Corpora in World Englishes

SANDRA C. DESHORS AND STEFAN TH. GRIES

Introduction

Over the past 50 years, the field of World Englishes (WEs) has undergone substantial methodological and theoretical developments that have gone hand-in-hand. A particularly influential methodological development has been the widespread turn toward corpora (singular, *corpus*), that is, collections of naturally occurring spoken or written text in electronic format. This adoption of corpus methods has been accompanied by a growing adoption of modern statistical approaches, which have informed studies of concrete phenomena (e.g., grammatical variation across Englishes) or the development of theoretical models of WEs. As Deshors and Bernaisch (2019, p. 85) note, through methodological advancements "not only have WE researchers managed to better understand the linguistic forces that drive the development of Englishes worldwide, but corpus-based research of world Englishes has become its own academic sub-field." Two very broadly defined kinds of studies can be distinguished: those that are more theoretical and structural in nature and those whose focus is more applied and anchored in sociolinguistic and cultural frameworks. In what follows, we discuss these two trends in that order and provide methodological commentary. In the final part of our chapter, we briefly discuss current developments in corpus-based WEs research and identify the main desiderata for future work.

More Theoretical Kinds of Research

WEs Scholars with more theoretical research agendas have used corpora to explore not only a wide range of linguistic areas including language structure (especially phonology, morpho-syntax, lexicogrammar) but also semantics and cross-variety linguistic variation. Much such work is anchored in usage-based frameworks (e.g., Construction Grammar), which assume that speakers' knowledge of linguistic items is correlated with the items' distributional characteristics in authentic language (Goldberg, 2006; Langacker, 1987). Rautionaho et al. (2018) apply this perspective to the progressive versus non-progressive alternation by exploring three main types of Englishes from Kachru's (1989) classification of Englishes:

- English as second language (ESL): Indian, Singaporean, and Nigerian Englishes from the *International Corpus of English* (ICE);
- English as foreign language (EFL): Finnish, French, and Polish learner Englishes from the *Louvain International Database of Spoken English Interlanguage* and the *International Corpus of Learner English*;
- native Englishes (ENL): British and American Englishes from ICE and the *Santa Barbara Corpus of Spoken American English*.

To explore how the alternation is conditioned by different contexts across English types, the authors use state-of-the-art statistical approaches (cluster analysis and logistic regression

The Encyclopedia of Applied Linguistics. Edited by Carol A. Chapelle
(Section Editor: Eric Anchimbe).
© 2025 John Wiley & Sons, Ltd. Published 2025 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781405198431.wbeal20127

modeling) and find that individual varieties conform well to the traditional Kachruvian model: With non-progressives especially, there is a clear-cut divide between ENLs, ESLs, and EFLs. However, they caution that the validity of the ENL–ESL–EFL continuum can be influenced by scholars' choice of quantitative approach, thereby underscoring the tight connection between theory and method in WEs' corpus-based research.

Also exploring the classification of Englishes via morpho-syntax, Szmrecsanyi and Kortmann (2011) sets high methodological standards for the field. Using cluster-analytic techniques, they demonstrate the impact that sophisticated quantitative corpus-based methods can have on our understanding of the typology of Englishes worldwide and how structural (dis-)similarities across Englishes are best explained based on the combined effects of geography, type of English (i.e., EFL, ESL, ENL), and linguistic features. This type of approach helps identify typological issues/features (e.g., lower/higher degrees of morpho-syntactic or structural complexity) across foreign- and second-language varieties, which may have important, applied pedagogical implications for second-language learners, instructors, and users.

Schützler (2020) is a good example of how sophisticated statistical techniques (here, Bayesian mixed-effect logistic regression modeling) to corpora of WEs can lead to insightful discoveries on inter-varietal differentiation in the area of lexicogrammar. He explores the variation of the clausal positioning of 1,259 *although*-constructions (final vs. non-final) based on language production mode (speech vs. writing), cognitive processing, and semantics across British, Canadian, New Zealand, Nigerian, Indian, and Philippine English (from the ICE). Across all varieties, *although*-constructions prefer final positioning in speech and in dialogic contexts and "cognitive, processing, and production-based constraints are strong enough to keep inter-varietal differentiation in check" (Schützler, 2020, p. 455).

