentage for each overall normative category in brackets

Construction semantics	School essays	Matric exam essays	Undergraduate essays	Dissertation extracts	Unedited published academic writing	Edited published academic writing
Prototype/standard usage	(1%)	(17%)	(8%)	(31%)	(27%)	(33%)
Temporary	0.07	1.09	0.27	0.27	0.31	0.37
Time frame	0	0.11	0	0	0.01	0.01
Simultaneous	0	0.08	0.20	0.08	0	0.02
Extended standard usage	(18%)	(11%)	(17%)	(10%)	(24%)	(22%)
Future	0.07	0.23	0.20	0	0.03	0.02
Inceptive	0	0	0	0.04	0.02	0.02
Habitual	0.33	0.26	0.04	0.04	0.02	0.01
Iterative	0	0.04	0.02	0	0.1	0.11
Coerce	0.2	0	0.04	0	0.05	0.05
Stance	0.07	0.08	0.02	0	0	0
Pluractional	0.26	0.26	0.78	0.04	0.04	0.05
Borderline standard usage	(45%)	(47%)	(50%)	(41%)	(43%)	(42%)
Characterising	0.26	0.30	0.26	0.04	0.08	0.08
Contextual frame	0.39	0.64	0.44	0.08	0.06	0.04
Ongoing	1.58	2.64	2.42	0.35	0.41	0.39
Non-standard usage	(36%)	(19%)	(25%)	(17%)	(6%)	(3%)
Perfect reading	0.26	0.38	0.13	0	0.02	0.01
Perfective	1.12	0.87	0.95	0.16	0.06	0.02
Unlimited	0.39	0.19	0.47	0.04	0	0
Indeterminate	0 (0%)	0.41 (5%)	0.02 (0%)	0 (0%)	0 (0%)	0 (0%)
Total	5.00	7.58	6.25	1.13	1.21	1.19

languages, which is historically derived from a verb form with the meaning 'stay' (Pretorius 1997: 145). Thus, in the earliest phases of acquisition, in line with the finding of Sharma (2009) for Hindi speakers learning English, Bantu language speakers match the progressive form in English to the semantic range of the persistitive aspect in their native languages. This extends frequently to usages that are not prototypical or regarded as standard in native varieties, such as unlimited time frames in the school essays, a usage that persists in the matric exam and undergraduate essays, as shown by (10), where the state of having AIDS, as an incurable disease, is temporally unlimited.

(10) Here we find that mine workers are mostly the people who are having HIV/AIDS and obviously no one want to work at mines. (Undergraduate essays, WGE008)

Also among the non-standard usages is the extension into the non-imperfective domain that is also attested at early stages, either as perfective events, exemplified by (8), or corresponding to a perfect aspect, as shown by (9). These non-standard extended usages do not conventionalise as a feature of BSAfE, however. The relative frequencies of these three usages are already lower in the matric exams and undergraduate essays, and are all but absent from advanced writing, as is visualised by Figure 3. By contrast, from the matric exam essays onwards, the native prototype meaning of temporariness occupies an increasing share of the overall proportion of uses. The most typical standard usages therefore become entrenched as part of the constructional representation of the BSAfE prototype, but as an extension of the ongoing meaning, which is prototypical for BSAfE, if less standard from the native-speaker perspective.

Habitual meanings, illustrated by (3) above, are relatively frequent only in the essays of the school learners and already declines in frequency with undergraduate student essays. The related category of pluractional meanings, illustrated by (4), holds a considerable share of all the progressive usages up to undergraduate student level. The characterising use, illustrated by (5), also maintains a reasonable frequency rate up to the undergraduate student essays. In advanced writing, however, these categories decline in normalised frequency as well as in their proportional share of the meanings. Unlike the non-imperfective usages, though, they do not disappear, and may still approximate the status of a conventionalised (but less salient) use of the progressive, especially the characterising use, with its very specific syntactic frame.

On the balance of the evidence, however, extensions to various subtypes of habitual meanings are not nearly as well entrenched as the extensions to the domain of ongoing states or activities. It is the ongoing use that is the truly innovative feature of BSAfE that has become entrenched, not simply as an extended usage, but as the prototypical meaning of the progressive form. As the proficiency of learners increase, they let go of the non-imperfective meanings, and gradually, but not to the same degree, also of the habitual meanings.

The comparative analysis of the unedited and edited corpus of published academic writing further supports our analysis. The acceptance of innovative uses of the progressive by editors may be regarded as the final, most stringent test of conventionalised status. Editors are typically native speakers from the same context, who act as gatekeepers of standard usage, and their acceptance of innovative usages signals a high degree of endonormative stabilisation. The findings of the comparative analysis are presented in detail in Table 4, and summarised in Figure 4. Editorial intervention plays a non-negligible further role in lowering the frequency of the most unusual usages, while increasing the most standard usages. Crucially, a large group of borderline cases enjoy acceptance among the editors, despite the fact that they are not BSAfE users themselves.

Table 4. Editorial intervention in published BSAfE academic writing

Construction semantics	Accepted	Removed	Added
Prototype/standard usage			
Temporary	80	0	15
Time frame	2	0	1
Simultaneous	1	0	3
Extended standard usage			
Future	5	2	0
Inceptive	5	0	0
Habitual	3	1	0
Iterative	28	0	0
Coerce	13	0	0
Stance	0	0	0
Pluractional	11	6	1
Borderline standard usage			
Characterising	19	1	2
Contextual frame	8	0	1
Ongoing	96	7	4
Non-standard usage			
Perfect reading	2	2	0
Perfective	6	9	0
Unlimited	0	1	0
Total	279	29	27

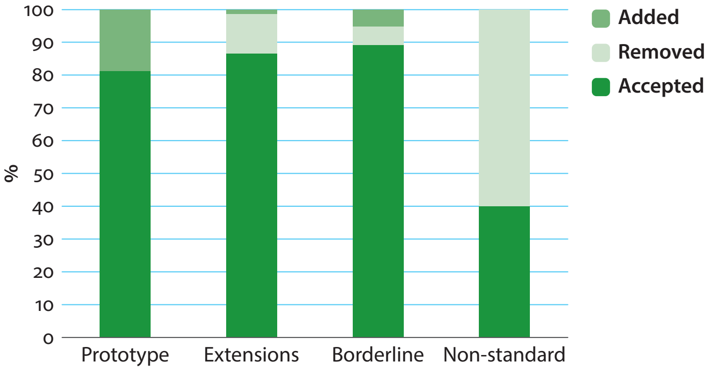


Figure 4. Editorial changes to the progressive, proportionally, by construction semantics

Editors never remove standard English progressive constructions, but do add them. The possibility exists that stigmatisation of the progressive stemming from feedback during graduate and postgraduate study may lead to a kind of hypercorrective avoidance of the progressive among advanced writers. Some opportunities to use the native prototype functions may therefore not be utilised by writers — with editors inserting the progressive where appropriate.[6] This kind of correction is illustrated in example (11):

(11) a. It was also found that there were non-success cases <u>who did not or were unable to implement</u> their learning from the FPD training. (Published writing, unedited, A098)

b. It was also found that there are unsuccessful SMTs <u>that are not implementing</u> or are unable to implement their learning from the FPD training. (Published writing, edited, A098)

Extended standard uses are minimally affected by editorial intervention, except the pluractional meaning, which is targeted for removal by the editors relatively frequently, as illustrated by (12) below. Together with the already low frequency of the habitual in published writing, the editors' intervention points to the lower degree of entrenchment of meanings related to habitual aspect in the construction network of the progressive form of advanced users of BSAfE and the lower degree of acceptability of this extended use.

[6] Editors intervene by changing texts minimally, as shown by example (13b) or more extensively, as shown by example (11b) and (12b). For the purposes of this paper, our focus is on the effect this has on the overall frequency of the progressive form, and the kinds of meanings that are selected to be added, removed or accepted. A more detailed analysis of how the extent of the editorial intervention interacts with the different kinds of meanings is presented by Kruger and Van Rooy (2017).

(12) a. We have a long way to go before we can understand clearly the complexity of how and why the pandemic is killing our children (Published writing, unedited, A090)

b. We have a long way to go before we can clearly understand the complexity of how and why the pandemic affects our children. (Published writing, edited, A090)

The ongoing meaning, together with characterising and contextual frame meanings, are all regarded as of borderline acceptability in standard English. The ongoing meaning, which, as pointed out above, is the most frequent meaning of the progressive across all corpora, is generally accepted by editors and seldom changed. Similarly, the characterising and contextual meanings attract few changes; overall there are almost equal numbers of removals and additions in the borderline category. The growing acceptance of these uses is very clearly signalled by the data, particularly in comparison with meanings related to the habitual use.

The non-conventionalised status of non-imperfective uses is evident in the fact that progressive forms with this meaning are most frequently removed (as exemplified in (13)), and never added.

(13) a. If the Department of Education's officials expect the educators to fill-in forms using specialized terminology, before attending to the learner, it will be very unfair to the learners who must still be attending school despite the apparent incompetence of educators in this regard. (Published writing, unedited, A090)

b. If the Department of Education's officials expect educators to fill out forms, using specialised terminology, before they are willing to attend to the learner, it will be very unfair to the learner who must still attend school despite the apparent incompetence of educators in this regard. (Published writing, edited, A090)

The analysis of the editorial changes to the progressive clearly demonstrates that non-BSAfE editors in the South African context tacitly accept the innovative prototypical meaning of the progressive (that of ongoing time), which signals the endonormative stabilisation of the innovation, and allows it to further diffuse in writing. However, there are limits to editors' acceptance: constructional extensions approaching the most non-prototypical perfect and perfective meanings are not allowed, and habitual uses are also frequently removed.

5. Conclusion

The results of this study enable us to trace the fate of the innovative uses of the progressive form from an early learner stage to the stage where the mature users of BSAfE retain one innovative use, but other candidate innovations gradually disappear, which confers on them the status of learner error. At the earliest stages, the persistitive meaning is transferred from the native languages to English. Learners extend this meaning to the domain of non-imperfective uses, but such extensions disappear with increased proficiency and do not become established new conventions of BSAfE. On the other hand, extensions in the direction of temporary activities or states increase as proficiency increases, and become entrenched as part of the constructional schema of the BSAfE progressive form. The original core meaning for early BSAfE learners, imperfectives with longer duration, which is at the borderline between standard and non-standard usage for native speakers, maintains its central position in the construction network even at the most advanced stages, and is the true innovation that conventionalises in BSAfE. Other developmental features that do not become established conventions are the quantitative overuse of the progressive and the omission of the auxiliary BE, while the extension to habitual meanings does not become very salient, although it does not disappear altogether either.

The finding can be explained with reference to a number of factors. Early learners draw on their native languages, which gives rise to innovative uses, but they also show evidence of the receptiveness that English as target language has for extended uses of the progressive (see Kranich 2010). Over time, feedback in the educational context that targets stigmatised uses (auxiliary omission, quantitative overuse and extensions to meanings that are recognisably non-standard) combine with input from standard language written texts to reinforce standard usage and extended uses that are not salient enough to be targeted for consistent feedback or normative correction. Feedback may engender an overreaction, such that in some semantic ranges, the construction is actually underused by advanced BSAfE speakers, but at this point editors step in to reinforce standard usage. The implicit feedback from continued exposure to the extended (but not completely non-standard) extended-time uses of the progressive form, reinforced by minimal editorial intervention, serves to further entrench the innovative uses associated with the ongoing use of the progressive. This reinforcement takes place against the background of spoken BSAfE, where the ongoing time use of the progressive is even more prominent (Siebers 2013, Van Rooy 2014).

The implication of our finding is that NE users display typical learner features at the early stage of language acquisition. Feedback from the educational system and normative interventions in the publication process serve to filter out many of the learner features that never become established and entrenched as conventionalised variants of the NE. These normative processes are particularly effective where the non-standardness of a particular usage is salient to the teachers and editors. However, innovations become entrenched, most likely at the border of standard language input, where the environment continues to reinforce such usage, and where local editors become sufficiently accustomed to the innovative uses to accept them and let them through to published texts.

In an environment where English is used extensively as spoken language, and where published texts are produced locally, the chances for some innovations to be reinforced and become conventionalised are therefore higher than in typical LE settings, where the published input is mainly produced by native speakers and the spoken environment is not characterised by such extensive indigenous input. Thus, if English input via the media and interaction with foreigners (be they native or non-native speakers) form a significant part of the input to the users of English, the likelihood of systematic reinforcement of innovative uses, which is a condition for conventionalisation, is lower. We therefore conclude that while it is not in principle impossible for an LE to develop into an NE, it will be much more difficult to achieve this in the usual LE settings. In other words, there is not so much a case of an unnecessary paradigm gap between NEs and LEs, but rather a case of genuine differences in the contexts of use that make it more likely for conventionalisation to be achieved in NE contexts than in LE contexts.

Acknowledgements

The financial support of the National Research Foundation (NRF) of South Africa is acknowledge with gratitude, but the views expressed in this chapter should not be attributed to the NRF. Feedback from the reviewers and editors contributed to the final chapter, but any remaining shortcomings are for our account.

References

Axelsson, M.W. & Hahn, A. 2001. "The use of the progressive in Swedish and German advanced learner English – A corpus-based study", *ICAME Journal* 25, 5–30.

Bamgbose, A. 1998. "Torn between the norms: Innovations in world Englishes", *World Englishes* 17(1), 1–14. https://doi.org/10.1111/1467-971X.00078

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.

Croft, W. 2012. *Verbs: Aspect and Causal Structure*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199248582.001.0001

Department of Higher Education and Training. 2014. *Statistics on Post-School Education and Training in South Africa* 2012. Pretoria: DHET.

Edwards, A. 2014. "The progressive aspect in the Netherlands and the ESL/EFL continuum", *World Englishes* 33(2), 173–194. https://doi.org/10.1111/weng.12080

Edwards, A. & Laporte, S. 2015. "Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels", *English World-Wide* 36(2), 135–169. https://doi.org/10.1075/eww.36.2.01edw

Gilquin, G. 2015. "At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared", *English World-Wide* 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Gilquin, G. & Granger, S. 2011. "From EFL to ESL: Evidence from the International Corpus of Learner English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 55–78. https://doi.org/10.1075/scl.44.04gra

Goldberg, A.E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Götz, S. & Schilk, M. 2011. "Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English for advanced German learners". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 79–100. https://doi.org/10.1075/scl.44.05sch

Gries, S. Th. & Deshors, S.C. 2015. "EFL and/vs. ESL? A multi-level modeling perspective on bridging the paradigm gap", *International Journal of Learner Corpus Research* 1(1), 130–159. https://doi.org/10.1075/ijlcr.1.1.05gri

Huddleston, R. 2002. "The verb". In R. Huddleston & G.K. Pullum (Eds.), *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, 72–211.

Hundt, M. & Vogel, K. 2011. "Overuse of the progressive in ESL and learner Englishes – fact or fiction?" In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 145–165. https://doi.org/10.1075/scl.44.08vog

Kirsner, R.S. 2014. *Qualitative-Quantitative Analyses of Dutch and Afrikaans Grammar and Lexicon*. Amsterdam: John Benjamins. https://doi.org/10.1075/sfsl.67

Kortmann, B. & Lunkenheimer, K. (Eds.). 2012. *The Mouton World Atlas of Variation in English*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110280128

Kranich, S. 2010. *The Progressive in Modern English: A corpus-Based Study of Grammaticalization and Related Changes*. Amsterdam: Rodopi.

Kruger, H. & Van Rooy, B. 2017. "Editorial practice and the progressive in Black South African English," *World Englishes* 36(1), 20–41. https://doi.org/10.1111/weng.12202

Laporte, S. 2012. "Mind the gap! Bridge between world Englishes and learner Englishes in the making", *English Text Construction* 5(2), 264–291. https://doi.org/10.1075/etc.5.2.05lap

Lunkenheimer, K. 2012. "Typological profile: L2 varieties." In B. Kortmann & K. Lunkenheimer (Eds.), *The Mouton World Atlas of Variation in English*. Berlin: Mouton de Gruyter, 844–873. https://doi.org/10.1515/9783110280128

Mair, C. 2006. *Twentieth-Century English: History, Variation, and Standardization*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486951

Meierkord, C. 2007. "Standards and norms in interactions across second language Englishes: The case of South Africa." In S. Kolk-Birke & J. Lippert (Eds.), *Anglistentag 2006 Proceedings*. Trier: Wissenschaftliche Verlag Trier, 331–340.

Meierkord, C. 2012. *Interactions Across Englishes: Linguistic Choices in Local and International Contexts*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139026703

Mesthrie, R. 2012. "Black South African English". In B. Kortmann & K. Lunkenheimer (Eds.), *The Mouton World Atlas of Variation in English*. Berlin: Mouton de Gruyter, 493–500. https://doi.org/10.1515/9783110280128

Mesthrie, R. & Bhatt, R.M. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511791321

Minow, V. 2010. *Variation in the Grammar of Black South African English*. Frankfurt/Main: Peter Lang.

Nesselhauf, N. 2009. "Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties", *English World-Wide* 30(1), 1–26. https://doi.org/10.1075/eww.30.1.02nes

Paulasto, H. 2014. "Extended uses of the progressive form in L1 and L2 Englishes", *English World-Wide* 35(3), 247–276. https://doi.org/10.1075/eww.35.3.01pau

Piotrowska, C.M. 2015. *A Diachronic Analysis of the Progressive Aspect in Black South African English*. Unpublished M.A. dissertation. North-West University.

Pretorius, E.J. & Mampuru, D.M. 2007. "Playing football without a ball: Language, reading and academic performance in a high-poverty school", *Journal of Research in Reading* 30(1), 38–58. https://doi.org/10.1111/j.1467-9817.2006.00333.x

Pretorius, R.S. 1997. *Auxiliary Verbs as Subcategory of the Verb in Tswana*. Unpublished PhD thesis. Potchefstroom University.

Ranta, E. 2006. "The 'attractive' progressive - Why use the -ing form in English as a Lingua Franca?", *Nordic Journal of English Studies* 5(2), 95–116.

Schneider, E. 2007. *Post-Colonial Englishes: Varieties Around the World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618901

Schneider, E. 2012. "Exploring the interface between world Englishes and second language acquisition – and implications for English as a lingua franca", *Journal of English as a Lingua Franca* 1(1), 57–91. https://doi.org/10.1515/jelf-2012-0004

Scott, M. 2015. *Wordsmith Tools 6*. Liverpool: Lexical Analysis Software.

Sharma, D. 2009. "Typological diversity in new Englishes", *English World-Wide* 30(2), 170–195. https://doi.org/10.1075/eww.30.2.04sha

Siebers, L. 2013. *Morphosyntax in Black South African English: A Sociolinguistic Analysis of Xhosa English*. Tübingen: Narr.

Van Rooy, B. 2006. "The extension of the progressive aspect in Black South African English", *World Englishes* 25(1), 37–64. https://doi.org/10.1111/j.0083-2919.2006.00446.x

Van Rooy, B. 2014. "Progressive aspect and stative verbs in Outer Circle Varieties", *World Englishes* 33(2), 157–172. https://doi.org/10.1111/weng.12079

Towards a process-oriented approach to comparing EFL and ESL varieties

A corpus-study of lexical innovations

Marcus Callies
Universität Bremen

This paper adopts a process-oriented approach to comparing EFL and ESL varieties and examines to what extent they are driven by general cognitive processes of language acquisition and production. A comparative corpus-study of lexical innovations in derivational morphology brings to light two general types of innovations: 1) interlingual, L1-based innovations, resulting from cross-linguistic influence, and 2) intralingual, L2-based innovations, resulting from various other processes. While the first type is virtually absent in ESL varieties, it is in the second type where similar types of innovations in EFL and ESL varieties can be observed. The paper argues that these innovations can be explained in terms of several underlying cognitive processes that serve to create and maximise morphological transparency and increase explicitness of form-meaning relations.

Keywords: lexical innovation, word-formation, cognitive processes, back-formation, (over-)regularisation, overaffixation, isomorphism, explicitness, transparency

1. Introduction

In the context of current research in the field of English corpus linguistics that challenges the traditional division between foreign language / learner varieties of English (EFL) and institutionalised second-language varieties of English (ESL) (the so-called "paradigm-gap", see Sridhar & Sridhar 1986), this paper presents a comparative corpus-study of lexical innovations in derivational morphology. The study is based on the observation that despite the manifold differences between EFL and ESL (summarised e.g. in Gilquin 2015 and Laporte 2012), and although the two types of varieties have traditionally been examined in different research paradigms (EFL in Second Language Acquisition research, ESL in

https://doi.org/10.1075/bct.98.05cal

research on World Englishes), there are a number of similarities that warrant a comparative perspective. Both are 'non-native' varieties, are acquired in institutionalised settings as foreign or second languages in language contact situations (Hundt & Mukherjee 2011: 2), and, most importantly for the present context, have been assumed to be subject to similar cognitive processes of language acquisition and production (see Schneider 2012; Sharma 2012; Williams 1987). This paper thus adopts a process-oriented approach to comparing EFL and ESL varieties and examines to what extent they are both driven by such processes.

The testing ground for doing so is word-formation, a major mechanism for the expansion of the vocabulary in a language that involves knowledge of the combinatory properties of affixes and bases. Surprisingly, there is still comparatively little research on EFL learners' productive use of derivational morphology when compared, for instance, to research on the acquisition of L2 vocabulary (see e.g. Callies 2015 for review). It appears that studies on word-formation in ESL varieties are equally sparse. Early work provided descriptive overview accounts of individual ESL varieties, discussing a whole set of word-formation processes but mostly presenting anecdotal evidence that lacked a broad empirical basis and quantitative documentation (see e.g. Baumgardner 1998 and Görlach 1989). More recently, however, Biermeier (2008, 2009, 2014) has examined word-formation in a broader range of eight Asian and African ESL varieties (India, Hong Kong, Singapore, the Philippines, Kenya, Tanzania, Nigeria and Ghana) on the basis of quantifiable corpus data from the *International Corpus of English* (ICE; Greenbaum 1996). Biermeier's work largely focuses on neologisms and provides ample evidence for the productive use and creative potential of ESL varieties to give rise to lexical innovations based on the rule-governed application of the major word-formation processes of English.

In this study, lexical innovations are considered forms that are unattested (or infrequent/rare) in the main standard varieties (British and American English) and are products of morphological regularity (i.e. by applying word-formation rules) or creativity in that new words are formed by either adapting L1 elements to fit L2 forms, or by recombining L2 elements. Thus, forms considered as innovations result from a productive process and can be considered systematic. The analysis therefore includes forms that may in other, more normative approaches be interpreted as errors or performance phenomena, i.e. mistakes. In line with the process-oriented approach to comparing the two types of varieties advocated here, however, such forms are explicitly included.

The present paper aims to compare EFL and ESL varieties and sets out to answer the following research questions: What types of lexical innovations in derivational morphology can be observed in the two types of varieties? What are the possible underlying processes that give rise to these innovations? Are there qualitative and quantitative differences and/or similarities between EFL and ESL varieties as to the types and number of innovations to be observed?

2. Corpus study: Data and methodology

The corpus study presented here draws on data from the *International Corpus of Learner English* (ICLE; Granger et al. 2009) that contains written texts (mostly argumentative essays) produced by EFL learners from 16 different mother tongue backgrounds. The learners are university students of English who learnt English as a foreign language in a classroom setting with formal instruction. Despite its explicit design, the ICLE is a fairly mixed collection of texts when considering the context in which they were produced and the proficiency level of the learners. Some essays were produced under exam conditions with a set time frame and no access to reference tools such as grammars and dictionaries, while others were written as homework assignments in the students' own time with access to reference tools. Moreover, in the compilation of the ICLE the learners' proficiency level was assessed globally by means of external criteria, i.e. students were considered advanced because of their institutional status as "university undergraduates in English (usually in their third or fourth year)" (Granger et al. 2009: 11). However, the results of human rating of twenty essays per ICLE-subcorpus according to the proficiency levels of the *Common European Framework of Reference for Languages* (CEFR) (Granger et al. 2009: 12) showed that the proficiency level of the learners represented in the ICLE actually varies between (higher) intermediate to advanced.

Taking advantage of the learner and textual metadata that are available for all corpus texts, five homogenous and comparable subcorpora of texts produced by Russian, Turkish, German, Italian and Spanish EFL learners were compiled (see Table 1). Since cross-linguistic influence has been shown to be an important factor in the acquisition of L2 lexis and morphology (e.g. Jarvis & Pavlenko 2008: Chapter 3), it was deemed important here to select learner groups from typologically different L1 backgrounds. The vocabulary of English is highly mixed because it has heavily borrowed words and derivational morphemes from Romance languages. Therefore, the set of learner corpora to be examined contains learners from Romance (Italian and Spanish), Germanic (German), Slavonic (Russian) and Turkic (Turkish) languages.

The five subcorpora comprise argumentative essays written by students who had seven to eight years of instruction in English at school and who had not spent more than six months in an English-speaking country. In addition, based on the assumption that in a high-stakes context such as a timed exam students would not try out ‘riskier’ strategies in case of problems in lexical search (because these may lead to errors which are usually penalised), only those texts were considered that were not produced under exam conditions and for which the learners were not set a time limit and had access to reference tools.

Table 1. ICLE-subcorpora examined in the present study

Corpus	Writers’ L1	Professional status	Genre	# texts	# words
ICLE_RUS	Russian	student	argum. essay	57	46,534
ICLE_TUR	Turkish	student	argum. essay	75	53,329
ICLE_GER	German	student	argum. essay	58	37,976
ICLE_ITA	Italian	student	argum. essay	32	20,702
ICLE_SPA	Spanish	student	argum. essay	52	33,594

In the procedure outlined above homogeneity and comparability were established at the expense of corpus size. However, since lexical innovations are ad-hoc, non-institutionalised formations, they cannot be assumed to occur in high frequencies. Thus, a second learner corpus also containing written texts was considered to enlarge the database. The *International Corpus Network of Asian Learners of English* (ICNALE; Ishikawa 2013) is a large corpus of controlled English essays written by several groups of ESL and EFL learners from Asian countries, i.e. learners from China and Hong Kong, Taiwan, South Korea, Japan, Singapore, Indonesia, Pakistan, the Philippines and Thailand. It currently consists of 5,600 essays (1.3 million words) produced by learners who were assigned to various proficiency levels on the basis of standardised language tests.[1] Because of its large size and multi-L1 design, the ICNALE has great potential for learner corpus studies. The usefulness of the data is somewhat limited for the present study because the texts were produced in a tightly controlled setting.[2] However, this controlled compilation process offers valuable methodological insights with regard to the effects of task setting on writers’ use of non-canonical forms / innovations (see also the paper by Van Rooy and Kruger (this volume), on the influence of editorial practices on innovations). Each student who contributed to the corpus had to write two essays on set topics (“It is important for college students to have a part time job” and “Smoking should be completely banned at all the restaurants in the country”). While this facilitates comparability across the subcorpora and enables the study of development across proficiency levels, it also limits the degree of lexical varia-

[1] See http://language.sakura.ne.jp/icnale/about.html for details (accessed 23 September 2015).

[2] See “The ICNALE-Written: Instructions for Participants” at http://language.sakura.ne.jp/icnale/instruction.html (accessed 23 September 2015).

tion among the texts found in the corpus. Moreover, essays were strictly timed (the writing time for one essay was set to 20–40 minutes) and had to be between 200 and 300 words in length. The use of reference works was disallowed. Students had to key in the essays into a word processing software and also had to perform an obligatory spell-check before submission. It can be assumed that during this obligatory spell-check the word processor identified and highlighted non-standard, hence potentially innovative forms to provide corrections which the large majority of students then followed. To conclude, it is very likely that this step led to the fact that most of the potentially used innovations were lost because of the controlled setting.

A subset of the ICNALE was examined to check its suitability as a supplementary database for the present study. Since the metadata are limited (e.g. it is not indicated whether or not the students spent time studying abroad) and not directly linked to the corpus texts, learners' proficiency level was applied to compile a subset that would match the ICLE data. Therefore, only texts written by learners who had been assigned to the B2 proficiency level according to the CEFR were selected as this is the most frequently attested proficiency level in the Asian, i.e. the L1-Chinese and L1-Japanese, components of the ICLE (Granger et al. 2009: 12). The subcorpus thus compiled from the whole corpus, i.e. considering all L1-backgrounds, consisted of a total of 464 texts / 131,000 words with an average text length of 244 words. When compared to the ICLE data, however, only very few instances of innovations were identified, and it can be assumed that this is largely because of the reasons given above. Nevertheless, the innovations found will be used in the discussion below for means of further exemplification, but will not be considered in the quantitative analysis presented in Section 3.

The EFL data are compared to similar ESL data from the ICE. Six different ESL varieties were selected (see Table 2). To ensure comparability to the EFL corpora, only texts from the written, non-printed, and non-professional sections were considered. However, as the ICE corpora contain only a small section of student writing, not only the student essays but also the exam texts had to be included for want of a sufficiently large database. This means that parts of the ESL data are not exactly comparable to the EFL data because they were produced under exam conditions and thus, may be less rich in lexical innovations for the reasons pointed out above.

For the same reasons mentioned previously in the context of the EFL corpora, the ESL database was enlarged by using the recently released *Global Web-based English Corpus* (GloWbE; Davies 2013-; Davies & Fuchs 2015), a web-derived corpus composed of 1.9 billion words from 1.8 million web pages in 20 different English-speaking countries (ESL and native English). The texts in the corpus mostly consist of informal blogs (about 60 per cent) and other written texts harvested

Table 2. Components of ICE examined in the present study

Corpus	Professional status	Genre	# texts	# words
ICE-East Africa (EA) (Kenya)	student	essay/exam	21	40,037
ICE-Hong-Kong (HK)	student	essay/exam	20	49,436
ICE-Philippines (PHI)	student	essay/exam	20	46,477
ICE-Singapore (SIN)	student	essay/exam	20	46,343
ICE-India (IND)	student	essay/exam	20	41,162
ICE-Nigeria (NIG)	student	essay/exam	23	29,695

from the Internet, such as newspapers, magazines, and company websites (Davies & Fuchs 2015: 3). Because of the sheer size of this corpus it was impossible to search it exhaustively for lexical innovations. Thus, several types of innovations identified in both the ICLE and ICE data were searched for in the entire GloWbE to obtain further evidence for their occurrence in other ESL varieties. Again, these data were not considered in the quantitative analysis presented in Section 3.

Table 3 lists the mixed set of derivational affixes that was investigated. The set includes negative prefixes and suffixes used to form verbs, nouns, and adjectives, among them native (of Germanic origin) and non-native affixes (of Latinate origin).

Table 3. Affixes examined in the present study

Type of affix	Germanic	Non-Germanic
prefixes	*un-*	*in-**, *de-*, *dis-*
suffixes		
– verbal		*-ify*, *-ate*, *-ize/-ise*
– nominal abstract	*-ness*, *-ment*, *-hood*, *-ship*	*-ity*, *-ism*, *-(ific)(at)ion*
– adjectival	*-ful*	*-able/-ible*, *-ive*, *-(ic)al*

* including the phonologically conditioned allomorphs *il-*, *im-*, and *ir-*.

While it would be desirable to examine a wider set of derivational affixes, in the present study the aim was to use a balanced but mixed set of native (i.e. of Germanic origin) and non-native affixes (i.e. of Latinate origin) as these differ in their morphophonological properties and ease/difficulty of acquisition. Most non-native suffixes integrate into the prosodic structure of the base and cause morphophonological changes leading to stem allomorphy and morphophonological opacity (e.g. *curious* > *curiosity* or *decide* > *decision*). They are also more restricted in combinability because they prefer non-native bases (e.g. **mindal* vs. *mental* or

**unpossible* vs. *impossible*). Therefore, for L2 learners derivatives with non-native suffixes can be assumed to be 'irregular', inconsistent and structurally less transparent, thus more marked and more difficult to acquire and use. On the other hand, native suffixes do usually not trigger mutations in the base and their derivatives are more transparent than those created by non-native suffixes. They are also less restricted in combinability because they are usually indifferent to the etymology of the base. In sum, they can be considered less marked when compared with non-native ones.

The ICE corpora contain a special mark-up with normative corrections that proved useful for automatically tracing further potential innovations. However, it appears that this mark-up (described in the ICE tagging manual in a section on "Normalizing the text") has not been applied consistently across all ICE components.[3] Exhaustive lists of all words carrying the affixes listed in Table 3 were retrieved from the corpora and then examined manually with false positives being discarded. For the verb- and noun-deriving suffixes the search included all inflected forms. All forms unfamiliar to the author were considered potential candidates for innovations and were thus checked against the *Oxford English Dictionary* online (OED 2015) and large reference corpora of Present-Day English, the *British National Corpus* (BYU-BNC; Davies 2004-) and the *Corpus of Contemporary American English* (COCA; Davies 2008-). Only forms not attested in either the dictionary or the corpora, or marked as obsolete or rare in the OED (which were then very often infrequent or not attested in the corpora) were included in the examination. Forms marked by the OED as having fallen out of use in the standard variety (i.e. obsolete or rare words) are included in this study because they are well-formed and were in use in English in earlier periods of time.[4] From the point of view of L2 users / learners (and also most native speakers), who form words on the basis of their (often implicit) knowledge of the word-formation rules of English, it is irrelevant if a respective word has fallen out of use as these linguistically naive users usually do not have access to historical information. In the case of co-existing forms, i.e. when two forms, an innovation and an established form, were attested in the OED and found in the corpora (e.g. *unmoral* vs. *immoral*, *destruct* vs. *destroy*) then the innovative form was clearly the dispreferred variant in terms of frequency of use (assessed by means of frequencies of occurrence in the corpora).

In a handful of cases it was difficult on the basis of the production data alone to reconstruct the process that may have given rise to a specific form. For instance,

[3] The manual can be retrieved from http://www.ice-corpora.net/ice/written.doc (accessed 29 September 2015).

[4] See Baumgardner (1998: 212; 224) for similar observations on Pakistani English.

the form *experimentating*, v. could be analysed as either being back-formed from the noun *experimentation* or as being derived by means of attaching the verbal suffix -*ate* to the noun *experiment*. Such cases illustrate the competing motivations that are at play here: either forming a new verb on the basis of an already existing complex form and retaining the paradigmatic relation of the two forms (= back-formation), or adding derivational suffix to an existing simpler base (*experiment*, n.) to clearly mark the new form as more complex, thereby setting it apart from the nominal base (= overaffixation). In these cases it was decided to take the general corpus frequency of the two competing base forms as an indicator of their availability in terms of cognitive activation. In other words, the form with the higher frequency of use was considered as the underlying base form unless contextual factors suggested otherwise, e.g. if the less frequent form was used and thus activated in the immediately preceding context. This is in line with psycholinguistic models of word storage and processing which assume that frequent words are more easily stored as wholes and accessed than infrequent complex words (see e.g. Baayen & Schreuder 2003). In principle, however, it seems impossible to decide which process was actually at play. For a comprehensive analysis, experimental data are needed. Ultimately, however, this does not impact the argument that is made here because both processes serve the same purpose, namely to create or maximise morphological transparency and increase the explicitness of form-meaning relations as will be shown in Section 3.

3. Results and discussion

The data evidence two general types of innovations: 1) interlingual, L1-based innovations that result from cross-linguistic influence (these will be discussed first further below), and 2) intralingual, L2-based innovations that are the product of various other production principles that will be exemplified and discussed in turn in the rest of this section.[5] Williams puts emphasis on economy of production, which includes regularisation and the production of redundant markers (1987: 169), and a tendency towards what she calls "hyperclarity", which serves the reduction of ambiguity and consists of the subprinciples of "maximum transparency" and "maximum salience" (1987: 178). Schneider (2012) discusses several (partially overlapping) processes that he suggests offer a processing advantage,

[5] In this paper, the term "cognitive process" is used for a relatively wide and partially overlapping set of principles of language perception, processing and production, some of which zoom in on particular morphological processes. For the sake of terminological clarity, all of these will be referred to as processes in this paper.

among them are simplification, (over-) regularisation, redundancy, analogy and isomorphism, i.e. the tendency towards a more explicit marking of categories as to a one-to-one matching of form and meaning.

3.1 Overview and exemplification of cognitive processes

a. Cross-linguistic influence (CLI)

CLI is conceived of here as "the influence resulting from similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired" (Odlin 1989: 27). It is broader in scope than the traditionally used term "transfer" and includes the various manifestations and outcomes of CLI in terms of interference (or negative transfer), positive transfer (or facilitation), avoidance and overproduction (see e.g. Ellis 2008: 354ff.). Some of the outcomes of CLI show that it is not only the differences, but also the similarities between two languages that can matter when explaining learners' L2 production. One major (psycho-)linguistic factor that inhibits or promotes transferability is cross-linguistic similarity (Jarvis & Pavlenko 2008: 213). Cross-linguistic similarity (or language distance) refers to the (perceived) relationship or degree of congruence between L1 and L2. It is also a central concept in Kellerman's (1983) re-evaluation of transfer as a cognitive process that is, among other things, subject to an individual learner's perception of the distance between L1 and L2 (what Kellerman calls "psychotypology").

For the EFL varieties, CLI was established by consulting for each L2 form found in the data a monolingual dictionary of the respective learner's L1.[6] If an L1 counterpart of the L2 form in question was available and could be clearly identified as the base form, then the L2 form was considered a product of CLI. For example, in (1), the form *refugiated* (not attested in standard English) found in the Spanish component of the ICLE was most likely formed on the basis of the Spanish verb *refugiarse*. By contrast, the L2 form *brutify*, also produced by a Spanish EFL writer, was analysed as a product of recombining L2 elements, not CLI, because although the Spanish cognate *brutalizar* is available, it was apparently not used to form the standard English form *brutalize*, see Example (4).

