New Englishes, New Methods: Focus on Corpus Linguistics

1. Introduction

British (and American) colonialism has pushed the global spread of English, has brought English in contact with different indigenous languages, and has led to the emergence of a vast array of English varieties in colonized territories, such as in Trinidad or the Philippines. Most of these colonies have become independent and have retained English as their official language. Colonialism has also led to lasting global and local linguistic inequalities: a colonial hegemonic ideology privileges English over indigenous languages and positions the colonial standard variety, British English (and also American English), as the only legitimate and correct norm, while local appropriations of English by the colonized are marginalized (Ashcroft et al. 1989, 7).

Language users, activists, and academics in postcolonial countries (and elsewhere) have been active in decolonizing the English language. Two approaches from the mid-1980s have been seminal for the decolonization of the academic study of the English language. Ngũgĩ's Decolonising the Mind (1986) programme advocates the promotion of indigenous African languages primarily through their use in literature in contrast to English and takes an anti-imperialist stance against the dominant position of English (and other colonial languages) in African societies. The World Englishes (WE) paradigm – with Kachru's (1985) model of WE as the central theory – dismantles the marginalization of postcolonial varieties of English by acknowledging different Englishes worldwide as national standard varieties. Kachru (1985) models this global heterogeneity of English in three circles: (1) inner circle Englishes, spoken as a native language (ENL), such as British or American English; (2) outer circle Englishes, which are learned as a second/subsequent language (ESL), such as Indian or Philippine English (PhiE); and (3) Englishes of the expanding circle, where English is learned as a foreign language (EFL), such as Chinese or German English. Kachru's model focuses strongly on the Englishes of the outer circle, thus giving legitimacy to Englishes that have arisen out of colonial contact and are emerging standard varieties developing their own norm. These so-called New Englishes (Mufwene 1994) become bearers of a new national identity, non-Western values, and decolonizing processes in postcolonial societies (Schneider 2007, 48-52). The ways in which New Englishes systematically differ from British (or American) English as the former colonial standard are not rejected as mistakes but are appreciated as innovations. The WE paradigm has sparked a vast amount of research, leading to the establishment of a number of specialized journals (e.g. World Englishes, English World-Wide), book series (e.g. Varieties of English Around the World), and associations (e.g. the International Association for World Englishes).

Despite this impact, WE and Kachru's (1985) circles model in particular have been strongly criticized from a postcolonial perspective (Bruthiaux 2003; Pennycook 2007, 21-24; Tupas 2004). The two most relevant points of criticism for this study are (1) the framework's reduction of sociolinguistic complexity in postcolonial speech...
communities and (2) lingering Western-centrism. First, as the WE framework focuses on institutionalized national standard varieties, it ignores non-standard varieties, pidgins and creoles, as well as hybrid Englishes. This focus also entails an underlying assumption of homogenous national standard varieties of English, disregarding internal variability, which, for example, has been attested by Westphal (2017, 218-221) for Standard Jamaican English. Second, Kachru’s (1985) circles model positions ENLs as the norm-providing centre, the underlying implication being that the inner circle varieties are in some way superior to other Englishes (Rajadurai 2005, 114). Although this was clearly not intended by Kachru (1985), this inequality of Englishes is apparent in much WE research: New Englishes are often defined negatively in contrast to British or American English, which function as yardstick varieties (Saraceni 2015, 79).

In this paper, we address these two points of criticism. As a point of departure, we employ and advocate an extended understanding of New Englishes, which includes not only national standard varieties learned as second/subsequent languages or dialects but a wider range that includes multilingual-dialectal language use. Second, we present a postcolonial criticism of research methods used for the analysis of New Englishes that highlights lingering hegemonic biases in the field. In particular, this article addresses the Western bias in corpus-linguistic studies with a focus on the actual linguistic phenomena (or variables) studied. We show the benefits of analyzing the canonical forms and local innovations of a specific variable simultaneously. These two goals (i.e. showing variability in new standard varieties and overcoming a Western bias in corpus-linguistic research on New Englishes) are united in a corpus-linguistic study of question tags (QTs) in different text types of the Trinidad & Tobago (ICE-T&T) and Philippine (ICE-PHI) component of the International Corpus of English (ICE).

The remainder of the article is structured as follows: the next section describes the sociolinguistic situation in Trinidad and in the Philippines. The third section provides the theoretical background and discusses the Western bias in methods used for research on New Englishes, zooming in on previous research on QTs. In the fourth section, we present the data and method of our study. The fifth section illustrates the variation between variant and invariant QTs, regional differences between PhiE and Trinidadian English (TrinE) with regard to invariant QT forms, and variation within the two corpora with regard to local (i.e. Tagalog and Trinidadian English Creole) QTs. In the last section, we discuss these results with regard to the wider research context and advocate a widening of the canon of variables used for corpus-linguistic research on New Englishes.