Corpus-based theoretical studies in phonology are relatively rare. However, Hay et al. (2018) show the potential of corpora to unveil, understand, and theorize the phonological patterns that characterize and distinguish Englishes worldwide. In two small corpus-based case studies they explore the usage patterns of linking- and intrusive-*r* usage in (nonrhotic) New Zealand English (NZE) to assess whether the patterns characteristic of early NZE are still observed today, based on both spontaneous speech and speech from a reading passage. The mixed-effects logistic regression statistical approach identifies (a) developmental patterns of linking- and intrusive-*r* through time, (b) influential patterns of morpheme and word boundaries on the production of *r*-sandhi, and (c) different usage patterns of *r*-sandhi by men and women.

Corpora have also allowed scholars to explore semantic questions in innovative ways. Mehl (2019) is a study on variation focusing on the three light verbs *make*, *take*, and *give* and light verb constructions (LVCs) in Singapore, Hong Kong, and British English (from the ICE). What stands out in the approach is the unusual direction in which form and meaning are studied, namely the onomasiological perspective from-meaning-to-form as opposed to from-form-to-meaning. Mehl (2019) finds that semantic patterns are constant across Englishes and notes, "there is no evidence for unique or innovative LVCs in the three corpora [and] there is remarkable similarity across the three corpora as well: all varieties, in most cases, prefer the related verb over the LVC in both speech and writing" (Mehl, 2019, p. 77).

Finally, in the area of language variation, corpora have helped identify ways in which individual dialects of native English can, overtime, influence the development of a non-native English variety. For example, Borlongan (2021) is a diachronic (1960s vs. 2000s) study of the understudied variety of Japanese English (in the Diachronic Corpus of Expanding Circle Englishes) and its use of putative American variants of spellings, single words, compound lexical items, lexical endings, suffixes in compounds, verb morphology, contractions, article usage, constituent sequence, verb–noun collocations, preposition choices, collective noun concord, and phraseology. Overall, the study portrays Japanese English as "overwhelmingly American in its choices of variants across the categories under investigation" (Borlongan, 2021, p. 54).

In sum, while the above overview is necessarily brief, it is clear that corpus-linguistic analysis has an increasingly greater impact on just about all areas of WEs research.

More Applied Kinds of Research

Sociolinguistic, pragmatic, or applied linguistics research has similarly benefited from the rise of corpus methods in linguistics and in WEs research in particular. The interconnectedness of linguistic, social, pragmatic, cultural, multilingual, and communicative aspects of language use is recognized as "a driving force beyond the structural development of Englishes" (Deshors & Gilquin, 2018, p. 287). A first case in point is van Rooy and Kruger (2018). Although that study is primarily a theoretical discussion on how to expand on current theoretical models of WEs, it also shows how multilingual digital repertoires can help us address empirical challenges resulting from the globalization of Englishes such as multilingualism, hybrid varieties, online communication, and complex identities. Based on a corpus of interactive user comments online that accompany daily summaries of the content of the most popular South African television soap operas, van Rooy and Kruger (2018) show how the core of online interactions consists of a shared pool of English (lexical) resources and global as well as local nonstandard English forms complemented by forms from South African languages. Ultimately, van Rooy and Kruger (2018) showcase the importance of accounting for WEs sociolinguistically and in ways that are ecologically valid for developing theoretical models of Englishes worldwide.

Funke and Bernaisch (2022), Revis and Bernaisch (2020), and Schneider (2018) are also located at the interface of culture and corpora and allow us to better understand who the speakers of WEs are and how their identities stand out systematically in corpus data. Starting with Schneider (2018), the study asks, how and to what extent traces of cultural impact can be detected in corpora. The study uses five components of the ICE representing (a) English as spoken in these countries and (b) major cultural traditions (specifically, Great Britain for a western culture; Hong Kong, Singapore, and India for Asian cultures; Nigeria for West African cultures). ICE data were occasionally supplemented with data from the Corpus of Global Web-Based English and quantitatively explored along three aspects: linguistic forms referring to objects/artefacts, expressions of cultural dimensions (e.g., collectivism vs. individualism, power distance, social relations), and linguistic constructions. Maybe unsurprisingly, it emerged that terms for concrete objects, artefacts, and notions appear in regional corpora (mainly locally) but cultural dimensions, while recognizable in corpus data, "vary from one dimension to another and by indicator terms, and they tend to be graded rather than absolute" (Schneider, 2018, p. 128). However, with linguistic constructions, traces of cultural influence are less clear: "such influences exist and have an impact, but if so, they are clearly more indirect and somewhat abstract" (Schneider, 2018, p. 128). Overall, the study shows how pervasive and diverse cultural information can be in corpora WEs and how informative such data is for WEs research.