In the present context, CLI emerges as foreignising, i.e. the morphophonological modification of an L1 form to adapt to the structure of the L2. In (1)–(3), L1

[6] For the present purposes, CLI was operationalised narrowly as morphophonological modification of an L1 form to adapt to the structure of the L2. In the ESL varieties the L1 backgrounds are typologically (esp. in terms of lexicon and derivational morphology) relatively far removed from English that CLI is highly unlikely and thus, was not assessed.

bases are used to create complex and well-formed but unattested (hence innovative) L2 forms.[7]

(1) People instead of worrying about their problems, they **refugiated** and was subyugated by religion … [*take refuge*; from Span. *refugiarse*, v. 'take refuge'] (ICLE_SPA)

(2) Lately the situation has deteriorated so much that **ambiental** risks represent a serious problem not only for nature … [*environmental*; from Ital. *ambientale*, adj. 'environmental'] (ICLE_ITA)

(3) The main causes of crime are, in fact, **emargination**, oppression, segregation … [*marginalisation*; from Ital. *emarginazione*, n. 'marginalisation'] (ICLE_ITA)

b. Word coinage

This process involves the formation of neologisms by recombining L2 elements. For instance, *unmerciful* and *twentyhood* were coined by combining existing English base forms (*merciful*, *twenty*) with derivational affixes to form new words which are well-formed but blocked by other forms that are already in use (*merciless*, *twenties*).[8]

(4) Money **brutify** the persons, it makes grow up in them the ambition for having more and more money. [*brutalize*][9] (ICLE_SPA)

(5) … poor people are the most cruel and **unmerciful** in this eternal fight for property. [*merciless*] (ICLE_RUS)

(6) For those of us (I assume this 'us' would include naïve young women on the brink of **'twentyhood'**) … [*twenties*] (ICE-PHI W1A-002)

(7) It enables the researchers to know the prevalence of **exceptionary** child and may be look for ways of minimising this. [*exceptional*] (ICE-EA W1A-021K)

[7] In the unaltered corpus examples of innovations provided in this paper, near-equivalents in standard English (if they exist) are given in brackets. Sometimes, L1 cognates and relevant word-formation processes are added.

[8] Further examples from the ICNALE are ***dangerousness*** (kor_smk_282_b2_0), ***teenage-hood*** (sin_ptj_094_b2_0), ***respirational*** [*respiratory*] (hkg_smk_007_b2_0), and ***unhealthful*** [*unhealthy*] (chn_smk_163_b2_0).

[9] Interestingly, the availability of the Spanish cognate *brutalizar*, v. did not trigger positive transfer to produce the standard verb form *brutalize*.

c. (Over-)Regularisation

This well-known process of language acquisition emerges in various forms in the data. First, there are innovations that involve cases where stem allomorphy applies in standard English, see (8) and (9).[10] Apparently, these forms are driven by the need to increase morphological transparency so that the morphophonologically unaltered base is preserved ('regularised') in the formation of a complex noun.

(8) … peace and world cooperation aimed at better **solvation** of numerous problems of humanity. [*solve*, v. → *solution*, n.] (ICLE_RUS)

(9) … casual music **consumation** … [*consume*, v. → *consumption*, n.] (ICLE_GER)

Overregularisation can also be observed in cases such as (10) and (11). Here, the etymological constraint on the combinability of affixes and bases such that Latinate affixes are preferred with Latinate bases is overridden and the Germanic prefix *un-* is attached, hence overgeneralised. Interestingly, in (10) this regularisation process applies even though L1-L2 similarity should have a facilitating effect.[11]

(10) … enrich its powerful and **unmoral** pockets? [*immoral*; cf. Ital. *immorale*, adj.] (ICLE_ITA)

(11) Pokuwaa made so many sacrifices to the gods due to her **unfertility** so that the gods will bless her with children. [*infertility*] (ICE-NIG ex_03)

Finally, regularisation also seems to affect cases of competition between the adjective-forming suffixes *-ic* and *-ical*, which in some cases results in different meanings as in the case of *economic* 'relating to trade, industry, and the management of money' and *economical* 'using money, time, goods etc. carefully and without wasting any'. Here, the competition is often resolved in favour of *-ical* as the default, more explicit variant which may actually result in the choice of a dispreferred, infrequent or rare variant of a rival pair as in (12) and (13).

(12) A person who took secondary education in a **touristical** place can speak fluently … [*touristic*, *touristy*] (ICLE_TUR)

(13) Rather, it is obvious that we, as informed readers, are to evaluate this **fantastical** journey as well as to appreciate the parallels between … [*fantastic*] (ICE-SIN W1A-004)

[10] Callies & Szczesniak (2007) report a further example from the Polish component of the ICLE: ***suspection*** [*suspicion*].

[11] There was only one example of this process found in the ICNALE-subcorpus used in this study: ***uncomfort*** [*discomfort*] (kor_smk_227_b2_0).

d. Overaffixation

Overaffixation is conceived of as the redundant use of a derivational affix which creates "overexplicit" instances where an additional affix is not necessary because of conversion or subtractive processes. This process is very likely motivated by isomorphism / the principle of iconicity: More (abstract) meaning is marked by more linguistic material. Many of these cases, e.g. (14)–(18), are apparently formed on a more complex, paradigmatically related base (i.e. *inventate*, *transportise*, *opportunity*, and *contradiction*).

(14) … effects the dictatorial governments in Africa, in the Middle East, in the ex-USSR, where the personal dignity had been **scarified** several times … [*scare*, *threaten*] (ICLE_ITA)

(15) … the television is an **inventation** which has caused more myopia in a lot of boys and girls. [*invention*] (ICLE_SPA)

(16) … the **transportization** of values of the state … [*transportation*] (ICE-PHI W1A-018)

(17) The questions or concerns does Huckleberry Finn raise about values in Western Society are individualism, alienation and **opportunitism**. [*opportunism*] (ICE-HK W1A-018)

(18) They find a resemblances between the objects & constructs a **uncontradictionary** experians about the fact & their special characteristics. [*uncontradictory*] (ICE-IND W1A-015)

This pattern is also well attested in the GloWbE corpus, see examples (19)–(22).[12]

(19) … a positive effect on the performance of dairy cattle mainly attributed to better utilisation of crop residues by choosing and **supplementating** with urea-molasses mixture. [*supplementing*] (GloWbE, Kenia)

(20) Due to pass experience people were probably afraid to leave their homes fearing **vandalizism**. [*vandalism*; from *vandalize*](GloWbE, Jamaica)

(21) Instead Ghana boasts a new era in which **capitalizism** and democracy are the talk of the day … [*capitalism*; from *capitalize*] (GloWbE, Ghana)

[12] Callies & Szczesniak (2007) provide a further example from Polish-English interlanguage: ***contestating*** [*contesting*]. Flowerdew (2006: 92) reports two examples from Chinese EFL writing: ***expenstion*** [*expense*] and ***prospection*** [*prospect*]. The only example found in the ICNALE subcorpus used in the present study is ***touristism*** [*tourism*] (hkg_smk_015_b2_0).

(22) Both, SHG and MFI models are to carve out the **mechanisism** to serve the segment as a long term profitable business opportunities. [*mechanism*; from *mechanise*] (GloWbE, India)

e. Back-formation

These innovations can be analysed as back-formations from more complex, paradigmatically related forms, see (23)–(26), in particular complex nouns ending in -*ation* that result in back-formed verbs in -*ate*.[13]

(23) … but they had a mind to ge[t] a better and **organizated** life to them. [*organized*; from *organization*] (ICLE_SPA)

(24) … always looking for some ideal, dreaming about it **imaginating** it. [*imagining*; from *imagination*] (ICLE_RUS)

(25) … entrenching in discriminations between obstacles as **representating** more or less infringement of freedom. [*representing*; from *representation*] (ICE-HK W1A-002)

(26) … **destruct** and disturb the communication. [*destroy*; from *destruction*] (ICE-IND W1A-013)

This pattern is again also found in the GloWbE data, see examples (27)–(30).

(27) Experience in **implementating** 2D/3D Graphics is an advantage. [*implementing*; from *implementation*] (GloWbE, Hong Kong)

(28) See the < photo album > on our site **documentating** this event. [*documenting*; from *documentation*] (GloWbE, Bangladesh)

(29) … they have perhaps contributed to **liberalizating** values in a predominantly Muslim culture … [*liberalizing*; from *liberalization*] (GloWbE, Bangladesh)

(30) … through restructuring and **privatizating** public enterprises … [*privatizing*; from *privatization*] (GloWbE, Tanzania)

[13] It could be argued that some of these forms may have been construed in analogy to other verbs that end in -*ate*. However, these forms follow a clear pattern as for all of them a more complex, paradigmatically related form is available (no other case is attested in the data), hence back-formation was chosen as the preferred analysis. A further example found in the ICNALE is ***considerate*** [*consider*; back-formation from *consideration*] (jpn_smk_302_b2_0). Callies & Szczesniak (2007) report further instances produced by Polish and German EFL learners: ***applicated*** [*applied*; back-formation from *application*], ***expectating*** [*expecting*; back-formation from *expectation*], and ***consolate*** [*console*; back-formation from *consolation*]. An example from Pakistani English is reported by Baumgardner (1998: 224): ***renunciate*** [*renounce*; back-formation from *renunciation*].

f. Simplification

It is well known that many high-contact varieties are influenced by simplification, i.e. the reduction of form or formal complexity (see e.g. Sharma 2012). In the present context, only very few cases of simplification were found, e.g. simplification of syllable clusters as in (31). In contrast to the observation that there are many more cases in which innovations were formed on more complex, paradigmatically related forms, there is only one case that could involve the reduction of a base, see (32).[14]

(31) … she was severly **critized** from each point. [*criticized*] (ICE-IND W1A-011)

(32) … laissez faire economic policy and **non-intervenist** approach to the Chinese culture. [*non-interventionist*; possibly formed on the basis of *intervene*, v.] (ICE-HK W1A-012)

g. Analogy

Analogy is a general cognitive process that transfers specific information or knowledge from one instance or domain (variously called the analogue, base, or source) to another instance. The driving force for analogy is the desire to make conceptually related linguistic units similar (or identical) in form, motivated by economy of form. Only few examples of analogy could be found in the data, see (33)–(35).[15]

(33) The main problem is about the political and **economical** conditions of the country that make people sluggish. [*economic*] (ICLE_TUR)

(34) The Cold War, which occurred from 1950 to 1990, was the conflict between the two great superpowers, the United States and the Soviet Union, in ideological, political, **economical** and social terms. [*economic*] (ICE-PHI W1A-018)

(35) … **tyranncy** of majority will occur. [*tyranny*; cf. *regency*, *democracy*] (ICE-HK W1A-017)

3.2 Quantitative findings

Due to their very nature of being ad-hoc, non-institutionalised formations, the types of innovations discussed above are admittedly rare (they occur less than

[14] A second example that presents evidence for this process occurred in the ICNALE: ***addictness*** [*addictiveness*] (kor_smk_233_b2_0).

[15] In (33) and (34) the potential analogues are underlined, in (35) they are mentioned in brackets.

once per thousand words in the individual corpora) and thus, the envisaged quantitative analysis suffers from the general problem of data sparsity. Individual lexical instantiations may be rare indeed, mostly single instances, but follow a pattern and in sum present compelling evidence for the operation of a variety of underlying cognitive processes.

Figures 1 and 2 aim to quantify and visualise the findings per variety type. When comparing the distribution of the seven cognitive processes across variety types (Figure 1) it is striking that all but CLI occur in both EFL and ESL. Zooming in on the five EFL learner populations (Figure 2), it becomes clear that it is the EFL learners with a Romance L1 background (Spanish and Italian) whose innovations are mostly driven by CLI when compared to learners with a non-Romance L1 (German, Russian, and Turkish). This suggests that in these learner varieties, there is a comparatively strong influence of the L1, most probably because of the perceived high degree of cross-linguistic similarity with regard to the common Latinate lexis and derivational morphology of English and Spanish / Italian which seems to facilitate CLI (see also Balteiro 2011 for similar findings).

Considering the six ESL groups (Figure 3), the findings by Biermeier on cross-variety differences in ESL can only be partially confirmed. However, the present study only used a subset of the data used by Biermeier (2009, 2014), who only included "well-formed" neologisms based on word coinage. Moreover, Biermeier

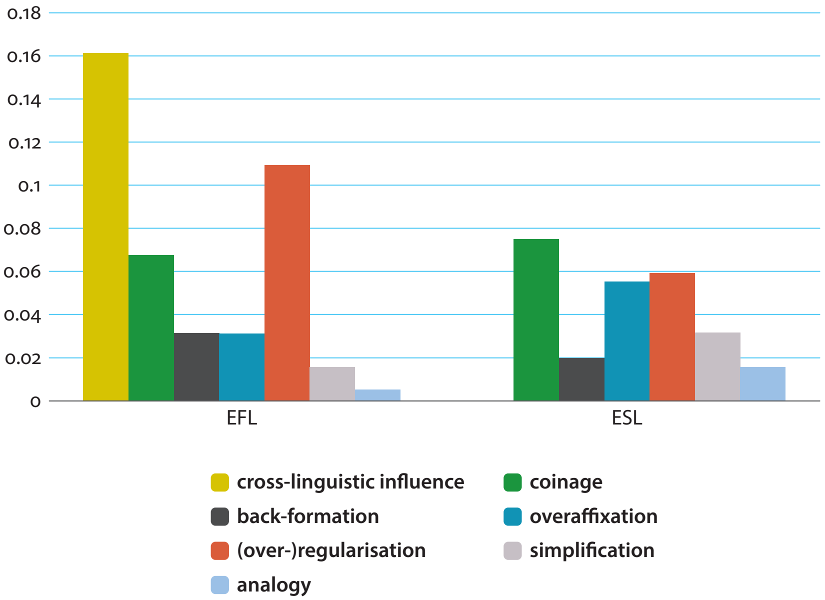


Figure 1. Distribution of seven cognitive processes across EFL and ESL varieties as represented in the ICLE and ICE subcorpora (frequencies per thousand words)

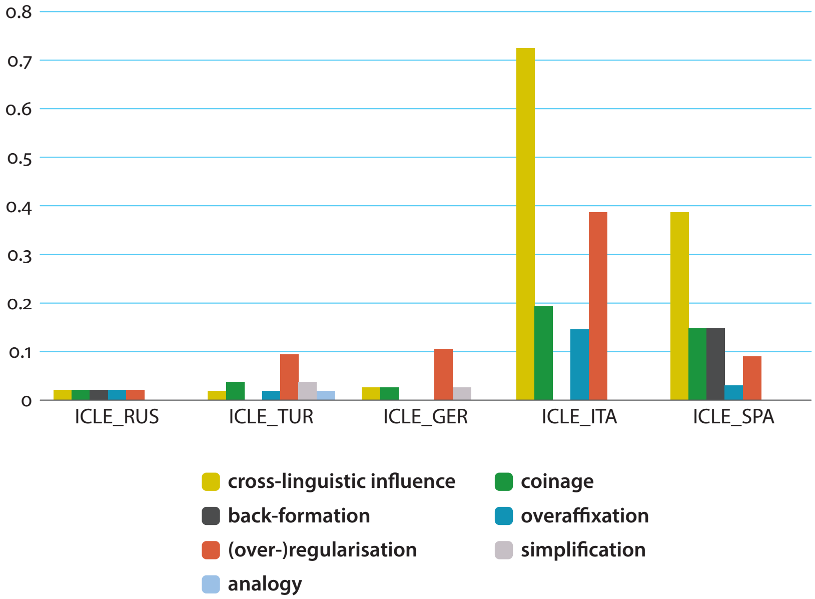


Figure 2. Distribution of seven cognitive processes across five EFL varieties as represented in the ICLE subcorpora (frequencies per thousand words)

looked at a wider set of word-formation processes (including compounding) than is the case here. In other words, Biermeier's approach was different as he did not consider unconventional forms, i.e. errors or performance phenomena. In sum, it has to be admitted that in view of data sparsity for the different ESL varieties the picture is still inconclusive as to the question if there are similarities and/or differences between them that could be related to their development in terms of the evolutionary phases described in Schneider's (2003) Dynamic Model. Future detailed analysis of the vast GloWbE data is needed to bring clear patterns and differences to light.

What the data clearly show, however, is that the many instances of (over-)regularisation, overaffixation and, to a limited extent, back-formation suggest that similar cognitive processes are at play in the two variety types. These processes serve to create or maximise morphological transparency and increase explicitness of form-meaning relations. In (over-)regularisation this is evident in the preservation of morphophonologically unaltered, hence more transparent base forms in the formation of complex nouns, or the loosening of combinatory restrictions. In overaffixation, more or more abstract meaning is marked by more, often redundant, linguistic form, and in back-formation, novel verbs are derived from more complex, paradigmatically related nouns.

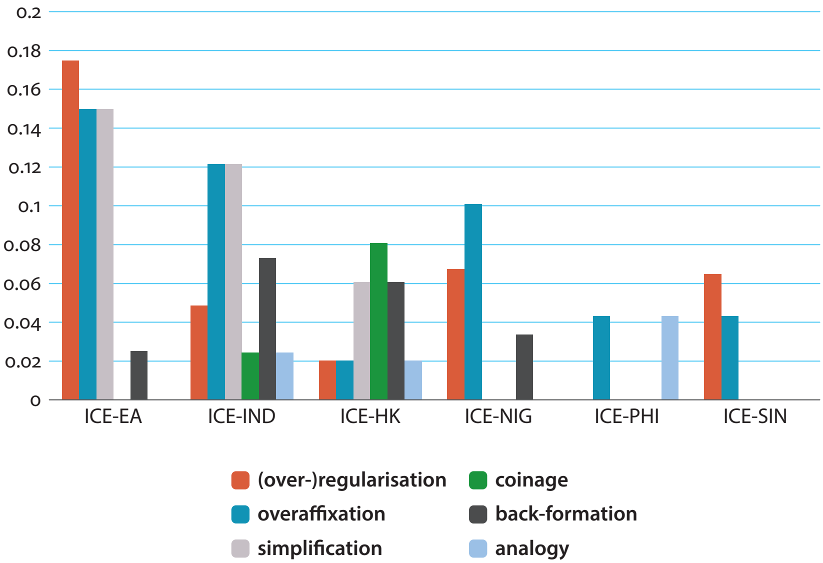


Figure 3. Distribution of six cognitive processes across six ESL varieties as represented in the ICE subcorpora (frequencies per thousand words)

In support of these findings it can be noted that similar observations have been made for users of English as a Lingua Franca (ELF). For example, Pitzl et al. (2008) carried out a study of lexical innovations in a 250,000 word spoken subcorpus of the *Vienna-Oxford International Corpus of English* (VOICE; VOICE 2009), a general corpus of naturally-occurring ELF interactions. They report structures and identify word-formation patterns (as well as their underlying cognitive processes serving to increase clarity and explicitness) as those suggested in the present paper,[16] see (36)–(39).

(36) back-formation: ***pronunciate*** [*pronounce*; from *pronunciation*], ***devaluated*** [*devalue;* from *devaluation*], ***examinates*** [*examine*; from *examination*], ***fragmentated*** [*fragmented*; from *fragmentation*]

(37) (over-)regularisation: ***unformal*** [*informal*], ***characteristical*** [*characteristic*], ***linguistical*** [*linguistic*]

(38) overaffixation: ***increasement*** [*increase*], ***supportancy*** [*support*], ***creativitly*** [*creatively*], ***innovatiations*** [*innovations*], ***controversity*** [*controversy*], ***opportunality*** [opportunity], ***pragmatistic*** [*pragmatic*]

(39) simplification: ***manufacters*** [*manufacturers*], ***contination*** [*continuation*], ***diversication*** [*diversification*]

[16] See Schneider (2012: 73–74) for further parallels between ESL, EFL and ELF.

More generally, the findings of the present study also confirm several previous observations that cognitively motivated processes functioning to maximise transparency and increase explicitness are at play in both ESL and EFL varieties. Generally speaking, Szmrecsanyi (2009: 331) in a study on grammatical analyticity versus syntheticity in varieties of English emphasises that "high-contact speaker communities put a premium on explicitness and transparency". Laporte (2012) studied complementation patterns with causative *make* and found *to*-infinitive complements of the type *making readers* ***to*** *laugh* (what Mesthrie (2006) refers to as "antideletion" of infinitive markers) that are not attested in Standard English but that could frequently be observed in some EFL and ESL varieties (more so in ESL varieties). In these cases, the infinitive marker is preserved to explicitly mark the clausal relationship between the verb and the complement clause. Also in the field of complementation, Callies (2008) found that English raising constructions of the type *We expect them to come back soon* and *This paper is difficult to read* are significantly underrepresented in EFL writing because of their high degree of typological markedness, their functional and semantic complexity, and comparatively little transparency and explicitness in terms of form-function relations. In general, it emerges that even fairly advanced EFL learners avoid loose-fit, i.e. less explicit and semantically opaque constructions even when they do exist in the L1. For ESL, Steger & Schneider (2012) make similar observations on the use of overt complementisers and finite complementation instead of non-finite or even raised complementation patterns because the former display a higher degree of isomorphism (*I want him to do that* vs. *I want that he should do that*). They also find instances of self-monitoring and self-correction, and even the avoidance of non-finite patterns and raised structures, which they interpret as evidence for users' insecurity in the choice of complementation pattern.

In the field of lexico-grammar, both Nesselhauf (2009) and Gilquin (2015) report instances of phrasal/prepositional verbs that include a semantically redundant particle used to make the direction that is implicitly expressed in motion verbs more explicit, e.g. *enter into*, *return back*, *approach to* (Nesselhauf 2009: 20), and *surface up*, *complete up*, *rise up* (Gilquin 2015: 105), a process that Mesthrie (2006) refers to as "insertion of redundant forms". In addition, the studies by Nesselhauf (2009) and Gilquin (2015) suggest that there is evidence for the influence of analogy on the creation of innovations. Nesselhauf (2009) observes that the complementation pattern *have* + *intention to* V frequently found in ESL and EFL instead of the standard *have* + *intention of* V*ing* is most likely based on analogy to the complementation patterns shown by the related verb (i.e. *intend to* V) and the noun (*intention to* V). She also argues that in some of the innovative prepositional verbs found in her data the preposition is used in the meaning assigned to it in similar constructions, i.e. semantically similar verbs, for example the use of *discuss about* in analogy to *talk about* and *speak about* (Nesselhauf 2009: 19–20.). Similarly, Gilquin (2015: 106) reports innovative phrasal verbs not listed in dictionaries, e.g. *cope up with*, which she interprets as possibly formed in analogy to phrasal-prepositional verbs like *come up with*, *meet up with* and *put up with*.

4. Conclusion and outlook

This paper has presented a corpus-study of lexical innovations in derivational morphology to compare EFL and ESL varieties by adopting a process-oriented approach. The study was motivated by the sparsity of research on word-formation in both ESL and EFL and the assumption that despite some fundamental differences between the two types of varieties, they are partially driven by similar production principles and cognitive processes of language acquisition and use.

The data have shown that interlingual, L1-based innovations resulting from cross-linguistic influence are found in EFL varieties and that there are clear L1-effects. L2-based innovations show strong parallels between EFL and ESL varieties, in particular as to three processes: (over-)regularisation, overaffixation and backformation. These have been interpreted as driven by the need to create or maximise morphological transparency and increase the explicitness of form-meaning relations. In the discussion, further parallels were drawn to similar findings for ELF users. Furthermore, it was possible to link the outcomes of the present study to previous studies comparing EFL and ESL in the field of verb complementation and lexico-grammar.

Admittedly, on account of the comparatively small database and the infrequency of the innovations examined here, the findings of the present study are preliminary and await confirmation on the basis of a much larger database. Still, it appears that there is now increasing evidence for the view that cognitively motivated processes to maximise transparency and explicitness are at play in EFL and ESL varieties. By contrast, it seems that, at least for the domains of language use studied so far, subtraction of form seems to be dispreferred ("antideletion").

Finally, and despite the very similar underlying processes that give rise to the innovative forms discussed in this paper, it is still unlikely that they will receive similar recognition as innovations in the two types of varieties. Bamgbose (1998: 3) proposed five interrelated measures that decide on the status of an innovation and if it will spread and eventually become institutionalised: 1) demographic (How many acrolectal speakers use it?), 2) geographical (How widely has it spread?), 3) authoritative (What is the social status of those who use it?), 4) codification (Where is the usage sanctioned?), and 5) acceptability (What are the attitudes of users and non-users towards this usage?).

In view of these measures, innovations are unlikely to gain acceptance in EFL settings for three main reasons. First, forms not codified in reference materials and textbooks, even if well-formed and conceptually possible, are regularly sanctioned as deviations and errors by teachers in educational settings due to the pervasive exonormative orientation adopted in foreign language teaching. Second, there is a tendency of many EFL learners to respond to this exonormative orientation and aim for an idealised native-speaker norm as this often carries the highest prestige. The third reason is closely related to the first: in ESL settings the social context of use provides ample opportunities for an innovation to spread, catch on and eventually become conventionalised. In EFL contexts, however, the opportunities for communicative situations between speakers to arise are limited outside of educational settings. Such opportunities are more easily created in Internet-based forms of communication which are discussed by Li (2010: 627ff.) who suggests a sixth factor in addition to Bamgbose's (1998) five measures: the popular choice of acrolectal English-L2 users in cyberspace. Li argues that because political and geographical boundaries are actually rendered obsolete on the Internet " 'geography' and 'demography' as measures of English users' perception of the correctness of a local usage have become comparatively less significant", a development which has "considerable impact on our perceptions of what counts as an error (i.e. the form is an unintended violation of some Standard English norm), as opposed to a linguistic innovation (i.e. the form is intended as a carrier of a new, probably culture-specific meaning with a local or glocal character)" (Li 2010: 628). This then provides a promising direction for future research on innovations in ESL and EFL varieties on the basis of already existing (e.g. the GloWbE) or yet to be compiled (for EFL) databases of English used on the Internet.

Acknowledgements

I would like to thank the guest editors of this volume and the two anonymous reviewers for helpful comments on an earlier version of the paper.

References

Baayen, H. & Schreuder, R. (Eds.). 2003. *Morphological Structure in Language Processing*. Berlin & New York: Mouton de Gruyter. https://doi.org/10.1515/9783110910186

Bamgbose, A. 1998. "Torn between the norms: Innovations in World Englishes", *World Englishes* 17(1), 1–14. https://doi.org/10.1111/1467-971X.00078

Balteiro, I. 2011. "Awareness of L1 and L2 word-formation. Mechanisms for the development of a more autonomous L2 learner", *Porta Linguarum* 15, 25–34.

Baumgardner, R.J. 1998. “Word-formation in Pakistani English”, *English World-Wide* 19(2), 205–246. https://doi.org/10.1075/eww.19.2.04bau

Biermeier, T. 2008. *Word-Formation in New Englishes: A Corpus-based Analysis*. Berlin: Lit Verlag.

Biermeier, T. 2009. “Word-formation in New Englishes. Properties and trends”. In T. Hoffmann & L. Siebers (Eds.), *World Englishes – Problems, Properties and Prospects: Selected papers from the 13th IAWE conference*. Amsterdam: John Benjamins, 331–349. https://doi.org/10.1075/veaw.g40.20bie

Biermeier, T. 2014. “Compounding and suffixation in World Englishes”. In S. Buschfeld, T. Hoffmann, M. Huber & A. Kautzsch (Eds.), *The Evolution of Englishes: The Dynamic Model and Beyond*. Amsterdam: John Benjamins, 312–330.

Callies, M. 2008. “Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety”. In G. Gilquin, M.B. Diez-Bedmar & S. Papp (Eds.), *Linking Contrastive and Learner Corpus Research*. Amsterdam: Rodopi, 201–226.

Callies, M. 2015. “Effects of cross-linguistic influence in word formation. A comparative learner-corpus study of advanced interlanguage production”. In H. Peukert (Ed.), *Transfer Effects in Multilingual Language Development*. Amsterdam: John Benjamins, 127–143.

Callies, M. & Szczesniak, K. 2007. Investigating productive word formation in advanced L2 acquisition. The potential of learner corpora. Paper presented at the 19th International Conference on Foreign and Second Language Acquisition, 16–19 May 2007, Szczyrk/Poland, (Available at http://www-user.uni-bremen.de/~callies/talks/Szczyrk2007.pdf).

Davies, M. 2004-. BYU-BNC. Based on the *British National Corpus* from Oxford University Press. Available at http://corpus.byu.edu/bnc/.

Davies, M. 2008-. The *Corpus of Contemporary American English*: 450 million words, 1990-present. Available at http://corpus.byu.edu/coca/.

Davies, M. 2013-. *Corpus of Global Web-Based English*: 1.9 billion words from speakers in 20 countries. Available at http://corpus.byu.edu/glowbe/.

Davies, M. & Fuchs, R. 2015. “Expanding horizons in the study of World Englishes with the 1*.9 billion word Global Web-based English Corpus (GloWbE) ”, English World-Wide* 36(1), 1–28.

Ellis, R. 2008. *The Study of Second Language Acquisition*. (2nd ed.). Cambridge: Cambridge University Press.

Flowerdew, J. 2006. “Use of signalling nouns in a learner corpus”, *International Journal of Corpus Linguistics* 11(3), 85–102.

Gilquin, G. 2015. “At the interface of contact linguistics and second language acquisition research. *New Englishes and Learner Englishes compared”, English World-Wide* 36(1), 90–123.

Görlach, M. 1989. “Word-formation and the ENL: ESL: EFL distinction”, *English World-Wide* 10(2), 279–313. https://doi.org/10.1075/eww.10.2.04gor

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *The International Corpus of Learner English. Version 2 (Handbook + CD-ROM)*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Greenbaum, S. (Ed.). 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Hundt, M. & Mukherjee, J. 2011. “Introduction: Bridging a paradigm gap”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 1–5. https://doi.org/10.1075/scl.44.01muk

Ishikawa, S. 2013. "The ICNALE and sophisticated contrastive interlanguage analysis of Asian Learners of English". In S. Ishikawa (Ed.), *Learner Corpus Studies in Asia and the World*, Vol. 1. Kobe: Kobe University Press, 91–118.

Jarvis, S. & Pavlenko, A. 2008. *Crosslinguistic Influence in Language and Cognition*. London: Routledge.

Kellerman, E. 1983. "Now you see it, now you don't". In S.M. Gass & L. Selinker (Eds.), *Language Transfer in Language Learning*. Rowley: Newbury House, 112–134.

Laporte, S. 2012. "Mind the gap! Bridge between World Englishes and Learner Englishes in the making", *English Text Construction* 5(2), 265–292. https://doi.org/10.1075/etc.5.2.05lap

Li, D.C.S. 2010. "When does an unconventional form become an innovation?" In A. Kirkpatrick (Ed.), *The Routledge Handbook of World Englishes*. London and New York: Routledge, 617–633.

Mesthrie, R. 2006. "Anti-deletions in an L2 grammar: A study of Black South African English mesolect", *English World-Wide* 27(2), 111–145. https://doi.org/10.1075/eww.27.2.02mes

Nesselhauf, N. 2009. "Co-selection phenomena across New Englishes", *English World-Wide* 30(1), 1–26. https://doi.org/10.1075/eww.30.1.02nes

Odlin, T. 1989. *Language Transfer. Cross-Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524537

Oxford English Dictionary (OED). 2015. *Online version*. Oxford: Oxford University Press. Available at http://www.oed.com.

Pitzl, M.-L., Breiteneder, A. & Klimpfinger, T. 2008. "A world of words: processes of lexical innovation in VOICE", *Vienna English Working Papers* 17(2), 21–46.

Schneider, E.W. 2003. "The dynamics of new Englishes: From identity construction to dialect birth", *Language* 79(2), 233–281. https://doi.org/10.1353/lan.2003.0136

Schneider, E.W. 2012. "Exploring the interface between World Englishes and Second Language Acquisition – and implications for English as a Lingua Franca", *Journal of English as a Lingua Franca* 1(1), 57–91. https://doi.org/10.1515/jelf-2012-0004

Sharma, D. 2012. "Second language varieties of English". In T. Nevalainen & E. Traugott (Eds.), *The Oxford Handbook of the History of English*. Oxford: OUP, 582–591.

Sridhar, K.K. & Sridhar, S.N. 1986. "Bridging the paradigm gap: Second language acquisition theory and indigenized varieties of English", *World Englishes* 5(1), 3–14. https://doi.org/10.1111/j.1467-971X.1986.tb00636.x

Steger, M. & Schneider, E.W. 2012. "Complexity as a function of iconicity: The case of complement clause constructions in New Englishes". In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: de Gruyter, 156–191.

Szmrecsanyi, B. 2009. "Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English", *Language Variation and Change* 21(3), 319–353. https://doi.org/10.1017/S0954394509990123

VOICE. 2009. *The Vienna-Oxford International Corpus of English* (Version 1.0 online). Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Marie-Luise Pitzl. Available at http://voice.univie.ac.at.

Williams, J. 1987. "Non-native varieties of English: A special case of language acquisition", *English World-Wide* 8(2), 161–199. https://doi.org/10.1075/eww.8.2.02wil

In case of innovation

Academic phraseology in the Three Circles

Alison Edwards and Rutger-Jan Lange
Leiden University / Erasmus University Rotterdam

This paper addresses the equivalence often drawn between labels such as *ESL*, *New Englishes* and *Outer Circle* on the one hand, and between *EFL*, *Learner Englishes* and *Expanding Circle* on the other. It argues that this mapping takes insufficient account of both intra-varietal variation and inter-varietal similarities. We compare the two non-native varietal types with each other and with native English on the basis of 'user' data from the International Corpus of English and the Corpus of Dutch English, focusing on three-word clusters in academic writing. Quantitative analyses reveal no clear grouping per circle, but rather a regional East Africa grouping. Case studies of four specific clusters (*in case of*, *due to the*, *the fact that* and *the other hand*) mostly show a native/non-native divide. Characteristics of both ESL and EFL, including innovative processes as well as learner strategies, are shown to be at play in the Outer and Expanding Circle alike. The findings are consistent with the notion of neither a strict divide between varietal types, nor a continuum.

Keywords: Learner Englishes, New Englishes, Outer Circle, Expanding Circle, non-native innovation

1. Introduction

The workshop on linguistic innovations in non-native Englishes at ICAME 36, on which this volume is based, raised questions about the distinction between error and innovation in World Englishes (WEs); in particular, "are EFL users doomed to be mistaken rather than creative?" (Deshors et al. 2015). To address this we must first define (i) what we mean by error and innovation and (ii) who we understand to be EFL users. With respect to (i), in this context an error is seen as an individual/idiosyncratic 'mistake' (a learner feature), whereas an innovation can be linked to structural nativisation, i.e. linguistic adaptation resulting in the development of

https://doi.org/10.1075/bct.98.06edw

local linguistic patterns (stable features) (Kachru 1982: 62; Schneider 2007: 5–6). By implication, therefore, in this paper we are interested in those innovations that appear to have passed through the first stage of Croft's (2000: 3–5) model of language change (*actuation*, i.e. the initial entrance of a marked feature into the pool of available variants)[1] and that are well into the second stage (*propagation*, i.e. frequent selection leading to conventionalisation). Such structural nativisation encompasses both "entirely new […] forms and structures" and "quantitative differences between varieties of English in the use of forms and structures […] shared by all Englishes" (Mukherjee & Gries 2009: 28).

Turning to (ii), EFL users are, in much WEs literature, often equated with the speech communities of the Expanding Circle in Kachru's (1985) Three Circles model. In this model, the Inner Circle countries include the UK, the US and those former settler colonies where English is the dominant first language used in all societal domains (e.g. Australia); the Outer Circle encompasses postcolonial societies such as India and Singapore; and the Expanding Circle refers to countries where English has traditionally been taught and used only for purposes such as trade and international communication (e.g. Brazil, Japan, Russia). Frequently mapped onto these three circles is the classification into English as a native (ENL), second (ESL) and foreign (EFL) language (Quirk et al. 1972: 3–4), respectively. This has implications for the capacity for structural nativisation accorded to different speech communities. While marked features arise in all language varieties, in WEs their propagation/diffusion has been specifically linked to the Outer Circle, whereby it has been asserted that conventionalisation

> is much more likely to occur in New Englishes in the Outer Circle than in Foreign Language English contexts in the Expanding Circle. […] On average, the contexts in which New Englishes arise provide much more regular opportunity to use the language [and] this may in part give more individual opportunity for entrenching certain forms. (Van Rooy 2011: 193–4)

The implication is that if the Outer Circle is where innovation and nativisation by ESL users lead to the emergence of dynamic New Englishes, the Expanding Circle is home to the performance varieties of EFL learners (Learner Englishes) where deviations from imposed external norms are to be analysed as errors in

[1] Croft (2000: 4) refers to this first stage as "innovation or actuation". We have opted for *actuation* so as to avoid potential confusion with the somewhat different sense of innovation referred to in the previous sentence of this paper. It should be noted that linking the term *innovation* to structural nativisation is a more demanding definition than that implied by Croft's first stage — indeed, it associates innovation instead with Croft's second stage — yet this is what seems to be implied when WEs scholars refer to the dichotomy between error and innovation (see e.g. Bamgbose 1998: 2; Kachru 1982: 62).

traditional second-language acquisition (SLA). However, despite the "popular but incorrect reduction" of the Three Circles to ENL–ESL–EFL that persists in much WEs literature (Hilgendorf 2015), many scholars point out that this mapping is not clear cut. A number of countries have, despite their non-postcolonial and thus Expanding Circle status, frequently been reported as transitioning from EFL- to ESL-using societies (such as the Netherlands and the Scandinavian countries, e.g. Berns 1995: 8; Graddol 1997: 11; see also Edwards 2016 for in-depth discussion on the status of English in the Netherlands). Seeing that, as Van Rooy's (2011) quote above suggests, it is regular use of the language rather than Outer Circle/postcolonial status *per se* that underpins change, there appears to be no compelling reason why, given frequent enough use, structural nativisation could not also occur in the Expanding Circle.