2. English in the Philippines and Trinidad

Though the sociolinguistic conditions of the emergence of PhiE are very different from TrinE, both postcolonial standard varieties need to be analyzed within a broader perspective on New Englishes. PhiE and TrinE can only be adequately described with acknowledgment of and in relation to other language varieties with which they coexist. The Philippines is a highly multilingual archipelago with roughly 106.7 million inhabitants located in Southeast Asia: Ethnologue lists 183 living languages (Eberhardt et al. 2019), English and Tagalog being the principal languages (e.g. Tayao 2008, 292-293). English was introduced to the Philippines through American colonization starting...
in 1889, when the USA annexed the Philippines from Spain. English was spread as a pan-ethnic lingua franca throughout the islands via a large-scale education program and was made the official language of the colonial administration. Filipino nationalism, which resulted in the country’s independence in 1946, has pushed standardized Tagalog (called Filipino) as the national language and has made it the second official language alongside English. Tagalog is predominantly spoken in the northernmost island, Luzon, which is also home to the economic and political capital Manila. The official language policy aims toward English-Filipino bilingualism but the other indigenous languages besides Tagalog, such as Cebuano or Ilocano, were declared auxiliary languages. This has led to a highly multilingual situation and many Filipinos use different languages in their daily life. In Metropolitan Manila, the combination of English with Tagalog is prevalent in many spheres of social life, such as informal conversations, the media, or advertising (Thompson 2003, 9-12), which has resulted in the hybrid Taglish. The different Filipino languages and especially Tagalog have a strong effect on the local standard PhiE, which needs to be viewed in this multilingual environment (Tupas 2004, 47-48; Tayao 2008, 293-295).

Trinidad is the second most populous island in the Anglophone Caribbean with roughly 1.26 million inhabitants. It is located in the Southeastern Caribbean and is the bigger island of the twin island state of Trinidad and Tobago. English arrived in Trinidad in 1797 when the British captured the island from the Spanish. During the British colonial era until 1962, an English lexicon-creole, Trinidadian English Creole (TEC), and a local standard variety of English (i.e. TrinE) developed in a linguistically and ethnically diverse environment (e.g. Deuber 2014, 28-29; Wilson 2014, 7-14). The two varieties are typically used in different situations and fulfil distinct functions (e.g. Youssef 2004): Standard TrinE is typically targeted in formal domains, such as education or administration, and is viewed as a highly desirable social goal. In contrast, Creole is mainly characteristic of informal contexts and functions as an identity marker. TEC was long stigmatized as ‘bad English’ but attitudes toward Creole have eased (Mühleisen 2001) as TEC has pushed into domains formerly reserved for Standard English (Wilson 2014, 33-39). Youssef (2010) argues that in most situations in Trinidad the mixing of TrinE and TEC is the unmarked choice. Thus, the distinction between the two varieties is blurred and TEC is gradually integrated into TrinE (e.g. Deuber 2014, 200-201). This leads to a situation where it is often difficult to distinguish between the two varieties unambiguously (e.g. Youssef 2004). Winer (2009) captures this complex linguistic relationship via the term Trinidad English/Creole (TE/C), which encompasses all forms of English and English-Creole spoken in Trinidad (Wilson 2014, 12-13).

In this study, we focus on Standard English usage in the Philippines and Trinidad but illustrate the effects of their local linguistic embedding on Standard English: we address the variability in PhiE and TrinE with regard to the integration of Tagalog and TEC, respectively. The study shows how the mix of English and Tagalog or English and Creole-associated QTs, respectively, varies in different speech contexts, operationalized as different text types in the ICE corpora: i.e. (face-to-face) conversations, phone calls, class lessons, and legal cross-examinations.

---

1 The sociolinguistic situation in Tobago is different from Trinidad. Texts featuring Tobagonian speakers are excluded from the analysis.
3. Western Bias in Research Methods

Research on New Englishes uses a battery of methods to describe varieties spoken in postcolonial societies around the world. While these methods are firmly rooted in linguistic practice, they mostly have their origin in research on monolingual ENL speech communities and are then transferred to multilingual-/dialectal contexts characteristic of New Englishes. Without critical assessment and adaptation, the methods may fail to adequately describe variation in New Englishes. Smakman and Heinrich describe a similar Western bias for sociolinguistic theory-making: although "[t]he incompatibility of several dominant sociolinguistic theories with those outside their Western domain is obvious and undisputed" (2015, xvi), they are nevertheless applied to, for example, New Englishes contexts. Smakman and Heinrich's (2015) critical perspective on Western biases in linguistics is applied to the methods used in research on New Englishes. In this paper, we advocate a postcolonial reevaluation of the methods used in corpus-linguistic research on New Englishes. We first present a critical assessment of the corpora and linguistic phenomena studied in corpus-linguistic research on New Englishes and with our own analysis of QTs show how methods can be altered and improved to fit the sociolinguistic settings of New Englishes.

The most widely used linguistic corpora (i.e. large electronic databases of authentic spoken and/or written language) for the analysis of New Englishes are the ICE corpora (Greenbaum 1996). The ICE project started with ICE Great Britain and now includes more than ten completed national corpora that sample educated or Standard English from ENL and ESL contexts. All corpora follow the same design: each corpus consists of 500 texts (300 spoken; 200 written) of 2,000 words each. These texts are sampled from a total of 28 (fifteen spoken; thirteen written) different text types encompassing formal (e.g. legal cross-examinations or academic writing) and informal situations (conversations or personal letters). With a total size of 1,000,000 words, the ICE corpora are by today's standards relatively small but due to their diverse design and the matching national components, they provide a rich tool for cross-variety comparisons and analyses of variety-internal variation with regard to text type.

