Corpus-Based Approaches to World Englishes: Applications and Challenges

Jiale Ding

University College London, London WC1E 6BT, United Kingdom

Abstract: This paper investigates the use of corpora in studying World Englishes (WEs), focusing on their role in analysing language variation, morpho-syntactic patterns, and sociolinguistic changes. It also addresses key challenges, such as the lack of representativeness, limited spoken data, and difficulties in managing Creole and standard English varieties. The study highlights the need for more comprehensive corpora to better understand the complexities of WEs.

Keywords: World Englishes; Corpora; Language Variation; Creole; Standard English

1. Introduction

Corpora are essential tools for examining the linguistic diversity in WEs. This essay explores how corpora help analyse different aspects of WEs, such as grammatical and sociolinguistic variation, as well as changes over time. Despite their benefits, current corpora face limitations in scope, particularly in spoken data and representation of Creole-English dynamics. This paper emphasises the importance of enhancing corpora to improve the study of WEs.

2. Discussion

Corpora are valuable tools for studying English varieties worldwide. The International Corpus of English (ICE) was specifically developed to provide comprehensive spoken and written language samples from various countries and regions. Following ICE, the Global Web-Based English (GloWbE) corpus was created, collecting most of its texts from the internet. These corpora provide rich, diverse datasets for analysing the usage and evolution of English in various global contexts, offering crucial insights into WEs.

2.1 Corpora in studying World Englishes

Corpora are essential for studying WEs by testing models of language variation and change. Through examining different varieties of English, researchers can refine existing theories. For example, corpora have been used to test Schneider's Dynamic Model and the hypothesis that new English varieties may develop regional standards. Corpora can also assess linguistic epicentres. Heller, Bernaisch, and Gries (2017) analysed English genitive alternation across six ICE sub-corpora to evaluate Indian and Singapore English as potential epicentres in Asia. This analysis, supported by extensive data, shows how corpora can effectively describe and model linguistic variation, providing statistical evidence for epicentre status.

Corpora enable researchers to examine morpho-syntactic variations in English usage across countries. By analysing linguistic features in diverse texts, researchers can identify patterns and differences in vocabulary and grammar, including specific lexical and syntactic features of different English varieties. Biermeier (2014) compared compounding and suffixation across twelve sub-corpora of the ICE, focusing on distinctions between native and non-native English varieties in Asia and Africa. ICE's diverse range of samples allows researchers to compare linguistic features across countries. This study employed ICE data to explore and compare the frequency and creativity of Word-formation strategies. By analysing the lexical richness present in various texts, researchers can characterise regional or national-specific morphological features.

Moreover, parsed corpora are highly valuable for studying syntactic variation, as shown by Collins (2008) in his analysis of the progressive aspect across nine ICE corpora. With grammatical structures and word classes already annotated, scholars can perform faster data analysis without needing manual annotation. The study revealed regional differences in progressive usage, with Australian and New Zealand English being the most innovative. Additionally, it compared the frequency of progressive forms in speech and writing, highlighting stylistic

variations between these contexts. This shows how corpora can uncover linguistic pattern differences across various forms of language.

Corpora facilitate the analysis of sociolinguistic and pragmatic variation in WEs. By examining social variables and pragmatic features, researchers can investigate language function and meaning in discourse. Collins (2020) used the GloWbE corpus to analyse comment markers (CMs) across twenty varieties, uncovering differences in their frequency and distribution. GloWbE's large size enables the retrieval of numerous CM tokens, and its informal nature makes it particularly suited for studying CMs, which are common in informal texts. While the resemblance between GloWbE data and spoken language is debated, it remains a valuable resource for studying pragmatic variation, often preferred over the smaller ICE corpus for such research.

Additionally, annotated corpora like ICE are valuable for sociolinguistic research, offering detailed metadata such as age and gender. Suárez-Gómez and Seoane (2023) used ICE-IND and ICE-PHI to explore grammatical variation in WEs based on these factors. These corpora provide rich, comparable datasets, enhancing sociolinguistic studies by allowing analysis across different demographic groups. Their design makes them particularly useful for studying sociolinguistic and pragmatic patterns in WEs.

Corpora serve as valuable historical records, allowing researchers to trace the evolution of English over time. They are especially useful for long-term studies on WEs by providing authentic language samples. For example, Rossouw and Van Rooy's (2012) analysis of the South African English corpus (SAfE), covering the 19th to late 20th centuries, showed a consistent use of modality until a decline in the latter 20th century. This reveals how historical corpora can uncover linguistic trends that contemporary data may miss. SAfE also connects linguistic shifts to socio-cultural and contextual changes, such as the decline in modality linked to shifts in register preferences and norms. Overall, historical corpora are crucial for diachronic studies, offering insights into how language evolves in response to social influences.