That stable linguistic innovations may arise in the Expanding Circle is acknowledged in the English as a Lingua Franca (ELF) paradigm, a separate but adjacent scholarly field to WEs. In ELF, ESL users and EFL learners are defined, regardless of their location in the Three Circles, depending on context. That is, a distinction is made between language acquisition and language use, and the respective learner and user identities are associated with whether one is in the language-learning classroom or outside it (Mauranen 2011: 157–159, 2012: 4–7). The site of innovation is, rather than being linked to the postcolonial Outer Circle, located at the level of discourse community, e.g. academia, where "spontaneous norms" arise that can stabilise and spread beyond their own borders (Mauranen 2012: 6). Working from this assumption, this paper explores the notion of innovation in academic discourse at the phraseological level. We are interested in the characteristics ascribed to ESL and EFL, how these characteristics map across the Outer and Expanding Circles, and how they relate to the notion of error versus innovation.

2. The paradigm gap and phraseological research

A number of recent studies in WEs have attempted to bridge the paradigm gap (Sridhar & Sridhar 1986) between non-native varietal types by comparing corpora from both the Outer and Expanding Circles, broadly arriving at one of two conclusions (Gries & Deshors 2015: 154). Some emphasise parallels across varietal types, identifying shared innovations and similar cognitively motivated processes in various areas of lexicogrammar and phraseology (e.g. Biewer 2011; Callies 2015; Edwards 2014; Edwards & Laporte 2015; Gilquin 2015; Nesselhauf 2009) and suggesting that individual varieties can be placed at different points on an ESL–EFL continuum. Others report markedly different tendencies across varietal types (e.g. Gries & Deshors 2015; Hundt & Vogel 2011; Szmrecsanyi & Kortmann 2011; Van Rooy 2006), implying that the conceptual and terminological divide between them should be upheld.

For instance, a number of studies link structural nativisation exclusively with Outer Circle varieties; consider the "stretched" use of the progressive aspect reported for stative contexts in e.g. Singapore English (Hundt & Vogel 2011) and persistitive contexts in Black South African English (Van Rooy 2006). Other studies do not focus on innovations *per se*, but emphasise the different quantitative tendencies of the two non-native varietal types. These include studies closely connected to the kind of phraseological research we focus on here. Learner corpus research in phraseology has consistently linked more advanced non-native proficiency to more native-like frequency and usage, and lower proficiency to a more restricted phraseological repertoire and more errors (Oksefjell Ebeling & Hasselgard 2015: 216). In this context, Ellis et al. (2015: 373) draw attention to the different non-native acquisition settings, pointing out that the amount and type of language exposure (i.e. ESL vs EFL) plays an important role;[2] thus "there is a need for more studies […] which compare the phrasicon in foreign and second language varieties of English" (Paquot & Granger 2012: 143). Two such comparative studies are of particular relevance here, broadly finding that their Outer Circle data are characterised by a high number of cluster types (creative licence) while Expanding Circle data are associated with a narrower repertoire and over-reliance on known clusters (a learner strategy).

First, Götz & Schilk (2011) compared (i) data from the German component of the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010) with the comparable *Louvain Corpus of Native English Conversations* (LOCNEC), and (ii) several spoken sections of the Indian component of the *International Corpus of English* (ICE; Greenbaum 1991) with the comparable British ICE sections. They found that the German speakers used fewer three-word clusters in terms of both types and tokens than the British reference corpus, whereas Indian speakers showed the opposite trend. Next, Gilquin (2015) took the innovative approach of aggregating multiple corpora per circle, contrasting Learner Englishes and New Englishes — on the basis of components of the International Corpus of Learner English (ICLE; Granger et al. 2009) representing 16 different first-language backgrounds and the student writing sections of six Outer Circle components of ICE, respectively — with the academic section of the

[2] It should be noted, as an anonymous reviewer pointed out, that Ellis et al. (2015) were referring to ESL in a traditional SLA context, i.e. in the sense of immigrants acquiring English in an Inner Circle country, as opposed to in a WEs context, i.e. for local use in an Outer Circle country.

British National Corpus (BNC) Baby edition. She found that the Learner Englishes underused more types of three-word clusters compared to the British data than did the New Englishes, and overused certain clusters such as teaching-induced expressions (e.g. *in order to*, *first of all*). She attributed this restricted range of expression to the Expanding Circle acquisitional environment, with less exposure to English in wider society and proportionally more classroom drilling than in Outer Circle settings.

3. Aims and hypotheses

The studies described above clearly demonstrate that the investigation of lexical bundles is an intriguing and fruitful way of exploring labels such as ESL and EFL across circles and the implications of these labels for the capacity of the respective speaker populations to effect structural innovation. We seek to build on Gilquin (2015) in particular in two ways. As a preliminary case study of three-word clusters across varietal types, Gilquin's (2015) investigation did not allow for in-depth consideration of inter-corpus variation within varietal types, which may have implications for the respective characteristics of English in the Outer and Expanding Circles, or for individual variation within corpora, which, as we shall see, can have implications for the distinction between error and innovation (G. Gilquin, personal communication).[3] Further, like most studies comparing directly across circles to date, it is (unavoidably) restricted to student writing. Such data are undoubtedly interesting, as the academic phrasicon is an area of language that needs to be purposefully acquired by native speakers (NSs) and non-native speakers (NNSs) alike. Moreover, until recently student writing was the only available data allowing for comparison across all three varietal types. In WEs, while broad-based, comparable corpora encompassing multiple genres and registers have long been available for the Inner and Outer Circles (e.g. ICE), data from the Expanding Circle has largely been limited to learner writing (e.g. ICLE), which reinforces the equivalence frequently drawn between the labels *EFL*, *Learner Englishes* and *Expanding Circle*. Therefore, we take advantage of a new, multi-genre Expanding Circle corpus of English in the Netherlands that we hope will, in combination with comparable data from a range of Inner and Outer Circle ICE corpora, help us to explore whether the above findings from student writing in the different varietal types also hold in expert academic writing.

[3] It should be noted that Gilquin (2015: 101), despite inter-corpus variation in her results, writes that "on the whole the general tendencies arguably remain valid for a majority of the varieties and speakers considered".

As is common in corpus-based phraseological research, Gilquin (2015) and Götz & Schilk (2011) took a bottom-up approach focusing on recurrent multi-word sequences known as lexical bundles (Biber et al. 1999: 990).[4] We follow those two studies in focusing on three-word clusters, also known as 3-grams, which along with four-word clusters are known to play an important role in academic writing (e.g. Biber & Barbieri 2007; Biber et al. 1999; Hyland 2008; Staples et al. 2013). We aim to explore the mapping of a number of ESL and EFL characteristics in the Outer and Expanding Circle varieties under investigation. Based on the findings of the literature discussed above, 'ESL characteristics' are taken to encompass a wide variety of types of three-word clusters plus evidence of stable innovations, be they "entirely new" clusters or "quantitative differences between varieties of English" in the use of shared clusters (cf. Mukherjee & Gries 2009: 28, see Introduction). By contrast, 'EFL characteristics' can be expected to manifest themselves as a restricted repertoire of cluster types and overuse of common clusters. The hypotheses are as follows:

- H0: Outer and Expanding Circle Englishes alike will display ESL- and EFL-like characteristics, rendering them structurally indistinguishable as regards the use of three-word clusters.
- H1: Outer and Expanding Circle Englishes have distinct structural tendencies in their use of three-word clusters, that is:
- H1(a) The Expanding Circle data will show more EFL-like characteristics (a restricted number of cluster types and high average use of common clusters), whereas
- H1(b) The Outer Circle data will show more ESL-like characteristics (a wider variety of cluster types cf. the Expanding Circle and structural innovation cf. the Inner Circle).

4. Data and methods

This study makes use of the academic writing sections of various national ICE components and the Corpus of Dutch English (NL-CE, Edwards 2016) (Table 1). Each subcorpus contained approximately 80,000 words in four broad areas (humanities, social sciences, natural sciences and technology), with each text being

[4] This is in contrast to a top-down approach working from a selection of predefined phraseological items. Although the bottom-up lexical bundle approach can yield clusters that do not necessarily form coherent semantic units, it nevertheless forms "an excellent starting point" to identify interesting phraseological units that merit further investigation (Oksefjell Ebeling & Hasselgard 2015: 210).

Table 1. Academic subcorpora used in the present study

Variety	Size (words)*	No. authors
Inner Circle		
ICE-GB	84,662	40
ICE-USA	84,055	40
Outer Circle		
ICE-SIN	79,777	40
ICE-IND	81,719	40
ICE-HK	102,969	40
ICE-PHI	88,823	40
ICE-KEN	80,343	40
ICE-TAN	80,890	37†
Expanding Circle		
NL-CE	79,655	40

* Automatic word counts were performed after stripping text files of markup by means of regular expressions (see Edwards 2016: 133). It is unclear why the word count for ICE-HK exceeds the recommended ICE size.

† Due to data scarcity, the Tanzanian component has slightly fewer texts than the other ICE corpora (Hudson-Ettle & Schmied 1999: 9).

approximately 2,000 words in length. To represent the Inner Circle both the British and American components of ICE were used, as in our view the heterogeneity of ENL varieties makes the use of a single NS yardstick inadvisable.[5] Representing the Outer Circle were six components of ICE, chosen for their geographical range — India (South Asia), Singapore, Hong Kong and the Philippines (Southeast Asia) and Kenya and Tanzania (East Africa) — and for maximum comparability with previous studies (e.g. Gilquin 2015). For the Expanding Circle, the NL-CE was used. This corpus has conceptual similarities with ELF corpora (e.g. ELFA 2008; VOICE 2013) in that it is explicitly profiled as representing Expanding Circle 'users' (as opposed to learners) of English. However, it can be regarded as falling under the umbrella of WEs in that its contributors (unlike those to the ELF corpora) share the same mother tongue. Moreover, it is based on the design of the written components of ICE (see further Edwards 2016: 114-125), and is thus ideally suited

[5] ICE-CAN and ICE-NZ were initially included as well, but as they did not yield insight that could not already be gained from ICE-GB and ICE-USA, the results are not reported here.

for comparative studies of WEs.[6] The academic writing section of the NL-CE comprises extracts from journal articles and book chapters written by Dutch academics at all career stages and not edited or translated by NSs. Naturally, it would have been desirable to include more Expanding Circle countries, but the lack of comparable data at this time means this limitation is unavoidable.

The WordList function in Wordsmith Tools (version 6; Scott 2015) was used to identify all three-word clusters in the corpora as a whole. We focused on those with a minimum of 50 occurrences (n = 201), ensuring that each cluster appeared in a minimum of five different texts in at least one corpus.[7] As the focus of this analysis is on general phraseology, topic-dependent clusters were excluded (n = 4).[8] This left 197 clusters with a total of 13,783 occurrences for inclusion in the analyses. Results are reported based on averages per author. We consider this to be a more natural unit for the present study than normalisation by an arbitrary number of words. First, as the corpora are designed such that each text file is approximately the same length, the per-author results broadly equate to normalisation per 2,000 words. The main exception is the somewhat larger ICE-HK, in which each author appears to contribute around one quarter more text. Assuming

[6] It is not an *a priori* given that the ICE model is the most appropriate to represent the uses of English in Expanding Circle countries. Other models are in the process of being developed elsewhere, e.g. for Sweden and Finland loosely based on the Brown model and taking into account the text types actually available on the ground (Laitinen 2011, 2016). For the NL-CE, since all ICE text types turned out to be feasible to collect, the ICE model was followed in order to facilitate comparison (see further Edwards 2016: 114–125).

[7] The rationale behind the minimum frequency and dispersion criteria is as follows. With regard to frequency, a minimum of 50 occurrences equates to 70 occurrences per million words. Although this is a very conservative inclusion threshold compared to the established threshold in the literature of 10 (Biber et al. 1999) or 20 (Hyland 2008; Cortes 2004) occurrences per million words, it breaks down to an average of 5.6 occurrences per subcorpus, which in these relatively compact corpora we considered sufficiently frequent to explore variety-specific usage. With respect to dispersion, to guard against individual idiosyncrasy, the established dispersion criterion in previous literature is three to five texts (Chen & Baker 2010; Biber et al. 1999) or 10% of the number of corpus files (Hyland 2008; Pérez-Llantada 2014). We decided to use the relatively high cut-off of five texts, but in at least one corpus as opposed to in all corpora so as not to exclude clusters that may be widespread in some corpora but not in others. For instance, it seems worthy of note that the cluster *a lot of* (n = 53) appears in five or more files in three of the NNS corpora (ICE-KEN, ICE-TAN and the NL-CE) but not in the NS corpora; this may point to issues of register variation that are worth exploring.

[8] The excluded topic-dependent clusters were *in Hong Kong*, *the Hong Kong*, *in New Zealand* and *Dar es Salaam*. *The United States* was not excluded as its 78 hits were not exclusive to ICE-USA, but spread across all corpora, whereas for the other topic-dependent clusters at least 97% of the occurrences were in one corpus only.

that academic writers choose their phrasing carefully, however, a longer text may not necessarily result in a proportional increase in number of word clusters (and indeed this potential issue concerning ICE-HK turns out to have no bearing on our results).[9] Second, probing the data at the individual level allows us to explore the effect of intra-corpus variation in a more in-depth way than is usual. As our results will show, this helps us to gain insight into the notion of stable varietal features versus learner characteristics. Statistical analyses were performed using one-way analysis of variance (ANOVA) with the Scheffé post-hoc test[10] to identify pairwise frequency differences between corpora; F-tests for the (in)equality of variances to identify differences in intra-corpus variation; and chi-square tests[11] for distributional differences at the level of specific clusters. Results were considered significant at the $p < 0.05$ level.

5. Results

Table 2 lists the top 10 most frequent three-word clusters per corpus, showing a number of interesting differences as well as considerable overlap (indeed, of the top 197 clusters, 79% appeared in all corpora). The table also indicates the raw number of three-word clusters in each corpus, ranging from 1,089 in ICE-USA to over 2,000 in ICE-KEN and ICE-TAN. To explore frequency differences across corpora, Figure 1 shows the average frequency of use per author of the top 197 clusters under investigation and associated 95% confidence intervals. One-way ANOVA revealed a highly significant difference between corpora ($F(8,348) = 14.8$, $p = 1.1\text{e-}16$), with Scheffé post-hoc tests confirming that the ICE-KEN and ICE-TAN authors used significantly more three-word clusters than those in all other corpora.[12] In short, no neat grouping emerges with respect to varietal type; with

[9] The potential inflation in cluster tokens due to this longer text length may explain why ICE-HK is the only corpus whose average frequencies are not significantly lower than those of ICE-TAN in Figures 1 and 2; however, this one pairwise difference does not detract from the main points of our results or conclusions.

[10] We ran both Scheffé and Tukey post-hoc tests (cf. Muñoz 2000: 175) but report only the Scheffé results; this test is more conservative and thus reduces the chance of false positives. The Tukey test did not affect the main thrust of our results; in almost all cases it merely increased the reported significance levels.

[11] In one case Fisher's exact test was substituted due to low expected frequencies. This is indicated in the relevant place in the text.

[12] All pairwise differences were at the $p < .01$ level. The exception was ICE-TAN vs ICE-HK, which were not significantly different.

Table 2. Total number and most frequent three-word clusters per corpus

	ICE-GB	ICE-USA	ICE-IND	ICE-SIN	ICE-HK	ICE-PHI	ICE-KEN	ICE-TAN	NL-CE
	N = 1431	N = 1089	N = 1369	N = 1311	N = 1641	N = 1354	N = 2316	N = 2004	N = 1268
1	per cent of	*as well as*	the case of	*in order to*	*one of the*	*the use of*	most of the	*as well as*	*as well as*
2	a number of	the number of	in the case	*as well as*	*as well as*	*as well as*	per cent of	*in order to*	*in order to*
3	*in terms of*	some of the	*as well as*	*the use of*	*the use of*	*in terms of*	*as well as*	*on the other*	due to the
4	there is no	*in order to*	*on the other*	*on the other*	*in order to*	*on the other*	the fact that	the other hand	based on the
5	it is not	end of the	*one of the*	based on the	the development of	the number of	some of the	in the country	*one of the*
6	*as well as*	part of the	part of the	there is no	*on the other*	*one of the*	*one of the*	*the use of*	part of the
7	*one of the*	the end of	*in order to*	*in terms of*	there is a	the other hand	the majority of	the development of	the number of
8	some of the	analysis of the	is to be	the number of	*in terms of*	it is the	*the use of*	*in terms of*	*the use of*
9	*in order to*	be able to	most of the	the other hand	the other hand	a number of	as a result	the fact that	according to the
10	*the use of*	most of the	in the present	there is a	as a result	based on the	a number of	the process of	based on a

Note: clusters that appear in the top 10 of over half of the corpora (5/9) are italicised.

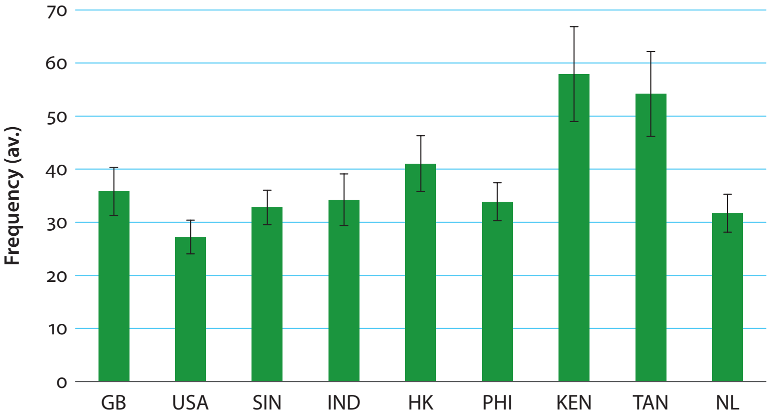


Figure 1. Average frequency of use of three-word clusters per author with 95% confidence intervals

the exception of the two East Africa corpora there are no statistically significant differences across the Inner, Expanding and remaining Outer Circle corpora.

The frequencies presented in Figure 1 are a product of the number of types of three-word clusters used per author (i.e. unique clusters) and the number of times they use those clusters (i.e. repetition per cluster). These two factors are separated out in Figures 2 and 3, with boxplots indicating the full range, interquartile range and median per corpus.

Figure 2 presents the distribution of the number of types of three-word clusters used by each author per corpus. This distribution closely resembles that in the previous figure, indicating that the variation in overall frequency across corpora is attributable largely to the breadth of writers' repertoires (and not, as we shall see below, to the number of times they use each cluster). ICE-KEN and ICE-TAN have the highest number of cluster types per author, with a mean of over 35 each; around one third more than the Inner Circle corpora. One-way ANOVA showed that the difference across corpora was highly significant ($F(8,348) = 17.6$, $p = 1.1\text{e-}16$), with post-hoc tests confirming that ICE-KEN and ICE-TAN were significantly different to all other corpora.[13] Hence, the two East African corpora, but not the other Outer Circle corpora, show the predicted wider variety of clusters than the Expanding Circle corpus, the NL-CE. The NL-CE does not show the hypothesised restricted number of types but instead falls in between the two native corpora. Interestingly, pairwise F tests showed that ICE-KEN and ICE-TAN also have significantly higher intra-corpus (i.e. individual) variation in terms of the number of types used per author than all other corpora (except ICE-GB). This seems to

[13] All pairwise differences were at the $p < .01$ level. The exception was ICE-HK, which was not significantly different to ICE-TAN but was significantly different to ICE-USA ($p < .01$).

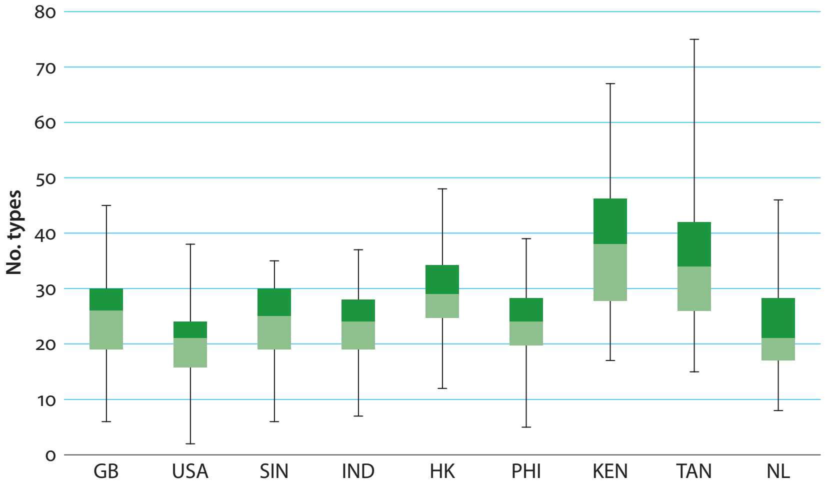


Figure 2. Variation across authors in number of types of three-word clusters used

indicate that the use of a wide variety of clusters is not stable across all authors in these corpora; we will return to this point in Section 6.

Figure 3 turns to the distribution across corpora of the number of times authors use three-word clusters. The information presented here is the same that would be provided by the commonly used type/token ratio (TTR) (it is formally equivalent to 1/TTR), but has the advantage of being intuitively easier to interpret. From this we can confirm that average use per cluster is not the driver behind the different distributions observed in Figure 1. Authors tend to repeat clusters between 1.2 and 1.6 times on average, with a mildly significant difference across

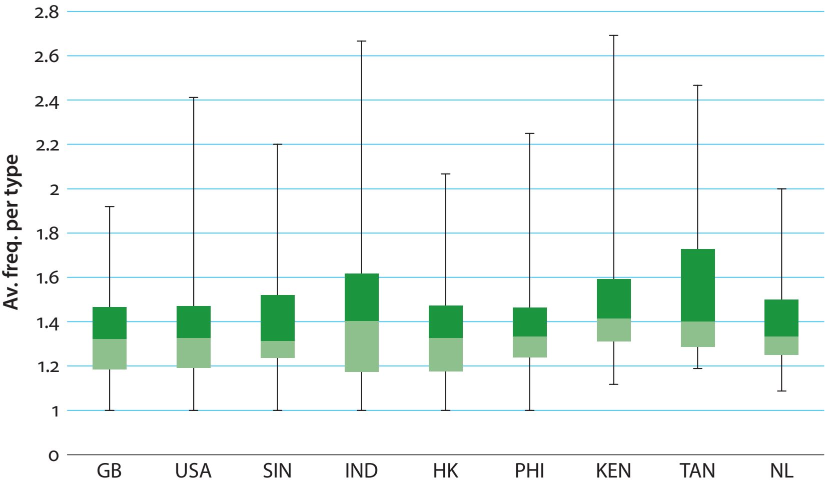


Figure 3. Variation across authors in average frequency of use per three-word cluster

corpora (one-way ANOVA F(8,348) = 2.4, $p = .02$), but no significant pairs according to the post-hoc test. The NL-CE, therefore, does not show the hypothesised overreliance on the clusters used; in fact there is no statistically significant overuse in terms of mean frequency of use per cluster type in any NNS corpus. However, pairwise F tests again reveal significantly higher intra-corpus variation in two Outer Circle corpora, this time ICE-IND and ICE-TAN, compared to all other corpora (except ICE-KEN and ICE-USA).

The above analyses have shown that differences between corpora in the number of types of three-word clusters mainly reflect a regional East Africa grouping (ICE-KEN and ICE-TAN), and differences in terms of average frequency of use are negligible. What such results do not reveal is whether there are differences in *which* specific clusters are popular per variety; we now turn to that here. A chi-square test of the raw frequencies of the most common clusters[14] revealed, unsurprisingly, significant differences across corpora ($\chi^2 = 1054.56$, df = 368, $p < 0.001$, Cramér's V = 0.15), with post-hoc pairwise chi-squares using the Holm correction showing that the only non-significant differences are between ICE-SIN and ICE-USA/ICE-PHI. In other words, distributional differences can be found even between the Inner Circle corpora as well as among and across the Outer and Expanding Circle corpora.

To explore these differences in depth, in the sections below we present a more detailed analysis of a number of specific three-word clusters. As indicated earlier, Table 2 listed the 10 most frequent clusters per corpus. Due to space limitations, we restrict ourselves to reporting on three of those clusters, namely *the fact that*, *the other hand* and *due to the*, which turned out to be particularly striking in terms of quantitative variation, one of the facets of structural nativisation set out by Mukherjee & Gries (2009: 28). In addition, we discuss a fourth cluster, *in case of*, which was too infrequent to occur in Table 2 but which helps to shed further light on the notion of innovation.[15]

[14] Only the 50 most frequent clusters were used here in order to ensure that the expected frequencies were all above 5. Of those, three clusters with frequency of zero in one or more corpora were also excluded.

[15] The cluster *in case of* was identified in a pilot analysis run prior to the present study. Following Gilquin (2015), the corpora were aggregated per circle and all three-word clusters with a minimum of three occurrences were extracted using the WordList function in Wordsmith Tools (version 6). Next, these lists were compared using the KeyWords function to identify under- and overused clusters in the Inner vs Outer Circle data and the Inner vs Expanding Circle data. *In case of* was the only topic-independent cluster found to be significantly overused in both the Outer and Expanding Circle cf. the Inner Circle corpora.

i. *the fact that*

The fact that is a common lexical bundle in academic discourse (e.g. Biber & Barbieri 2007). It appears in the present dataset a total of 155 times, with a highly significantly different distribution across corpora ($\chi^2 = 36.47$, $df = 8$, $p < 0.001$, Cramér's V = 0.05).

Figure 4 plots the number of authors per corpus (dispersion) who used *the fact that* against the average frequency of use in each corpus (repetition). The coloured markers indicate the average frequency and the vertical bars indicate the minimum and maximum frequencies per corpus. In terms of dispersion, ICE-KEN and ICE-TAN stand out: *the fact that* is used by considerably more authors than in the other corpora. In ICE-KEN in particular, it is also used more often on average (twice per author, as opposed to just once in e.g. ICE-GB). This higher average, however, can be attributed to repeated use by a select number of authors. In ICE-GB, *the fact that* is used by 12 authors with only one repetition; that is, one author who uses the cluster twice. A similar pattern holds for most of the other corpora. By contrast, of the 24 authors who use *the fact that* in ICE-KEN, half do so multiple times, two of them as many as six times each. This suggests that the quantitative variation (i.e. high average use) of this cluster in ICE-KEN compared to the other corpora is probably less indicative of structural nativisation, i.e. a stable feature of academic writing in Kenyan English, than it is of a sort of phraseological crutch for just some individuals in the absence of other text-structuring devices. An illustration is provided in (1), where *the fact that* appears in four out of five consecutive sentences within a single text.

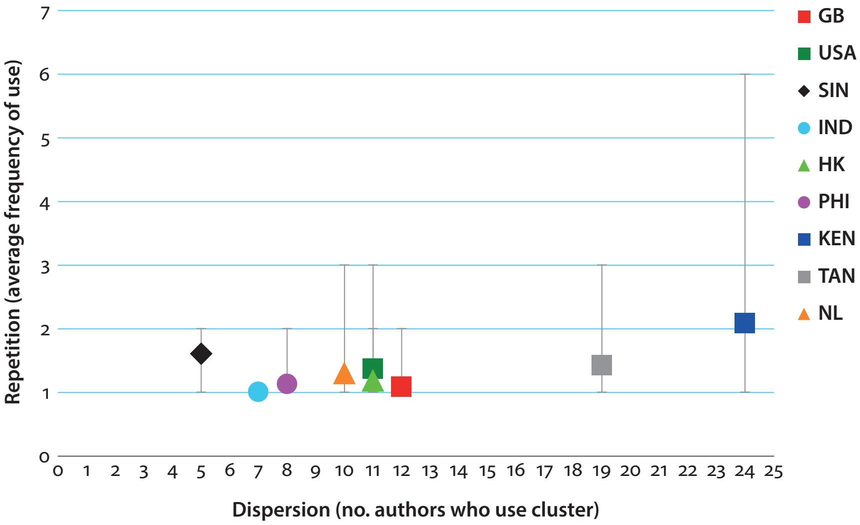


Figure 4. Dispersion and repetition of *the fact that*

(1) Though the proportions may differ, all studies have testified to **the fact that** most abortions occur late. […] This may be supported further by **the fact that** most of the patients reported not having planned or wanted the index pregnancy. If this is the case, it points to **the fact that** induced abortion is perhaps more prevalent that hitherto asserted, and that the actual magnitude may never be known.
While this may be so, we should not also lose sight of **the fact that** a good proportion of abortions […] are being done privately and fairly early […] (ICE-KEN W2A-028)

ii. *the other hand*

On the one hand … on the other (hand) is a common cohesive device in English academic writing, whose various constituent parts appear among the most frequent three-word clusters in five of the seven NNS corpora, although not in the NS corpora (Table 2). The cluster *the other hand* appeared a total of 161 times in the dataset, with a highly significantly different distribution across corpora ($\chi^2 = 29.66$, df = 8, $p < 0.001$, Cramér's V = 0.05).

Figure 5 reveals a clear split whereby *the other hand* shows considerably greater dispersion across authors in the Outer Circle corpora compared to the other corpora.[16] In terms of repetition, while in the NL-CE and the native corpora no individual author uses this cluster more than twice, select authors in ICE-IND and ICE-PHI use it up to four times, and one author in ICE-TAN uses it eight times in

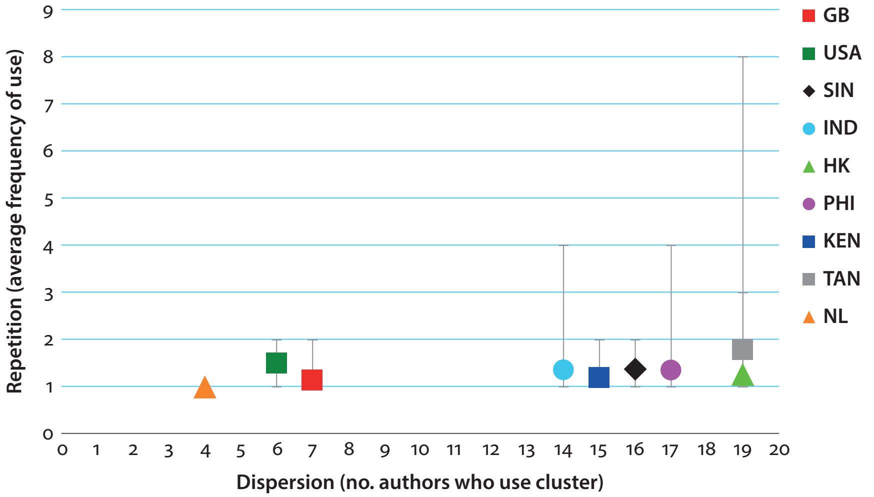


Figure 5. Dispersion and repetition of *the other hand*

[16] This pattern holds even when taking into account the realisation of *on the other* (without *hand*).

a single text (ten times when including *on the other*). Thus, like *the fact that*, high use of *the other hand* seems to be less of a stable feature than a learner strategy, with some individuals relying heavily on this particular cohesive device. At the level of qualitative innovation, in the NNS corpora *the other hand* occasionally appeared with an additional contrast marker such as *yet* or *while* ((2)–(3)), which was not observed at all in the NS corpora. Although low in frequency, such usages may tie in with the reported NNS tendency towards hyper-explicitness in grammatical or logical relations (see e.g. Seidlhofer 2004 on "black colour"; Edwards & Laporte 2015 and Nesselhauf 2009 on the tendency to make directionality expressed in prepositions into more explicit; or Callies 2015 on "hyperclarity" in derivational morphology).

(2) On the one hand it is universal or common to all societies and yet on **the other hand** it is unique to the set of socio-historical circumstances associated with its community of speakers. (ICE-SIN W2A-005)

(3) [T]he educated respondents must have gone out of their cultural and religious beliefs to embrace new values which can help them. While on **the other hand**, those who have a lower level of education, it could be that they have been confined to the rural areas where most adhere to their traditional and religious values […] (ICE-KEN W2A-017)

iii. *due to the*

The cluster *due to the* is subject to variation even in NS English, as evidenced by the amount of discussion dedicated in usage guides and on online grammar forums to its 'correct' usage. For instance, one popular reference source rails against its use in place of *because of* or *owing to* (Kumar 2010: 128), while another suggests that this extension is already well underway:

> *Due to* and *owing to* mean just what *because of* means. […] *Owing to* fought and won its way to respectability a good while ago, and now *due to* has almost won its battle, although there is a residue of conservative unhappiness over it when it does not follow a linking verb, as in *He arrived late, due to a flat tire*. (Wilson 1996: 160)

Given this variation between *due to* and related variants in NS varieties, we may expect to see variable usage among NNSs too. In the present dataset there were 120 occurrences of the lexical bundle *due to the*, with a highly significantly different distribution across corpora ($\chi^2 = 27.94$, df = 8, $p < 0.001$, Cramér's V = 0.05). Figure 6 shows a clear split between the NS and the NNS varieties, with the latter corpora showing much higher dispersion. The highest individual variation could be seen in the NL-CE, with one author repeating *due to the* as many as six times.

Looking at the concordances, the expansion in meaning of *due to* is apparent in all corpora, but more advanced in the NNS corpora. ICE-GB and ICE-USA adhered at least half of the time to the prescribed adjectival usage, "[modifying] a noun or pronoun directly preceding it in the sentence or following a form of the verb *to be*" (Coghill & Garson 2006: 11), e.g. *Cutbacks* ***due to*** *increased funding […]*. By contrast, the NNS varieties adhered to the prescribed use less than half of the time, with ICE-SIN and the NL-CE having the highest proportion of non-standard uses ((4)–(5)). This apparently more flexible conception of the contexts in which *due to* can be used in academic writing may be further accelerated by the fact that *because of* is typically viewed (and indeed often explicitly taught) as less formal than *due to*; consider e.g. Bruckfield's (2012: 281) claim that "*due to* is more formal than its 'cousin' *because of*".

(4) A modular approach has to be adopted to ensure the success of CIM implementation **due to the** complexity in both the technical and organizational aspects. (ICE-SIN W2A-035)

(5) We saw that it does not matter which path is chosen, but choosing the linear path, **due to the** simple integration scheme, is by far the cheapest and easiest choice. (NL W2A-028)

This semantic expansion is apparent in varying degrees in all corpora under investigation. By contrast, certain innovations at the lexicogrammatical level surrounding *due to the* seem to be exclusive to the NNS varieties. In the NS data, all occurrences of *due to the* in conjunction with a modal verb use *may*, whereas in the NNS data other modals, such as *could*, also co-occur (6). Further, in the NNS data modal uses are occasionally combined with an additional hedging word (7),

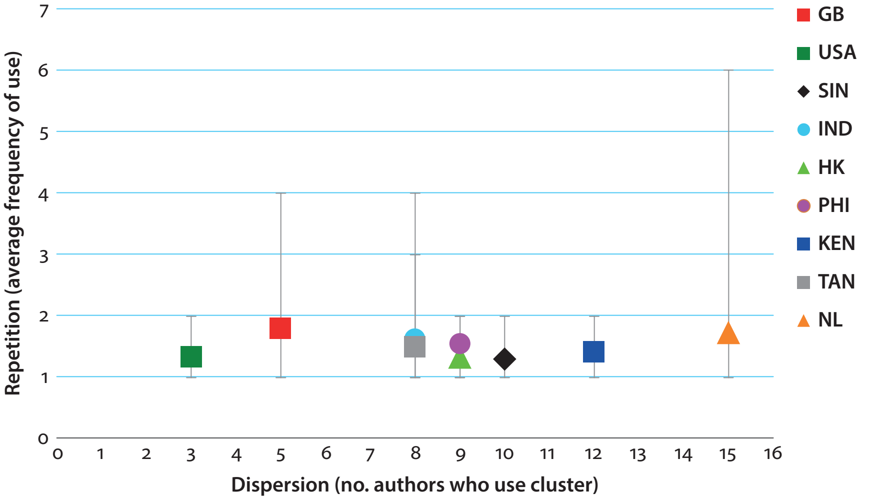


Figure 6. Dispersion and repetition of *due to the*

which does not occur at all in the NS data. Like the additional contrast markers with *the other hand* discussed under (ii) above, this may point to a trend towards hyper-explicitness or redundancy in NNS varieties.