The focus on educated or standard speech across all text types, which is defined by people who speak it (i.e. 18 or over, formal education in English to the completion of secondary education; Greenbaum 1996), is problematic for New Englishes contexts. In many postcolonial societies like Trinidad or the Philippines, the exclusive use of Standard English is associated with very formal contexts/text types (e.g. Schneider 2007, 141; Youssef 2004). This predominance of Standard English in highly formal situations is evident in the two ICE corpora used for the purposes of this paper: for example, in a legal cross-examination from ICE-PHI (S1B-068), a lawyer is ordered to testify in English and to avoid using Tagalog. However, informal contexts are marked by frequent inclusion of Creole and Tagalog, respectively. To avoid too strong incursions of TEC, informal conversations in ICE-T&T are often artificially constructed: for example, teachers were asked to talk about a particularly serious topic (Deuber 2010, 36). In order to avoid too much code-switching to Tagalog, several comments made by speakers in ICE-PHI indicate that the informants were told to stick to English: for example, in text S1A-041, a speaker reminds their interlocutor that they are supposed to speak in English; or in S1B-015, a speaker apologizes for having said
something in Tagalog. These biases towards monolithic standard speech in corpus compilation ensure comparability across corpora but restrict local variability and mirror a narrow definition of New Englishes detached from the actual sociolinguistic complexes in which Standard Englishes in postcolonial speech communities emerge. Researchers are of course aware of the influence of local languages on New Englishes formation: for example, ICE includes the markup &lt;indig&gt; to give credence to lexical code-switching. Nevertheless, Mair states that the failure to recognize the multilingual embedding of these new Standard Englishes is the most "glaring lacuna" (2011, 234) in corpus-based research on New Englishes and advocates the compilation of multilingual corpora.

Despite these problems, ICE-based research has contributed greatly to the description of New Englishes (e.g. Bautista 2004; Deuber 2014) and has shown variation across Englishes (e.g. Hundt and Gut 2012). However, many studies take a "deficiency approach" (Saraceni 2015, 87) describing New Englishes negatively in contrast to British or American English and paying less attention to local dynamics of language use (Hansen 2018, 49). In addition, many of the phenomena used in analyses are biased in two ways: (1) most ICE-based research has focused on morpho-syntax and lexicon/phraseology, while phonology and pragmatic phenomena remain under-researched. (2) The phenomena studied in New Englishes often are adapted from research on ENLs that has shown significant variation for the feature(s) under analysis. However, just because a feature shows meaningful variation in an ENL variety does not entail that it is also salient (i.e. locally meaningful and significant) for variation in a New English speech community. Local meaningful features may be overlooked in such an approach to New Englishes. These two biases together have created a canon of variables frequently used for the description and investigation of New Englishes and patterns of use found for New Englishes are interpreted in reference to British and American English: for example, modals (Hansen 2018, 16-45) or progressives (Gut and Fuchs 2013, 245-248) have been extensively studied for ENLs and there is a growing body of comparative work on New Englishes.

This study focuses on QTs as one specific set of discourse pragmatic features (DPFs), which means that they are syntactically optional and mainly serve a pragmatic function (Pichler 2013, 4). Speakers append QTs to statements in order to receive a confirmation from their interlocutors, to integrate other participants in the conversations, or to emphasize their statements (e.g. Kimps 2018, 14-27; Wilson et al. 2017, 732-734). There is a wide range of different forms that can function as QTs. On the one hand, there are variant, also called canonical, QTs, whose structure depends on the main clause they append to, such as do you or isn't it. On the other hand, there are invariant QTs, which do not change their form in relation to the statement to which they are attached. Invariant QTs include single words, such as right, multi-word units, such as you know, or phones (or phonetic sequences), such as eh. Reference grammars (e.g. Biber et al. 1999, 208-210; 1089) discuss variant QTs at length and thus canonize them as the norm, while invariant QTs are only discussed fleetingly and are thus marginalized.

There is a wide body of research on QTs but most studies have analyzed QTs in ENL varieties and have mainly focused on variant forms, i.e. the "canonical form" (Algeo 2006, 293): for example, Barron et al. (2015) study variant QTs in British and Irish English, Tottie and Hoffmann (2006) compare British and American English, and
Kimps (2018) analyzes variant QTs and *innit* in British conversations. Most studies on QTs in New Englishes also focus on variant QT forms, including invariant uses of *isn’t it* and *is it*: for example, Wong (2007) studies variant QTs in Hong Kong English; Borlongan (2008) analyzes their use in PhiE; and Parviainen (2016) investigates QTs in four Asian New Englishes. All three studies use the ICE corpora and interpret their findings in comparison to variant QT use in British and American English. Wong (2007) highlights that invariant uses of *is it* are characteristic of Hong Kong English, whereas Borlongan (2008) and Parviainen (2016) conclude that PhiE and the four Asian Englishes, respectively, are characterized by a high frequency of invariant uses of *isn’t it*. Other invariant QT forms, including local forms (e.g. Tagalog QTs in the Philippines; Lim and Borlongan 2011), are excluded from these analyses. The conclusions on nativization in the different Asian New Englishes are thus based on ENL-canonized features and on a negative definition in contrast to British and American English as "points of reference" (Parviainen 2016, 98). In contrast, previous research that has focused on invariant forms (e.g. Bautista 2011; Columbus 2010) has attested the high frequency and salience (i.e. local meaningfulness and significance) of local invariant forms for New Englishes. Thus, the Western bias described for the three studies on QTs in Asian Englishes results in a whitewashing of New Englishes as non-canonical local forms are excluded.

