June 2025: Author's accepted manuscript,
forthcoming in *English World-Wide*

The YouTube Corpus of Singapore English Podcasts

Steven Coats, University of Oulu, Finland; Carmelo Alessandro Basile, Sorbonne Nouvelle University, France; Cameron Morin, Paris-Cité University, France; Robert Fuchs, University of Bonn, Germany

Running head
SINGAPORE ENGLISH PODCAST CORPUS

Abstract
Recent advances in streaming protocols and automatic speech recognition (ASR) have enabled large-scale spoken language corpora, yet research on Singapore English remains constrained by small or text-based datasets. The YouTube Corpus of Singapore English Podcasts (YCSEP) addresses this gap with 620 hours of transcribed, diarized speech from over 1,300 podcast episodes by Singapore-based content creators. YCSEP supports the empirical analysis of phonetics, morphosyntax, and discourse, enabling the study of low-frequency features like discourse particles and reduplication. The dataset reflects informal, spontaneous speech from diverse speakers and facilitates investigation into nativization and endonormative stabilization processes in postcolonial English. Built using a pipeline of yt-dlp, WhisperX, and Pyannote, YCSEP offers robust empirical grounding for linguistic features such as verb complementation and modality. It also contributes to broader theoretical discussions on areal norms and construction grammar in World Englishes.

Keywords
Singapore English, Corpus Linguistics, Automatic Speech Recognition, Discourse Particles, Whisper, Grammar, Phonetics, New Englishes

Acknowledgements
The authors wish to thank Jakob Leimgruber for his comments on a draft version of the text and Finland's Centre for Scientific Computing (https://csc.fi) for access to computational resources.

1. Introduction

Over the past 15 years, access to online audio and video content, standardization of video and audio streaming protocols, and improvements in automatic speech recognition (ASR) accuracy have greatly facilitated the collection of curated speech datasets, opening new perspectives for empirical language research. Descriptive, corpus-based research on Singapore English, a recently nativized postcolonial variety, has hitherto been undertaken mostly on the basis of relatively small datasets or corpora of written language, data which are not always suitable for capturing the full range of lexical, grammatical, and syntactic variation in the variety. In this article, we introduce the *YouTube Corpus of Singapore English Podcasts* (YCSEP), comprising ASR transcripts and audio content for 620 hours of naturalistic Singapore English speech from

over 1,300 individual podcast episodes by popular Singapore-based podcasters.[1] Transcripts were generated with state-of-the-art ASR technology and include automatic speaker diarization, enabling analysis of turn-based interaction.

The corpus provides a rich source of data for the analysis of the phonetics and phonology, lexis, morphosyntax, and pragmatics of Singapore English, including linguistic and interactive phenomena that have been well documented in the existing research literature, but not yet sufficiently attested in empirical data due to their rarity. As an increasingly 'mature' English variety that, according to theoretical accounts, has entered the "endonormative stabilization" stage (Schneider 2007: 153–161) and is increasingly acquired as a native language at home (Tan 2014), Singapore English may additionally exhibit innovative features that have not yet been systematically described on the basis of naturalistic speech data, pertaining to (e.g.) verb complementation, reduplication, or modality (see, e.g. Wee 2004; Bao 2010a,b). YCSEP may serve as a proving ground for the discovery and analysis of such features, which tend to emerge "at the interface between lexis and grammar" (Schneider 2007: 83).

From a theoretical perspective, the YCSEP represents data on the basis of which conjectures pertaining to the typology of postcolonial English varieties may be testable – for example, the emergence of "areoversal norms" (Kortmann 2019) in Southeast and South Asia or construction grammar accounts of nativization (Hoffmann 2021); the data may also be amenable to analysis using computational construction grammar techniques (Dunn 2024).

The rest of the article is organized as follows: In Section 2, the pipeline-based methods for data collection, transcription, and processing are described and information about the sampled podcasts is provided. Section 3 discusses a few of the principal features of Singapore English (SgE), research into their properties in previous corpus-based accounts, and the possibilities that YCSEP offers for more in-depth investigation; in addition, evidence is provided for features that have until now not been considered to be part of the inventory of SgE. Section 4 discusses the applicability of the corpus for studies of the phonetics and prosody of the variety. Section 5 discusses several limitations of the underlying data and caveats for working with the corpus, followed by a brief outlook for future work.