(6) The differences could be **due to the** various sources of plants or extractives, formulations used, and the species and level of pest infestations. (ICE-PHI W2A-028)

(7) These spectral changes may perhaps **be due to** the formation of the higher valent manganese complex (as shown in the scheme in figure (6). (ICE-IND W2A-021)

iv. *in case of*

In case of is a much lower frequency cluster than the others discussed so far, with just 32 occurrences in total with a highly significantly different distribution across corpora (Fisher's exact test $p < 0.001$; Figure 7). However, as one of few clusters that appear numerous times in the NNS corpora but not at all in the NS corpora, it represents an interesting case of potential innovation. The concordances reveal that it is used interchangeably in the NNS corpora with *in the case of*, especially in the NL-CE and ICE-IND[17] ((8)–(9)). Variable article usage has long been attested in New Englishes (e.g. Y. Kachru 2003; Sharma 2005; Wahid 2013) and it may be

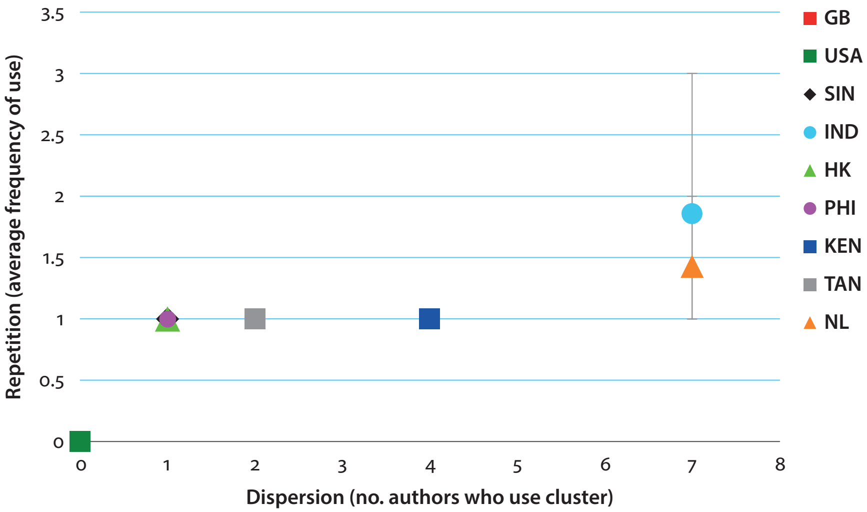


Figure 7. Dispersion and repetition of *in case of*

[17] The relatively high frequency of *in case of* in ICE-IND is all the more striking given that *in the case* and *the case of* are the top two most frequent three-word clusters in ICE-IND (Table 2). While this suggests that *in case of* is not the preferred variant, it is clearly one option in the pool of variants for some ICE-IND authors.

that, here, pressure to omit the article is strengthened by the analogy with highly salient standard English (StdE) expressions such as *in case of emergency* or *in case it rains*; note for instance that in (9) there are up to four other places where an article may have been expected in StdE (e.g. *in Ø 20th century*).

(8) Differences in sensory perception were put aside as either a personal talent — as **in case of** artists — or an individual, neuro-psychological deficiency, or both. (NL W2A-014)

(9) According to Vern Bullough […] there hasn't been a change in male sexual pattern in 20th century whereas premarital sex rate doubled **in case of** women by 1970s and rose to new peak by 1976. (ICE-IND W2A-005)

6. Discussion and conclusion

Broadly, our frequency results revealed a regional East Africa grouping (ICE-KEN and ICE-TAN) with negligible other differences, while the case studies of specific clusters mostly showed a NS–NNS divide. We are therefore unable to reject the null hypothesis; contrary to H1, the results revealed no clear distinction between NNS varietal types or distinct mapping of ESL and EFL characteristics onto the Outer and Expanding Circles respectively. Specifically, (i) the Expanding Circle data, represented by the NL-CE, did not show a restricted repertoire of cluster types or overreliance on common clusters (H1a); (ii) the East African corpora displayed a wider variety of cluster types than the NL-CE, but the other Outer Circle corpora did not (H1b); and (iii) similar innovations (discussed in more detail below) could be seen across the two NNS varietal types. This complex interplay of characteristics may indicate that 3-grams are simply too fine-grained and variable to reflect varietal type (thus extending Gries & Mukherjee's 2010 similar finding for New Englishes), or that categorisations such as ESL and EFL may be more relevant when exploring sociocultural aspects such as identity, but less so for individual structural features (cf. Davydova 2012; Hundt & Vogel 2011; Werner 2013). In any event, the present results are inconsistent with the notion not only of a divide between varietal types, but also a continuum. A continuum implies that the respective varieties are largely discrete, internally coherent and located closer to one end of the ESL–EFL spectrum than the other; essentially a fuzzy rather than a strict divide. The present results do not sit well with this notion, instead supporting several other recent studies that find what Deshors (2014: 298) calls "intermingled" results across varietal types (e.g. Deshors 2014; Edwards 2014; Laporte 2012).

The grouping of the two East Africa varieties warrants further mention. Geographic proximity has emerged in a number of studies as the most important factor in lexicogrammatical patterning, superseding the groupings expected on the basis of not only the Three Circles model but also Schneider's (2007) Dynamic Model of the Evolution of Postcolonial Englishes (e.g. Fuchs 2015). At the same time, there are considerable differences between the history and use of English in Kenya and Tanzania, making them what Hudson-Ettle & Schmied (1999: 4) refer to as "strange bedfellows". What the two contexts do share is data sparsity in terms of English academic texts in the natural sciences and technology; thus, the corpus compilers were forced to substitute texts in areas such as agriculture and environmental development and to include somewhat less formal academic texts (Hudson-Ettle & Schmied 1999: 9). This more liberal approach to the academic categories may have contributed to the wider variety of types of three-word clusters seen in the ICE-KEN and ICE-TAN data. Further, the high levels of intra-corpus variation and frequent repetition by certain authors of specific three-word clusters may be attributable to somewhat more variable discourse-related proficiency in these corpora.

Examples of innovation found across all NNS varieties (but not in the NS varieties) included the tendency towards (i) variable modal usage, such as the use of *can* or *could* (rather than just *may*) with *due to the*, (ii) the use of an additional hedging word such as *perhaps* alongside modals, and (iii) the use of an additional contrast maker like *yet* or *while* in conjunction with *on the other hand*. The latter two seem to be motivated by shared cognitive processes among NNSs, and as previously noted may be linked with the trend towards hyper-explicitness or redundancy reported elsewhere (e.g. Callies 2015; Edwards & Laporte 2015; Gilquin & Granger 2011; Nesselhauf 2009; Seidlhofer 2004). Another example was the use of the cluster *in case of* — by 23 different authors across all NNS corpora — where StdE would typically require *in <u>the</u> case of*. This appears to be an instance of what Mukherjee & Hoffmann (2006: 161–166) call "semantico-structural analogy", or the attribution of new meanings to expressions based on analogy with existing forms. In this case the meaning of *in the case of* is imparted through a form based on *in case (it rains)/in case (of emergency)*, expressions that conventionally signal precautionary action. Given that, in Dutch at least, there is an exactly equivalent structure using the definite article (*in het geval van*), this tendency to blend existing target language patterns even seems to override the possibility of positive transfer.

As the above innovations were common to both NNS varietal types, this raises questions about the labels error and innovation in the Outer and Expanding Circles. Referring back to the two-stage process of language change (cf. Section 1), innovations such as *in case of* — including their occurrences in the Expanding

Circle data — may arguably be regarded as having reached a stable stage 2 state, since they are not only dispersed across numerous authors but have also made their way into highly normative expert academic writing; as Mauranen (2012: 26) writes, "[w]hen an innovation has diffused sufficiently widely to be observable, change has already taken place." This leads us to conclude that, insofar as certain discourse communities are concerned, structural nativisation is no less likely to occur in the Expanding than in the Outer Circle, and that we should be wary of drawing unequivocal links between the terms *EFL/Learner Englishes* and *Expanding Circle*.

One further innovation was attested among NSs as well as NNSs (but still more pronounced among the latter). Namely, our results testify to the encroachment of the expression *due to* into the semantic space of *because of*. Close inspection of the concordances revealed a sliding scale in upholding the traditional division (see e.g. Coghill & Garson 2006: 11) from ICE-GB (in the majority of cases) to ICE-USA (exactly half of the cases) down to a low of around one third of cases in the NL-CE. Indeed, going hand in hand with this in the NL-CE is a markedly low use of the alternate *because of*; it would be interesting to further pursue this division of labour in future research. As noted previously, it may be that register considerations play a role here, with NNSs perceiving *because of* as less formal. In any event, it seems plausible that the lead taken by NNSs in effecting such quantitative shifts could serve to accelerate change already underway in the native varieties. This may hold in particular for academic writing, as investigated here, given that NNSs are increasingly numerous in this domain not only as contributors but also as gatekeepers (journal editors, reviewers).

Three advantages of our methodological approach are worth commenting on. First, the decomposition of the overall frequencies allowed us to take a more nuanced look at what otherwise might be too readily termed 'overuse'. The high frequencies of three-word clusters in the East African varieties turned out to be attributable to a high number of different types, while the high frequencies of specific clusters (e.g. *the other hand*, *due to the*) in the NNS corpora were attributable to their use by more authors than in the NS corpora. In other words, the frequent use of various three-word clusters among the NNSs in general and the East African varieties in particular resulted from the use of a wider variety of types and wider dispersion across authors, not from authors repeating each cluster more frequently (except in certain individual cases; see below).

The second and third methodological points concern inter- and intra-corpus variation. While the present study was partly inspired by Gilquin (2015), we elected not to follow her practice of aggregating data per circle. Indeed, our results showed considerable inter-corpus variation, illustrating that any such aggregation should be done with caution. Further, our reporting of results per author rather than on the basis of traditional normalised frequency counts allowed us to identify large intra-corpus variation at the level of individual authors. As Gries (2008) points out, dispersion across authors is rarely satisfactorily addressed in corpus-based analyses. This ought to be rectified, as it has important implications for distinguishing between individual error and stable innovation. Intuitively, high individual variation in any one corpus would seem to go against the notion of stable norms. It could be argued to represent a speech community in flux, with some but not all individuals shifting towards new norms. However, the relevant concordances in our Outer and Expanding Circle data alike revealed that this variation tended to manifest itself in repeated use of certain clusters by a limited number of individuals (although not enough, as we have seen above, to significantly drive up the average frequency of use). This points rather to an interlanguage strategy, whereby learners latch onto particular phraseological crutches to compensate for the absence of more varied linguistic resources to signal contrast, express causality and so forth (the "teddy bear" effect (Hasselgren 1994)).

To explore these issues further, the inclusion of a wider range of genres and more Expanding Circle data would be desirable. Several multi-genre Expanding Circle user corpora are currently in the pipeline (Laitinen 2011, 2016)[18] and will make a valuable contribution to the existing stock of corpora available for comparative analyses. Given sufficient data for the study of lexical bundles, even more fine-grained results could be obtained by conducting a discourse-functional or structural analysis (cf. Biber & Barbieri 2007; Biber et al. 1999; Chen & Baker 2010) or investigating differences across disciplinary lines (cf. Cortes 2004; Hyland 2008).

Acknowledgements

We would like to thank the editors of the volume and three anonymous reviewers for their comments on an earlier version of the manuscript. The usual disclaimers apply.

References

Bamgbose, A. 1998. "Torn between the norms: Innovations in World Englishes", *World Englishes* 17(1), 1–14. https://doi.org/10.1111/1467-971X.00078

[18] Indeed, we considered including the academic sections of the Corpus of English in Finland (FIN-CE) and Corpus of English in Sweden (SWE-CE) in the present study; however, at the time the analyses were conducted the data collection process was still ongoing (see Laitinen 2011; 2016 and Laitinen & Levin 2016 for initial results).

Berns, M. 1995. “English in the European Union”, *English Today* 11(3), 3–11. https://doi.org/10.1017/S0266078400008348

Biber, D. & Barbieri, F. 2007. “Lexical bundles in university spoken and written registers”, *English for Specific Purposes* 26(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken And Written English*. Harlow, UK: Longman.

Biewer, C. 2011. “Modal auxiliaries in second language varieties of English: A learner’s perspective”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 7–33. https://doi.org/10.1075/scl.44.02bie

Bruckfield, A. 2012. Prepositions: The Ultimate Book. *Mastering English Prepositions for International Students* (Rev. ed.). Oak Publishers.

Callies, M. 2015. Towards a process-oriented approach to comparing EFL and ESL varieties: A corpus-study of lexical innovations. Paper presented at the pre-conference workshop *Corpus Linguistics and Linguistic Innovations in Non-native Englishes, ICAME 36*, Trier, 27 May 2015.

Chen, Y. & Baker, P. 2010. “Lexical bundles in L1 and L2 academic writing”, *Language Learning & Technology* 14(2), 30–49.

Coghill, A.M. & Garson, L.R. 2006. *The ACS Style Guide: Effective Communication of Scientific Information*, Michigan: American Chemical Society. https://doi.org/10.1021/bk-2006-STYG

Cortes, V. 2004. “Lexical bundles in published and student disciplinary writing: Examples from history and biology”, *English for Specific Purposes* 23(4), 397–423. https://doi.org/10.1016/j.esp.2003.12.001

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. Harlow, UK: Longman/ Pearson.

Davydova, J. 2012. “Englishes in the outer and expanding circles: A comparative study”, *World Englishes* 31(3), 366–385. https://doi.org/10.1111/j.1467-971X.2012.01763.x

Deshors, S.C. 2014. “A case for a unified treatment of EFL and ESL - A multifactorial approach”, *English World-Wide* 35(3), 277–305. https://doi.org/10.1075/eww.35.3.02des

Deshors, S.C., Götz, S. & Laporte, S. 2015. Corpus Linguistics and Linguistic Innovations in Non-Native Englishes: A thematic introduction. Thematic introduction to the pre-conference workshop *Corpus Linguistics and Linguistic Innovations in Non-native Englishes, ICAME 36*, Trier, 27 May 2015.

Edwards, A. 2014. “The progressive aspect in the Netherlands and the ESL/EFL continuum”, *World Englishes* 33(2), 173–194. https://doi.org/10.1111/weng.12080

Edwards, A. 2016. *English in the Netherlands: Functions, forms and attitudes. Varieties of English around the World (VEAW)*, vol. G56. Amsterdam & Philadelphia: John Benjamins.

Edwards, A. & Laporte, S. 2015. “Outer and Expanding Circle Englishes: The competing roles of norm orientation and proficiency levels”, *English World-Wide* 36(2), 135–169. https://doi.org/10.1075/eww.36.2.01edw

ELFA. 2008. *The Corpus of English as a Lingua Franca in Academic Settings*. Director: Anna Mauranen. Available at http://www.helsinki.fi/elfa/elfacorpus.

Ellis, N.C., Simpson-Vlach, R., Römer, U., O’Donnell, M. & Wulff, S. 2015. “Learner corpora and formulaic language in second language acquisition research”. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Cambridge Handbook of Learner Corpus Research*. Cambridge, UK: CUP, 357–378. https://doi.org/10.1017/CBO9781139649414.016

Fuchs, R. 2015. The frequency of the present perfect in World Englishes. Paper presented at the *21st IAWE conference*, Boğaziçi University/Turkey, 8-10 October 2015.

Gilquin, G. 2015. "At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared", *English World-Wide* 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Gilquin, G., De Cock, S. & Granger, S. 2010. *The Louvain Database of Spoken English Interlanguage*. (Handbook + CD-ROM). Louvain-La-Neuve: Presses universities de Louvain.

Gilquin, G. & Granger, S. 2011. "From EFL to ESL: Evidence from the International Corpus of Learner English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 55–78. https://doi.org/10.1075/scl.44.04gra

Götz, S. & Schilk, M. 2011. "Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced German learners." In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 79–100. https://doi.org/10.1075/scl.44.05sch

Graddol, D. 1997. *The Future of English: A Guide to Forecasting the Popularity of the English Language in the 21st Century*. London: British Council.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (Eds.). 2009. *International Corpus of Learner English*. Version 2 (Handbook + CD-ROM). Louvain-la-Neuve: Presses universitaires de Louvain.

Greenbaum, S. 1991. "ICE: The International Corpus of English", *English Today* 7(4), 3–7. https://doi.org/10.1017/S0266078400005836

Gries, S. Th. 2008. "Dispersions and adjusted frequencies in corpora", *International Journal of Corpus Linguistics* 13(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. Th. & Deshors, S.C. 2015. "EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap", *International Journal of Learner Corpus Research* 1(1), 130–159. https://doi.org/10.1075/ijlcr.1.1.05gri

Gries, S. Th. & Mukherjee, J. 2010. "Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes", *International Journal of Corpus Linguistics* 15(4), 520–548. https://doi.org/10.1075/ijcl.15.4.04gri

Hasselgren, A. 1994. "Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary", *International Journal of Applied Linguistics* 4(2), 237–258. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x

Hilgendorf, S.K. 2015. The Expanding Circle, transnational media, and linguistic localization. Plenary lecture at the *21st IAWE conference*, Boğaziçi University/Turkey, 8-10 October 2015.

Hudson-Ettle, D. & Schmied, J. 1999. *Manual to accompany the East African component of the International Corpus of English (ICE-EA): Background information, coding conventions and lists of source texts*. Chemnitz: Chemnitz University of Technology.

Hundt, M. & Vogel, K. 2011. "Overuse of the progressive in ESL and learner Englishes – fact or fiction?" In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 145–166. https://doi.org/10.1075/scl.44.08vog

Hyland, K. 2008. "As can be seen: Lexical bundles and disciplinary variation", *English for Specific Purposes* 27(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Kachru, B.B. 1982. "Models for non-native Englishes". In B.B. Kachru (Ed.), *The Other Tongue: English Across Cultures.* Chicago: University of Illinois Press, 48–74.

Kachru, B.B. 1985. "Standards, codification and sociolinguistic realism: The English language in the outer circle". In R. Quirk & H. Widdowson (Eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge, UK: CUP, 11–30.

Kachru, Y. 2003. "On definite reference in World Englishes", *World Englishes* 22(4), 497–510. https://doi.org/10.1111/j.1467-971X.2003.00315.x

Kumar, S. 2010. *English Usage for the CAT* (2nd ed.). Chandigarh, Delhi and Chennai: Pearson.

Laitinen, M. 2011. "Contacts and variability in international Englishes: Compiling and using the Corpus of English in Finland", *Studies in Variation, Contacts and Change in English* 6. Available at www.helsinki.fi/varieng/series/volumes/06/.

Laitinen, M. 2016. "Ongoing changes and advanced L2 use of English: Evidence from new corpus resources". In M. José López-Couso, B. Méndez-Naya, P. Núñez-Pertejo & I.M. Palacios-Martínez (Eds.), *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora*. Amsterdam and New York: Brill/Rodopi, 59–84.

Laitinen, M. & Levin, M. 2016. "On the globalization of English: Observations of subjective progressives in present-day Englishes". In E. Seoane & C. Suárez-Gómez (Eds.), *World Englishes: New Theoretical and Methodological Considerations*. Amsterdam: John Benjamins, 229–252.

Laporte, S. 2012. "Mind the Gap! Bridge between World and Learner Englishes in the making", *English Text Construction* 5(2), 264–291. https://doi.org/10.1075/etc.5.2.05lap

Mauranen, A. 2011. "Learners and users - Who do we want corpus data from?". In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.45

Mauranen, A. 2012. *Exploring ELF: Academic English Shaped by Non-native Speakers*. Cambridge, UK: CUP.

Mukherjee, J. & Gries, S. Th. 2009. "Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English", *English World-Wide* 30(1), 27–51. https://doi.org/10.1075/eww.30.1.03muk

Mukherjee, J. & Hoffmann, S. 2006. "Describing verb-complementational profiles of New Englishes", *English World-Wide* 27(2), 147–173. https://doi.org/10.1075/eww.27.2.03muk

Muñoz, C. 2000. "Bilingualism and trilingualism in school students in Catalonia". In J. Cenoz & U. Jessner (Eds.), *English in Europe: The Acquisition of a Third Language*. Clevedon, UK: Multilingual Matters, 157–178.

Nesselhauf, N. 2009. "Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties", *English World-Wide* 30(1), 1–26. https://doi.org/10.1075/eww.30.1.02nes

Oksefjell Ebeling, S. & Hasselgard, H. 2015. "Learner corpora and phraseology". In S. Granger, G. Gilquin & F. Meunier (Eds.), *Cambridge Handbook of Learner Corpus Research*. Cambridge, UK: CUP, 207–230. https://doi.org/10.1017/CBO9781139649414.010

Paquot, M. & Granger, S. 2012. "Formulaic language in learner corpora", *Annual Review of Applied Linguistics* 32, 130–149. https://doi.org/10.1017/S0267190512000098

Pérez-Llantada, C. 2014. "Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage", *Journal of English for Academic Purposes* 14, 84–94. https://doi.org/10.1016/j.jeap.2014.01.002

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1972. *A Grammar of Contemporary English*. London: Longman.

Schneider, E.W. 2007. *Postcolonial English: Varieties Around the World*. Cambridge, UK: CUP. https://doi.org/10.1017/CBO9780511618901

Scott, M. 2015. *WordSmith Tools*. Version 6. Liverpool: Lexical Analysis Software.

Seidlhofer, B. 2004. "Research perspectives on teaching English as a Lingua Franca", *Annual Review of Applied Linguistics* 24, 209–239. https://doi.org/10.1017/S0267190504000145

Sharma, D. 2005. "Dialect stabilization and speaker awareness in non-native varieties of English", *Journal of Sociolinguistics* 9(2), 194–224. https://doi.org/10.1111/j.1360-6441.2005.00290.x

Sridhar, K.K. & Sridhar, S.N. 1986. "Bridging the paradigm gap: Second language acquisition theory and indigenized varieties of English", *World Englishes* 5(1), 3–14. https://doi.org/10.1111/j.1467-971X.1986.tb00636.x

Staples, S., Egebert, J., Biber, D. & McClair, A. 2013. "Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section", *Journal of English for Academic Purposes* 12(3), 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Szmrecsanyi, B. & Kortmann, B. 2011. "Typological profiling: learner Englishes versus indigenized L2 varieties of English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 168–187.

Van Rooy, B. 2006. "The extension of the progressive aspect in Black South African English", *World Englishes* 25(1), 37–64. https://doi.org/10.1111/j.0083-2919.2006.00446.x

Van Rooy, B. 2011. "A principled distinction between error and conventionalized innovation in African Englishes". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 189–207. https://doi.org/10.1075/scl.44.10roo

VOICE. 2013. *The Vienna-Oxford International Corpus of English* (version 2.0 online). Director: Barbara Seidlhofer. Available at http://voice.univie.ac.at.

Wahid, R. 2013. "Definite article usage across varieties of English", *World Englishes* 32(1), 23–41. https://doi.org/10.1111/weng.12002

Werner, V. 2013. *The Present Perfect in World Englishes: Charting Unity and Diversity*. PhD dissertation, University of Bamberg.

Wilson, K.G. 1996. *The Columbia Guide to Standard American English*. New York: Columbia University Press.

Innovative conversions in South-East Asian Englishes

Reassessing ESL status

Stephanie Horch
University of Freiburg

Singapore English and Hong Kong English started out as contact varieties and developed into ESL varieties belonging to the Outer Circle (Kachru 1985). Both varieties show a similar contact ecology (Chinese), but differ in their socio-institutional status in the Dynamic Model (Schneider 2003, 2007). By analyzing innovative verb-to-noun conversion in these two varieties, and comparing them to British English, this study shows that despite the obvious similarities in substratum, the usage frequency of conversion in both varieties differs considerably. These findings, similar to — most recently — Deshors (2014) and Gilquin (2015), call into question the established notion of ESL in general and the status of SgE and HKE as ESL varieties in particular. In order to accurately reflect contemporary language use, it is reasonable to conceptualize the notion of ESL as a continuum and to situate HKE and SgE at opposite ends.

Keywords: conversion, English as a second language (ESL), paradigm gap, World Englishes, corpus analysis

1. Introduction

Despite early calls for a unification of World Englishes and second language acquisition (SLA) research (Sridhar & Sridhar 1986, Williams 1987), it is only recently that linguists have started to compare SLA to the acquisition of English as a second language (ESL) typical of postcolonial contexts like Hong Kong or Singapore. This "paradigm gap", as it has been referred to (Sridhar & Sridhar 1986, taken up in e.g. Mukherjee & Hundt 2011), was found to be less 'unbridgeable' than thought of, with various studies highlighting features and processes that foreign-language

https://doi.org/10.1075/bct.98.07hor

(EFL) and second-language varieties[1] of English share (e.g. Biewer 2011, Deshors 2014, Edwards & Laporte 2015, Gilquin 2015, Gilquin & Granger 2011). These findings have led researchers to rethink the traditional classification of English into foreign-language and second-language contexts and to assume a continuum instead, with categories blending into each other rather than constituting clearly delimitable classes (e.g. Biewer 2011: 28, Deshors 2014: 298, Gilquin & Granger 2011: 76).

While previous studies (e.g. Biewer 2011, Deshors 2014, Gilquin & Granger 2011) have often compared ESL varieties to EFL varieties, this study compares corpus data of two Asian ESL varieties, Hong Kong English (HKE) and Singapore English (SgE). It aims to show that two varieties that are seemingly closely related as regards their colonial history as well as the contact ecology out of which they have arisen do, in actual fact, offer a very different picture concerning the socio-institutional status of English in general and — as a result — the innovation under study in particular, thus further challenging the clear-cut nature of the EFL–ESL–ENL distinction and, more specifically, the notion of *ESL variety* itself. The linguistic phenomenon investigated in the present study is verb-to-noun conversion (VNC).

The structure of the paper is as follows. Section 2 gives a brief overview of common classifications of varieties of English as well as an overview of HKE and SgE, and points out in how far they can be viewed as ESL varieties. Section 3 describes VNC as the innovation that is at the core of the comparative analysis of the South-East Asian (SEA) varieties. In Section 4, the database as well as the methodology for the study are outlined. Section 5 offers a quantitative analysis of the corpus data as well as a qualitative analysis of select examples in their discourse-pragmatic context. This is followed by a discussion in Section 6, which addresses the usefulness of categories such as ESL in the light of the corpus findings and provides concluding remarks.

2. Background

2.1 Classifying World Englishes

Two of the most common models of classifying varieties of English are the three-way classification into EFL (English as a foreign language) — ESL (English as a second language) — ENL (English as a native language) and the Kachruvian Three Circles model (Kachru 1985). The classification of English into EFL–ESL–ENL

[1] The notion "variety" is used to designate variants of the English language that can clearly be distinguished from one another on the grounds of linguistics features. Varieties in many cases are, but need not be, associated with geographical regions; they can, for example, also be characteristic of specific social groups.

was originally conceived by Strang (1970) and later taken up by Quirk et al. (1972), with Kachru's circles mapping onto it: Speakers within the Inner Circle are native speakers of English (ENL). The Outer Circle is comprised of mostly postcolonial regions, where English serves as (co-)official language and is learnt in addition to another language, hence ESL. In the Expanding Circle, English is used as a foreign language but not as an official language (EFL).

One fairly recent classification of varieties of English is the Dynamic Model (DM), proposed by Schneider (2003, 2007). The main idea of the model is that the "emergence of individual Postcolonial Englishes" is due to an underlying "uniform process". Similar developments across varieties are "grounded in specific, cross-culturally parallel sociolinguistic conditions in colonization" (Schneider 2014b: 10). According to Schneider's (2007) account, varieties of English undergo up to five developmental stages in each of which the identity of settlers and indigenous people is negotiated anew, leading to linguistic changes and innovations.

i. The first phase, **foundation**, is characterized by limited borrowing and pidginization.
ii. The phase of **exonormative stabilization** sees some grammatical transfer and a strong orientation towards British English as the standard.
iii. The **nativization** phase, which mostly coincides with political independence, is crucial to the development of a new variety. This phase is marked by linguistic innovations in all domains of language.
iv. Codification of the newly independent variety and increased linguistic creativity are the hallmarks of the fourth phase, **endonormative stabilization**.
v. The final stage of **differentiation** is defined by the emergence of dialects and sociolects within the new, postcolonial variety.

2.2 HKE and SgE as ESL varieties

HKE and SgE are generally classified as ESL varieties, that is, as belonging to the Outer Circle. Both varieties, which are briefly introduced in this subsection, show a similar contact ecology in that they both emerged from the sustained contact of Chinese dialects with British English (BrE) during the colonial rule. For a broader overview of the development of English in Asia, see Schneider (2014a).

Hong Kong was a British colony from 1841 until 1997. In the first century of colonial rule, bilingualism in English and Chinese was restricted to select individuals. In the late 1970s, with the introduction of free and compulsory primary education in English, the English language spread and the number of bilinguals increased (Bolton 2000: 269). After the Handover of Hong Kong to the Chinese in 1997, English remained an official language, next to Cantonese and Mandarin, two

Chinese dialects. The official language policy is one of trilingualism, i.e. fluency in Cantonese, English, and Mandarin (Bolton 2012: 228). However, despite the official policy, English is mainly restricted to public contexts such as administration, the legal system, business, and higher education (Evans 2010: 165). In more private settings, Cantonese prevails (Census and Statistics Department, Hong Kong SAR 2013). Due to the different functions English, Cantonese, and Mandarin serve, Pang (2003: 17) describes the situation in Hong Kong as "increasingly triglossic": English is of high instrumental value but the emotional attachment to the language is fairly low, while Cantonese is the language of family and friends (Gisborne 2009: 152–153). This division of labor between English and Chinese is only possible because of the great homogeneity of the Hong Kong population (over 90% have Cantonese as native language, cf. Census and Statistics Department, Hong Kong SAR 2013). Due to this situation, the status of HKE as a variety in its own right has repeatedly been questioned (e.g. Johnson 1994: 182, Pang 2003), but newer studies (e.g. Bolton 2003: 197–225, Setter et al. 2010: 8) agree that HKE is becoming increasingly nativized and should be considered an independent variety of English.

In Singapore, on the other hand, the situation is very different, even though the historical context is similar. Singapore was a British colony from 1819 to 1963 (Leimgruber 2013a: 1–6). After independence, the reason to make English the official language of the country was readily at hand: In contrast to all other languages in Singapore, English was the only one that could serve as an interethnic lingua franca to unite the different ethnicities present (Wee 2013: 105–109). Additionally, three mother tongues, corresponding to the ethnic groups, were recognized as official languages: "Mandarin for the Chinese, Malay for the Malays, and Tamil for the Indians" (Leimgruber 2013a: 12). The assignment of languages to ethnic groups did not necessarily reflect actual language use and led to a considerable strengthening of Mandarin, which has replaced other Chinese dialects in Singapore (Lim 2010: 30). Contemporary SgE can thus be assumed to be largely influenced by Mandarin (Ansaldo 2004: 135). Contrary to the situation in Hong Kong, the English language is pervasive in all domains in Singapore. Tan (2014: 334) shows that particularly among the youngest Singaporeans, English "has overtaken [all] other languages as the main language in the domains of home, leisure, intimacy and self", which is indicative of the on-going language shift towards English. As

Tan (2014: 319) claims, "English can and should be thought of as a mother tongue for Singaporeans".[2]

After briefly outlining the profiles of HKE and SgE, it becomes evident that while both Asian varieties have emerged from a similar language contact ecology, they differ in how evolved they are in the DM. HKE is generally classified as a variety at the stage of nativization (phase III), whereas SgE is categorized as an endonormatively stabilized variety (phase IV). The difference in the degree of institutionalization is likely to be reflected in language use, as the DM predicts and as previous studies (e.g. Edwards & Laporte 2015) suggest. Table 1 offers a comparison of HKE and SgE.

Table 1. South-East Asian varieties investigated

	HKE	**SgE**
time of British occupation	1841–1997	1819–1963
official languages	Cantonese, Mandarin, English	Mandarin, Malay, Tamil, English
language policy	trilingualism (Cantonese, English, Mandarin) and biliteracy (Chinese, English)	English and one ethnic mother tongue (either Malay, Tamil, or Mandarin)
areas of use of English	administration, law, business, higher education	all domains
most important contact language	Cantonese	Mandarin
developmental stage in DM	III	IV

The overarching aim of the present study is thus to analyze how structural innovations emerge and develop in HKE and SgE, two ESL varieties of English that are potentially characterized by influence from their Chinese substrate. Specifically, the phenomenon of innovative VNCs, described in more detail in Section 3, is studied. It is further explored whether and how the (near-)absence of inflectional and derivational morphology in the substratum (Sun 2006: 64, 73) and its tendency to liberally allow for verbs in nominal contexts (see below) can be traced in

[2] A comparison of HKE and SgE on the basis of their similar (historical) contact ecologies might be considered problematic, seeing that Malay and Tamil are also contact languages of English in Singapore. Notwithstanding, the number of L1 speakers of these languages is comparatively small, as the 2010 census indicates (12.2% Malay, 3.3% Tamil; Wong 2011). Furthermore, as Leimgruber (2013b: 236) and also Tan (2014: 333) stress, particularly the community of Tamil speakers is readily switching to English as its dominant language. As regards the structural feature analyzed in the present study, VNC, no significant distinction between HKE and SgE is expected to arise from the typology of the contact languages: In Chinese dialects as well as in Malay, an agglutinating language, "there are hardly any morphological processes to speak of" (Ansaldo 2009: 139).

innovative VNCs in these varieties. In addition, the interaction between substratum influence and the degree of institutionalization as operationalized by the DM is assessed. This is achieved by way of analyzing corpus data from HKE and SgE. Data from BrE, the parent variety of both varieties, serve as a basis of comparison.

3. Verb-to-noun conversion as innovation

In the present study, VNC is drawn on to illustrate potential effects of the Chinese substratum as well as potential differences between HKE and SgE due to their developmental stage. Conversion is defined as the change of word class without overt morphological marking (Plag 2003: 107–116):[3]

(1) Click the link below **to access** the searchable database. (GloWbE-US)

(2) because at the end of the day it wasn't that big of an **ask** (GloWbE-US)

The most common direction of conversion in English is from noun to verb (Don et al. 2000: 949), as illustrated in (1). The other direction, from verb to noun, as in (2), is less common in English, but highly frequent and formally unconstrained in Mandarin and Cantonese: "any verb in Cantonese can appear in subject and object positions without change in form" (Matthews & Yip 1994: 55; cf. Po-Ching & Rimmington 2004: 16 for Mandarin).

It has often been assumed that transfer from the substratum is one of, if not the most, important mechanism/s in shaping contact varieties, next to the sociolinguistic context. Following Bao (2005, 2009, 2010a, 2010b), the productivity of a transferred feature is assumed to depend crucially on the structural convergence of substratum and lexifier language,[4] with structural convergence leading to successful transfer and a higher productivity of the transferred feature. In the present case, it can be hypothesized that the Chinese substratum will lead to a higher productivity of VNC in the Chinese-substratum varieties of English, given the very high productivity of the feature in Chinese and the structural convergence of Chinese and English when it comes to VNC. Considering that VNC is — if not highly at least mildly — productive in English, substrate influence might, however, be difficult to recognize as such. It is therefore necessary to focus on

[3] This phenomenon is also known under the name of zero-derivation (e.g. Kastovsky 1982: 172–175, Marchand 1960: 293–308). However, the notion of conversion is preferred here. For a detailed account of the problematic nature of conceptualizing conversion as zero-derivation see Balteiro (2007: 25–32).

[4] The terminology of *substratum* and *lexifier language* is adopted from Bao (2005, 2009, 2010a, 2010b). In other works on emergent Englishes, these languages are also called *contact languages*.

innovative conversions, that is, those that have not been attested in native varieties of English.[5]

When studying innovations, it is necessary to tease apart narrow and broad definitions of the concept. Traugott & Trousdale (2013: 2), for example, define innovations in the narrow sense as "feature[s] of an individual mind". By replication "across populations of speakers", these *ad hoc* innovations can become conventionalized. Or, as van Rooy (2011: 192–193) describes it in contact linguistic terms, *ad hoc* innovations enter the feature pool, may be selected and then evolve as innovative features of a contact variety. What distinguishes *ad hoc* innovations from innovative features is, first, the degree of systematicity with which they are used and, second, the degree of acceptability of the innovation (van Rooy 2011: 195). Both systematicity as well as acceptability are gradient notions, with *ad hoc* innovations ranging low on systematicity as well as acceptability. Conventionalization involves the spread of the innovation, i.e. an increase in systematicity, and with that also (supposedly) an increase in acceptability. Operationalizing innovations by assessing their systematicity and acceptability is suitable for a study within the usage-based paradigm, as the two concepts can be assessed empirically comparatively easily with corpus-analytic and experimental methods.

The innovative VNCs analyzed in this study are hypothesized to range low on systematicity and probably also low on acceptability in native varieties[6] and can thus be thought of as *ad hoc* innovations, that is, as "feature[s] of an individual mind", rather than conventionalized features. However, in the SEA varieties, the influence of Chinese might have induced conventionalization, resulting in a higher systematicity of use and a higher acceptability of the phenomenon.

Conventionalization is operationalized as codification in the present study. That is, an instance of VNC will count as an instantiation of the innovation as long

[5] Previous research on conversion in Asian Englishes has, to my knowledge, exclusively relied on word lists which had been compiled before conducting the analyses. These studies therefore did not detect truly innovative lexical features. Cases in point are Biermeier (2008) and Evans (2015).

[6] For the acceptability of VNC in varieties of English, see Horch (2017: 195–243). In a web-based experiment, speakers of native varieties were found to judge sentences containing innovative VNCs significantly worse than speakers of HKE and SgE, and also significantly worse than sentences containing features attested in other varieties of English, such as the use of *fi* for *to* in Jamaican English *(we need fi tell them fi put down the gun)*. Within the two Asian varieties, HKE participants rated the sentences containing VNC higher, presumably because of their higher systematicity of use of the process.

as it lacks an entry in the *Oxford English Dictionary* (online edition).[7] All those converted forms that have already been codified are defined as pertaining to the native variety, that is, the standard or norm, and are not analyzed.