In this study, we analyze variant and invariant QTs in TrinE and PhiE. We do not define these two New Englishes negatively but instead highlight differences between the two New Englishes and analyze the local variability of QT forms with regard to different text types of the ICE corpora. In particular, we address the following research questions:

- What is the relation of variant to invariant QT forms in TrinE and PhiE?
- What are categorical and preferential differences in QT forms between the two New Englishes?
- How are local QT forms distributed across the four text types in the two corpora?

The results of these research questions are then the basis for the final discussion on analyzing forms beyond the canon of corpus-linguistic variables.

4. Data and Method

We analyze the use of QTs in four text types of ICE-PHI and ICE-T&T: conversations, phone calls, class lessons, and legal cross-examinations. These text types differ in their level of formality and communicative settings. Phone calls are restricted to verbal communication and two participants, whereas conversations include non-verbal means of communication and potentially more people. Both are informal private dialogues, while class lessons and legal cross-examinations are public text types and have a higher level of formality – legal cross-examinations being the most formal.

The ICE corpora restrict the analysis to standard language use in the two postcolonial speech communities but the choice of the variable allows studying variation in the overall use of and the overlap with Creole and Tagalog QTs in these four contexts. Table 1 gives an overview of the size of the data, which totals 170,000 words per variety.
The analysis takes into account the full range of QT forms in 85 texts from each corpus. Similar to most pragmatic phenomena there is a form-function mismatch for DPFs, including QTs (Pichler 2013, 6). This means, a top-down design which is based on a range of forms that potentially function as QTs is problematic as searches for individual forms generate many tokens that do not function as QTs, and other forms not included in the study design are excluded. Thus, the current study takes a corpus-based approach, which is time-consuming but does not restrict the researchers to canonized forms: the researchers read through all texts of the corpora and identified QTs in context. The decision whether a specific form functions as a QT is grounded on the following criteria: QTs are DPFs (i.e. they are syntactically optional) and are attached to utterances. QTs are not: fillers (i.e. forms surrounded by repetitions or other fillers, such as *uh*, are excluded), entire utterances on their own (e.g. *right* used as a backchannel), items used in their full literal sense (e.g. *the right choice*), or part of fixed expressions (e.g. *right now*). Crucially, they fulfil an informative (i.e. signalling that a confirmation is needed), facilitative (i.e. integrating the interlocutors into the discourse), and/or punctuational function (i.e. adding emphasis). Excerpts (1) to (3) illustrate QTs that meet these criteria.

(1) <#>: Cara’s your bridesmaid *no* (ICE-PHI:S1A-070)
(2) <#>: Victor cool it *will you* (ICE-PHI:S1A-088)
(3) <#>: I am asking you *right* (ICE-T&T:S1B-007)

This qualitative identification of each token is the basis for a quantitative analysis that (1) illustrates the overall distribution of variant vs. invariant QTs across the four text types, (2) highlights the differences in forms between the two varieties, and (3) looks at the constraints of variation of individual local (i.e. Tagalog and TE/C) QT forms. The decision whether a QT is Tagalog is based on the Filipino transcribers’ marking of Tagalog tags as indigenous (<indig>), but we excluded QTs in utterances which are completely in Tagalog. The classification of QTs in ICE-T&T as Creole-associated is based on Winer’s (2009) dictionary of TE/C and Wilson et al.’s (2017) discussion of QTs in TrinE. For the comparison of frequencies across different text types, the results are normalized to tokens per million words due to the unequal sizes of the four text types. The results of all three parts of the analysis are first presented with descriptive statistics and then further analyzed via binary logistic regression modelling: form is the dependent variable (reduced to binary distinctions, such as *right* vs. all other forms); variety (PhiE vs. TrinE) and text type (conversations vs. phone

---

2 The analysis is restricted to the first 45 conversations (S1A-001 to S1A-045) in ICE-T&T and to a random selection of 45 texts in ICE Philippines.
3 <#> marks a new utterance.
calls vs. class lessons vs. legal cross-examinations) are fixed predictor variables; speaker is inserted as a random factor (random intercept) to ensure that the results are not biased by idiosyncratic variation. Statistical analyses were carried out with Rbrul (Johnson 2009). Rbrul reports whether a factor has a significant effect on the distribution via p-values and the direction and effect size of the individual levels of a predictor variable via centred factor weights, which range from 0 to 1: factor weights above 0.5 indicate a preference for the application value and below 0.5 a dispreference.

5. Results: QTs in PhiE and TrinE

The first part of the analysis gives an overview of general QT frequencies and focuses on variation between variant and invariant QTs. Despite slightly more QTs in the TrinE corpus, the frequency of variant and invariant QTs and their distribution across the four text types is very similar in the two varieties. For PhiE, 40 variant and 1072 invariant QTs were identified, and for TrinE seventeen variant and 1182 invariant QTs. Figure 1 shows the (normalized) QT frequencies for the two varieties for each text type as tokens per million words. Invariant QTs dominate by far over variant ones in both corpora and for all text types. A binary logistic regression model with form (variant vs. invariant) as dependent variable, variety and text type as predictor variables, and speaker as a random factor shows that neither variety (\( p=0.086 \)) nor text type (\( p=0.265 \)) have a significant effect on the distribution of variant QTs. Thus, variant forms are equally marginal for all text types in both corpora.