Finally, and as for discourse and pragmatics, Revis and Bernaisch (2020) investigate how corpora inform discussions on pragmatic nativization across Englishes, a topic that remains today rarely explored. They focus on filled (e.g., *uh* or *uhm*) and unfilled pauses (i.e., silence) by speakers of Indian and Sri Lankan English (in the ICE) and explore whether (a) variety-specific differences exist in the use of filled/unfilled pauses and (b) what factors influence speakers' choice of pauses—specifically, whether speakers are influenced by the internal structure of texts and whether their choices are speaker-related, genre-related, or related to speakers' English variety. They include a wide variety of structural and socio-biographical factors to their analysis (e.g., word class, dialogue vs. monologue, scripted vs. unscripted text, speaker's age, gender, and English variety) and, therefore, use advanced predictive modeling techniques (conditional inference trees and generalized linear mixed-effects models). Their results show that (a) pauses cluster, (b) their choices are influenced by their frequencies in texts, and (c) filled pauses occur less in monologues. Funke and Bernaisch (2022) is a methodologically similar study—also adopting a multifactorial predictive modeling method (this time, random forest)—to explore how socio-biographic and pragmatic factors contribute to the use of intensifiers and downtoners also in Indian and Sri Lankan Englishes. Broadly, across varieties, intensifiers and downtoners are more similar than different and they are used more frequently by younger female in informal conversations.

As before, our discussion had to be selective, but hopefully still illustrates the wide variety of corpus methods and statistical applications in this part of WEs research.

Desiderata for the Future

By definition, corpus-linguistic research on WEs is where varieties research and general corpus linguistics intersect, which means its desiderata are motivated by both these disciplines. As for the former, the discipline of WEs needs to reconcile how multiple forces are pulling it into different directions. On the one hand, the field needs to address what its models are trying to do, what counts as important phenomena, and what counts as evidence. For instance, how much predictive capability do we require our models to have and how much do we allow them to be falsified from different kinds of data? (See Bernaisch et al., 2022 for discussion and critique of several common assumptions.) Relatedly, do we focus on, to use their terminology, linguistic butterflies (often infrequent surface structure deviations from, e.g., British English) or linguistic ants (often frequent structural choices with higher functional loads)? And, finally, how do we integrate usually current linguistic data with often diachronic sociocultural information (e.g., attitudinal and/or sociolinguistic information) and how do we reconcile different explanations for the same empirical findings?

On the other hand, there are many open methodological questions, some of them involving statistical, others involving resource availability. For example, how do we study what we study—do we rely on observed (relative) frequencies across varieties including comparisons with the presumed source variety or do we rely (more) on multifactorial statistical modeling and exploratory tools? And if we do the latter, what kinds of predictors are most relevant? Not all of the field has realized that even just for statistical reasons alone, level-1 predictors (i.e., observation-level predictors describing each individual speaker choice) need to be included to be able to for any generalizations regarding level-2 predictors (e.g., speaker-specific predictors) or level-3 predictors (e.g., variety differences) to be valid (see Gries, 2023, Section 2). Lastly, we need better resources; most importantly, we need

- more corpora with more diverse register/genre coverage to better study the dispersion of features—ants and butterflies—in and across varieties;
- more diachronic corpora for research on phenomena such as epicentral influence, because Gries et al. (2018) have demonstrated that the currently predominant apparent-time approach to indigenization or nativization phenomena cannot deliver the results it has been claiming to deliver.

With such additions to our methodological toolkit and corpus resources and a renewed focus on what theoretical models can and should do, corpus-based research on WEs will continue to evolve in promising ways.

Q1 wbeal20523 wbeal0033 wbeal0258.pub3 wbeal20158 wbeal20137

SEE ALSO: Bilingualism and Multilingualism - An Overview of the Field; Corpus Linguistics: Overview; Corpus Linguistics: Quantitative Methods; Edgar W. Schneider; The Dynamic Model of Postcolonial English; World Englishes and the Native Speaker

References

Bernaisch, T. J., Gries, S. T., & Heller, B. (2022). On the relation between theoretical models and statistical modeling: The case of linguistic epicentres. *World Englishes*, *41*, 333–346.

Borlongan, A. M. (2021). A new American-lineage English? Proportions of American variants in Japanese English. *Asian Englishes*, *23*(1), 51–61.

Deshors, S. C., & Bernaisch, T. (2019). Corpus approaches to World Englishes: A bird's eye view. In P. de Costa, D. Crowther, & J. Maloney (Eds.), *Investigating World Englishes: Research methodology and practical applications* (pp. 21–43). Routledge.