In the present study, the contrast of norm vs. innovation is operationalized by a comparison of an influential native variety of English, BrE, and two New Englishes, SgE and HKE. Thus, VNCs present in the BrE standard are assumed to constitute the norm, and VNCs that are not attested for BrE are classified as innovations. The choice of BrE as a reference variety is based on that, first, BrE is the historical parent variety of the SEA varieties in question, and, second, that it is an "inner circle 'super-variet[y]'" (Collins & Yao 2013: 479), with the potential to influence other varieties around the world. From the perspective of HKE and SgE, the choice of BrE as a basis of comparison seems suitable, considering that in many Asian varieties which have not developed a codified standard yet, English language teaching, e.g. in schools, still orients towards BrE and US English, the two most influential native varieties (Kirkpatrick 2012: 17).

The hypotheses that guide the analysis can be summarized as follows. Firstly, the ESL varieties (HKE, SgE) are expected to show a higher productivity of VNC than the native variety (BrE) due to extensive influence from the Chinese substratum. Secondly, the SEA varieties are assumed to differ according to the socio-institutional status of English in these regions, that is, due to the fact that SgE is further evolved than HKE along the cline of developmental stages in the DM. Thirdly, in line with a usage-based approach to language (e.g. Bybee 2010), verbs of different frequencies of occurrence are expected to be processed differently within varieties. These processing differences are likely to affect the productivity of conversion. Verbs that are highly frequent will show "increased morphological stability" and will consequently convert to a lesser degree than less frequent verbs. (This so-called conserving effect of frequency (Bybee 2010: 24–25) is also observed for, among others, highly frequent irregular verbs such as *burn*, which generally resist a regularization of the paradigm (*burnt* > *burned*) longer than less frequent irregular verbs such as *spill* (*spilt* > *spilled*)).

[7] A major corpus of a native variety or frequency data obtained on the basis of such a corpus could also serve as reference, as a reviewer points out. Yet, the reliability of these data crucially hinges on the reliability of the corpus itself. Seeing that dictionaries are carefully edited, they can be considered a more objective record of the language. Nonetheless, the editing process is time-consuming, so that dictionaries generally 'lag behind' actual language use. As a compromise, the online edition is preferred here.

4. Methods and data

For many studies concerned with low-frequency language phenomena such as VNC, the *International Corpus of English* (ICE; The ICE Project 2014), the benchmark corpus in World Englishes research, has proven too small to conduct reliable statistical analyses (e.g. Biermeier 2008: 198). Therefore, the present study is based on data from the *Corpus of Global Web-based English* (GloWbE; Davies 2013). For a detailed description of the corpus and the compilation procedure, see Davies & Fuchs (2015). GloWbE comprises 1.9 billion words stemming from webpages hosted in twenty different English-speaking countries (Davies & Fuchs 2015: 2–3). The sections for HKE and SgE comprise around 40 million words, the section for BrE roughly 390 million words (Davies & Fuchs 2015: 6). The data were sampled in 2012 and 2013 (see http://corpus.byu.edu/glowbe). For every variety, GloWbE offers data from informal blogs (roughly 60%) and from other websites of a supposedly more formal register (roughly 40%). The aim of this distinction is to emulate the 60% to 40% ratio of spoken to written language that is also present in the ICE corpora (Davies & Fuchs 2015: 3–4). Nevertheless, as Mair (2015: 30–31) and also Peters (2015: 41–42) point out, the questions of whether blogs constitute a genre and whether they are comparable to spoken language at all remain unclear. Also, the reliability of web-corpora in general has met with skepticism due to their heterogeneity as regards "types of speakers" and "language variants" (i.e. basi-/meso-/acrolectal, Mukherjee 2015: 35).[8] Notwithstanding these points of criticism, Heller & Röthlisberger (2015) could show that the results of a quantitative corpus study on the dative and genitive alternation were largely independent of whether the data came from ICE or GloWbE, thus allowing for the tentative conclusion that ICE and GloWbE can be expected to yield comparable results.

In order to evaluate the success of VNC in Asian varieties of English, verbs fulfilling the following criteria were selected out of two frequency bins (high and low), each containing 100 verbs: The eligible verbs have a corresponding and near-synonymous deverbal noun formed by derivation (e.g. *examination*) but have not been converted yet, that is, show no attested lexicalized deverbal converted form (e.g. **an examine*) according to the OED.[9] Out of the resulting group of 46 verbs, twenty were randomly selected. (For a list, see Table 3 in the Appendix.) For these

[8] A more detailed discussion of the benefits and potential drawbacks of GloWbE can be found in a recent issue of *English World-Wide* (Vol. 36, No. 1) as well as in Horch (2017: 69–80).

[9] This eliminates verbs such as *play*, which does not have a corresponding deverbal noun (**playation*, **playment*), or *estimate*, for which a converted form is attested in the OED.

potentially innovative conversions, random samples of size 1,000[10] were drawn from GloWbE for both the potential singular (corresponding to the infinitive, e.g. *examine*) as well as the potential plural (corresponding to the third person singular form, e.g. *examines*). The automatic part-of-speech tagging provided in GloWbE is not sufficiently accurate to rely on it in a study of conversion,[11] therefore, the samples were manually coded for part-of-speech. The normalized (to the corpus size) token frequencies of the converted forms were input into a logistic regression model, together with the normalized token frequency values for the verb as well as the corresponding deverbal noun (e.g. *examination*).

The fact that the twenty verbs under scrutiny were selected randomly as well as the circumstance that a random sample was drawn might limit the possibility to generalize the tendencies encountered for these verbs to other verbs. For instance, the sample does not allow for any generalizations on the semantic level. Yet, focussing on individual verbs or types of verbs is not the scope of the present study (which is why verbal lexemes were included as a random effect in the regression model, see below). However, considering that a total of roughly 120,000 forms was manually tagged for part-of-speech for this study, it can be assumed that the results are robust for the phenomenon of VNC in general, even if they might be less reliable for individual verbs.

The statistical method of a logistic regression establishes the relationship between a binary dependent variable and various independent variables. (For a detailed account of the benefits of logistic regressions, particularly in the field of linguistics, see Gries (2015).) In this case, the logistic regression calculates the odds of conversion (i.e. nominalization of *examine* as *examine/s: he passed a nation-wide* ***examine*** *in law*) versus the odds of derivation (i.e. nominalization of *examine* as *examination/s: he passed a nation-wide* ***examination*** *in law*) depending on three independent variables: the variety of English (HKE, SgE, BrE), the frequency of the verb in the corpus (frequencyVerb), and the frequency of the derived, near-synonymous form in the corpus (frequencyDeriv). Furthermore, the effects of verb frequency and the derived, near-synonymous form could play out differently for each variety, so that an interaction between these two predictors and variety is specified. The lexeme itself should not influence the results, which is why it is included as a so-called random effect. By excluding individual lexemes

[10] The number of 1,000 was chosen for reasons of feasibility.

[11] For example, a search for nominal *examine*, [examine].[nn*], yields no hits in GloWbE. Yet, as the example in (3) clearly demonstrates, *examine* does occur in unambiguously nominal contexts in this corpus.

from the calculation, general tendencies can be established independently of the verb lexemes.[12]

5. Results

5.1 Quantitative perspective

The results for the fixed effects of the logistic regression model are displayed in Table 2.[13] The “Estimate” column gives the predicted change in log odds (abbreviated *B*) and the column “Std. Error” gives the standard error. *P* values, indicating the statistical significance of the estimates, are listed in the “*p* value” column.

Table 2. Fixed effects of generalized linear mixed model for conversion in SEA Englishes vs. BrE. BrE set as the reference level

	Estimate	**Std. Error**	***p* value**
(Intercept)	−6.91	0.244	< 0.001 ***
varietyHK	1.73	0.117	< 0.001 ***
varietySG	0.79	0.122	< 0.001 ***
frequencyDeriv	−0.55	0.148	< 0.001 ***
frequencyVerb	0.29	0.199	0.149
varietyHK:frequencyDeriv	−0.07	0.075	0.326
varietySG:frequencyDeriv	−0.06	0.096	0.546
varietyHK:frequencyVerb	−0.31	0.079	< 0.001 ***
varietySG:frequencyVerb	−0.19	0.108	0.079 .

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The results show that conversion has a higher chance of occurring in the Chinese-substratum varieties, that is, in HKE and SgE compared to BrE (cf. Figure 1).[14] However, the chances (given in logarithmic odds) are not equal in HKE and SgE, but considerably higher in HKE ($B = 1.73$) than in SgE ($B = 0.79$).

[12] This results in the following model equation: glmer(conv, deriv ~ (1|lexeme) + variety * (frequencyDeriv + frequencyVerb), family = “binomial”,…), whereby (1|lexeme) is the notation for including the lexeme as a random effect. The model was calculated using the glmer() function from the lme4 package (Bates et al. 2014) for R (R Core Team 2014).

[13] For the random effect, see Table 4 in the Appendix.

[14] BrE was set as the reference level, so that all estimates are given in reference to the native variety.

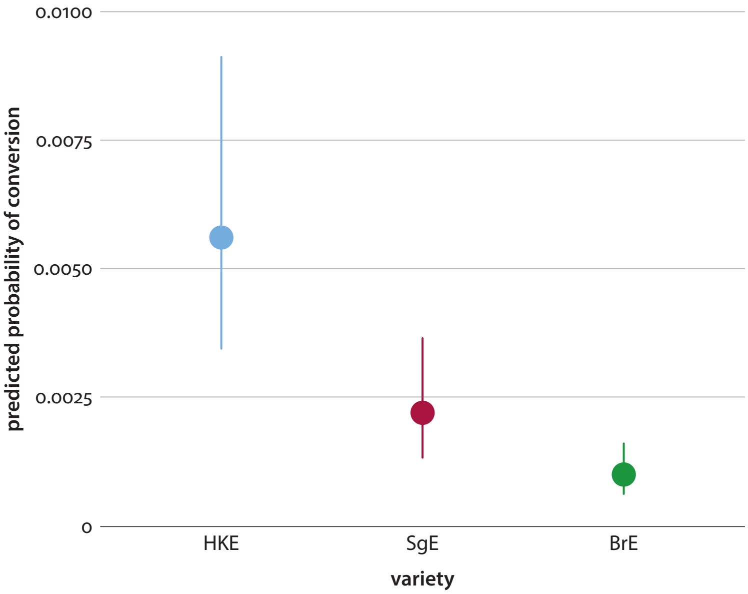


Figure 1. Predicted probabilities of conversion per variety

Furthermore, the frequency of the competing, near-synonymous derived form turns out to be a highly significant predictor of VNC ($B = -0.55$, $p < 0.001$). The negative estimate indicates an indirect relation: the more frequent the derived form is, the lower the probability that the corresponding verb is converted to a noun. This tendency holds not only for the native variety but also for the Asian varieties.

As far as the frequency of the verb is concerned, it does not significantly influence the likelihood of VNC in BrE ($p > 0.1$) or in SgE ($p > 0.05$). Yet, in HKE, the frequency of the verb emerges as a highly significant predictor of VNC ($B = -0.31$, $p < 0.001$).

These results suggests that a shared substratum does not necessarily lead to similar usage patterns of transferred features in the contact language. It rather seems that the degree of institutionalization is crucial in the formation of a contact variety. The less advanced variety, HKE, shows a usage pattern that is further away from that of the native variety. This tallies with findings from SLA that indicate that conversion is favored in the early stages of language learning and that morphologically more complex processes become more frequent with increasing target language proficiency (Pavesi 1998: 226).

The importance of the socio-institutional status of English is further corroborated by learner-like tendencies present in HKE, but not in SgE ($p > 0.05$ for varietySG:frequencyVerb) or BrE ($p > 0.1$ for varietyGB:frequencyVerb). In HKE,

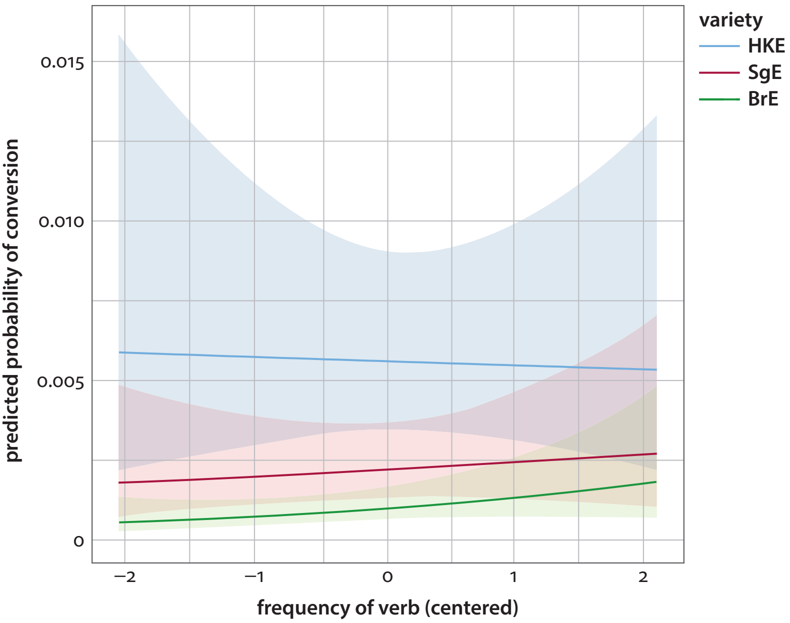


Figure 2. Effect for verb frequency. In HKE, odds become higher for less frequent verbs compared to BrE and SgE. The shaded areas indicate the confidence intervals (0.95) for each line.

contrary to the other varieties, the frequency with which a verb occurs in the entire corpus is of relevance to the odds of that verb being converted to a noun. The negative estimate of –0.31 ($p < 0.001$) indicates an inverse relation: The less frequently a verb occurs, the higher the odds of conversion become; the more frequent a verb is, the lower the odds that it is converted to a noun. The effect is visualized in Figure 2.[15]

Thus, in HKE, in the high-frequency range, there is little creativity and speakers prefer highly frequent nominal forms well-known to them, i.e. nouns formed by derivation. In the low-frequency range, on the other hand, when unable to retrieve a nominal form, speakers form a new word adhering to the easiest process, namely conversion, a process that involves no additional morphological material. This *shortest path principle* (Wald 1993: 68) is summarized by Biewer (2011: 14) as follows: "[I]f the rules of the target language allow for variation, one variant will be selected, and the selected variant will be the one that 'correspond[s] most closely' to the L1 feature". In the case of HKE, conversion is the process that corresponds

[15] It has to be acknowledged that the confidence intervals indicated in Figure 2 are fairly high, particularly for HKE. However, the results for BrE vs. HKE appear to be robust (almost no overlap of confidence intervals).

most closely to the L1 (Chinese), which explains why more transfer is observed in the low-frequency range.

Notwithstanding this comparatively unconstrained use of VNC in HKE, there is a mechanism that applies across all varieties, working against VNC: The frequency of the deverbal noun significantly shapes the odds of VNC, with a higher frequency of the deverbal noun (e.g. *examination*) blocking the conversion of a verb to a noun. This blocking constraint (cf. Aronoff 1976) serves the avoidance of synonyms and applies to the SEA Englishes to the same degree as to the native variety (the estimates for HKE and SgE are not significantly different from that for BrE, $p > 0.1$ in both cases), despite the influence of the substratum.

Summarizing these findings, it is obvious that the influence of the substratum can only insufficiently explain the differences in VNC between the SEA Englishes. The degree of institutionalization of English determines the degree and range of transfer from the substratum. Where English is less institutionalized, there is more transfer. It is therefore necessary to combine the two explanatory approaches, substrate transfer and degree of institutionalization, so as to understand the usage patterns of VNC. Following Bao's approach (2005, 2009, 2010a, 2010b), the substratum contributes the structural features to the emergent variety and the superstratum moderates their productivity. The more compatible grammatical features are, the more frequently they are expected to occur. Nonetheless, the evolutionary stage of the respective variety of English plays a vital role in shaping the frequency pattern that is contributed by the superstratum language. As far as VNC is considered, it seems that the more evolved the variety is, the more the frequency pattern of the contact variety approximates that of the native variety. This tallies with Edwards & Laporte's (2015) findings that the most institutionalized ESL varieties show patterns most similar to native varieties, but contrasts with the wide-spread assumption that a greater norm orientation might result in a higher similarity of EFL varieties to ENL varieties, and might lead ESL varieties to show patterns dissimilar to ENL varieties. This hypothesis is pursued by e.g. Mukherjee & Gries (2009), who show that verb-complementational profiles of some Asian varieties reflect the evolutionary stages at which the varieties are located, with the least institutionalized variety exhibiting the profile closest to the native variety. These apparently contradictory results thus seem to imply that "the evolution of World Englishes does not necessarily have the same impact on all linguistic features" (Laporte 2012: 286), leading Bernaisch (2015: 214–218) to conclude that different innovative features are affected by ongoing institutionalization in different ways, some showing endonormative and some exonormative tendencies. The quantitative results suggest that the more institutionalized variety (SgE) is closer to the native variety as regards the usage profile of VNC, while the less institutionalized variety (HKE) displays a distinct, and different usage profile.

Generally, the results further reveal that VNC is an elusive phenomenon in the varieties investigated. The predicted probability for VNC is low in all varieties (cf. Figure 1). This illustrates the need for large corpora when researching this innovation. As far as the individual verbs are concerned, it can tentatively be concluded that some verbs apparently display a higher inclination to convert to nouns than others. Of all the verbs listed in Table 3, *choose*, *require*, *examine*, and *continue* are among the verbs for which most corpus evidence is available across varieties. In the SEA varieties, *improve* is also fairly frequently encountered in nominal contexts. Examples are given in (3) through (5).

(3) From this, I learn to make **choose**. (GloWbE-HK)

(4) Try to make a decision whether the net space is adequate for your **requires** or not. (GloWbE-HK)

(5) This may be particularly helpful if you not too long ago installed a hardware **improve** to the now defunct system. (GloWbE-SG)

Yet, as this result rests on random samples of size 1,000 of a group of randomly chosen verbs, a meaningful interpretation is neither easily found nor within the scope of this study.

5.2 Qualitative perspective

The quantitative data are complemented by a qualitative analysis of select examples in their discourse-pragmatic context.[16] The aim of such a close reading is to assess whether differences between the SEA varieties extend beyond mere frequency.

(6) The Nanchang Bayi trade union was clandestinely set up on 14 August 2006. The chair, Gao Haitao, was elected by popular vote. Since then he had fought against Wal-Mart management over one issue after another. It is significant that he had studied law on his own while supporting himself by working at Wal-Mart part-time. In 2005 he passed a nation-wide **examine** in law and decided to stay on in Wal-Mart as a full-timer. His legal knowledge became his main weapon to fight against Wal-Mart. (GloWbE-HK)

Excerpt (6) is taken from a website hosted in Hong Kong called *China Labor News Translations*, which offers "English translations of Chinese-language reports, commentaries and blogs on labor issues". The Chinese texts are translated

[16] The analysis provided here cannot replace an exhaustive qualitative investigation. Yet, the examples have been chosen in such a way that they represent what appear to be general trends in these varieties.

by volunteers who are "former Chinese labor activists now residing outside China, the foreign media, or foreign scholars, NGOs, trade unions". The website is directed towards non-Chinese (native) speakers of English, encouraging them to "build a more nuanced understanding of […] Chinese labor issues" (China Labor News Translations n.d.). Consequently, it is not surprising that the text is of a formal register, showing an elaborate style characterized by infrequent lexical items (*clandestinely*), hypotactic sentences (*while supporting…*), the passive voice (*was elected*) and metaphors (*His legal knowledge became his main weapon*). However, regardless of the formality of the text, the author-translator of the text uses conversion in one instance (*examine*). Considering that this excerpt is a rendition of a text in (presumably) the translator's L1, it is highly probable that this instance of VNC is the result of direct transfer from the L1. The omission of the article in *he had fought against Ø Wal-Mart management* also points to transfer from the translator's assumed L1, Chinese.[17]

> (7) You can say that I'm easily contented, no **deny** about that. I made a quick decision to be a Stay-At-Home-Mum five years ago and I went ahead to start an online business on my hobby two years back. Besides having the gut feeling and full support from my family, my positive attitude and optimism put me through those rollercoaster rides through these years. I do have my downtimes and bad hair days, but I've learn to pick myself up fast and keep moving forward. (GloWbE-SG)

Excerpt (7), taken from a website hosted in Singapore, illustrates that VNC is used in different contexts in SgE. While excerpt (6) is markedly formal in register, (7) is a very informal text. It is part of a blog entry by a Singaporean woman who describes her life as a mother. Overall, her writing is very close to Standard English (except for the omission of a past-tense marker in *I've learn*). Nonetheless, in the first sentence, she converts the verb *deny* to a noun. Contrary to excerpt (6), conversion could in this case be due to analogy, potentially inspired by the [*no* N *about*] construction (e.g. *no doubt about*). As corpus evidence suggests, *deny* can be expected to occur in the form of a present participle (e.g. in *no denying (the fact) that*). Yet, due to the (substrate-induced) non-standard patterns of morphological marking that SgE shows (e.g. Gut 2009), it is reasonable to assume that the *-ing* form is dispreferred here. Therefore, the [*no* N *about*] construction is chosen, which helps avoid the use of a verb altogether. The fact that *learn* in the last sentence of (7) is simplified (*I've learn to pick myself up*) is consistent with the conversion of *deny* in that both could be seen as indicators of a general tendency to

[17] I thank an anonymous reviewer for pointing this out.

avoid bound morphemes, potentially owing to the analytic nature of the Chinese substratum.[18]

Accordingly, it appears that innovative VNCs occur not only more frequently in HKE than in SgE, but also in different contexts. In SgE, VNC seems to be restricted to (very) informal contexts, while in HKE it also extends to more formal contexts. Considering that in native varieties conversion is associated with informal registers (e.g. Cannon 1985), the usage profile of VNC in SgE approximates that of BrE to a larger extent than that of HKE. Furthermore, as excerpt (7) indicates, VNC is often facilitated by analogy in SgE. The modelling on analogical formations is presumably not of equal relevance in HKE, again hinting at the comparatively unconstrained nature of the process in HKE. This shows that HKE speakers, despite an exonormative orientation towards the BrE standard (Pang 2003), do not show the same awareness of VNC as speakers of SgE. This is in line with Chui's (2010: i) findings that Hong Kong students displayed an "inadequate sensitivity to mode difference in English", with written texts often exhibiting features typical of spoken language and *vice versa*. Yet, this qualitative analysis, as it rests on a limited number of examples, does not allow for firm conclusions but can merely provide preliminary indications. The claims made would need to be substantiated by a larger number of examples.

6. Discussion and conclusion

The results presented herein suggest that in contact variety formation, effects of the substratum lessen with an increasing degree of institutionalization. A high degree of institutionalization reduces both the quantity and range of transfer, regardless of the status of English as an official language in a region/country. VNC has higher odds of occurrence in Chinese-substratum varieties, which can most likely be ascribed to the influence of the substrate. The qualitative analysis corroborates this assumption, particularly (6), in which VNC is used in a translation from Chinese to English. In (7), VNC also seems to result from a tendency in the variety that is due to the non-morphemic nature of the substratum.

Nonetheless, substrate influence is moderated by the degree of institutionalization of English, as both analyses show. Not only are the odds of conversion lower in SgE, the more advanced variety, there are furthermore tendencies characteristic of learner varieties to be found in HKE. First, there is a statistically significant

[18] The omission of bound morphemes could also result from a general simplification tendency observed in L2 varieties. See e.g. Seoane & Suárez-Gómez (2013: 11) or Werner (2014: 330) for tense and aspect marking.

preference for the non-morphemic word-formation process of VNC (compared to the native variety), which is also observed for EFL speakers. Second, the preference for VNC is particularly pronounced for verbs of low frequency, indicative of the *shortest path principle*. Also, an effect of register is apparent from the excerpts in 5.2. While converting a verb to a noun can occur in formal and informal contexts in HKE, this process is restricted to the more informal registers in SgE, again illustrating that institutionalization is crucial in shaping contact varieties.

As has already been mentioned, the precise effects of register and mode would need to be corroborated by further research, also drawing on large corpora of spoken language. Furthermore, this study has drawn on GloWbE and thus on data from innovative web registers. While the results are plausible, it would be worthwhile to scrutinize the same process in other, more 'traditional' registers. Additionally, it might be rewarding to focus on specific groups of verbs, e.g. with similar semantics. Also, a comparison of the Chinese-substratum varieties with other varieties with largely synthetic substrata (such as Indian English) could yield interesting insights into the nature of VNC as a process transferred from an analytic substratum and/or as a general learner process.

Notwithstanding this outlook on methodology and further points of interest, the present, GloWbE-based study demonstrates that it is not only the boundary between EFL and ESL varieties that is blurry, it is the notion of ESL in itself. While HKE and SgE have both been termed ESL varieties, VNC plays out very differently in these varieties, with the degree of institutionalization assuming a key role. Subsuming SgE and HKE under the heading of ESL varieties does not truthfully represent contemporary language use, as it suggests similarities, when, in actual fact, HKE might be more similar to EFL varieties than to the highly institutionalized SgE (cf. Edwards & Laporte 2015).

Yet, the fact that theoretical categories almost never hold out to empirical analysis hardly comes as a surprise. Nonetheless, this is by no means a plea for rejecting all classification (in this I side with Buschfeld (2013) and Biewer (2011)), but rather a call for a more careful interpretation of the theoretical notions (cf. Biewer 2011: 11). While there can be no perfect model — after all, models are only a simplified representation of reality — a combination of the EFL–ESL–ENL distinction and the DM as proposed in Buschfeld (2013: 75–76) can be assumed to approximate actual language use in HKE and SgE as evidenced in the corpus data. What Buschfeld (2013: 75) suggests is a modification of the EFL–ESL–ENL distinction by, firstly, interpreting it as a continuum and, secondly, "explicitly tying in the development of identity constructions and the linguistic effects, namely the degree of nativization and subsequently institutionalization". In the case of HKE and SgE, it is precisely the degree of nativization that determines the probability of VNC. While HKE and SgE may diachronically have started out as contact varieties emerging from similar settings, the sociolinguistic contexts now differ considerably. The current investigation has provided indications that the less advanced variety, HKE, shows trends and processes typical of acquisition settings, such as the *shortest path principle* or a stronger preference for the less complex word-formation process in general, so that it can be assumed that ESL varieties located at an early stage of the DM are similar to learner varieties. However, the picture seems to change with increasing institutionalization. For VNC, SgE, an advanced ESL variety, does not resemble learner varieties but rather displays a usage pattern more similar to the native variety. The present study thus supports the view that an increase in institutionalization can instigate change, more precisely, a transition from an EFL-like to an ENL-like variety (cf. Buschfeld 2013: 191). This ultimately necessitates that the notion of ESL be understood as (part of) a continuum and that HKE and SgE are to be located at opposite ends. As a consequence, "varietal types should be approached in an integrated fashion" (Edwards & Laporte 2015: 163), encouraging the investigation of similar processes and innovations in EFL/ESL and ESL/ENL varieties (as in e.g. Deshors 2014). This paper has hopefully shown how this aim is achieved by studying an innovation and how it is affected by ongoing institutionalization in two SEA varieties of English.

Acknowledgements

For their comments on an earlier version of this article, I am indebted to Christian Mair and two anonymous reviewers. This work was carried out while the author was part of the Research Training Group "Frequency effects in language" at the University of Freiburg, Germany, funded by the German Research Foundation DFG (GRK 1624).

References

Ansaldo, U. 2004. "The evolution of Singapore English. Finding the matrix". In L. Lim (Ed.), *Singapore English. A Grammatical Description*. Amsterdam: John Benjamins, 127–149.

Ansaldo, U. 2009. "The Asian typology of English: Theoretical and methodological considerations", *English World-Wide* 30(2), 133–148. https://doi.org/10.1075/eww.30.2.02ans

Aronoff, M. 1976. *Word Formation in Generative Grammar*. Cambridge: MIT Press.

Balteiro, M. 2007. *The Directionality of Conversion in English. A Dia-Synchronic Study*. Bern: Peter Lang.

Bao, Z. 2005. "The aspectual system of Singapore English and the systemic substratist explanation", *Journal of Linguistics* 41(2), 237–267. https://doi.org/10.1017/S0022226705003269

Bao, Z. 2009. "*One* in Singapore English", *Studies in Language* 33(2), 338–365. https://doi.org/10.1075/sl.33.2.05bao

Bao, Z. 2010a. “A usage-based approach to substratum transfer: The case of four unproductive features in Singapore English”, *Language* 86(4), 792–820.

Bao, Z. 2010b. “*Must* in Singapore English”, *Lingua* 120(7), 1727–1737. https://doi.org/10.1016/j.lingua.2010.01.001

Bates, D., Maechler, M., Bolker, B. & Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package Version 1.1–7. Available at http://cran.r-project.org/package=lme4 (accessed March 2016).

Bernaisch, T. 2015. *The Lexis and Lexicogrammar of Sri Lankan English*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g54

Biermeier, T. 2008. *Word-Formation in New Englishes. A Corpus-based Analysis*. Münster: Lit.

Biewer, C. 2011. “Modal auxiliaries in second language varieties of English: A learner’s perspective”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 7–34. https://doi.org/10.1075/scl.44.02bie

Bolton, K. 2000. “The sociolinguistics of Hong Kong and the space for Hong Kong English”, *World Englishes* 19(3), 265–285. https://doi.org/10.1111/1467-971X.00179

Bolton, K. 2003. *Chinese Englishes. A Sociolinguistic History*. Cambridge: CUP.

Bolton, K. 2012. “Language policy and planning in Hong Kong. The historical context and current realities”. In E. Low & A. Hashim (Eds.), *English in Southeast Asia. Features, Policy and Language in Use*. Amsterdam: John Benjamins, 221–238. https://doi.org/10.1075/veaw.g42.18bol

Buschfeld, S. 2013. *English in Cyprus or Cyprus English. An Empirical Investigation of Variety Status*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g46

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511750526

Cannon, G. 1985. “Functional shift in English”, *Linguistics* 23, 411–431. https://doi.org/10.1515/ling.1985.277.1.411

Census and Statistics Department, Hong Kong SAR. 2013: online. Thematic Household Survey – Report No. 51. Use of Language in Hong Kong. Utilisation of Child Health and Family Planning Services Provided by Maternal and Child Health Centres. Available at http://www.statistics.gov.hk/pub/B11302512013XXXXB0100.pdf (accessed March 2016).

China Labor News Translations. n.d. About. Available at http://www.clntranslations.org/about/ (accessed February 2015).

Chui, S. 2010. “Sensitivity to differences between speech and writing: Hong Kong students’ use of syntactic features in English”. PhD thesis, The Chinese University of Hong Kong.

Collins, P. & Yao, X. 2013. “Colloquial features in World Englishes”, *International Journal of Corpus Linguistics* 18(4), 479–505. https://doi.org/10.1075/ijcl.18.4.02col

Davies, M. 2013. *Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries*. Available at http://corpus2.byu.edu/glowbe/ (accessed March 2016).

Davies, M. & Fuchs, R. 2015. “Expanding horizons in the study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (GloWbE)”, *English World-Wide* 36(1), 1–28. https://doi.org/10.1075/eww.36.1.01dav

Deshors, S. 2014. “A case for a unified treatment of EFL and ESL: A multifactorial approach”, *English World-Wide* 35(3), 277–305. https://doi.org/10.1075/eww.35.3.02des

Don, J., Trommelen, M. & Zonneveld, W. 2000. “Conversion and category indeterminacy”. In G. Booij, C. Lehmann & J. Mugdan (Eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung*, Vol. 1. Berlin: De Gruyter, 943–952.

Edwards, A. & Laporte, S. 2015. "Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels", *English World-Wide* 36(2), 135–169. https://doi.org/10.1075/eww.36.2.01edw

Evans, S. 2010. "Business as usual: The use of English in the professional world in Hong Kong", *English for Specific Purposes* 29(3), 153–167. https://doi.org/10.1016/j.esp.2009.11.005

Evans, S. 2015. "Word-Formation in Hong Kong English: Diachronic and synchronic perspectives", *Asian Englishes* 17(2), 116–131. https://doi.org/10.1080/13488678.2015.1036510

Gilquin, G. 2015. "At the interface of contact linguistics and Second Language Acquisition research. *New Englishes and Learner Englishes compared", English World-Wide* 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Gilquin, G. & Granger, S. 2011. "From EFL to ESL. Evidence from the International Corpus of Learner English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 55–78. https://doi.org/10.1075/scl.44.04gra

Gisborne, N. 2009. "Aspects of the morphosyntactic typology of Hong Kong English", *English World-Wide* 30(2), 149–169. https://doi.org/10.1075/eww.30.2.03gis

Gries, S. 2015. "The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models", *Corpora* 10(1), 95–125. https://doi.org/10.3366/cor.2015.0068

Gut, U. 2009. "Past tense marking in Singapore English verbs", *English World-Wide* 30(3), 262–277. https://doi.org/10.1075/eww.30.3.02gut

Heller, B. & Röthlisberger, M. 2015. "Big data on trial. Researching syntactic alternations in GloWbE and ICE". Paper presented at *From data to evidence. Big data, rich data, uncharted data*, University of Helsinki, 19–22 October 2015.

Horch, S. 2017. *Conversion in Asian Englishes. A usage-based account of the emergence of new local norms*. Freiburg: Albert-Ludwigs University. <https://freidok.uni-freiburg.de/data/12910> https://doi.org/10.6094/978-3-928969-68-0

Johnson, R. 1994. "Language policy and planning in Hong Kong", *Annual Review of Applied Linguistics 1993*/1994 (14), 177–199. https://doi.org/10.1017/S0267190500002889

Kachru, B. 1985. "Standards, codification and sociolinguistic realism: The English language in the Outer Circle". In R. Quirk & H. Widdowson (Eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: CUP, 11–30.

Kastovsky, D. 1982. *Wortbildung und Semantik*. Düsseldorf: Schwann-Bagel.

Kirkpatrick, A. 2012. "Theoretical issues". In E. Low & A. Hashim (Eds.), *English in Southeast Asia. Features, Policy and Language in Use*. Amsterdam: John Benjamins, 13–31. https://doi.org/10.1075/veaw.g42.04kir

Laporte, S. 2012. "Mind the gap! Bridge between World Englishes and Learner Englishes in the making", *English Text Construction* 5(2), 264–291. https://doi.org/10.1075/etc.5.2.05lap

Leimgruber, J. 2013a. *Singapore English. Structure, Variation, and Usage*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139225755

Leimgruber, J. 2013b. "The management of multilingualism in a city-state: Language policy in Singapore". In P. Siemund, I. Gogolin, M. Schulz & J. Davydova (Eds.), *Multilingualism and Language Diversity in Urban Areas. Acquisition, Identities, Space, Education*. Amsterdam: John Benjamins, 227–256. https://doi.org/10.1075/hsld.1.12lei

Lim, L. 2010. "Migrants and 'Mother Tongues': Extralinguistic forces in the ecology of English in Singapore". In L. Lim, A. Pakir & L. Wee (Eds.), *English in Singapore. Modernity and Management*. Hong Kong: Hong Kong University Press, 19–54. https://doi.org/10.5790/hongkong/9789888028436.003.0002

Mair, C. 2015. "Response to Davies and Fuchs", *English World-Wide* 36(1), 29–33. https://doi.org/10.1075/eww.36.1.02mai

Marchand, H. 1960. *The Categories and Types of Present–Day English Word-Formation. A Synchronic-Diachronic Approach*. Wiesbaden: Otto Harrassowitz.

Matthews, S. & Yip, V. 1994. *Cantonese. A Comprehensive Grammar*. London: Routledge.

Mukherjee, J. 2015. "Response to Davies and Fuchs", *English World-Wide* 36(1), 34–37. https://doi.org/10.1075/eww.36.1.02muk

Mukherjee, J. & Gries, S. 2009. "Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English", *English World-Wide* 30(1), 27–51. https://doi.org/10.1075/eww.30.1.03muk

Mukherjee, J. & Hundt, M. (Eds.). 2011. *Exploring Second-Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.44

OED Online. 2015. Oxford: OUP. Available at http://www.oed.com (accessed March 2016).

Pang, T. 2003. "Hong Kong English: A stillborn variety?", *English Today* 19(2), 12–18. https://doi.org/10.1017/S0266078403002037

Pavesi, M. 1998. "'Same word, same idea': Conversion as a word formation process", *International Review of Applied Linguistics in Language Teaching* 36(3), 213–231.

Peters, P. 2015. "Response to Davies and Fuchs", *English World-Wide* 36(1), 41–44. https://doi.org/10.1075/eww.36.1.02pet

Plag, I. 2003. *Word–Formation in English*. Cambridge: CUP. https://doi.org/10.1017/cbo9780511841323

Po-Ching, Y. & Rimmington, D. 2004. *Chinese. A Comprehensive Grammar*. London: Routledge.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1972. *A Grammar of Contemporary English*. London: Longman.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Version 3.2.0. Wien: R Foundation for Statistical Computing. Available at http://www.r-project.org (accessed March 2016).