The variant QT forms do not always agree with the verb in the main clause and thus do not correspond completely to the canonized form: in PhiE 35.0% and in TrinE 29.4% of variant QT forms do not agree with the verb in the main clause; these are mostly *isn't it* and *is it* (not). Furthermore, in TrinE the seventeen tokens are produced by only nine different speakers; one speaker is even responsible for six variant QT tokens. Thus, the canonized variant forms are even more of an idiolectal phenomenon for both New Englishes than the already small token frequencies suggest.
The second part of the analysis focuses on differences between the two corpora with regard to specific invariant QTs and shows that there are substantial regional differences. Table 2 summarizes the token frequencies of individual forms in both varieties and presents the percentage of use of each form per variety. On the one hand, there are forms used by speakers from both varieties (left side in Table 2: shared QTs), and, on the other hand, there are QT forms exclusively used by speakers from one variety (right side in Table 2: variety-exclusive QTs). *eh/*e* is an exceptional case because there is a difference in spelling: it is marked as Tagalog in ICE Philippines and is listed by Winer (2009, 325) as a typical feature of TE/C. However, *eh* has been attested as an invariant QT in a range of different varieties of English (e.g. Columbus 2010) and *e*/*eh* is mainly used emphatically in both corpora (see excerpts 4 and 5). Thus, *e*/*eh* is in a sense both global and local.

(4) <#> Uh I like to sing e (ICE-PHI:S1A-062)
(5) <#> Don’t befriend me eh (ICE-T&T:S1A-031)
Table 2: Shared QT forms in PhiE and TrinE

<table>
<thead>
<tr>
<th>form</th>
<th>PhiE</th>
<th>TrinE</th>
<th>PhiE</th>
<th>TrinE</th>
</tr>
</thead>
<tbody>
<tr>
<td>alright</td>
<td>10 (0.9%)</td>
<td>210 (17.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eh/e</td>
<td>71 (6.4%)</td>
<td>99 (8.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OK</td>
<td>237 (21.3%)</td>
<td>121 (10.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>or what</td>
<td>8 (0.7%)</td>
<td>9 (0.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>182 (16.4%)</td>
<td>399 (33.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yeah</td>
<td>10 (0.9%)</td>
<td>19 (1.6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>2 (0.2%)</td>
<td>20 (1.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>you know</td>
<td>107 (9.6%)</td>
<td>191 (15.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>variant_QT</td>
<td>40 (3.6%)</td>
<td>17 (1.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>other$^5$</td>
<td>20 (2.0%)</td>
<td>30 (2.7%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Results of regression models on the effect of variety on shared QTs

<table>
<thead>
<tr>
<th>form</th>
<th>PhiE</th>
<th>TrinE</th>
<th>p</th>
<th>FW.-PhiE</th>
<th>FW.-TrinE</th>
<th>R²-total</th>
</tr>
</thead>
<tbody>
<tr>
<td>alright</td>
<td>10 (0.9%)</td>
<td>210 (17.5%)</td>
<td>0.033</td>
<td>0.25</td>
<td>0.75</td>
<td>0.92</td>
</tr>
<tr>
<td>OK</td>
<td>237 (21.3%)</td>
<td>121 (10.1%)</td>
<td>0.039</td>
<td>0.60</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td>right</td>
<td>182 (16.4%)</td>
<td>399 (33.3%)</td>
<td>&lt;0.001</td>
<td>0.37</td>
<td>0.63</td>
<td>0.47</td>
</tr>
<tr>
<td>you know</td>
<td>107 (9.6%)</td>
<td>191 (15.9%)</td>
<td>&lt;0.001</td>
<td>0.39</td>
<td>0.62</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 2: Shared QT forms in PhiE and TrinE

The analysis of individual QT forms shows that the majority of QT forms are shared by speakers in both corpora: the shared forms account for 93.5% of QT occurrences in the TrinE corpus and for 62.7% of QTs in the PhiE corpus. Thus, Filipinos use a higher frequency of exclusively local QTs. Although such a high proportion of QTs are shared, there are preferential differences in the frequencies of several QTs. Binary logistic regression models with form as a dependent variable, variety as predictor variable, and speaker as a random factor show that there are significant regional differences in the frequency of alright, OK, right, and you know. Table 3 provides an overview of the regression models where variety turned out as a significant factor for the distribution. There is a preference for OK in the PhiE corpus, while there is a preference for alright, right, and you know among the TrinE speakers.

In each corpus, speakers use QT forms which are exclusive to their variety of English (right side in Table 2). The remainder of the analysis focuses on the use of these local QTs. Filipino speakers integrate various Tagalog particles as QTs into PhiE. Tagalog ‘no’ is the contracted form of the invariant QT ano (Bautista 2011; Lim and Borlongan 2011, 70), and with 263 occurrences it is the most frequent QT form in the PhiE corpus, while the full form ano has a much lower frequency of only thirteen occurrences. $^5$ For

---

$^4$ Shared forms with a token frequency fewer than fifteen are summarized under the category other.

$^5$ See Bautista (2011) for an in-depth discussion of PhiE ‘no’ in contrast to global no.
example, in excerpt (6) a lecturer uses the QT 'no' to check whether his students are following his explanations.

(6) <#>So if you were to define entrepreneurship you will say something like uh it is the ability 'no' (ICE-PHI:S1B-004)

_Ha_ is another Tagalog interrogative particle (Rubino 2002, 123) used in PhiE: 48 occurrences of _ha_ as a QT were identified in the corpus. In excerpt (7), the speaker uses _ha_ to soften the blow of his face-threatening act to his interlocutor and thus to highlight their humorous intention.