2. Existing SgE resources

2.1 Written and mixed corpora

There are several corpora that provide written, or written and spoken, data on SgE. These corpora vary in size and composition (see Table 1). Some, like the ICE Singapore corpus, follow a carefully designed multi-genre design, while others, such as *GloWbE*, contain various text types without categorisation. Some specialised corpora, such as *CoSEM*, comprise a single text type, such as mobile text messages.

[1] Version 0.1 of the searchable online corpus is available at https://ycsep.corpora.li.

Table 1: Written and mixed corpora of SgE

Corpus Name	Text Types	Size (words, rounded)	Reference
International Corpus of English (ICE) Singapore	Various spoken and written text types, structured	1 million	Nelson 2002
Flowerpod Corpus	Internet forum and blog content	700,000	Ziegeler 2014
Corpus of Global Web-Based English (GloWbE) Singapore	Various text types from the internet, unstructured (split into blogs and general)	43 million	Davies and Fuchs 2015
News on the Web (NOW) Singapore	Newspapers and magazines	Not specified, but probably larger than GloWbE Singapore	Davies 2016–
Corpus of Singapore English Messages (CoSEM)	Mobile text messages from WhatsApp	6.9 million	Gonzales et al. 2023
HardWareZone Corpus	Blogs	525,000	Basile 2024
Salary Corpus	Blogs	480,000	Basile 2024

2.2 Speech corpora

Several spoken SgE corpora or audio resources have been created, including the *NIE Corpus of Spoken Singapore English* (NIECSSE; Deterding and Low 2001), the *Grammar of Spoken Singapore English Corpus*, (GSSEC; Lim 2001; Lim and Foley 2004) the Singapore component of the *International Corpus of English* (ICE-SIN), and the *Singapore English Computer-Aided Language Learning* (CALL) resource (Chen et al. 2010). As of early 2025, NIECSSE audio and transcripts are available online; the availability of the audio recordings for the GSSEC (which was later incorporated into the spoken component ICE-SIN), the audio for ICE-SIN, and the audio and transcripts for the CALL resource are unclear. The most comprehensive corpus of SgE speech is the Singapore English *National Speech Corpus* (NSC; Koh et al. 2019), described as having been created “for automatic speech recognition (ASR) research and speech-related applications”. The original NSC, released in 2019, comprises approximately 3,000 hours of speech from reading and conversational contexts, including use of words and

phrases specifically relevant for Singapore. The conversational data was manually transcribed. An additional three parts, containing codeswitching conversations, debates, finance discussions, emotional speech, and simulated call center calls, were released in 2021. Also noteworthy is the Oral History Interviews project of the Singapore National Archives,[2] a searchable collection of transcripts and audio from interviews conducted with Singaporeans, which has also been utilized for linguistic analysis (Li et al. 2022).

Corpus Name	Text Types	Size (words, rounded)	Reference
NIECSSE	Excerpts of interview transcripts	< 20,000	Deterding and Low 2001
GSSEC	Dyadic and group conversations	~ 60,000	Lim 2001
Singapore English CALL resource	Read sentences	?	Chen et al. 2010
National Speech Corpus Parts 1-3	Phonetically balanced sentences, random sentences containing Singapore-relevant words, dyadic conversations	~ 13 million	Koh et al. 2019
National Speech Corpus Parts 4-6	Codeswitching dyadic conversations, debates, finance discussions, emotion recordings, simulated call center calls	~ 7 million	Koh et al. 2019

2.3 Evaluation of existing corpora and databases

As the brief overview above shows, there are several corpora and databases of SgE available to the research community. Some of these corpora contain informal written or informal spoken data, but are comparatively small, enabling research on high-frequency linguistic phenomena. Other corpora are large, but contain mostly formal, written data. Additional datasets containing larg(er) amounts of informal spoken data would enable researchers to analyze less frequent features of colloquial SgE and might also enable data-driven sociolinguistic research. The YouTube Corpus of Singapore English addresses this gap.