Schneider, E.W. 2003. "The dynamics of New Englishes: From identity construction to dialect birth", *Language* 79(2), 233–281. https://doi.org/10.1353/lan.2003.0136

Schneider, E.W. 2007. *Postcolonial English. Varieties Around the World*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511618901

Schneider, E.W. 2014a. "Asian Englishes – into the future: A bird's eye view", *Asian Englishes* 16(3), 249–256. https://doi.org/10.1080/21639159.2014.949439

Schneider, E.W. 2014b. "New reflections on the evolutionary dynamics of World Englishes", *World Englishes* 33(1), 9–32. https://doi.org/10.1111/weng.12069

Seoane, E. & Suárez-Gómez, C. 2013. "The expression of the perfect in East and South-East Asian Englishes", *English World-Wide* 34(1), 1–25. https://doi.org/10.1075/eww.34.1.01seo

Setter, J., Wong, C. & Chan, B. 2010. *Hong Kong English*. Edinburgh: Edinburgh University Press.

Sridhar, K. & Sridhar, S. 1986. "Bridging the paradigm gap: Second language acquisition theory and indigenized varieties of English", *World Englishes* 5(1), 3–14. https://doi.org/10.1111/j.1467-971X.1986.tb00636.x

Strang, B. 1970. *A History of English*. London: Routledge.

Sun, C. 2006. *Chinese. A Linguistic Introduction*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511755019

Tan, Y. 2014. "English as a 'Mother Tongue' in Singapore", *World Englishes* 33(3), 319–339. https://doi.org/10.1111/weng.12093

The ICE Project. 2014. *International Corpus of English*. Available at http://ice-corpora.net/ice/index.htm (accessed July 2014).

Traugott, E. & Trousdale, G. 2013. *Constructionalization and Constructional Changes*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199679898.001.0001

Van Rooy, B. 2011. "A principled distinction between error and conventionalized innovation in African Englishes". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 189–207. https://doi.org/10.1075/scl.44.10roo

Wald, B. 1993. "On the Evolution of *would* and other modals in the English spoken in east Los Angeles". In N. Dittmar & A. Reich (Eds.), *Modality in Language Acquisition*. Berlin: De Gruyter, 59–96.

Wee, L. 2013. "Governing English in Singapore. Some challenges for Singapore's language policy". In L. Wee, R. Goh & L. Lim (Eds.), *The Politics of English. South Asia, Southeast Asia and the Asia Pacific*. Amsterdam: John Benjamins, 105–124. https://doi.org/10.1075/wlp.4.09wee

Werner, V. 2014. *The Present Perfect in World Englishes. Charting Unity and Diversity*. Bamberg: University of Bamberg Press.

Williams, J. 1987. "Non-native varieties of English: A special case of language acquisition", *English World-Wide* 8(2), 161–199. https://doi.org/10.1075/eww.8.2.02wil

Wong, W. 2011. *Census of Population 2010. Statistical Release 1: Demographic Characteristics, Education, Language and Religion*. Singapore: Department of Statistics, Ministry of Trade and Industry, Republic of Singapore.

Appendix

It has to be noted that this list was compiled prior to the corpus analysis. It might therefore be the case that some verbs change category from one variety to the other. Nonetheless, as frequency is modeled as a continuous predictor, this does not interfere with the significance of the results.

Table 3. Randomly selected verbs and corresponding deverbal nouns

high frequency		**low frequency**	
verb	**deverbal noun**	**verb**	**deverbal noun**
allow	allowance	approve	approval
choose	choice	calculate	calculation
consider	consideration	deny	denial
continue	continuation	distribute	distribution
create	creation	examine	examination
develop	development	expand	expansion
improve	improvement	imagine	imagination
provide	provision	possess	possession
refer	reference	satisfy	satisfaction
require	requirement	specify	specification

Table 4. Random effect of generalized linear mixed model

Groups	Name	Variance	Standard Deviation
lexeme	(Intercept)	1.071	1.035

Number of observations: 60, groups: lexeme, 20

The fate of linguistic innovations

Jersey English and French learner English compared

Anna Rosen
University of Freiburg

Drawing on spoken corpus data, this study traces the emergence and development of Norman French-influenced innovations in the nativised L2 variety of Jersey English and compares them to features in the speech of French-speaking learners of English. The comparison shows that such innovations do not differ from errors in a learner variety on a formal linguistic level and that they arguably result from the same processes as are present in foreign language acquisition, such as transfer or simplification. The paper therefore argues that innovations can only be identified reliably in retrospect, once they are more widely accepted in the speech community. It also points to the social factors that are crucial in shaping the use and probable fates of former innovations in Jersey English and suggests a typology of innovations according to their developments.

Keywords: linguistic innovation, error, Jersey English, French learner English, transfer

1. Introduction

Jersey English (JersE), a small and lesser-known variety[1] on the periphery of continental Europe, has emerged over recent centuries in a linguistic and cultural contact setting between Norman French, standard French and various varieties of (mainly) British English. Jersey's somewhat remote geographical location, in

[1] This paper uses a wide definition of 'variety' as an umbrella term for any form of native or non-native English that has a typical feature pool used by a group of speakers (see also van Rooy 2011: 290 for a similar definition). It will usually be specified whether the term refers to a native or a learner variety, to English as a Foreign Language (EFL) or English as a Second Language (ESL). Note that no attempt is made here to differentiate between terms such as 'variety' or 'dialect' for the JersE context nor to solve the problem of which new forms of a language truly constitute an independent variety in their own right or should simply be seen as variants.

https://doi.org/10.1075/bct.98.08ros

combination with diverse linguistic influences and its economic transformation from an agricultural society into a centre of international finance, make the English variety spoken there an interesting showcase for the origin and development of linguistic innovations in non-native Englishes. This in turn provides a useful starting point for a more general discussion of such notions as 'innovation', 'error' and 'norm'.

'Innovations' in JersE are here defined as features that have arisen in the specific contact situation of the island, differ from a southern British spoken norm in terms of grammar or pragmatics and have been conventionalised in such a way that they occur in the speech of more than individual speakers in a relatively small corpus. In this sense, the definition follows Kachru's (1982: 45) notion of 'deviation':

> it is different from the norm in the sense that it is the result of the new 'un-English' linguistic and cultural setting in which the English language is used; it is the result of a productive process which marks the typical variety-specific features; and it is systemic within a variety, and not idiosyncratic.

The definition of innovation used in this paper also agrees with van Rooy's (2011: 189) proposal of the criteria "grammatical stability" and "acceptability" used in classifying some form as a conventionalised innovation.

Based on sociolinguistic interview and archive data, compiled into a 350,000-word corpus of spoken JersE analysed in detail in Rosen (2014), this paper offers a historical perspective on linguistic innovations in a formerly non-native English variety. It traces the probable origin and development of contact-induced grammatical features of JersE, in particular a Verb-*and*-Verb construction in which the second verb always appears in the infinitival form, existential *there's* with time reference and the pragmatic particle *eh*, as illustrated in examples (1), (2) and (3), and compares these to data from advanced French-speaking learners of English taken from the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin et al. 2010).

(1) I went out and see him. (JIC28m1979)

(2) There's sixty years we're married. (JIC12f1935)

(3) It's good for the cattle, eh? (JIC06f1933)

In doing so, the paper has three central aims. First, it intends to show that both innovations and what have usually been termed errors in EFL research probably result from identical cognitive and psycholinguistic processes (see also Biewer 2011: 13; Mesthrie & Bhatt 2008: 159–167; van Rooy 2011: 193) and are difficult, if not impossible, to distinguish on a purely formal linguistic level. Many JersE features also occur in the spoken language of French-speaking learners of English, in which setting they would traditionally be considered errors. Thus, the difference between the notions of (not yet conventionalised) innovations on the one hand and errors on the other seems to be terminological and attitudinal — a matter of perspective and norm-orientation rather than a linguistic difference. To a certain extent, then, this paper is in the same vein as the contributions to Mukherjee & Hundt (2011), in which learner Englishes and second-language varieties are described and compared on an empirical basis. The study also demonstrates that, despite obvious differences between EFL and ESL in terms of learning settings, goals, motivation and norm orientation (see Biewer 2011: 9–13; Kachru & Nelson 2006: 26; Mukherjee & Hundt 2011: 212–213; Sridhar & Sridhar 1986), there are enough parallel features to make us reconsider a simple distinction between labels such as 'innovations' on the one and 'errors' on the other hand.

The second aim of the paper is to show that conventionalised innovations can only be identified reliably in hindsight — once they have caught on and are consistently used in the speech community. Unlike in other domains of life, where innovations would perhaps be defined as new creative forms intended as such, it is often very difficult to determine, at least from transcribed corpus material only, if a speaker intended to be innovative and create a new form or if he or she simply missed the appropriate form in the target language and made a mistake. The JersE corpus data allow us to establish, in retrospect, which linguistic phenomena deviating from British norms can be considered JersE innovations rather than errors made by individual speakers in a group learning process, as they became conventionalised and widely accepted in the JersE speech community.

At the same time, the corpus data combined with a qualitative approach that relies on intimate background knowledge of participants and linguistic material can be used to identify — and this is the third aim of this paper — those processes and social factors which are important in shaping the use and fate of these former innovations. A typology of JersE innovations according to their paths of development will be suggested. As will be seen, such conventionalised innovations can develop very differently in terms of speaker recognition or identity-creating potential and their chances of survival. Some of these features go unnoticed by speakers of the JersE speech community, some can be used consciously to signal local identity and some are assigned (often affectionately) to older traditional islanders only.

After a brief introduction to JersE and its sociolinguistic setting, the data and methods used in this study will be presented. In Section 4, conventionalised innovations in JersE will be presented one by one and compared systematically to data from LINDSEI. For this comparison, all features that comply with the definition of innovation in this paper have been extracted from the complete morpho-syntactic repertoire of contact-induced features of JersE established in Rosen (2014: 176–177), which is based on the spoken JersE corpus and on feature lists in Viereck

(1988) and Jones (2010). The three features which will be examined in more detail in this section have been selected as they illustrate a) different possible outcomes of such a comparison and b) different pathways of development reflected in varying sociolinguistic distributions and potentials for creating identity among speakers of JersE today. The paper concludes with a discussion of the results in the light of the notions of innovations and errors and suggests a preliminary classification of JersE innovations according to possible 'fates'.

2. Jersey English and its sociolinguistic setting

With 99,500 inhabitants and 118 km^2, Jersey is the largest of the Channel Islands, which are positioned between England and France in the Bay of St. Malo and have been associated with the English Crown since 1066. Today, the Islands have a special status as dependencies of the British Crown in Europe and are practically self-governing, apart from matters of diplomatic representation and defence. Although there has been increasing Anglicisation since the 19th century, Norman French was still widely spoken, particularly in the rural parishes, until the beginning of the 20th century. The transformation into a monoglot English society is by now, however, almost complete, with only 2 to 3% of the population still speaking insular Norman. All of the latter are bilingual speakers of Norman French and English, the vast majority of them using English in most domains of everyday life. The English variety spoken on Jersey can therefore be described today, following Mesthrie's (1992) terminology, as a nativised L2 variety.[2]

In general, however, it is not easy to place JersE within models of World or New Englishes. For example, it is neither formally captured by Kachru's (1985) Three Concentric Circles model nor by Schneider's (2003, 2007) Dynamic Model, as Jersey is not a nation-state and does not qualify for the label of 'postcolonial English' given its political status and history. Although Jersey used to be a trading outpost and had English troops garrisoned there for some centuries, its political autonomy (despite proximity to Britain) and its migration structures within a European context differentiate it from other places where New Englishes have developed. A significant continuing influx of immigrants from Britain, Poland and Madeira, as well as the island's transformation into an international finance centre, have led to ongoing dialect and language contact. The politically and historically oriented definitions of Inner and Outer Circle varieties in Kachru's model, as described for instance in Kachru & Nelson (2006), do not admit easy assignment of

[2] For more detailed background information on Jersey English see Rosen (2014: 25–42); on the Channel Islands in general see Jones (2010).

JersE. Others have pointed out similar problems in classifying varieties according to Kachru's model (see e.g. Bongartz & Buschfeld 2011: 50; Bruthiaux 2003: 159, 166; Mukherjee & Hundt 2011: 210–211). Bruthiaux (2003: 162) also criticises how the model does not take into account variation within varieties. This is also relevant for JersE, which is not a very uniform dialect. Instead, as in many other dialect regions, speakers born on Jersey can be positioned along a continuum of more standard and more traditional dialect users (Rosen 2014: 206). Krug & Rosen (2012) argue for two opposing sources of pressure exerted upon JersE speakers: an exonormative British English standard that guides speakers' choices in more formal situations and local norms that are linked to informal spoken encounters and specific speaker networks. These local norms are in no way formally codified.

Schneider's (2003, 2007) Dynamic Model, although it is strictly speaking only geared towards postcolonial Englishes, draws from the frameworks of identity theory, language contact and accommodation theory and takes the speech community rather than the nation-state as its basis. It is thus more easily applicable to the situation of JersE, which emerged in a similar contact situation to some postcolonial varieties.[3] According to this model (for a brief synopsis, see Schneider 2007: 56), JersE has clearly completed phase 3 ('nativization') and touches on the next two phases ('endonormative stabilization' and 'differentiation') as regards identity construction and linguistic attitudes. The island of Jersey is perhaps too close to the UK, geographically, culturally and in its education system, for the codification endeavours typical of phase 4 ('endonormative stabilization'). As mentioned before, JersE is not very homogeneous and is probably too small for proper dialectal differentiation within the variety — yet group-specific linguistic behaviour can be found, which is prototypical of Schneider's phase 5 ('differentiation'). Thus, the contact and language-shift situation in Jersey, where speakers of English as a native language and (to a much lesser degree) as a second language live side by side, makes the variety spoken there a good test case for the investigation of linguistic innovations.

3. Data

The corpus of JersE was originally compiled and analysed for a larger project on grammatical variation and change in JersE (Rosen 2014). During fieldwork in 2008, 40 sociolinguistic interviews with speakers who were born and grew up in

[3] Schneider (2003: 235) himself emphasises that what is important for his model to apply is "not the colonial history or the former colonial status", but the specific type of contact situation that typically arises in such circumstances.

Jersey were conducted and then transcribed. The choice of participants forms a balanced sample of speakers from three age-groups, 20 male and 20 female speakers, with the oldest speaker group being equally divided into 10 monolingual and 10 bilingual, i.e. Norman French and English, speakers (see Table 1). All subgroups contain speakers from various educational and occupational backgrounds and from different parts of the island. To also capture diachronic variation, recordings from two oral history projects with 20 speakers born another twenty to thirty years earlier than those in the oldest group of the 2008 data were collected and transcribed. The recordings of both projects are stored in the Jersey Archive and differ in formality of speech, which is why they are divided into the more informal *Jersey Archive Corpus* (JAC) with 13 speakers and a more formal additional component (JACa) with 7 speakers. The components of the JersE corpus comprise a total of approximately 350,000 words: the *Jersey Interview Corpus* (JIC) includes 267,845 words, the JAC contains 39,790 words and its additional component JACa contains 46,537 words.

Table 1. Speaker numbers of JIC, JAC(a) and LINDSEI-FR

	JIC				**JAC(a)**	**LINDSEI-FR**
	monolingual			**bilingual**		
age group	**20–39**	**40–59**	**60+**	**60+**		
male	5	5	5	5		
female	5	5	5	5		
total	10	10	10	10	20	46

In addition, questionnaire data from an acceptability study with the 2008 interview participants and data taken from the spoken portion of the *British National Corpus* (BNC) and from the *International Corpus of English* (ICE) are drawn on, in order to gauge speaker's reactions and attitudes towards typical JersE features and to determine the extent to which JersE features differ quantitatively or qualitatively from those in other English varieties.[4]

The database for the present study therefore fulfils, at least partially, the desiderata put forward in other studies addressing the paradigm gap between EFL and ESL. First, as it also contains older oral history data, there is a diachronic component to it that helps to trace the development of features in JersE, although it would be even better for an exploration of innovations to have some language data from the earliest stages of the new variety (see Szmrecsanyi & Kortmann 2011: 185 on

[4] Detailed information on the data used for this study and how they were collected can be found in Rosen (2014: 43–68).

missing corpora of early stages of L2 varieties). Second, the JersE corpus also has the advantage of containing very detailed information on the individual speakers and their social and linguistic background, something that is wanting in most corpora of New Englishes, as Gilquin (2015: 18) criticises. And last, the addition of attitudinal data obtained in the acceptability study helps us to shed further light on the difference between conventionalised innovations and errors, just as Mukherjee & Hundt (2011: 216–217) suggest.

In order to examine the extent to which former JersE innovations also feature in the speech of advanced French-speaking learners of English, the data of native speakers of French who also use French at home were extracted from LINDSEI. The learner contributions from 46 such speakers, all of them university students who, on average, had 4.6 years of English at school and 3.75 years of English at university, amount to 83,589 words. This dataset will be called LINDSEI-FR in what follows. Note that regarding corpus comparability (see also Table 1), the JersE corpus and the LINDSEI data are not ideally suited, be it in corpus size, speaker characteristics, type of spoken language or — most importantly perhaps — in the linguistic background of Norman French versus (standard) French speakers (see also Davydova 2012: 383–384 and Gilquin 2015: 118 on the problems of comparing ESL to learner corpora). It should thus be borne in mind that this comparison only provides a first approach to the claims investigated in this paper. At the same time, support for the claims will be even more convincing if a comparison of these corpora nevertheless reveals parallel features in both the learner and the JersE varieties.

Both the JersE corpora and LINDSEI-FR were searched for the features under investigation using the concordance program *WordSmith Tools 5* (Scott 2008). Where necessary, result lists were sorted manually to include only the features in question. To enable meaningful comparisons between the corpora and between speaker groups of the JIC, frequencies were usually converted into normalised (per 10,000 words) frequencies. Results were then analysed using descriptive statistics and, as far as the frequencies of occurrence allowed it, also tested for statistical significance and significant correlations.

4. Contact-induced innovations in JersE: Origins, developments and parallel features in LINDSEI-FR

As outlined above, this section compares contact-induced conventionalised innovations in JersE, all of them part of its morpho-syntactic repertoire for at last a century, probably much longer, with French learner data. It focuses in particular on the verb-*and*-verb construction FAP, existential *there's* followed by a time reference and the discourse particle *eh*. Feature by feature, their probable origin and modern-day usage will be summarised, based on more detailed analyses in Rosen (2014), and compared with equivalent structures in LINDSEI-FR. The section is rounded off with a list of less frequent contact-induced JersE features, their likely fate and, if existing, parallel forms in LINDSEI-FR.

4.1 The verb-*and*-verb construction FAP

One of JersE's distinct features compared to other English varieties[5] is a coordinated verb construction, as in *I went out and see him*. Barbé (1993: 110–138, 1995) was the first to carry out a thorough analysis of this feature in Guernsey English, which she termed FAP (**F**irst verb plus the conjunction *And* followed by the **P**lain infinitive). Unlike the standard English construction, FAP allows the use of an inflected verb form in the first position followed by an infinitive. Otherwise, FAP shares the same syntactic and semantic characteristics as its standard equivalent (Rosen 2014: 113–115). It has been argued that FAP emerged as a result of the specific contact-situation in the Channel Islands (Rosen 2014: 120–123). Norman French-speaking learners of English, probably influenced by syntactic patterns in their first language (which correspond to standard French *aller* + infinitive, *être en train de* + infinitive and *venir de* + infinitive), reanalysed and simplified standard coordinated verb structures and transformed them into FAP. Such a reanalysis may have been supported by the existence of standard patterns such as *I went to see him* and *I went and set/put/cut/etc.* with a potential for misreading. FAP structures would not have distorted communication dramatically, but may have been more convenient in referring to past events (see also Barbé 1993, 1995). This would also support Kachru & Nelson's (2006: 90) view that

> [t]he innovations in lexicon, grammar and discourse inspired by the primary languages [i.e. substrate languages, AR] contribute to building the learners' communicative competence in the target language on the one hand, and acculturation of the target language to the local context on the other.

In this sense, FAP can be considered "the result of a productive process", which is, according to Kachru (1982: 45), characteristic of deviations, and can thus be classified as a former innovation which is now conventionalised in JersE. Data from the acceptability questionnaire and from metalinguistic comments during fieldwork have shown, however, that islanders are not consciously aware of this feature and do not identify it as typically Jersey.

[5] For an analysis of ICE to confirm this, see Rosen (2014: 119).

Further support for the contact-induced origin of FAP comes from occurrences of similar structures used by French-speaking learners of English in LINDSEI. Although instances do not relate to the past and do not contain the most frequently used first verbs in JersE, namely forms of *go* and *come*, the four FAP-like forms from LINDSEI-FR in (4)–(7) do apparently indicate that such structures are difficult even for advanced learners and can emerge in simplification processes, perhaps inspired by verb patterns in French varieties.

(4) he tries and . . make it up (LINDSEI-FR 004)

(5) the woman stands up and say (LINDSEI-FR 013)

(6) he tries . and and do his be= and . to do his best (LINDSEI-FR 028)

(7) they're just . walking around and say okay . here (LINDSEI-FR 040)

Pauses between the verbs and the conjunction *and* as in (4) and (6), indicated by dots, and the self-repair in (6) additionally point to the learners' struggling with this construction.

Just as in the LINDSEI data, FAP structures occur only rarely in the JersE corpus, 25 times in the JIC and 9 times in the JAC. Yet, despite these low frequencies, a surprisingly clear sociolinguistic pattern surfaces (see Table 2): FAP occurs predominantly in the oldest speaker group and this seems to be a trend that is confirmed by the archive data, in which FAP structures are used more than twice as often as in the data recorded some 20 years later. Younger women do not use FAP at all in the JIC. Instead, it is mostly used by bilingual speakers and by monolingual speakers who move in close-knit rural networks where some members still speak Norman-French. All FAP-users share the same socio-economic background, mostly farming, and all grew up — and for the most part still live — in one of the rural parishes of Jersey.

Table 2. Distribution of FAP by speaker groups in JIC and JAC (absolute figures, normalised frequencies per 10,000 words in brackets; adapted from Rosen 2014: 115)

	JIC				**JAC**
	monolingual			**bilingual**	
age group	**20–39**	**40–59**	**60+**	**60+**	
male	1	2	2	7	
female	0	0	1	12	
total	25 (0.93)				9 (2.26)

The only younger speaker in whose data FAP occurs comes from a traditional Jersey family, has grandparents who still speak Jersey French and used to work on his parents' farm after he left school. Interestingly, another male speaker from the same age group with exactly the same kind of background does not use FAP or any other contact-induced features. In their interviews, the two speakers reveal quite different attitudes towards Jersey: Whereas the FAP-user is very attached to rural life in Jersey and cannot imagine living anywhere else, the non-user identifies less strongly with Jersey and would prefer to live elsewhere, were it not for his elderly parents. On the other hand, younger speakers with a strong sense of local identity but with different network structures and/or a higher socio-economic background do not use FAP either. A strong local identification with, and positive attitudes towards, Jersey then seem to be necessary prerequisites for the continued use of this former innovation, but have to be accompanied by further social constellations. Although frequencies in the corpus are low, this sociolinguistic distribution can also be found with other contact-induced features of JersE such as existential *there's* + time reference, described immediately below.

4.2 Existential *there's* with time reference

In JersE, existential *there's* can be combined with a time reference and, typically, a sentence in the present tense as illustrated in examples (2) and (8)–(9), all taken from the JIC.

(8) there's quite a few years now that there's not much growing (JIC10m1928)

(9) And believe it or not, there's only about three weeks now that the dog went home (JIC04f1904)

This is clearly a syntactic calque from Norman French where *y'a* is followed by a time reference (as *il y a* in standard French) to refer to the period of time that has elapsed since the occurrence of an event (e.g. Jones 2001: 168). There are no tokens of this or a similar feature in LINDSEI-FR. Simple explanations for this could be the small corpus size and the topics of discussion or the advanced stage of the learners where long years of instruction override any impulse to use such a structure. It could arguably also testify to the possibility that this construction is a unique innovation of Channel Island English.

Metalinguistic comments after the interviews by some speakers reveal the feature to be seen as typically JersE and predominantly associated with traditional island speech. Corpus data confirm this. Nearly all 15 occurrences of this feature in the JIC are produced by bilingual speakers while the two monolingual users of this construction both had Norman French-speaking parents, still move in dense rural networks and profess a strong local identity.[6]

Thus, the overall sociolinguistic distribution is very similar to that of FAP. Although both features are evidently still in use today, their usage is in all likelihood drastically declining and only preserved within specific, i.e. rural and traditional, networks and apparently among speakers with a strong sense of Jersey identity.

4.3 The discourse particle *eh*

By contrast, the discourse particle *eh*, the third conventionalised innovation presented here in more detail, illustrates different trends with regard to its overall development and its sociolinguistic distribution. It can occur in a number of syntactic contexts and fulfil diverse pragmatic functions. Mainly, it serves as a tag after questions and statements asking for agreement, confirmation or an opinion from the listener, see example (10), as an emphasising device, see (11), or it helps the speaker to hold the floor or to establish a connection to the listener, see (12). The discourse particle *eh* is undoubtedly the feature of which speakers of JersE are most aware.

(10) It's a crazy world we live in, eh? (JIC09m1935)

(11) We're far better than the Guernsey ones, eh! (JIC14m1926)

(12) In the next bedroom was the cabin, eh, where we used to wash (JIC04f1935)

While *eh* is used in most, if not all, spoken English varieties, it still seems to be much more prominent in some varieties, such as Canadian, New Zealand and Scottish English. It can be shown, however, that there are quantitative and qualitative differences in the use of *eh* in JersE in comparison to other varieties, especially British English (Rosen 2014: 93–96), justifying its label as a JersE innovation. Innovations in the use of discourse particles have also been commonly found in other New Englishes, as discussed, for instance, in Mesthrie & Bhatt (2008: 136–140).

In Channel Island English, the high frequency of *eh* has been explained by contact with Norman French (Jones 2001: 169; Ramisch 1989: 106; Rosen 2014: 87–91) as Norman French *hé* and *hein* are frequently found in the same syntactic and pragmatic contexts.[7] Unfortunately, in LINDSEI brief filled pauses are transcribed

[6] There are also two instances of this feature in a diary written by Nan Le Ruez, a native of Jersey and bilingual in Jersey French and English, during the Second World War, which shows that this innovation had probably been conventionalised even then. Her diary was published in 1994 under the title *Jersey Occupation Diary.*

[7] Interestingly, some researchers have ascribed the use of *eh* in Canadian English, at least in part, to the influence of Canadian French (cf. Gold 2008). Yet, to date, no corpus study has been carried out to investigate this.

as <eh>, probably with a different pronunciation than the one for the JersE discourse particle, which is realised as /e/ or /eɪ/. Without the accompanying audio recordings, not available in the published version of the LINDSEI database, such fillers cannot be disambiguated from discourse particle *eh*. Example (13) from LINDSEI-FR, for instance, might be read as either form.

(13) it's an interesting town (eh) but I I wouldn't like to live there (LINDSEI-FR 035)

An analysis of the transcriptions, based on syntactic position, pragmatic context and co-occurrence of pauses and self-repairs, suggests that most of the <eh>-occurrences in LINDSEI are filled pauses, possibly realised as /ɜː/ or /ʌ/. 173 out of 1155 <eh>-occurrences, however, could potentially also represent the discourse particle *eh*. In addition, there is one instance of French *hein* in LINDSEI-FR as can be seen in (14).

(14) There are two coasts in Italy <foreign> hein <foreign> (LINDSEI-FR 015)

Generally, it has been shown that the use of discourse markers in learner Englishes can vary in frequency from native Englishes (see, e.g. De Cock et al. 1998; Gilquin 2015: 114) and that the French component of LINDSEI displays a more frequent use of *in fact* than other LINDSEI subcorpora, which has been attributed to transfer from French where the parallel discourse marker *en fait* exists (Gilquin et al. 2010: 60–61). In addition, Gilquin (2015: 116) states that 12 percent of foreign words occurring in LINDSEI are discourse markers. This again underlines the likelihood for an innovation of JersE *eh* due to the contact situation with Norman French and for similar processes being at work in both EFL and ESL varieties.

The JIC contains a total of 278 *eh*-tokens. When it comes to their sociolinguistic distribution, Rosen's (2014: 77–87) analysis shows three statistical tendencies. First, *eh*-use correlates significantly with age: the older the speakers, the more they use *eh* in their speech and the archive data confirm this trend. Second, *eh*-use also correlates significantly with education and occupation: *eh* occurs least frequently among speakers with a high level of education and a higher occupational background. And third, bilingual speakers in the JIC use *eh* five times more often than monolingual speakers. As the use of *eh* by individual speakers varies greatly, however, the results show no statistical significance in this case. Although the observed trends may well point to change in progress, a comparison with data from the spoken component of the BNC suggests that the particle *eh*, despite its decline, is still used more often in JersE, even among younger speakers (Rosen 2014: 93–96).

Comments during the interviews for the JIC, results from the acceptability study and an examination of the use of *eh* by Jersey journalists, comedians and singers equally suggest that *eh* can be used to signal — both subconsciously and consciously — a Jersey identity (Rosen 2014: 96–100). At the same time, *eh* is usually associated with the speech of traditional, older islanders and its use is only accepted as polite in informal and/or local contexts.

In sum, the particle *eh*, unlike the structures of FAP or *there's* followed by a time reference, is an innovation that the speech community is well aware of and can be considered a relatively stable feature of JersE although its overall development seems to suggest a gentle decline in its use, especially among younger and more educated speakers.

4.4 Further contact-induced innovations in JersE and parallel features in LINDSEI-FR

In what follows, a greater selection of contact-induced innovations in JersE, as established in Rosen (2014: 176–177), is checked against the French component of LINDSEI to discover possible similar features in both varieties. These JersE features have all been transferred or at least reinforced by transfer from Norman French, where parallel structures exist. In the realm of prepositional usage, transfer comes into play as the Norman French prepositions *à*, *en* and *siez* can express both destination and position so that JersE *to*, *in* and *at* can also indicate motion and location. All innovations but emphatic *that one* are obsolescent in today's JersE and are only used by JIC speakers above the age of 60. It should be noted, however, that frequencies are generally quite low and that the corpora are not ideally matched (see Section 3), so that this overview can only offer a first idea of the extent of parallelisms in JersE and an EFL variety.

At first glance, Table 3 displays some striking similarities between both datasets, with all but one feature in the domains of emphasis, prepositional usage and definite article use being attested in both corpora, albeit with varying frequencies. The exception is emphatic pronouns, which cannot be found in LINDSEI-FR, although numerous instances of repetitions of subject pronouns do occur such as *we we*, *she she*, *they they*. Without accompanying audio material, however, it is not possible to judge whether some of these instances could be similar in phenomenon to the emphatic use of pronouns in JersE, a stereotypical if obsolescent feature in Jersey, or if all of these occurrences are simply performance phenomena in Biber et al.'s terms (2002: 436). Prepositional usage in JersE and in the learner variety, on the other hand, seems to be surprisingly similar. Examples (15) and (16) illustrate these correspondences for the use of destinational *in* and *at*:

(15) we went in Italy (LINDSEI-FR 028)

(16) they come at the university at twenty (LINDSEI-FR 032)

Table 3. Overview of contact-induced innovations in JIC and JAC(a) and equivalent features in LINDSEI-FR (absolute frequencies; normalised frequencies per 10,000 words in brackets)

feature and JersE example from JIC	JIC + JAC(a)	LINDSEI-FR
emphatic pronoun use *But **me, I** didn't.*	14 (0.4)	0
emphatic *that one* *So he can talk for hours, **that one**.*	9 (0.25)	1 (0.12)
positional *to* *The girls stayed **to** Gorey*	3 (0.08)	1 (0.12)
destinational *in* *We went uh **in** France*	13 (0.37)	40 (4.79)
destinational *at* *She used to go **at** another farm*	12 (0.34)	3 (0.36)
locational *at* *we were sleeping **at** the bedroom*	2 (0.06)	1 (0.12)
on for 'in' *I didn't know what was **on** the letter.*	5 (0.14)	16 (1.91)
non-standard use of definite article *At home, we used to speak **the** Jersey French.* *And years ago, your potatoes all went to **the** town.*	41 (1.16)	23 (2.75)

A closer look at individual features and tokens, however, also reveals noteworthy differences. For instance, the frequency of occurrence of destinational *in* is significantly higher in LINDSEI-FR ($\chi^2 = 109.05$, df = 1, $p < 0.001$). While in JersE the use of *on* for 'in' only occurs in expressions with nouns like *letter*, *(news)paper*, *book*, where Norman French would have *sus* 'on', in LINDSEI-FR, only 12 out of 16 instances of *on*-use instead of 'in' are explicable by direct transfer, as in example (17) as opposed to example (18). In any case, this illustrates that prepositions present a formidable problem for the language learner.

(17) He seems to have (er) made (erm) . improvements on the on the painting (LINDSEI-FR 018)

(18) There was always some= somebody coming (er) on the[i:] evening (LINDSEI-FR 037)

In a similar vein, the use of the definite article in LINDSEI-FR not only differs from JersE in terms of frequency ($\chi^2 = 11.75$, df = 1, $p < 0.001$) but also partly in terms of context. Whereas the non-standard use of the definite article with adverbials of direction and position and with generic reference (including to institutions) can be found in both varieties, the data in LINDSEI-FR do not include definite article use in combination with names of languages. Thus, although the first impression of the overview presented in Table 3 suggests clear parallels between JersE innovations and features in a French-influenced learner variety of English, it is necessary to take the context of each occurrence into consideration; the low frequencies, too, only allow for a very cautious interpretation of the data. Divergences could also be anchored simply in differences between corpus size and content and between mainland French and insular Norman French.

5. Discussion

The JersE innovations presented in Section 4 all emerged in the specific contact situation of the Channel Islands and are arguably the result of similar processes as those found in foreign language acquisition, such as transfer, simplification and generalisation. A comparison of these innovative JersE features with French learner data from LINDSEI indeed gives an overall picture of close parallels between a former second language variety and a foreign language variety, despite some differences. Since Sridhar & Sridhar's (1986) influential paper on the paradigm gap between research on second language acquisition (SLA) and indigenised varieties of English, similarities and differences between EFL and ESL varieties have only recently been explored, often with an identification of numerous common forms and developments in both varieties that justifies a more unified treatment of EFL and ESL research (see, for instance, Biewer 2011: 28; Deshors 2014; Gilquin 2011; 2015: 116; Hilbert 2011: 141; Hundt & Vogel 2011: 161; Laporte 2012; Nesselhauf 2009). The present paper offers, with a comparison of JersE and LINDSEI data, another piece of evidence for the view that the gap between EFL and ESL research should be bridged: The results show that it is impossible to distinguish between innovations emerging in a new variety and so-called errors in EFL on a purely linguistic level. Linguistic structures by themselves, as could be seen for FAP-constructions or prepositional usage for example, are identical. It seems that it is attitudes, acceptability and norm-orientation which lie at the heart of the differences between the two. This is in line with Gut's (2011: 120–121) and Mukherjee & Hundt's (2011: 215) assessment. Learners' awareness of and orientation towards the external norm also manifests in self-repairs accompanied by hesitations and pauses in LINDSEI-FR (see examples in Section 4).

The three JersE features presented in more detail further demonstrate that a comparison between ESL and EFL varieties influenced by the same contact language might not always be straightforward. Whereas some features, such as FAP-constructions, seem to occur alike in JersE and a French learner variety, others, such as discourse particle *eh*, merely show that there are similar tendencies to be found in both varieties within a certain linguistic domain, in this case the one of discourse markers. The feature of existential *there's* with a time reference additionally illustrates that sometimes parallels cannot be attested in the data. This might be due to the nature and size of the corpora or, just as likely, it might be the result of the specific sociolinguistic and historical context on the one hand and of instruction on the other, which prevent the development of identical or similar structures in both varieties.

If, however, new forms which deviate from an external norm have a common starting point both in learner varieties and new Englishes and are usually not distinguishable on a linguistic level, it is only in retrospect that we can identify a feature as an innovation proper, i.e. once it has been accepted within the speech community and become conventionalised. The data from JersE allow us to filter innovative forms, which, as argued above, have emerged in group learning processes similar to individual SLA processes, because they spread within the speech community and are therefore used even today by more than individual speakers.