(7) <#>Oh what do you mean uh you mean I'm so <O>laughter</O> I'm not a good person _ha_ (ICE-PHI:S1A-091)

_Ba_ is the shortened form of the "formulaic yes-no question hindi _ba_ or _di ba_" (Lim and Borlongan 2011, 61) and all three related forms function as invariant QTS. With 82 tokens, the long form _hindi/di ba_ is the second most frequent Tagalog QT in the corpus, while the shortened form _ba_ is rather rare with only nine occurrences. In excerpt (8), the speaker uses _di ba_ to emphasize that they like a particular movie.

(8) <#>The movie Dying Young <#>So good _di ba_ (ICE-PHI:S1A-064)

Trinidian speakers use several forms that are associated with Creole (ent, _nah_, _eh_, or _what_ or are typical for formal language use in Trinidad (_not so_) but are not necessarily exclusive to the variety. _Eh_ (Winer 2009, 325) and _or what_ (Winer 2009, 655) are associated with TEC but are also used by Filipino speakers in ICE-PHI. _Ent_ is strongly associated with TEC but with only three occurrences it is very infrequent in spoken Standard English represented by ICE-T&T. In contrast, the QTS _nah_ (49 occurrences) and _not so_ (30 occurrences) are used more frequently. The latter is not associated with Creole but taught as a polite alternative to _ent_ by Trinidian parents. In excerpt (9), the speaker uses _nah_ to emphasize that an issue is political, and in (10) an attorney uses _not so_ to receive a confirmation that there was no relationship between the witness and the accused person.

(9) <#>That real political _nah_ (ICE-T&T:S1A-014)
(10) <#>There was nothing between you and the accused at that point in time _not so_ (ICE-T&T:S1B-062)

Thus, Filipino and Trinidian speakers use a range of invariant local forms in the corpora, which represent Standard English in the two postcolonial speech communities. The differences in the frequencies of the individual forms highlight that not all forms are incorporated equally.

Moreover, the distribution of the most frequent Tagalog and TE/C forms across the different text types in the two corpora ( _hindi/di ba, 'no, ha, nah, not so, and e/eh_ ) was investigated. The results show that text type has a strong effect on the distribution of the selected local QTS in the corpora but there is striking variation between the individual forms. Figures 2 and 3 show the distribution of Tagalog and TE/C QTS, respectively, in the four text types as percentages by text type based on the normalized frequencies of each QT to account for the unequal text type sizes.

---

6 _Ent_ is spelled _ain't_ in ICE-T&T.
The descriptive statistics highlight that in general Tagalog QTs are extremely rare in legal cross-examinations, i.e. the most formal text type: there are only two ‘no tokens and one occurrence of e. E, ha, and hindi/di ba are mainly used in the two informal text types of public dialogues. However, Tagalog QTs are not absent from class lessons: hindi/ba occurs twelve times, ha twice, and e once. In addition, this text type shows the highest proportion of ‘no with 213 occurrences. ’No is used very frequently by teachers and lecturers in utterance final position to check whether their students are still following their explanations and have understood their line of argumentation as illustrated in excerpt (6).

TE/C QTs are present in all text types but the particular distribution differs between the three forms and if a form is very infrequent in a particular text type, it also serves a particular function. Nah is mainly used in informal dialogues – with a particularly high rate in phone calls (twenty-one occurrences) – though it is also present in class lessons (five occurrences) and legal cross-examinations (four occurrences). However, one attorney alone uses nah three times. Both teachers and attorneys use nah in an emphatic and very often antagonistic way: for example, in excerpt (11), the attorney rebukes a witness and orders them to tell the truth.

(11) <#>Oh gorm <#>But the majority of times Sir <#>Isn't it true <#>Talk the truth nah
(ICE-T&T:S1B-070)

Similar to nah, eh is also very frequent in private dialogues and rather infrequent in legal cross-examinations, but there is a markedly higher frequency than nah in class lessons (30 occurrences). The functions of eh in the two public text types are very
similar to *nah*: teachers often use *eh* with directives geared at students (excerpt 12), and attorneys use *eh* to pressure witnesses to disclose information (excerpt 13).

(12) <#>Denzel if you come without me homework done tomorrow you have to pay five dollars *eh* (ICE-T&T:S1B-006)
(13) <#>You remember you you've taken an oath *eh* (ICE-T&T:S1B-062)

*Not so* has a particularly high frequency in legal cross-examinations (twenty occurrences) and is rarely used in conversations (two occurrences) and phone calls (one occurrence). This high prevalence in legal cross-examinations substantiates the formal connotations of *not so*. *Not so* is commonly used by attorneys if they want witnesses to confirm certain information, as in excerpts (10) and (14). If teachers use *not so*, they mostly do so turn-finally and expect their students to answer. For example, in excerpt (15), a teacher is checking whether the students have understood the word 'sceptic' in the poem *Night of the Scorpion*.

(14) <#>Those packages which were contained in the Adidas bag were plastic wrapped packages *not so* (ICE-T&T:S1B-065)
(15) <$Teacher><#>You already have the meaning of sceptic *not so*<$Students><#>Yes Miss (ICE-T&T:S1B-004)

Figure 3: Distribution of TE/C QTs across text types [%-by text type]

These descriptive statistics show the variation in the samples, but they are problematic for drawing general conclusions as some of the token frequencies are very low. Regression models with form as dependent variable, text type as fixed explanatory predictor, and speaker as random factor are capable of illustrating the effects of text type more reliably as they take idiosyncratic variation into account (Johnson 2009, 363-365).
Table 4 shows the results of the regression models where text type has a significant effect on the variation.