Deshors, S. C., & Gilquin, G. (2018). Modeling world Englishes in the 21st century: New reflections on model-making. In S. C. Deshors (Ed.), *Modeling World Englishes in the 21st Century: Assessing the interplay of emancipation and globalization of ESL varieties* (pp. 281–294). John Benjamins.

Funke, N., & Bernaisch, T. J. (2022). Intensifying and downtoning in South Asian Englishes: Empirical perspectives. *English World-Wide*, *43*(1), 33–65.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Gries, S. T. (2023). Corpus-linguistic and computational methods for analyzing communicative competence: Contributions from usage-based approaches. In M. H. Kanwit & M. Solon (Eds.), *Communicative competence in a second language: Theory, method, and applications* (pp. 115–131). Routledge.

Gries, S. T., Bernaisch, T. J., & Heller, B. (2018). A corpus-linguistic account of the history of the genitive alternation in Singapore English. In S. C. Deshors (Ed.), *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties* (pp. 245–279). John Benjamins.

Haselow, A. (2021). Dealing with trouble in conversation in English-speaking cultures: Conversational repair in global varieties of English. *English World-Wide*, *42*(3), 324–349. Q2

Hay, J., Drager, K., & Gibson, A. (2018). Hearing R-sandhi: The role of past experience. *Language*, *94*(2), 360–404.

Kachru, B. B. (1989). World Englishes and applied linguistics. In M. L. Tickoo (Ed.), *Languages and standards: Issues, attitudes, case studies* (pp. 178–205). SEAMEO Regional Language Centre.

Kortmann, B. (2010). Variation across Englishes: Syntax. In A. Kirkpatrick (Ed.), *The Routledge handbook of World Englishes* (pp. 422–446). Routledge.

Kortmann, B., Lunkenheimer, K., & Ehret, K. (Eds.). (2020). *The electronic World Atlas of varieties of English*. Zenodo.

Kruger, H., & van Rooy, B. (2019). A multifactorial analysis of contact-induced change in speech reporting in written White South African English (WSAfE). *English Language & Linguistics*, *24*(1), 179–209.

Laitinen, M. (2020). Empirical perspectives on English as a lingua franca (ELF) grammar. *World Englishes*, *39*(3), 427–442.

Langacker, R. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (vol. 1). Stanford University Press.

Mehl, S. (2019). Light verb semantics in the International Corpus of English: Onomasiological variation, identity evidence and degrees of lightness. *English Language and Linguistics*, *23*(1), 55–80.

Rautionaho, P., Deshors, S. C., & Meriläinen, L. (2018). Revisiting the ENL-ESL-EFL continuum: A multifactorial approach to grammatical aspect in spoken Englishes. *ICAME Journal*, *42*, 31–68.

Revis, M., & Bernaisch, T. J. (2020). The pragmatic nativisation of pauses in Asian Englishes. *World Englishes*, *39*(1), 135–153.

Schneider, E. W. (2018). The interface between cultures and corpora: Tracing reflections and manifestations. *ICAME Journal*, *42*, 97–132.

Schützler, O. (2020). *Although*-constructions in varieties of English. *World Englishes*, *39*(3), 443–461.

Szmrecsanyi, B., & Kortmann, B. (2011). Typological profiling: Learner Englishes vs. Indigenized L2 varieties of English. In J. Mukherjee & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 167–188). John Benjamins.

van Rooy, B., & Kruger, H. (2018). Hybridity, globalisation and models of Englishes: English in South African multilingual digital repertoires. In S. C. Deshors (Ed.), *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties* (pp. 77–108). John Benjamins.

Suggested Readings

Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction* (2nd ed.). Routledge.

Schreier, D., Hundt, M., & Schneider, E. W. (Eds.). (2020). *The Cambridge handbook of World Englishes*. Cambridge University Press.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.

The abstract and keywords will not be included in the PDF or any printed version of your article, but are necessary for publication on Wiley's online publishing platform to increase the discoverability of your article.
If the abstract and keywords are not present below, please take this opportunity to add them now.
The abstract should be a short paragraph up to 200 words in length and keywords between 5 to 10 words.

Abstract: In this overview, we survey recent and current developments in corpus-based research on World Englishes. We exemplify current strands of research in both more theoretical and more applied parts of research on varieties of English and conclude with theoretical, methodological, and resource desiderata.

Keywords: corpora; varieties research; nativization; epicenter research; statistical modeling