More importantly, additional information on the speakers recorded in the corpus and data taken from an acceptability questionnaire have proved essential in tracing the social factors which are important to the use and chances of survival of such former innovations, as well as in establishing a typology of JersE innovations according to patterns of development. The findings presented in the previous section clearly illustrate that former JersE innovations do not share the same fate. The current frequency distributions across age, social and linguistic background of JersE speakers suggest that the discourse particle *eh*, though socially stratified in its use, survives as a stereotypical and identity-signalling feature, whereas the use of both FAP and of existential *there's* followed by a time reference is decreasing and only occurs in very specific, i.e. rural and tight-knit, social networks. As shown in the analysis of individual speakers and their social settings, a strong sense of local identity seems to be a necessary precondition for the use of the latter innovations. Yet such identity alone does not necessarily lead to the use of local features. A higher educational and occupational background can usually override the identity factor. Unlike in some other speech communities (e.g. van Rooy 2011: 204), local norm-preservers are thus not educated speakers, as the educational domain in Jersey is heavily influenced by standards and teachers imported from the UK and islanders have to leave Jersey to enter higher education. On a more general level, the findings therefore also suggest that two pressures are exerted on the development and ultimate fate of these features: pride in and affection for local norms on the one hand, and on the other hand the pressure (especially in formal situations) for a standard which is also widely accepted outside the island. Figure 1 summarises this interplay

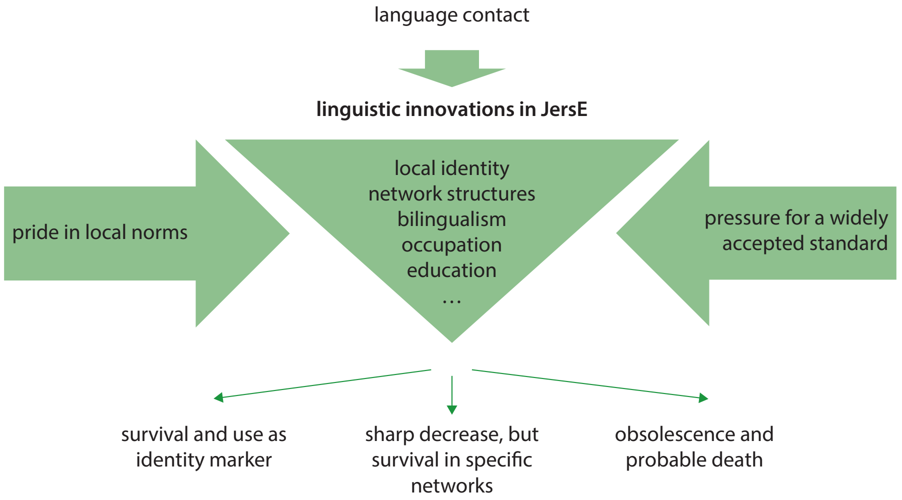

language contact
linguistic innovations in JersE
pride in local norms
local identity
network structures
bilingualism
occupation
education
…
pressure for a widely accepted standard
survival and use as identity marker
sharp decrease, but survival in specific networks
obsolescence and probable death

Figure 1. Factors determining the emergence and development of JersE innovations

of social factors, identity issues, linguistic awareness and network structures that determine the development of conventionalised innovations in JersE.

The term 'local identity' has here been used for islanders' identification with the way of life in Jersey and their Norman heritage and their wish to stay on the island and to belong to its community. Knowing about Jersey culture and knowing the local pronunciation of Norman words for family names, the island's toponymy and customs is one obvious way for islanders to show their belonging; another one is to continue to use innovative and therefore distinct forms of JersE. So far, however, there is no indication that any of the obsolescent features of JersE have been dusted off and recycled by a younger, better educated generation of speakers to signal their sense of local belonging. This has happened in other insular or contact varieties such as Cajun English, for example (Dubois & Horvath 2008).

The only feature at present that can be used in a conscious way to express identity by speakers of all ages and backgrounds is the discourse particle *eh*. This pragmatic marker is also the only innovation of JersE, among the ones presented here and more generally, that is comparatively frequently used, even among younger speakers, and serves as an identity marker. It is also the only innovation that all speakers of JersE are aware of and which comes first to their mind when asked about their way of speaking English. In Figure 2, an attempt to identify and exemplify types of grammatical and pragmatic innovations in JersE according to their developments, *eh* is therefore clearly in the category of surviving features.

The typology in Figure 2 represents a continuum of possible outcomes for contact-induced JersE innovations rather than clear-cut categories — with features that are surviving and used as identity markers at one end and those that

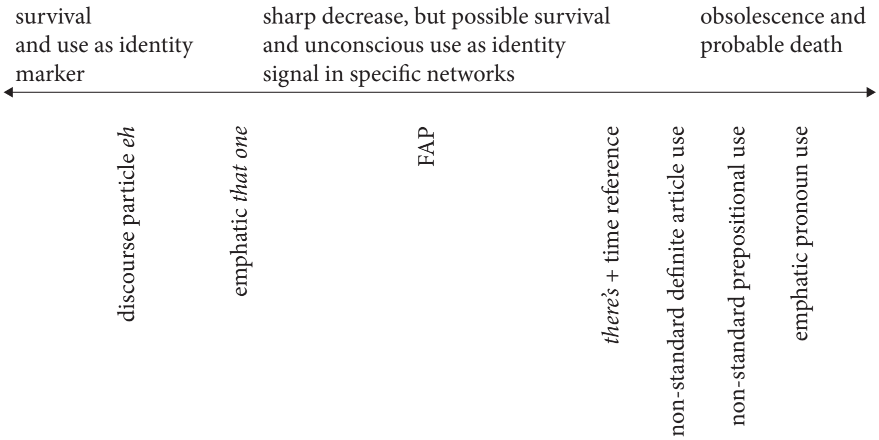

survival and use as identity marker
sharp decrease, but possible survival and unconscious use as identity signal in specific networks
obsolescence and probable death
discourse particle *eh*
emphatic *that one*
FAP
there's + time reference
non-standard definite article use
non-standard prepositional use
emphatic pronoun use

Figure 2. Types of contact-induced JersE innovations according to their development

are obsolescent and gradually dying out at the other end. It is noteworthy that the feature pool skews to the right and that not many former innovations can be said to be part of the repertoire of every JersE speaker today. Again, this might point to the growing pressure of an external norm that (especially well-educated) speakers face and can be taken as a visible indication that JersE is developing into a variety that is close to (southern) British English.

Although some social factors (such as education and occupational background, network structures and aspects of identity) can be filtered out in shaping the development of innovations, it is difficult to find decisive linguistic criteria that determine their fate. It seems that innovations that are also acceptable in other L1 varieties of English, such as *eh* or emphatic *that one*, have a greater chance of survival than features that involve the creation of a completely new structure, such as FAP or *there's* + time reference. Emphatic pronoun use, however, also occurring in the BNC as in "Me I'd rather try something different." (KB7 6893) or in Scottish English (Macaulay 1989), can be argued to show less deviation from a British norm than does FAP, yet it is undoubtedly obsolescent in JersE. Kachru (1982: 45–46, 48–49) and Nelson (1982) point out how difficult it is to define and apply the criterion of an acceptable extent of deviation from the norm, especially in terms of intelligibility. It is also a criterion that does not seem to be pivotal in the development of innovations in JersE. While a different use of prepositions might indeed render utterances more difficult to understand to an outsider of the speech community, there does not seem to be a great difference between emphatic pronoun use and FAP with regard to intelligibility, yet they must be placed at different points along the continuum. The explanation behind the different developments of innovations, at least in JersE, is thus more complex and probably lies in an intricate web of social aspects, attitudes, feature awareness and norm-orientations.

If linguistic factors, such as the extent of norm-deviation or intelligibility, seem to play a lesser role in the ultimate fate of these features, this again advocates that linguistic research should treat innovative features in ESL varieties on a par with so-called errors in EFL varieties. It also suggests, just as the general comparison of JersE and LINDSEI data does, that such research needs to focus more on the developmental, social and pragmatic context of innovative phenomena in ESL and EFL. For this, however, we would need data concerning early stages in the formation of ESL and EFL varieties (which also preserve the richness of metadata and information on individual speakers), data from larger-scale acceptability, awareness and intelligibility studies and, most importantly, directly comparable datasets for both types of varieties.

Acknowledgements

I wish to thank Luke Bradley, the editors and three anonymous reviewers for their helpful comments on an earlier version of this paper.

References

Barbé, P. 1993. *Exploring Variation in Guernsey English Syntax*. PhD dissertation, University of London.

Barbé, P. 1995. "Guernsey English: A syntax exile?", *English World-Wide* 16(1), 1–36. https://doi.org/10.1075/eww.16.1.02bar

Biber, D., Conrad, S. & Leech, G. 2002. *The Longman Student Grammar of Spoken and Written English*. Harlow: Pearson Longman.

Biewer, C. 2011. "Modal auxiliaries in second language varieties of English. A learner's perspective". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 7–33. https://doi.org/10.1075/scl.44.02bie

Bongartz, C. & Buschfeld, S. 2011. "English in Cyprus: Second language variety or learner English?". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 35–54. https://doi.org/10.1075/scl.44.03bus

British National Corpus. Version 4.3 (BNCweb CQP-edition). 2013. Developed by S. Hoffmann and S. Evert. Available at http://corpora.lancs.ac.uk/BNCweb/.

Bruthiaux, P. 2003. "Squaring the circles: Issues in modeling English worldwide", *International Journal of Applied Linguistics* 13(2), 159–178. https://doi.org/10.1111/1473-4192.00042

Davydova, J. 2012. "Englishes in the Outer and the Expanding Circles: A comparative study", *World Englishes* 31(3), 366–385. https://doi.org/10.1111/j.1467-971X.2012.01763.x

De Cock, S., Granger, S., Leech, G. & McEnery, T. 1998. "An automated approach to the phrasicon of EFL learners". In S. Granger (Ed.), *Learner English on Computer*. London: Longman, 67–79.

Deshors, S. 2014. “A case for a unified treatment of EFL and ESL: A multifactorial approach”, *English World-Wide* 35(3), 277–305. https://doi.org/10.1075/eww.35.3.02des

Dubois, S. & Horvath, B. 2008. “Cajun Vernacular English: Phonology”. In E. Schneider (Ed.), *Varieties of English 2. The Americas and the Caribbean*. Berlin: Mouton de Gruyter, 208–218.

Gilquin, G. 2011. “Corpus linguistics to bridge the gap between World Englishes and Learner Englishes”, *Communicación en el siglo XXI* 2: 638–642.

Gilquin, G. 2015. “At the interface of contact linguistics and second language acquisition research: New Englishes and learner Englishes compared”, *English World-Wide* 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Gilquin, G., De Cock, S. & Granger, S. (Eds.). 2010. *Louvain International Database of Spoken English Interlanguage* (CD-ROM + Handbook). Louvain-la-Neuve: Presses universitaires de Louvain.

Gold, E. 2008. “Canadian *eh*? From eh to zed”, *Anglistik* 19(2), 141–156.

Gut, U. 2011. “Studying structural innovations in New English Varieties”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 101–124. https://doi.org/10.1075/scl.44.06gut

Hilbert, M. 2011. “Interrogative inversion as a learner phenomenon in English contact varieties. A case of Angloversals?”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 125–143. https://doi.org/10.1075/scl.44.07hil

Hundt, M. & Vogel, K. 2011. “Overuse of the progressive in ESL and learner Englishes – fact or fiction?”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 145–165. https://doi.org/10.1075/scl.44.08vog

Jones, M.C. 2001. *Jersey Norman French: A Linguistic Study of an Obsolescent Dialect*. Oxford: Blackwell.

Jones, M.C. 2010. “Channel Island English”. In D. Schreier, P. Trudgill, E. Schneider & J. Williams (Eds.), *The Lesser-Known Varieties of English. An Introduction*. Cambridge: Cambridge University Press, 35–56. https://doi.org/10.1017/CBO9780511676529.004

Kachru, B. 1982. “Models for non-native Englishes”. In B. Kachru (Ed.), *The Other Tongue. English across Cultures*. Urbana, Chicago and London: University of Illinois Press, 31–57.

Kachru, B. 1985. “Standards, codification and sociolinguistic realism: The English language in the outer circle”. In R. Quirk & H. Widdowson (Eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.

Kachru, Y. & Nelson, C. 2006. *World Englishes in Asian Contexts*. Aberdeen and Hong Kong: Hong Kong University Press.

Krug, M. & Rosen, A. 2012. “Standards of English in Malta and the Channel Islands”. In R. Hickey (Ed.), *Standards of English – Codified Varieties around the World*. Cambridge: Cambridge University Press, 117–138. https://doi.org/10.1017/CBO9781139023832.007

Laporte, S. 2012. “Mind the gap! Bridge between World Englishes and Learner Englishes in the making”, *English Text Construction* 5(2), 265–292. https://doi.org/10.1075/etc.5.2.05lap

Le Ruez, N. 2003. *Jersey Occupation Diary. Her Story of the German Occupation*, 1940–1945. Bradford on Avon: Seaflower Books.

Macaulay, R. 1989. “He was some man him: Emphatic pronouns in Scottish English”. In T. Walsh (Ed.), *Synchronic and Diachronic Approaches to Linguistic Variation and Change*. Washington DC: Georgetown University Press, 179-187.

Mesthrie, R. 1992. *English in Language Shift. The History, Structure and Sociolinguistics of South African Indian English*. Cambridge: Cambridge University Press.

Mesthrie, R. & Bhatt, R.M. 2008. *World Englishes. The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511791321

Mukherjee, J. & Hundt, M. (Eds.). 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.44

Nelson, C. 1982. "Intelligibility and non-native varieties of English". In B. Kachru (Ed.), *The Other Tongue. English across Cultures*. Urbana, Chicago and London: University of Illinois Press, 58–73.

Nesselhauf, N. 2009. "Co-selection phenomena across new Englishes: Parallels (and differences) to foreign learner varieties", *English World-Wide* 30(1), 1–26. https://doi.org/10.1075/eww.30.1.02nes

Ramisch, H. 1989. *The Variation of English in Guernsey/Channel Islands*. Frankfurt am Main: Lang.

Rosen, A. 2014. *Grammatical Variation and Change in Jersey English* [Varieties of English around the World G48]. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g48

Schneider, E. 2003. "The dynamics of New Englishes: From identity construction to dialect birth", *Language* 79(2), 233–281. https://doi.org/10.1353/lan.2003.0136

Schneider, E. 2007. *Postcolonial English. Varieties around the World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618901

Scott, M. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Sridhar, K. & Sridhar, S. 1986. "Bridging the paradigm gap: Second language acquisition research and indigenized varieties of English", *World Englishes* 5(1), 3–14. https://doi.org/10.1111/j.1467-971X.1986.tb00636.x

Szmrecsanyi, B. & Kortmann, B. 2011. "Typological profiling: Learner Englishes versus indigenized L2 varieties of English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 167–187. https://doi.org/10.1075/scl.44.09kor

Van Rooy, B. 2011. "A principled distinction between error and conventionalized innovation in African Englishes". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 189–207. https://doi.org/10.1075/scl.44.10roo

Viereck, W. 1988. "The Channel Islands: an Anglicist's no-man's land." In J. Klegraf & D. Nehls (Eds.), *Essays on the English Language and Applied Linguistics on the Occasion of Gerhard Nickel's 60th Birthday*. Heidelberg: Groos, 468–478.

“It’s always different when you look something from the inside”

Linguistic innovation in a corpus of ELF Skype conversations

Marie-Louise Brunner[1], Stefan Diemer[1] and Selina Schmidt[2]
[1]Trier University of Applied Sciences & Saarland University / [2]Birmingham City University

The article discusses linguistic creativity in informal Skype conversations between university students from eight different European countries. The basis for the study is the *Corpus of Academic Spoken English* (CASE), a corpus of Skype conversations in an English as a Lingua Franca (ELF) context. With the help of qualitative examples, the article examines innovative language use and proposes a taxonomy for functionally accepted innovations, distinguishing instances of L1 influence, approximations and ad hoc innovation. Our findings point towards an assertive and creative perspective on language use, which seems to have a positive influence on the communicative setting, e.g. illustrated by code-switching in combination with laughter. CASE participants use non-standard forms and innovations freely, accommodating to each others’ language use. They also establish their own ephemeral communication strategies and showcase and emphasize their respective language and cultural backgrounds.

Keywords: linguistic innovation, English as a Lingua Franca (ELF), Skype, spoken corpus

1. Introduction

This article discusses linguistic creativity in informal Skype conversations between non-native speakers of English. On the basis of corpus data, it examines to what extent non-standard features can be identified as examples of language innovation in an international context and analyzes whether participants’ language use goes beyond the basic need to make themselves understood. The study uses mostly

https://doi.org/10.1075/bct.98.09bru

qualitative analyses to explore non-standard features with the aim of identifying, and in one case quantifying, and categorizing linguistic innovation, illustrating its use in a pragmatic context, and developing a taxonomy of linguistic innovation in English as a Lingua Franca (ELF).

The basis for the article is an investigation of English as it is used in ELF Skype conversations compiled by the *Corpus of Academic Spoken English* project (CASE; Diemer et al. 2018) at Trier University of Applied Sciences, Saarland University, and their partner institutions. The language documented in CASE is often called ELF. Various studies have established that as an international academic language, English has developed distinct features in lexis, syntax and pragmatics (e.g. Firth 1996; Meierkord 1996). ELF is now the language of choice in (on- and offline) communication between non-native students and researchers. Seidlhofer (2001: 133) argues that it should thus obtain "a central place in description alongside English as a native language". Several corpora have been compiled in this field, such as the corpus of *English as Lingua Franca in Academic Settings* (ELFA; Mauranen et al. 2008) and the *Vienna-Oxford International Corpus of English* (VOICE; Seidlhofer et al. 2013). However, CASE is, to our knowledge, one of the first compilations of spoken academic English in an international context that is not mainly composed of lectures and conference conversations, but informal conversations outside of course or project work. It focuses on the language used by advanced non-native speakers in a private and informal, but academic context. Informal here means that conversations do not take place in a controlled lab environment or in an official/public setting, while the academic setting is established by the participants (university students) as well as context (as part of a university course/project) and topics (topic prompts and naturally occurring topics). Conversations and recordings are conducted via Skype, so that CASE is also the first extensive Skype corpus to be compiled to date, offering valuable insights into how the medium affects academic spoken discourse. CASE consists of more than 250 hours of conversations. The first round of talks between students of English at Saarland University, Germany; Sofia University, Bulgaria; Bologna University (Forlì Campus), Italy, and the University of Santiago de Compostela, Spain, was recorded between October 2012 and November 2013. Further rounds of talks were concluded in January and July 2015, and February 2016, with additional partners at Helsinki University and Hanken School of Economics, Helsinki, Finland, Linnaeus University, Växjö, Sweden, and Université catholique de Louvain, Louvain-la-Neuve, Belgium. A native speaker component with participants from Birmingham City University, UK, and Boise State University, USA, has been compiled between August 2016 and February 2018. Twenty transcribed conversations from the CASE project are publicly available for research (see www.umwelt-campus.de/case), with more to follow.

2. ELF, EFL and ESL

In the context of this volume on *Linguistic Innovations*, it is useful to illustrate similarities and differences between ELF, English as a Second Language (ESL) and English as a Foreign Language (EFL). We follow Kachru's (1985) distinction between inner, outer and expanding circle, and the corresponding designation of English varieties[1] by Görlach (1991) into English as a Native Language (ENL), ESL, and EFL varieties.

Traditionally, the term ESL is restricted to outer circle varieties of English such as Indian English, where English serves "country-internal functions" (Jenkins 2015: 2), or, as Crystal (2003: 60) puts it, "has become part of a country's chief institutions and plays an important second-language role in a multilingual setting". ESL is used to communicate with other ESL speakers and also acquired in this context (Sridhar & Sridhar 1986: 5), which means that the notion of standard cannot only be established in the context of ENL varieties, but is increasingly negotiated between ESL speakers who use their variety as a "lingua franca for interethnic, *intra*national communication" (Meierkord 2012: 69) with its own conventions and characteristics. As a consequence, ESL varieties do not follow ENL standards but develop their own standards (cf. e.g. Schneider 2011).

EFL, by contrast, is situated in Kachru's expanding circle, where English does not have "a history of colonisation", and is not given any special administrative status (cf. Crystal 2003: 60). Foreign-language varieties have been the domain of English language learning and teaching research, which focuses on the goal of interaction with native speakers, resulting in the idealization of native speaker competence (cf. also Seidlhofer 2001: 133). This implied standard affects learners' language use, penalizing supposed non-standard usage.

The restriction of EFL to what has also been called performance or learner varieties (e.g. Hundt & Mukherjee 2011) has favored the traditional separation of research approaches into EFL and ESL, respectively, focusing on different acquisition and usage settings. In this context, Sridhar and Sridhar (1986: 5) observe a "paradigm gap" between existing English learning theories and ESL variation, and Mesthrie & Bhatt (2008: 156) further elaborate on this lack of common ground between the contrasting EFL and ESL perspectives.

Recently, scholars have started to study the two varieties in relation to each other, conducting various corpus studies using EFL and ESL data (e.g. Deshors 2014; Diemer 2013; Edwards 2014; Gilquin 2011; Götz & Schilk 2011) in order

[1] In this article, we use the term "variety" not only to refer to types of institutionalised English (ENL, ESL), but also to different realizations of English usage that are loosely characterised by various common features and strategies (EFL, ELF).

to investigate common and overlapping features. As Edwards (2014: 173) puts it: "[T]he two varietal types share a common acquisitional starting point, which results in similar strategies such as transfer, redundancy and regularisation".

In contrast, ELF "orients to achieving mutual comprehension" between speakers of different language and cultural backgrounds (Mauranen 2012: 7). It is the "preferred option for cross-cultural communication" and exhibits non-native-like features (Seidlhofer 2003: 9). Successful communication is the key objective, whereas the imitation of native speaker varieties does not play a central role (Hülmbauer 2013: 50–51; Jenkins 2015: 45). Hülmbauer (2013: 50–51) points out that the "tension between stability and flexibility […] is inherent to the concept of ELF and represented by its two constituents: E — the relatively more stable English code as a basis — and LF — its ever-changing, flexible lingua franca adaptation". Seidlhofer (2011: 10) proposes a functional conceptualization of ELF, not a formal one, and considers as ELF "any use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option" (Seidlhofer 2011: 10), specifically including ENL speakers in intercultural situations.

There are overlaps between ESL and ELF on the one hand and ELF and EFL on the other hand. ESL and ELF share similar purposes of communication, as Meierkord (2000) observes, who suggests that ESL varieties function as an INTRAnational lingua franca, producing "nativized second languages", while ELF functions as an INTERnational lingua franca whose speakers "need to be regarded as learners of a language". ESL and ELF are also related via language usage effects that create new strategies and differences from native varieties (cf. Schneider 2012). Schneider (2012: 57) suggests that "sociolinguistically stable ELF settings may be hypothesized to represent initial stages in a trajectory towards ESL formation". However, while it is true that ESL, similar to ELF, does not follow imposed ENL standards as the native speaker ideal is no longer prevalent (cf. e.g. Deshors 2014; Gilquin 2011), Widdowson (2015: 362) points out that in contrast to ESL, ELF does not adhere to the norms of any variety:

> Because what is clearly evident in the use of ELF is that communicative capability not only does not depend on conformity to Standard English norms — it does not depend on conformity to the norms of any other variety either. […] The study of ELF considers variability not in terms of variety at all but as the variable use of English as inter-community communication, as communication across communities.

Mauranen (2012) makes a similar argument, characterizing ELF as a set of strategies aimed at achieving mutual and situated comprehension. Any patterns created remain ephemeral (i.e. dependent on the situation), though they may recur with similar language backgrounds, creating what Mauranen (2012: 29) calls *similects*. ELF speakers also use some features which are typical for EFL learners: Widdowson (2015: 371) mentions

> [t]he resemblance, often noted, between the linguistic features of much ELF usage to that of learner language. English learners and the ELF users they will become, both naturally and instinctively put the linguistic resources at their disposal to pragmatic use and so act on their communicative capability.

Both EFL and ELF speakers use, for example, approximations of standard forms, as well as grammatical deviation (cf. e.g. Björkman 2008: 122; Mauranen 2012: 41–44).

However, Widdowson (2015: 371) continues that "[t]he difference [between EFL and ELF] of course is that learners are discouraged from doing this and forced into unnatural conformity". Jenkins (2015: 45) points out that in contrast to the EFL notion of standard, in ELF "differences from native English that achieve this [successful intercultural communication] are regarded not as deficiencies but as evidence of linguistic adaptability and creativity", with the imitation of native speaker varieties no longer as the ultimate goal. In fact, as Widdowson (2015: 366) puts it, "[i]n ELF interaction, the interlocutors cannot depend on shared linguacultural conventions and so they have to find common ground by developing their own local conventions in flight as it were, as appropriate to their own contexts and purposes".

3. Taxonomy of innovations

ELF can be considered a rich data source for language innovations, since, as Hülmbauer (2013: 50) observes, "[d]ue to the relative instability and unpredictability of speaker constellations, levels of proficiency and contextual aspects in ELF, its users are frequently forced to improvise". In our analysis, we use the term *standard* for language use that is generally accepted in the inner-circle varieties (ENL), i.e. established in a native-speaker environment. Conversely, non-standard language use or, in Hülmbauer's (2013: 47) words, "unconventional language", is language use that is not generally accepted in ENL varieties. These varieties include, but are not limited to *Standard American English*, *Standard British English* etc. This definition is based on the definition of standard in an EFL context (cf. Ellis 2008). Although we are aware that there are also standardized ESL varieties (such as *Standard Indian English*), these are unlikely to be taught as standard in an EFL context. The use of the plural form *furnitures* would therefore be considered to be non-standard use even though this form is accepted in several ESL varieties. Based on EFL research, non-standard language use can be distinguished into errors and mistakes, i.e. competence (knowledge-based) vs. performance (processing-based) issues (Corder 1967; Ellis 2008). In this context, we would argue that errors and mistakes can be further distinguished from innovations, which reflect a more assertive, purposeful and creative language use.

In contrast to Croft's (2000) and Van Rooy's (2011) distinction between errors and conventionalized innovations based on diffusion as a criterion, we argue for a more flexible approach, following Kachru's (2006: 247–48) distinction between errors and "functionally appropriate innovation". Pitzl (2012: 46) observes similar motivations in ELF users' creative language use where "the use of creative idioms and metaphorical expressions in ELF interactions seems to be functionally motivated and is always intended to serve the overall goal of achieving successful communication".

We thus propose to conceive of non-standard forms as being either non-innovative deviations or functionally accepted innovations. The boundary between these categories is not always clearly delimitable, and some cases might arguably fall in either category. As we take a very inclusive stance on the concept of innovation, we also count instances as innovations where an either conscious or unconscious strategy can be inferred. In some cases, this might lead to the inclusion of arguably non-innovative forms. However, this reduces the possibility of missing instances of innovation. There are some tendencies that can be observed regarding the distinction between non-innovative and innovative non-standard forms. We do not consider non-innovative deviations to be creative in the sense that the speaker is actively adapting language for his or her purposes. They may be minor slips of the tongue, result from carelessness, or from lack of knowledge. These deviations tend to occur on the pronunciation level and on the morphosyntactic level, especially with non-standard concordance (an exception to this are cases of regularization where a strategic purpose could be inferred). Innovations, according to our definition, are instances of (more or less conscious) strategic language use. They can occur in lexical, phraseological, syntactic, morphosyntactic (e.g. regularization), and even multimodal (e.g. gesture) environments. They are functionally accepted by both interlocutors in the context in which they occur, and tend to support successful communication through creativity and the creation of new forms (similar to what Schneider 2007: 102 calls "linguistic creativity" in ESL). In the absence of a systematic and generally elaborated taxonomy of innovations in ELF, we inductively developed our own taxonomy based on our corpus data, taking various researchers' notions of different cases of innovation into account (cf. "approximations", Mauranen 2012; "regularization", Schneider 2007; "idioms", Pitzl 2009). We also included L1 influences as innovative strategies, based on research that shows that particularly code-switching is used in ELF "to accommodate diversity and/or the interlocutor(s)" (Vettorel 2014: 211), and "allows for meaning making and greater nuances of expression" (Cogo 2009: 268). In addition, this strategy is, as

Cogo (2009: 266) observes, "performed with expertise, a certain nonchalance and playfulness", indicating a certain degree of creativity.

These innovations can be further subdivided into (1) L1 influence, (2) Approximations, and (3) *Ad hoc* innovation. The boundaries between the three categories are relatively fuzzy so that one phenomenon can belong to more than one category. Our taxonomy (cf. Figure 1) shows a decreasing influence of the respective L1 (and a correspondingly increasing influence of English language patterns) as well as an increasing degree of innovation and creativity (from left to right).

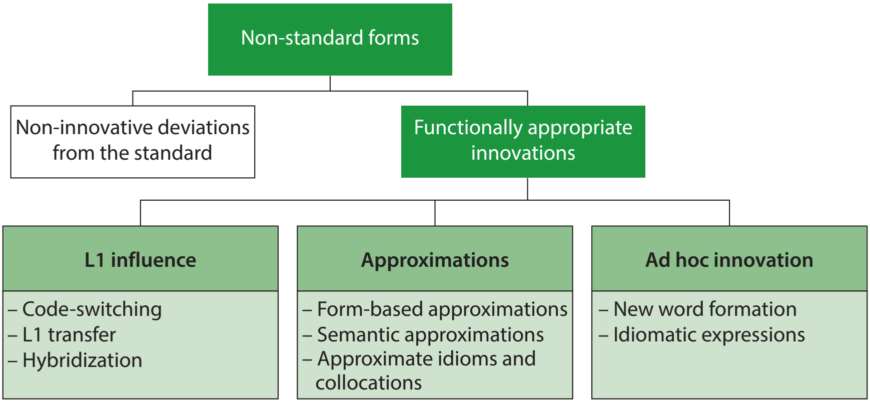

Non-standard forms
Non-innovative deviations from the standard
Functionally appropriate innovations
L1 influence
– Code-switching
– L1 transfer
– Hybridization
Approximations
– Form-based approximations
– Semantic approximations
– Approximate idioms and collocations
Ad hoc innovation
– New word formation
– Idiomatic expressions

Figure 1. Taxonomy of non-standard forms in CASE

1. L1 influence

The first category depicts an influence of the native languages and can be further divided into three subcategories: code-switching, L1 transfer, and hybridization. The first feature of this category, code-switching, refers to "the alternation of two languages within a single discourse, sentence or constituent" (Poplack 1980: 583), i.e. the meaningful integration of elements from other languages than English (the respective native languages or further L2s) into the mainly English conversation. L1 transfer occurs on a lexical, phraseological, or syntactic level when aspects of the utterance are directly translated from the respective native languages to English in the conversational setting. Hybridization is another form of L1 influence and also, arguably, more creative as it presupposes a conscious and more elaborate formation process. It occurs when participants create a hybrid form consisting of native language (L1 of speaker or of interlocutor) and English elements. These elements may be lexical, but they can, for example, also consist of syntactic or phraseological structures. The target form in this context can be both from the respective native language or from English.

2. Approximations

In characterizing the second category of innovation, approximation, i.e. the assertive use of similar but non-standard forms, we follow Mauranen's (2012) observation that conversation partners in ELF interactions use a range of approximations which, though not quite standard English, usually do not pose a communication hindrance and are understood and tolerated by both partners. According to Mauranen (2012: 108), approximations "are not arbitrary substitutions", but "have recognizable features in common with an item that would meet conventional expectations". She also points out that "[b]y approximating target forms well enough, speakers can contribute to communicative success" (ibid.: 41). It can be argued that approximations could additionally be interpreted as being reminiscent of the participants' EFL background as they seem to aspire to achieve a standard form but do not fully succeed (cf. also Mauranen 2012). This category can be further divided into form-based and semantic approximations. Form-based approximations can be defined as the use of slightly wrong forms (*assimilisation* for *assimilation*, Mauranen 2012: 102) or the use of existing but unsuitable words (*base* instead of *basis*, ibid.: 103).

Schneider's (2007) categorization for ESL varieties does, for the most part, not provide a sufficient template for an ELF context as the occurrence of many of his features is dependent on constant reiteration of these innovations (due to the diffusion criterion) in an INTRAnational environment, such as exaptation or grammaticalization. However, we have implemented his category of regularization as part of our taxonomy of innovations. Regularization as innovation strategy could, arguably, in this context be categorized as a subcategory of Mauranen's (2012) form-based approximations, as it also refers to slightly non-standard but still clearly recognizable forms. In this case, however, instead of being a more or less random deviation, it occurs when "for the expression of a consistent grammatical category the same formative (e.g. morpheme) is appended to all possible units (stems), while in a reference or source variety it is not used with certain words (which are then typically categorized as irregular)" (Schneider 2007: 103). We also find semantic approximations which are defined as being "English words, and all have some semantic components that allow the intended meaning to be deciphered. Some are near-synonyms, distinguished by their contextual properties (for instance *strength — power*, *in front of — ahead of*, or *normal — ordinary*)" (Mauranen 2012: 103).

Finally, we treat approximate idioms and collocations as part of the category of approximations. We conceive of both as referring to instances of "semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair 1991: 110). Mauranen (2012) points out that ELF users "latch on to salient features of a phraseological unit, which they use in its established sense, but without exactly reproducing the standard form" (Mauranen 2012: 230). Vettorel (2014: 1911) elaborates that "ELF users do not shy away from the use of figurative language, but rather employ idiomatic expressions often adapting them and creatively mingling their compositional elements with their (pluri)lingual resources". This is consistent with the use of idioms and collocations in CASE.

3. *Ad hoc* innovations

This last category comprises what we would argue are among the more creative instances of innovation: new word formations based on existing English word formation processes. *Ad hoc* word formations based on existing morphological templates have been researched, among others, by Van Rooy & Terblanche (2010) in an ESL setting. They concluded that because these formations do not meet their diffusion criterion they should not be counted as innovations. However, in an ELF setting these forms certainly qualify as innovations, in our opinion, as the diffusion criterion does not apply, but rather the criterion of functional acceptance. These *ad hoc* innovations exploit, as Widdowson (2003: 47) points out, "the morphological resources of English […] to bring new lexical items […] into existence" (he uses the example of *prepone* in analogy to *postpone*, a form that also exists in the ESL variety of Indian English, as Hülmbauer 2013: 49 points out), and thus exhibit a certain degree of awareness concerning English word formation processes on the part of the ELF speakers. Pitzl (2009: 298) observes that ELF users also make use of idiomatic expressions which she defines as "conventionalized preconstructed phrases", although in ELF "these expressions often display considerable non-conformity in reference to native speaker (NS) norms". She therefore describes this phenomenon as "re-metaphorization of idioms" or "idioms 'gone wrong'" (ibid.: 317). New idiomatic expressions also occur in our data and are treated as part of this third category displaying a high degree of creativity.

4. Data: CASE and BabyCASE

CASE consists of more than 250 hours of Skype conversations between university students from Germany, Bulgaria, Spain, Italy, Sweden, Finland, and Belgium, and conversations between British and American speakers for reference purposes. The coordination of the project takes place at Trier University of Applied Sciences, Germany. The transcription conventions for CASE (Brunner et al. 2017) include not only lexis, but also spoken language features, resulting in an orthographic transcription supplemented by an annotation of key discourse features. The annotation scheme that was adopted differentiates between prosody, paralanguage (such as laughter) and non-verbal cues (such as gestures and gaze). The transcription includes mark-up features, such as overlap, turns, laughter, prosody, gestures, and phonemic transcriptions (where salient, e.g. non-standard pronunciation leading to potential misunderstandings when two words become homophonous). A conversion tool developed by Matt Gee (Gee 2014) provides an XML version of the annotated CASE transcription and, additionally, extracts (by deleting all XML tags) a purely orthographic version without prosodic, paralinguistic, and non-verbal features suitable for Part-of-Speech (POS) tagging.

In our analysis, we use examples from BabyCASE (BabyCASE; Brunner et al. 2017), a corpus of 20 preliminary transcripts, in order to allow a quantitative analysis of the code-switching tag, see Section 6.1.1 Code-switching. BabyCASE was also used to facilitate extraction of individual instances of innovation using several extraction methods (see Section 5 Methodology). Figure 2 shows the general composition of BabyCASE.

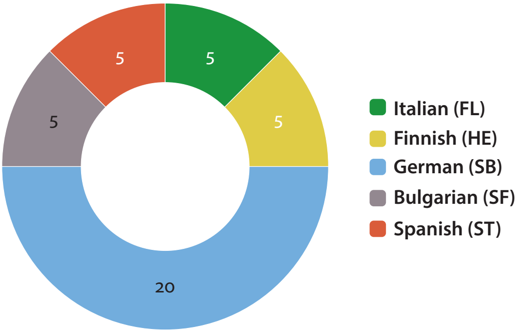


Figure 2. Composition of BabyCASE

BabyCASE contains 775 minutes (13 hours) of Skype conversations, comprising 115,000 words in the annotated version. It is composed of 20 conversations by 19 different speakers from four corpus components: CASE SB-SF (German and Bulgarian L1 speakers), SB-FL (German and Italian), SB-ST (German and Spanish), SB-HE (German and Finnish). The duration of the talks is between 30 and 45 minutes; various topic prompts are represented. The data represents the rich data of CASE itself, with 10 hours of video and 3 hours of audio recordings.

Transcription follows CASE transcription convention guidelines (see Brunner et al. 2017 for a detailed overview). Conversations are identified by a combination of topic[2] (first number), place of recording, and participant ID (SBxx = Saarbrücken

[2] For a complete topic list see the CASE website: https://www.umwelt-campus.de/ucb/index.php?id=11349#c35050.

participant number xx; SFyy = Sofia participant number yy). Abbreviations used in the examples are: FL (Forlì, Italy), HE (Helsinki, Finland), SB (Saarbrücken, Germany), SF (Sofia, Bulgaria), ST (Santiago de Compostela, Spain). Emphases are added for clarification purposes (in italics) in our examples.

5. Methodology: Finding linguistic innovations

Many existing studies of innovation in ELF focus on qualitative analysis based on corpus data and do not quantify the results (e.g. Pitzl 2009; Ranta 2009), as quantification of singular non-standard instances remains highly problematic in the context of a corpus which is not tagged for this specific purpose. As we encountered similar issues, the main focus of our study is on qualitative analysis. We do, however, include one instance of quantitative analysis in the context of code-switching (cf. Section 6.1.1 Code-switching) which was tagged in our corpus and could thus easily be quantified. We specifically address the extraction methods used to find individual instances of innovations in our data for qualitative analysis. These extraction methods support the researcher in finding some instances of innovative features more easily instead of having to search for innovations manually (by reading through the whole data set). These methods do not provide an exhaustive list and can therefore not account for all instances of innovation. They could thus not be used to make any quantitative claims. The extraction methods used in our analysis are Part-of-Speech (POS)-tag search, lexical comparison, low-frequency analysis, and *n*-gram analysis.