<table>
<thead>
<tr>
<th></th>
<th>conversations</th>
<th>phone calls</th>
<th>class lessons</th>
<th>legal cross-examinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p$</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>% FW</td>
<td>% FW</td>
<td>% FW</td>
<td>% FW</td>
</tr>
<tr>
<td>$e$</td>
<td>0.024</td>
<td>12.1</td>
<td>0.75</td>
<td>11.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.2</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.72</td>
</tr>
<tr>
<td>hindi/di ba</td>
<td>0.003</td>
<td>14.8</td>
<td>0.77</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>2.3</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.49</td>
</tr>
<tr>
<td>$'no$</td>
<td>0.001</td>
<td>7.3</td>
<td>0.37</td>
<td>12.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.38</td>
<td>41.7</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12.3</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.59</td>
</tr>
<tr>
<td>$eh$</td>
<td>0.002</td>
<td>13.8</td>
<td>0.70</td>
<td>6.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.51</td>
<td>5.6</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.4</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.22</td>
</tr>
<tr>
<td>not so</td>
<td>0.059</td>
<td>0.5</td>
<td>0.19</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.23</td>
<td>1.3</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>22.7</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 4: Results of regression models on the effect of text type on local QTs

The analysis shows that text type has a significant effect on the variation for $e$, hindi/di ba, and $'no$ in PhiE and for $eh$ in TrinE. The differences in the use of $ha$ ($p=0.13$) and $nah$ ($p=0.29$) do not reach the level of significance. The $p$-value for the effect of text type on not so is slightly above the threshold value (0.05) but as the analysis with speaker as a random factor is very conservative, the results for not so are included. For $e$, there is a strong preference in both private dialogues, a slight preference in legal cross-examinations, and a strong dispreference in class lessons. For hindi/di ba there is a preference in conversations and a dispreference for phone calls and class lessons. As there was no token in legal cross-examinations this text type was excluded from the analysis. For $'no$ there is a clear preference in class lessons, a slight preference in legal cross-examinations, and a dispreference in the two private dialogues. In ICE-T&T, there is a strong preference for $eh$ in conversations, no clear directionality in phone calls, a slight dispreference in class lessons, and stronger dispreference in legal cross-examinations. The picture is very clear for not so: there is a strong preference in legal cross-examinations and a dispreference in all other text types. These results substantiate the variation described in the descriptive part, but they also show that the low token frequencies impede further inferential statistical work.

Text type has a significant effect on the distribution of Tagalog and TE/C QTs: they are variably integrated into standard speech represented by the two corpora. For PhiE, Tagalog QTs are mainly restricted to informal private dialogues, with the exception of $'no$, which has a high prevalence among teachers and lecturers. Variation in the TrinE corpus is more fluid: formality does have an effect on the distribution of some local QTs but the distinction is not clear-cut as speakers in all text types employ Creole-associated QTs.

6. Conclusion

The analysis of QTs in ICE-T&T and ICE-PHI has shown that variant QTs are marginal in both New Englishes in all contexts examined. Thus, these canonized forms have very limited use for describing or distinguishing between PhiE and TrinE and are not salient markers of variation within these New Englishes. In contrast, speakers from both corpora use a wealth of invariant QTs. Though many QT forms are shared by speakers
in the two New Englishes, there are significant preferential differences between the two varieties which go beyond idiosyncratic variation: while OK is preferred in ICE-PHI, alright, right, and you know are used significantly more frequently by TrinE speakers in ICE-T&T. In addition, speakers from each country have their own local tags: Tagalog QTs in the Philippines and TE/C QTs in Trinidad. Thus, both shared and exclusive invariant QTs are informative for cross-variety differences. The analysis has also shown that local QTs are telling of variation within the two New Englishes, as their use is strongly constrained by text type. In contrast to canonized variant QTs, which have been the focus of most previous research (e.g. Borlongan 2008; Parviainen 2016; Wong 2007), invariant QTs offer the possibility to not only study cross-variety differences but also local dynamics of use in New Englishes.

The extent of the variation of QTs within the two corpora, which depict Standard English usage in the Philippines and Trinidad, reflects the multilingual/dialectal embedding of these two new standard varieties of English. The analysis of the full range of QTs in ICE-PHI shows that Tagalog has a strong linguistic influence on PhiE on this pragmatic level of variation. Although Tagalog QT forms are very frequent in the corpus, there seems to be a relatively clear divide between the contexts in which Filipino speakers insert Tagalog QTs into English and those where they do not use them. Most Tagalog QTs are very infrequent apart from informal private dialogues. However, no seems to be more versatile as it is employed in a wider range of contexts, corroborating Bautista’s (2011) observations. This shows that different local forms have different patterns of variation and thus have different salience for distinguishing between formal and informal language use. The interaction of English and Tagalog (or other L1s) is highly relevant for the understanding of Standard English in the Philippines but is often excluded in studies that focus on variables salient for variation and change in monolingual ENL communities.