However, there is a lack of multinational historical corpora for other WEs, which limits cross-varietal diachronic research. Although there are diachronic corpora available, they are specific to particular countries, like Corpus of Present-Day Spoken English, which focuses on British English. The limited focus restricts the comparative studies. Therefore, it is crucial to develop multinational historical corpora to enable diachronic analyses across English varieties.

2.2 Problems in Existing Corpora

A significant challenge in studying spoken varieties of WEs is the lack of audio recordings. Among existing WEs corpora, only ICE includes spoken data, and even then, not all countries represented in the ICE corpus have available spoken samples. Therefore, many studies must rely solely on written texts. For example, the analysis of ICE-Sri Lankan was restricted to written materials due to the absence of spoken data, emphasising the reliance on written corpora to study English evolution.

Text-only corpora are limited in capturing spoken language nuances, such as discourse markers and conversational interactions. Audio recordings provide insight into crucial elements like pronunciation and intonation, which are key to understanding communication. Without them, important aspects of spoken English may be overlooked, leading to incomplete analyses.

Existing corpora of WEs struggle with representativeness, mainly in countries where English is a Second Language (ESL). In Fiji, for example, collecting local texts is challenging due to the limited availability of newspapers and magazines. Moreover, the unstable political situation from 2006 to 2015 complicated data collection, restricting access to important registers like parliamentary debates.

The scarcity of news and parliamentary debate data raises concerns about the representativeness of WEs corpora in capturing Fiji's linguistic landscape. This limitation affects the ability to fully understand how English is used in various contexts, such as media and politics. The absence of these text types not only impacts the overall size of the corpus but also restricts insights into Fijian English usage in political, legal, and social spheres.

Corpus construction can affect New Englishes studies due to methods based on monolingual English-speaking countries. When constructing a corpus, sampling frames may fail to recognise the linguistic diversity in ESL countries, where English coexists with other languages. Although these frames suit English-speaking countries, they may not fully represent multilingualism in ESL contexts.

The corpus compilation process can introduce biases. For example, the ICE Trinidad and Tobago (ICE-T&T) corpus shows how this affects research on New Englishes. The ICE corpus equates "educated" English with "standard" English, favoring a specific linguistic norm. In

Trinidad and Tobago, while the ICE corpus includes educated English, the spoken component also features creolised English. As the boundaries between Trinidadian English Creole and Trinidadian English blur, it becomes harder to differentiate them. However, only a portion of this variation is classified as Standard English. By focusing on educated or standard varieties, the ICE corpus overlooks the full linguistic diversity in ESL countries, potentially limiting the understanding of variations in New Englishes.

Furthermore, corpus research struggles to represent the full range between standard English and Creole varieties. Bias in corpus compilation often prioritises standard English due to its formal status, leaving Creole varieties underrepresented. Since Creole was recognised as a distinct language in 1975, its usage has increased, mainly in education, where it is being integrated into curricula, challenging the traditional dominance of standard English.

Moreover, attempts to limit the influence of Creole, such as in the ICE-T&T, can result in less authentic dialogues. For example, participants were encouraged to discuss serious topics, and teachers were instructed to avoid excessive code-switching to Tagalog. These practices highlight the difficulties of accurately capturing Creole in corpora. Overall, representing both standard English and Creole in corpora remains challenging, showing the need to enhance the authenticity and inclusiveness of WEs corpora.

3. Conclusion

In summary, corpora are essential for analysing WEs, offering researchers vast linguistic data. However, challenges like limited representativeness, small size, and insufficient spoken material persist. To improve research, more comprehensive corpora are needed, with innovative data collection methods and a focus on Creole-Standard English hybridisation. Despite these obstacles, corpora remain invaluable for both cross-sectional studies of English varieties and longitudinal studies tracking language change. Addressing these issues will create new opportunities for research in corpus linguistics and WEs.

References

[1] Biermeier, T. (2014). Compounding and suffixation in World Englishes. In The Evolution of Englishes: The Dynamic Model and Beyond (pp. 312-330). John Benjamins.

[2] Collins, P. (2008). The progressive aspect in World Englishes: A corpus-based study. Australian Journal of Linguistics, 28(2), 225-249.

[3] Collins, P. C. (2022). Comment markers in world Englishes. World Englishes, 41(2), 244-270.

[4] Gries, S. T., Bernaisch, T., & Heller, B. (2018). A corpus-linguistic account of the history of the genitive alternation in Singapore English. Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties (pp. 312-330). John Benjamins.

[5] Rossouw, R., & Van Rooy, B. (2012). Diachronic changes in modality in South African English. English World-Wide, 33(1), 1-26.

[6] Suárez-Gómez, C., & Seoane, E. (2023). The role of age and gender in grammatical variation in world Englishes. World Englishes, 42(2), 327-343

About the author:

Jiale Ding (b. 2001), female, Han nationality, from Hequ County, Shanxi Province, China, is a graduate student at University College London, London, United Kingdom. Her research interests are in World Englishes and corpus linguistics.