POS-tag search: For the purposes of a POS-tag search, Matt Gee's CASE extraction tool (Gee 2014) was used to create an orthographic version (see also Section 4) which was POS-tagged with the *Constituent Likelihood Automatic Word-tagging System* (CLAWS; Garside & Smith 1997) tagger. The POS-tagged version of BabyCASE was used to find instances of phrasal verb transfer, as in Example (6), and, in combination with *n*-gram analysis, to find the syntactic ellipsis in Example (17).

Lexical comparison: Several of the innovation strategies that we are documenting result in non-standard lexical items. These can be found by lexical comparison of the CASE data with a reference corpus. We used the *wordandphrase.info* (Davies 2012) interface of the *Corpus of Contemporary American English* (COCA; Davies 2008–2015), which automatically marks non-standard (i.e. not occurring in COCA) forms. As some spoken-language features and proper names are also marked as non-standard, the resulting list of non-standard items had to be categorized manually, resulting in a final list of non-standard items. This method was used to find examples for lexical L1 transfer, see Example (5), and form-based approximation, see Example (10).

Low-frequency analysis: As most of the innovations we illustrate in this article are ephemeral and occur rarely or even just once, an analysis of low-frequency items — which we here use to include items occurring ten times or less — was used to find them. To perform a low-frequency analysis an inverted word (or item) list was created using AntConc (Anthony 2014). We then manually searched for non-standard items, which were then categorized and quantified. The cut-off limit for the present study was set at ten occurrences in order to include items that might recur in one or multiple conversations. Low-frequency analysis was performed to find examples for lexical L1 transfer (see Example (5)), phrasal verb L1 transfer (Example (6)), syntactic L1 transfer (Example (7)), and regularization of past tense (Example (11)).

N-gram analysis: Usually *n*-gram analysis relies on finding frequently occurring patterns (of the length *n*). However, like low-frequency analysis, it can be used, in inverted format, for finding rare patterns in relatively small corpora. We used this method to find instances of phrasal verb L1 transfer, see Example (6), and syntactic L1 transfer, in Example (7), by searching for rarely occurring, non-standard 2-, 3-, and 4-grams in the CASE data.

6. Analysis: Linguistic innovation in CASE

6.1 L1 influence

6.1.1 *Code-switching*

Code-switching, that is an alternation of English with other languages, is frequent in our data, which is concurrent with other studies of ELF data, for example by Pennycook (2010), Cogo (2009), or Vettorel (2014). Vettorel (2014: 211) emphasizes that code-switching "is commonly and effectively deployed in ELF interaction". In our data, CASE participants use code-switching frequently for various communicative aims, for example to further communication, to convey words that are untranslatable or unknown, or that have a particular cultural connotation, or strategically for emphasis of cultural identity, or even to create humour. Usually code-switched items are taken from the speaker's L1 (rarely the interlocutor's L1 or other L2s of either of the conversation partners). Code-switches were the only feature that could be quantified as part of the present study. Quantification is possible because code-switches are annotated in CASE. Annotated code-switches in CASE are transcribed in the following format: *Code-switch ((Language of code-switch (duration))), e.g. Wurst ((German (0.1))).* Examples (1) and (2) illustrate instances of code-switching in CASE. In Example (1), we see the use of a Spanish interrogative interjection *cómo* which seems to happen unintentionally as part of an ordinary conversational flow where it has an almost formulaic discursive function. The code-switch *Heilig Abend* in Example (2) illustrates the replacement of a missing lexical item, which occurs frequently in non-native language use.

(1) Code-switching to Spanish: Discursive
07SB54ST04:
ST04: yes well. {shrugs}
SB54: (1.2) are you a student as well?
ST04: *cómo* ((Spanish (0.5)))? {moves closer to screen}

(2) Code-switching to German: Replacing
06SB73ST14:
SB73: (2.4) so we set it up o:n uh:,
(till) the twenty fourth of uh December?
it's: *Heilig Abend* ((German (0.9)))?
ST14: [oh],
SB73: [yeah]?

A quantitative analysis of BabyCASE[3] (cf. Table 1) reveals that code-switches occur in seven of the 20 conversations. There are 56 code-switches in total, 15 into Spanish, 36 into German, and five into Italian. There is no code-switching into Bulgarian or Finnish. Twelve of the participants code-switch, most participants into their native language, only three into their interlocutor's native language (one Finnish, Bulgarian, and Italian participant each code-switch into German). Four conversations contain code-switching into more than one language, and two speakers code-switch to different languages in the same conversation.

Table 1. Frequency of code-switching in BabyCASE

Code- switching to	*n*	Per 10,000	Conversations	Participants	% reciprocal
Spanish	15	1.3	3	3	26.7%
German	36	3.1	7	9	39%
Italian	5	0.4	1	2	100%
All	**56**	**4.8**	7	**12**	**41.2%**

More than 40% of code-switches are reciprocal within one minute of each other (see Table 1), which seems to indicate that code-switching frequently prompts

[3] Numbers may still vary, as BabyCASE is based on the preliminary first version of the CASE transcripts.

more code-switching, signalling acceptance for this deviation from the established common language, which is then taken up by both interlocutors and used as a strategy also in the rest of the conversation. An analysis of the surrounding intonation units shows that the paralinguistic feature of laughter seems to be closely associated with code-switching. In 30.5% of the cases, the laughter is concurrent, i.e. occurs within the same intonation unit or in the following one. If we consider the five surrounding intonation units (five before and five after the code-switch), this percentage increases to 67.9% while there were only 32.1% of instances where there is no laughter in the five surrounding intonation units (see Table 2; Example (3)).

Table 2. Code-switching and laughter in BabyCASE

Code-switching (*n*=56) and laughter in BabyCASE	***n***	**% of total**
No laughter (5 surrounding intonation units)	18	32.1%
Concurrent laughter (same or next intonation unit)	17	30.5%
Contextual laughter (5 surrounding intonation units incl. concurrent)	38	67.9%

(3) Code-switching and laughter I
07SB49FL33:
SB49: uhm,
Knödel ((German (1.0))),
I don't know if there is a word in English for that,
[((chuckles))]
FL33: [((LAUGHS))]

A closer qualitative analysis of the position of laughter in relation to the code-switches also shows that the majority of laughter (28 of 38 instances of contextual laughter after the code-switch, 14 before, 4 both before and after, see also Table 3) occurs in the five intonation units after the code-switch.

Table 3. Code-switching and contextual laughter in BabyCASE

Code-switching and contextual laughter (*n*=38) in BabyCASE	***n***	**% of total**
Contextual laughter (5 following intonation units)	28	73.7%
Contextual laughter (5 preceding intonation units)	14	36.8%
Contextual laughter (both preceding and following intonation units)	4	10.5%

Code-switching is, in our data, frequently used in a humorous manner, resulting in co-occurring laughter. Sometimes, the sound of interlocutors' native languages used in the context of a purely English setting seems to be reason enough to warrant laughter by one or both participants due to the sheer unintelligibility and, in the interlocutor's eyes, unpronounceability (cf. Example (4)), or, in the speaker's eyes, untranslatability (cf. Example (3)). Laughter in these instances contributes to creating rapport between interlocutors (Spencer-Oatey 2000) by establishing common ground and reducing the situational awkwardness (cf. Chafe 2007). Code-switching as an innovative strategy also showcases participants' own language background, and allows them to emphasize their cultural identity in a playful way (cf. also Auer 2005).

(4) Code-switching and laughter II
06SB73ST04:
SB73: okay how is it called that day?
ST14: (1.3) uhm *dia das letras galegas*. ((Galician (1.3)))
SB73: okay, ((laughs))
ST14: [((ehh)) not] gonna try right? ((ehh))
SB73: [it's- a- a-] no no no. [((laughs))]
ST14: [((laughs))]

Laughter in the context of code-switching could also be explained as a means of mitigating a delicate situation, in this case for example embarrassment, by indicating non-seriousness (Chafe 2007). Non-seriousness is here used in the sense that the linguistic item in question is not perceived as being serious, thereby categorizing it as a "non-problematic item", where a potentially problematic situation is defused by means of laughter (Jefferson et al. 1987: 172). This particular laughter, *coping laughter* (Warner-Garcia 2014), mitigates and downplays situations of embarrassment, but also other salient incidents, like committing a minor transgression, a *faux pas*, or displaying one's own shortcomings. Laughter thereby functions as a safe "exit strategy" (Partington 2006: 94), inviting the interlocutors to show sympathy, which could again be interpreted as evidence of the creation of rapport (Spencer-Oatey 2000). Code-switching, in combination with laughter, thus generally seems to have a positive effect on the communicative setting, putting the partners at ease with each other.

6.1.2 *L1 Transfer*

L1 transfer occurs when aspects of the utterance are directly translated from the respective native languages to English in the conversational setting. It cannot be reliably quantified in CASE, as it is both rare and not annotated. We draw a distinction between lexical, phrasal and syntactic transfer. Examples for L1 transfer were extracted both manually (by looking through the data) and with the help of a sequence of lexical comparison and low-frequency analysis (see Section 5). For the identification of lexical L1 transfer, a lexical comparison was performed. Forms not found in COCA (using wordandphrase.info) were listed. In a next step, low-frequency items were searched manually to find possible transfer features from the languages of the participants.

Example (5) shows one instance of lexical L1 transfer from German, the use of the term *half-day job*, referring to a part-time job (German: *Halbtagsjob*).

(5) Lexical L1 transfer from German
06SB95HE21:
SB95: and it is *half-day* job,

Instances of phrasal and syntactic transfer were found with a combination of *n*-gram and low-frequency analysis (see Section 5), as a lexical comparison did not produce the desired result. We performed an *n*-gram analysis of verb-particle combinations in the POS-tagged corpus version (extracting all 2-grams, 3-grams, and 4-grams to include intervening items such as discourse markers) and manually searched through those with a low-frequency analysis.

An example from the results is the phrase *look something from the inside* (Example 6) from Bulgarian *погледнем нещо отвътре* (say: *poglednem neshto otvǔtre*) which literally translates as *to look something from the inside* — there is no preposition with *look* in the Bulgarian original, as *погледнем* is transitive.[4]

(6) Phrasal verb L1 transfer from Bulgarian
11SB14SF05:
SF05: [((ehh))] yeah,
it's sure.
it's always .. different when you look something from the INside?
and from the outside,

Syntactic transfer was found in a similar manner, by listing and then manually analyzing rare 2-, 3-, and 4-grams (with an occurrence below 10, see low-frequency analysis in Section 5) in the orthographic (instead of the POS-tagged) corpus version and then comparing them with patterns in the respective L1s of the CASE speakers. The examples for syntactic transfers that were found in CASE can be quite complex and long, as in the case of Example (7) *with your family together* which directly translates a German syntactic construction (*mit deiner Familie zusammen*).

(7) Syntactic L1 transfer from German
07SB54ST04:
SB54: … wonderful Christmas dinner you can have *with your family together*,

[4] In cases where the transfer L1 was not known to the researchers, native speakers, as well as reference books and dictionaries were consulted.

6.1.3 *Hybridization*

Hybridization occurs when participants create a hybrid form consisting of native language and English elements. The target form in this context can be both from the respective native language (cf. Example (8), *Christmas … Krippe* based on German *Weihnachtskrippe*) or from English (cf. Example (9), *Krimi […] shows* based on English *crime shows*). The L1 part of the hybrid form was qualified as code-switching and could thus be found by searching for code-switches using the CASE annotations.[5]

(8) Hybridization I
07SB54ST04:
SB54: … and uh in Germany we have uhm, {looks away}
… the uhm *Christmas … Krippe* ((German (0.6)))?_((ehh))
I don't know how to say it in English?
.h uhm there is Maria Joseph [the three] holy-

(9) Hybridization II
08SB106HE03:
HE03: and then we have we have all these uh,
like .. uh *Krimi* ((German (0.5))) *shows* we have,

In Example (8), the German student cannot recall the English lexical item *manger* and uses a hybrid innovation strategy by replicating the German term *Weihnachtskrippe* with a generic English (*Christmas*) and a German element (*Krippe*, *crib*). Both the pause and the elaboration in the following line point to an insecurity in the item's usage. A similar pause (.. *uh*) can be seen before the code-switched element *Krimi* in Example (9).

6.2 Approximations

6.2.1 *Form-based approximation*

Form-based approximations make up for missing items in the concrete contextual setting. They are indicative of either performance- or competence-based momentary processing issues and aid in overcoming these communication hindrances without interrupting the conversational flow (cf. also Mauranen 2012). Some examples for cases where approximation results in slightly non-standard lexical items or forms (such as *anomynous* for *anonymous* in our data) were identified via lexical comparison (creating a list of non-standard items with *wordandphrase.info* and then manually checking for instances of form-based approximation). This

[5] Instances of hybridization with a code-switched element were also counted as part of the code-switching study in 6.1.1.

method does not work where similar existing words are used, and we did not find a reliable alternative method of quantifying all instances of form-based approximation.

An example for form-based approximation with an existing form is *I was quite interesting* instead of *I was quite interested* (Example 10), where present and past participle are confused which is a typical learner feature. The use of the approximation in this case does not have any negative impact on the conversation, seems to be understood without problems and is functionally accepted.

(10) Form-based approximation
11SB14SF05:
SF05: =yeah=.
SB14: =well *I was quite interesting*,
when you said you had to send your students home.
so are you a teacher as well?
SF05: .. u:h .. YES,

Regularization as innovation strategy (cf. also Schneider 2007: 103 for regularization in ESL) can, as mentioned above, be categorized as a subcategory of Mauranen's (2012) form-based approximations, also referring to slightly diverging but recognizable forms, but here applied to the use of regular patterns (e.g. plural or past tense formation) in a non-standard manner. Examples for regularizations were found through a low-frequency analysis of the orthographic version of the corpus, manually checking for regular patterns that are used in a non-standard manner. We searched for non-standard usage of inflectional morphemes. We found examples for plural and past tense regularization, for instance *stuffs* as plural of the uncountable noun *stuff*, and *hearded* (11) as regularized simple past of *hear*, or *childrens* (a double plural, regularizing the irregular plural of *child*).

(11) Regularization of past tense
08SB106HE03:
HE03: he's from Germany.
and so he when he *hearded* about this case he said,
well we can just speak German for half an hour. ((laughs))

6.2.2 *Semantic approximation*

Instances for semantic approximations could only be found manually (by reading through the data), as none of the methods described in Section 5 would reflect nuances in meaning.

In Example (12), the German conversation partner uses *you are a bit silent* to mean that the volume of the conversation is not high enough, not that the other person is not saying much (which might have been the case later in the conversation but is rather unlikely right at the start). This might also be influenced by the L1, as the German equivalent to *silent*, *leise*, is a more general term to indicate a low volume.

(12) Semantic approximation
06SB16SF05:
SF05: … hi,
can you hear me?
SB16: uh yes,
you are a bit silent,
but .. yeah.

6.2.3 *Approximate idioms and collocations*

We found several examples for approximate idioms and collocations through *n*-gram analysis. However, this method was not quantifiable, and the main part of the search still had to be performed manually by listing low-frequency *n*-grams and checking them one by one for similarities to existing idioms and collocations. In (13), the Bulgarian student varies the idiom *look at things from different angles*, altering both sequence and lexis. The resulting approximation, *look at different angles on things*, is apparently close enough to the original form to be easily understood by the German conversation partner. In (14), the Spanish student uses the approximation *I have my French (a little bit) rusted* instead of the standard phrase *my French is a little bit rusty*.

(13) Approximate idioms and collocations I
08SB24SF02:
SF02: … m:h we do a lot of discussions and,
.h you can *look at different angles on things*.
SB24: yeah.

(14) Approximate idioms and collocations II
07SB54ST04:
ST04: (1.2) because I *have my French (a little bit) rusted*.
… *(I think)*.

6.3 *Ad hoc* innovation

6.3.1 *New word formation*

Examples for new word formations were identified by means of a low-frequency analysis of standard derivational patterns, (such as *-ous*, *-ment*, *-ion*), followed by a manual investigation of the results. This isolated non-standard items such as *healthious* (cf. Example (15)) or *installate*. In Example (16) the German speaker uses the phrasal particle *out* as a verb, meaning *get something out*, in combination with *out* in its original function: *out it out of it*, i.e. *getting the baking tray out of the oven*.

(15) New word formation: Derivation
07SB25SF06:
SF06: you're not bent,
and you're not uh_uhm at desk,
you do something (*healthious*),
you do something different,

(16) New word formation: Conversion
07SB54ST04:
SB54: (1.4) and then you *out it out of it* and you have a great ((/k/reat)) uh,
… wonderful Christmas dinner you can have with your family together:,
ST04: oh it's ni:ce,

In Example (17), the Spanish speaker replaces an unknown item with gesture, completely eliding the elusive verb and using the multimodal affordances of Skype to express the intended meaning. *n*-gram analysis of the POS-tagged corpus was used to identify this defective sentence pattern. While our analysis found several instances of derivation and conversion, examples for ellipsis are very rare in the corpus; the ellipsis in Example (17) was the only one that could be identified in BabyCASE.

(17) Ellipsis
07SB54ST04:
ST04: eggs ((e/k/s)) you have to- I have- I don't know how to:,
(1.2) to say this,
… when you,
(1.6) the eggs ((e/k/s))? *{mimics putting eggs in a bowl}*
… in a bowl,
you put the bo- box and mix them?
SB54: in a- [yeah],

6.3.2 *New idiomatic expressions*

Finally, newly created idiomatic expressions can be considered a strategy with a high degree of innovation. Like with semantic approximations, these instances could not be identified reliably except by manual search. Examples (18) and (19) show that the use of these innovations does not hinder the conversational flow, as they are easily understood in context. This is consistent with Pitzl's (2012: 46) observation that "formal variation of idioms does not seem to be a disrupting factor in ELF conversations". In Example (18), the Bulgarian conversation partner constructs a new simile, creating a vivid mental image of snowflakes falling densely *like sheet of paper*. Example (19) seems to create an idiosyncratic variation of the expression *kith and kin*, i.e. *a large variety and quantity of people*, that is *the kittens and the grandmas and everyone*. There does not seem to be any link to a corresponding item in the speakers' native languages. This instance would support Pitzl's view of re-metaphorization, where newly constructed idioms "fulfill a striking variety of communicative functions, such as providing emphasis, increasing explicitness, elaborating a point, talking about abstract concepts" (2012: 317).

(18) New idiomatic expressions I
07SB10SF15:
SF15: well .. yeah .. here it is below zero too,
and *you can see the snowflakes basically,*
falling like sheet of paper [a whole cloud],
SB10: [oh Jeez],
.. ((laughs))

(19) New idiomatic expressions II
01SB78HE04:
HE04: and you just go out in the street,
and just party like crazy,
and there is like,
the kittens and the grandmas and everyone,

This category in particular shows participants' assertive appropriation of English in a lingua franca context and portrays them as active and self-confident language users, reflecting Hülmbauer's (2013: 50) observation that ELF users "tend to approach unconventional language in a rather straightforward, undeterred fashion", and, thus, "are more flexible than the communicators within native speaker communities".

6.4 Temporarily persistent innovations

As Hülmbauer (2013: 63) points out, innovations in ELF tend to be "of an ephemeral nature, i.e. they might not be relevant beyond their immediate context of origin". However, the innovative conversation strategies summarized in our taxonomy may also be taken up and used by both participants in the ensuing conversation, creating a temporarily accepted item which may persist throughout the rest of the interaction. This is similar to Mauranen's (2012: 49) notion of *accommodation* where "alternative forms of a word are negotiated, and […] speakers accommodate and finally converge on one of them". We found two examples for temporarily persistent innovations in our data. In Example (20) the conversation partners negotiate the use of the non-standard term *Romanic languages*, to mean *Romance languages*, and also use it later in the conversation, without negative influences on the conversational flow. In Example (21) the Spanish student proposes a translation of the term *Reyes Magos*, *the magic kings* (directly translated from Spanish) for *the Three Magi / the Three Wise Men / the Three Kings*. This proposal is indirectly rejected by the German conversation partner, who instead suggests the option *the three holy kings* (a direct translation from German), which is immediately accepted by the Spanish speaker. It could be argued that this goes beyond a singular co-construction, or accommodation (Mauranen 2012: 49), as it is later taken up and repeatedly used in the rest of the conversation.

(20) Temporarily persistent innovation I
01SB32FL06:
FL06: what are they called,
Ro:manic languages?
SB32: yeah,
yes,
Ro- Romanic languages,

(21) Temporarily persistent innovation II
07SB54ST04:
ST04: … *Reyes Magos* ((Spanish (0.9))) do you know what the *three … magic-*
-*k_uh:m*,
(1.0) *kings* are or not.
SB54: the *three what*? {moves closer to screen}
ST04: ((hehe)) no this is well-
this is something similar to Santa Claus but with [the] *three uh queens,*
that go to:,
SB54: [mhm],
ST04: to (belen) to Bethlehem to give Jesus,
uh,
SB54: the *three holy holy KINGS* [you mean].
ST04: [ah yes] *holy kings*.
SB54: … mhm [yeah] I know them. […]
ST04: uh the *three holy kings* are similar to:,
>Santa Claus< in the sense that,
.h they also bring- .. uh toys to children.
SB54: m mhm? {nods}
ST04: but we prefer .. the *three holy ki:ngs*,

These newly established temporary forms are usually not perpetuated beyond the singular conversational event but remain idiosyncratic and unique. Conversation partners might be motivated to use these successful strategies again in another conversational setting, where they would, however, have to be re-negotiated.

6.5 Limitations to quantification of innovations in CASE

In order to fully document and analyze the various levels of innovation in ELF data, a combination of qualitative and quantitative analysis would be ideal. While the present study was originally intended to be both qualitative and quantitative, the researchers found that due to the complexity and relative scarcity of the innovations in ELF (as compared to conventionalized innovations in ESL), only few innovations that are not specifically tagged, such as phrasal L1 transfer, are at all countable, but it is by no means certain that the proposed method finds them all. In addition, the manual search and quantification process is very time-consuming even with those features. While features such as form-based approximations or semantic approximation, derivation, ellipsis and hybridization are, theoretically, quantifiable with the methods presented here, the instances we found are, in fact, so rare (at least in our data) and the procedure leaves so many potential instances unaccounted for, that a reliable interpretation is, in our opinion, not practical at this point. Several of the innovations presented here are impossible to quantify automatically — these include syntactic L1 transfer, form-based approximations in the shape of existing words, semantic approximations, conversions or idiomatic expressions. In view of the limited searchability of these innovative features as well as the ephemeral nature of ELF, implementing a quantitative aspect goes beyond the context of this study, though it is recommended for future studies in this field.

7. Linguistic innovations in context

The present study illustrates that CASE participants use a wide variety of functionally accepted innovations. The L1 influence is visible in code-switching, transfer, and hybridization. Approximation processes take place on a morphosyntactic, semantic and phraseological level, while *ad hoc* innovation produces new derivations and new idiomatic expressions.

ELF speakers thus use some features which are also typical for EFL contexts. With regard to innovation, ELF also shares strategies with ESL, although in ESL a key criterion for innovation is conventionalization, which is not the case with ELF, as it is inherently ephemeral. In the context of our CASE data, ELF cannot develop endonormatively, as was proposed by Schneider (2007) as a typical stage in the development of ESL varieties, since the ELF speakers in our data do neither attempt to create a separate variety, nor are they in a position to do so.

Seidlhofer (2004: 212) describes ELF as "an English that has taken on a life of its own". Our findings regarding ELF speakers in CASE equally point towards an assertive and creative perspective on language use, reflecting (inter)cultural influences and the international context, and emphasizing cultural identities. Language use in ELF thus goes beyond the basic need to make oneself understood and reflects the requirements of an increasingly international community of speakers who adapt English to their own purposes. Despite their learner background, CASE participants use non-standard forms and innovations freely. This correlation between innovation and self-confidence has also been commented on in written computer-mediated settings featuring ELF speakers (cf. Diemer 2013; Vettorel 2014). This is indicative of a lessened influence of the native-speaker standard model (Jenkins 2009) and an increasing focus on successful communication. The confident use of non-standard forms illustrates innovation patterns similar to ESL varieties (cf. Kachru 2006), such as transfer, semantic approximation, and innovative word formation or phraseologisms, as discussed for example in studies by Deshors (2014), Gilquin (2011), and Laporte (2012). Similar to both ESL and EFL, strategies in ELF seem to be dependent on language backgrounds. The influence of the respective interacting or native languages is reflected in what Mauranen (2012: 29) calls "similects" which we would describe as strategy bundles that are used by speakers with the same or similar language backgrounds. As there is no creation of (or intent to create) separate varieties with a new "standard", it could be argued that ELF reflects a combination of various communities of practice and situation-related idiosyncrasies. In this setting, participants accommodate to each other's innovative language use (cf. also Mauranen 2012), creating their own ephemeral communication strategies that can, in some cases, be used throughout the particular interaction.

Creative language use seems to have a positive influence on the communicative setting, as illustrated with the example of code-switching in combination with laughter, which seems to create rapport between interlocutors. With the documented combination of strategies, participants also showcase their respective language and cultural backgrounds, for example through code-switching, L1 transfer, or hybridization, as also observed by Auer (2005), Hülmbauer (2013), Pitzl (2012), Vettorel (2014), and Widdowson (2015). Creative language use exemplified by innovations allows ELF speakers to negotiate their own cultural identities in an international context.

This study develops a taxonomy of innovations in ELF and provides methods for finding and analysing innovations, using new data from a corpus of ELF Skype conversations, thus contributing to achieving a better understanding of the linguistic phenomenon of language innovation that is the main focus of this volume. The proposed qualitatively developed taxonomy will need to be further elaborated with the help of quantitative analyses (as far as possible, depending on the type of innovation), contributing to an even more complete picture of innovative language use in ELF.

References

Anthony, L. 2014. *AntConc*. Version 3.4.3. Tokyo: Waseda University. Available at http://www.laurenceanthony.net/ (accessed February 2018).

Auer, P. 2005. "A postscript: Code-switching and social identity", *Journal of Pragmatics* 37(3), 403–410. https://doi.org/10.1016/j.pragma.2004.10.010

Björkman, B. 2008. "'So where we are?' Spoken lingua franca English at a technical university in Sweden", *English today* 24(2), 35–41. https://doi.org/10.1017/S0266078408000187

Brunner, M.-L., Collet, C., Diemer, S. & Schmidt, S. 2017. *BabyCASE*. Birkenfeld: Trier University of Applied Sciences.

Brunner, M.-L., Diemer, S. & Schmidt, S. 2017. *CASE Project Transcription Conventions*. Available at: http://umwelt-campus.de/case-conventions (accessed February 2018).

Chafe, W. 2007. *The Importance of Not Being Earnest: The Feeling Behind Laughter and Humor*. Amsterdam: John Benjamins. https://doi.org/10.1075/ceb.3

Cogo, A. 2009. "Accommodating difference in ELF conversations: A study of pragmatic strategies". In A. Mauranen & E. Ranta (Ed.), *English as a Lingua Franca*: Studies and Findings. Newcastle: Cambridge Scholars, 254–273.

Corder, S.P. 1967. "The significance of learner's errors", *International Review of Applied Linguistics in Language Teaching* 5(1-4), 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Pearson Education.

Crystal, D. 2003. *English as a Global Language*. Oxford: Oxford University Press. https://doi.org/10.1017/CBO9780511486999

Davies, M. 2008-2015. *The Corpus of Contemporary American English: 450 million words, 1990-2012*. Available at http://corpus.byu.edu/coca/ (accessed February 2018).

Davies, M. 2012. WordAndPhrase.Info. Available at http://www.wordandphrase.info (accessed February 2018).

Deshors, S.C. 2014. "A case for a unified treatment of EFL and ESL: A multifactorial approach", *English World-Wide* 35(3), 279–307. https://doi.org/10.1075/eww.35.3.02des

Diemer, S. 2013. "The return of the prefix? New verb-particle combinations in blogs". In H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling (Eds.), *Corpus Perspectives on Patterns of Lexis*. Amsterdam: John Benjamins, 223–244. https://doi.org/10.1075/scl.57.16die

Diemer, S., Brunner, M.-L., Collet, C. & Schmidt, S. 2018. *Corpus of Academic Spoken English*. Birkenfeld: Trier University of Applied Sciences (coordination) / Saarbrücken: Saarland University / Sofia: St Kliment Ohridski University / Forlì: University of Bologna-Forlì / Santiago: University of Santiago de Compostela / Helsinki: Helsinki University & Hanken School of Economics / Birmingham: Birmingham City University / Växjö: Linnaeus

University / Lyon: Université Lumière Lyon 2 / Louvain-la-Neuve: Université catholique de Louvain / Boise: Boise State University. Available at http://umwelt-campus.de/case (accessed February 2018).

Edwards, A. 2014. "The progressive aspect in the Netherlands and the ESL/EFL continuum", *World Englishes* 33(2), 173–194. https://doi.org/10.1111/weng.12080

Ellis, R. 2008. *The Study of Second Language Acquisition* (2nd ed.). Oxford: Oxford University Press.

Firth, A. 1996. "The discursive accomplishment of normality. On 'lingua franca' English and conversation analysis", *Journal of Pragmatics* 26(2), 237–259. https://doi.org/10.1016/0378-2166(96)00014-8

Garside, R. & Smith, N. 1997. CLAWS part-of-speech tagger for English. UCREL. Available at http://ucrel.lancs.ac.uk/claws/ (accessed February 2018).

Gee, M. 2014. CASE XML Conversion Tool. Available at http://rdues.bcu.ac.uk/case (accessed February 2018).

Gilquin, G. 2011. "Corpus linguistics to bridge the gap between World Englishes and Learner Englishes", *Comunicación en el siglo XXI*, Vol. II. Santiago de Cuba: Centro de Lingüística aplicada, 638–642.

Görlach, M. 1991. *Englishes: Studies in Varieties of English, 1984-1988*. Amsterdam: John Benjamins.

Götz, S. & Schilk, M. 2011. "Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced German learners". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 79–100. https://doi.org/10.1075/scl.44.05sch

Hundt, M. & Mukherjee, J. 2011. "Introduction". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language and Learner Englishes: Bridging the Paradigm Gap*. Amsterdam: John Benjamins, 1–5.

Hülmbauer, C. 2013. "From within and without: The virtual and the plurilingual in ELF", *Journal of English as a Lingua Franca* 2(1), 47–73. https://doi.org/10.1515/jelf-2013-0003

Jefferson, G., Sacks, H. & Schegloff, E.A. 1987. "Notes on laughter in the pursuit of intimacy". In G. Button & J.R.E. Lee (Eds.), *Talk and Social Organization*. Clevedon: Multilingual Matters, 152–205.

Jenkins, J. 2009. "English as a lingua franca: Interpretations and attitudes", *World Englishes* 28(2), 200–207. https://doi.org/10.1111/j.1467-971X.2009.01582.x

Jenkins, J. 2015. *Global Englishes. A Resource Book for Students* (3rd ed.). London and New York: Routledge.

Kachru, B.B. 1985. "Standards, codification and sociolinguistic realism: The English language in the outer circle". In R. Quirk & H. Widdowson (Eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.

Kachru, B.B. 2006. "The English Language in the outer circle". In K. Bolton & B.B. Kachru (Eds.), *World Englishes – Critical Concepts in Linguistics 3*. London and New York: Routledge, 241–255.

Laporte, S. 2012. "Mind the gap! Bridge between World Englishes and Learner Englishes in the making", *English Text Construction* 5(2), 265–292. https://doi.org/10.1075/etc.5.2.05lap

Mauranen, A. 2012. *Exploring ELF: Academic English Shaped by Non-Native Speakers*. Cambridge: Cambridge University Press.

Meierkord, C. 1996. *Englisch als Medium der interkulturellen Kommunikation. Untersuchungen zum non-native-/non-native speaker-Diskurs*. Frankfurt am Main: Peter Lang.

Meierkord, C. 2000. "Interpreting successful lingua franca interaction. An analysis of non-native/non-native small talk conversations in English", *Linguistik Online* 5(1). Available at https://bop.unibe.ch/linguistik-online/article/view/1013/1673 (accessed February 2018).

Meierkord, C. 2012. *Interactions across Englishes: Linguistic Choices in Local and International Contact Situations*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139026703

Mesthrie, R. & Bhatt, R. 2008. *World Englishes*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511791321

Partington, A. 2006. *The Linguistics of Laughter: A Corpus-Assisted study of Laughter-Talk*. London and New York: Routledge.

Pennycook, A. 2010. *Language as a Local Practice*. Oxford: Routledge.

Pitzl, M.-L. 2009. "We should not wake up any dogs. Idiom and metaphor in ELF". In A. Mauranen & E. Ranta (Ed.), *English as a Lingua Franca: Studies and findings*. Newcastle: Cambridge Scholars, 298–322.

Pitzl, M.-L. 2012. "Creativity meets convention: Idiom variation and remetaphorization in ELF", *Journal of English as a Lingua Franca* 1(1), 27–55. https://doi.org/10.1515/jelf-2012-0003

Poplack, S. 1980. "'Sometimes I'll start a sentence in Spanish y termino en español'": toward a typology of code-switching", *Linguistics* 18(7/8), 581–618. https://doi.org/10.1515/ling.1980.18.7-8.581

Ranta, E. 2009. "Syntactic features in spoken ELF-learner language or spoken grammar". In A. Mauranen & E. Ranta (Eds.), *English as a Lingua Franca: Studies and Findings*. Newcastle: Cambridge Scholars, 84–106.

Schneider, E.W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618901

Schneider, E.W. 2011. *English around the World: An Introduction*. Cambridge: Cambridge University Press.

Schneider, E.W. 2012. "Exploring the interface between world Englishes and Second Language Acquisition – and implications for English as a Lingua Franca", *Journal of English as a Lingua Franca* 1(1), 57–91. https://doi.org/10.1515/jelf-2012-0004

Seidlhofer, B. 2001. "Closing a conceptual gap: the case for a description of English as a lingua franca", *International Journal of Applied Linguistics* 11(2), 133–158. https://doi.org/10.1111/1473-4192.00011

Seidlhofer, B. 2003. *A Concept of International English and Related Issues: From 'Real English' to 'Realistic English'*. Strasbourg: Council of Europe: Language Policy Division.

Seidlhofer, B. 2004. "Research perspectives on teaching English as a lingua franca", *Annual Review of Applied Linguistics* 24, 209–239. https://doi.org/10.1017/S0267190504000145

Seidlhofer, B. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Seidlhofer, B., Breiteneder, A., Klimpfinger, T., Majewski, S., Osimk-Teasdale, R., Pitzl, M.-L. & Radeka, M. 2013. The Vienna-Oxford International Corpus of English (version 2.0 XML). Available at https://www.univie.ac.at/voice/ (accessed February 2018). Vienna: University of Vienna.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Spencer-Oatey, H. 2000. "Rapport management: A framework for analysis". In H. Spencer-Oatey (Ed.), *Culturally Speaking: Managing Rapport through Talk across Cultures*. London: Continuum, 11–46.

Sridhar, K.K. & Sridhar, S.N. 1986. "Bridging the paradigm gap: Second-language acquisition theory and indigenized varieties of English", *World Englishes* 5(1), 3–14. https://doi.org/10.1111/j.1467-971X.1986.tb00636.x

Van Rooy, B. 2011. "A principled distinction between error and conventionalized innovation in African Englishes". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language and Learner Englishes: Bridging the Paradigm Gap*. Amsterdam: John Benjamins, 189–208.

Van Rooy, B. & Terblanche, L. 2010. "Complexity in word-formation processes in New Varieties of South African English", *Southern African Linguistics and Applied Language Studies* 28(4), 357–374. https://doi.org/10.2989/16073614.2010.548022

Vettorel, P. 2014. *English as a Lingua Franca in Wider Networking: Blogging Practices* (Vol. 7). Berlin: Walter de Gruyter.

Warner-Garcia, S. 2014. "Laughing when nothing's funny: The pragmatic use of coping laughter in the negotiation of conversational disagreement", *Pragmatics* 24(1), 157–180. https://doi.org/10.1075/prag.24.1.07war

Widdowson, H.G. 2003. *Defining Issues in English Language Teaching*. Oxford: Oxford University Press.

Widdowson, H.G. 2015. "ELF and the Pragmatics of Language Variation", *Journal of English as a Lingua Franca* 4(2), 359–372. https://doi.org/10.1515/jelf-2015-0027

Subject index

At a time when the paradigm gap (Sridhar & Sridhar 1986) between the EFL and ESL research areas is attracting much scholarly attention, the contributions in the current volume explore this gap from the perspective of linguistic innovations across the two different types of non-native Englishes. In this endeavour, this volume unveils the many facets of linguistic innovations in non-native English varieties and explores the fine line between learners' erroneous versus creative use of a target language. Adopting empirical, corpus-based approaches to portray linguistic innovations characteristic of EFL and ESL varieties, the contributions show how the interaction of linguistic and social forces influences the development of novel linguistic forms in both endonormative ESL contexts and exonormative EFL contexts. This volume is of relevance to linguists who are interested in the features of non-native English and who wish to gain a better understanding of the nature of innovations along the EFL – ESL continuum.

Originally published as a special issue of *International Journal of Learner Corpora Research* 2:2 (2016).

ISBN 978 90 272 0146 1

9 789027 201461

John Benjamins Publishing Company