The results for text type variation of local QTs in ICE-T&T reflect the language situation in Trinidad, where it is more difficult to distinguish between English and Creole than between English and Tagalog in the Philippines. Eh and nah, which are both associated with Creole, are present across all four text types, while ent as a QT is largely avoided by speakers in ICE-T&T. Nevertheless, the formality of a text type has a significant effect on the distribution of eh and not so, the latter being typical of ‘formal’ Standard English in Trinidad. Thus, the analysis of QTs is capable of showing fine-grained variation in TE/C. In analogy to the findings for the Philippines, the analysis shows that different forms have different meaning-distinguishing potential between English and Creole (Irvine 2008). However, if elite speakers use QTs that are associated with Creole in very formal situations they mostly employ these forms for specific functions: contrary to previous descriptions of Creole as powerless and signalling solidarity (Rickford and Traugott 1986), teachers and attorneys strategically employ Creole forms to ‘do power’ (i.e. exert authority). A more detailed and qualitative analysis of functional properties of QTs in the corpus has the potential to describe these sociolinguistic dynamics in specific discourses more precisely.

The analysis of QTs in ICE-T&T and ICE-PHI has offered a glimpse into the multilingual/dialectal linguistic complexes characteristic of postcolonial standard varieties. The variability in New Englishes that results from this sociolinguistic situation is often omitted by studies which rely on canonized linguistic variables, such as modals or progressives, which are primarily relevant for variation and change in
monolingual ENL communities. Thus, future research on New Englishes needs to move beyond this methodological bias toward ENL-based variables, which reflects a limited view of what variables are worth examining. One aspect of finding new methods is identifying and analyzing (under-researched) variables that are first and foremost salient for variation in New Englishes but not necessarily for variation in ENLs.

The analysis has shown that it is highly beneficial to widen the canon of phenomena used for corpus-linguistic analyses of New Englishes. A major criterion for selecting a variable should be its salience for variation in the variety under research. Ideally, variables include global and local forms and thus allow studying the multilingual/-dialectal embedding of Englishes. Lexical items seem a promising choice but the low frequencies in small corpora inhibit closer analyses. We propose DPFs as an area with a high potential, which we have illustrated in this study on QTs. Although DPFs are still under-researched corpus-linguistic phenomena (Pichler 2013, 10-16), especially in New Englishes, and are rarely included in descriptions of varieties, they have a high potential for studying cross-variety and variety-internal variation: first, they have a high frequency in spoken discourse. Second, in New Englishes contexts, the wide array of forms potentially includes English-based forms and forms derived from other language varieties of the speech community. Third, DPFs have a high functional diversity, which can also be addressed as an additional level of variation, as shown by Wilson et al. (2017). Fourth, many DPFs have a high indexical loading but are not stigmatized as stereotypes and thus function as markers of sociolinguistic variation. However, DPFs can be quite problematic to analyze due to the form-function mismatch (Pichler 2013, 6-9). This requires an analysis of each token in context and thus a close engagement with the details of the data. As highlighted above for the use of Creole forms by elite speakers in formal situations, this close reading can potentially reveal very important insights into fine-grained sociolinguistic dynamics, which are not captured by highly abstract approaches that rely mainly on counting individual forms across (sub)corpora. Thus, DPFs offer various opportunities for corpus-linguistic studies and the findings can be the basis for the creation of sociopragmatic profiles of individual DPFs, which highlight local patterns of use and can be included in future descriptions of New Englishes.

In addition, a stronger inclusion of language-external variables into corpus-linguistic analyses of New Englishes is crucial to illustrate further local levels of variation, as shown in this study with the focus on text type variation. The ICE data also provide speaker information, such as age and gender, which can be used beneficially to describe sociolinguistic variation in New Englishes as shown by Hansen (2018, 267-295) for modals. The ICE data might not be ideal for covering a wider range of multilingual/-dialectal variation envisioned with our broader understanding of New Englishes but there are ways to tap into this largely unexplored field. In this article, we have shown the potentials of a widening of the research canon to new variables that include local realizations, the focus on text type variation, and the close reading of the data for ICE-based research.

An approach that engages critically with the Western biases in WE research – including theories, methods, and the data used – is necessary to contribute to a more positive definition of New Englishes, focusing on local variability. Studies on New Englishes that analyze forms canonized in research on ENLs naturally compare the
results to existing research on these varieties and thus are prone to a deficit approach and a negative definition of New Englishes. To further decolonize the study of the English language, analyses of local linguistic variation in New Englishes need to more earnestly consider the multilingual-/dialectal complexes of New Englishes and thus there is a need for a wider notion of New Englishes, which in the Philippines covers a continuum from English to Tagalog and in Trinidad from English to Creole. The ICE corpora have some potential to realize this postcolonial research agenda but nevertheless are limited by their focus on standard Englishes. New multilingual-/dialectal corpora, such as a TE/C corpus or a corpus that covers English/Taglish/Tagalog, are a desideratum (Mair 2011). Such corpora could provide more accurate representations of the English that is spoken in New Englishes contexts, and allow more systematic, corpus-based investigations of New Englishes in relation to the other languages with which they come into contact.

Future research projects on New Englishes need to build on the postcolonial research agenda of the WE framework by critically addressing Western biases in established theories, methods, data, and data collection techniques. This paper has focused on corpus linguistics and has illustrated the potential of new corpus-linguistic variables for the study of New Englishes. Similar Western biases can also be found in other areas, such as phonetics, pragmatics, or language attitude research. The critical examination of established methods in these fields may eventually lead to the development of new original methods that have their origin in research on New Englishes.

Works Cited