[2] https://www.nas.gov.sg/archivesonline/oral_history_interviews

3. Corpus creation: Data and methods

The corpus content was drawn from Singapore podcasts available on the online video platform YouTube. While not all Singapore podcasters upload their content to YouTube (some distribute episodes via Spotify or other content providers), the country's most popular podcasts are typically accessible on multiple content platforms, including YouTube. Corpus content was harvested using the open-source Python package yt-dlp,[3] as of 2025 the most widely used scripting library for downloading content from YouTube and other websites that utilize popular streaming protocols such as DASH (Dynamic Adaptive Streaming over HTTP; Sodagar 2011) or HLS (HTTP Live Streaming; Pantos and May 2017). We selected podcast content available via YouTube not only to facilitate harvesting with yt-dlp, but also due to the accessibility of the platform. As of early 2025, YouTube remains fundamentally open; content is accessible without a subscription. Thus, analysts who wish to verify a particular utterance or exchange or to consider multimodal aspects of communication in the context of the original underlying video can do so on the YouTube platform.[4]

From the perspective of copyright, the corpus is made available under the provisions of Sections 190–199 of the Singapore Copyright Act of 2021, which explicitly permits the use of copyrighted material for purposes such as research and study. Use of copyrighted material for research purposes is also permitted under US and EU law.[5] Although the transcripts in YCSEP are diarized (see below), utterances are not linked in corpus metadata to real identities. YCSEP acknowledges individuals' rights to object to the processing or storage of any personal data. Where appropriate and feasible, personally identifying information may be removed from the dataset upon request by individuals whose speech is represented in the corpus materials.

3.1 Podcasts included in the corpus

The first release of the corpus (v1) contains content from six podcast channels: *Historyogi*, *NOTG* ("Nuggets On The Go"), *Randomly Relatable SG*, *The Daily Ketchup Podcast*, *Yah Lah BUT*, and *You Got Watch*. Podcast hosts include representatives of Singapore's main population groups (Chinese, Malay, and Indian), and episodes feature a diverse array of topical content, including discussion, commentary, and interviews and conversations with a wide range of individuals, from musical artists and online influencers to government ministers. As a result, the corpus represents a diverse sample of natural conversational English spoken in Singapore.

In order to illustrate the nature and breadth of the data, we briefly describe each of the podcasts: (i) The podcast *Historyogi* is described on its "About" page as "focused on Singapore history & heritage". The content, much of which consists of commentary on historical persons

[3] https://github.com/yt-dlp/yt-dlp.

[4] However, videos are not guaranteed to be available forever – they can be made private or removed by channel owners. In addition, YouTube/Google may limit access to content in the future.

[5] US Code Title 17 provides for "fair use" of copyrighted material; EU Directive 2019/790 permits "text and data mining for the purposes of scientific research".

and events, is generated by a 33-year-old Singaporean of Indian heritage. (ii) *NOTG* ("Nuggets On The Go") is a podcast by two well-known Singaporean realtors and businessmen, the brothers Melvin and Adrian Lim, who describe their podcast as "dishing out golden nuggets on real estate, finance, business, entrepreneurship, and investing, Singapore-style". The *NOTG* podcast is one outlet among many for the hosts, who have a significant online presence.[6] (iii) The popular podcast *Randomly Relatable SG* describes itself as "the Number 1 Youth Podcast in the World!"; the three principal hosts, of Malay background, describe their content as "We talk about random things you relate to and provide a platform for the unspoken to speak". The content of the podcast comprises discussions on various topics and interviews with guests. (iv) The *Daily Ketchup Podcast* is one of Singapore's most popular podcasts, as of early 2025.[7] Podcast episodes typically feature discussion by five regular hosts, of Chinese background, as well as interviews and discussions with a wide range of guests, including members of the Singapore Government. (v) *Yah Lah BUT*, hosted by two Singaporeans of Indian and of Chinese background, is also widely popular. The podcast features political commentary, as well as other content, in the form of interviews and discussions; it is described as "uncensored". (vi) *YouGotWatch's* YouTube page contains various types of content. The podcast, called "The Hop Pod", mainly features interviews with guests by the three main hosts, who are Singaporeans of Chinese background. The podcast's blurb is "we hop into different pods of spaces people find themselves in during life's transitions". Table 2 provides a breakdown of the corpus content by channel, number of episodes, number of utterances, and number of words.

Table 2: Corpus content by podcast channel, no. of episodes, utterances and words, and total length.

Podcast channel	**No. of episodes**	**Length (hours)**	**No. of utterances**	**No. of words**
Historyogi	43	15.93	16,358	173,605
NOTG	134	62.54	65,155	821,538
Randomly Relatable SG	82	17.20	26,005	227,103
The Daily Ketchup	367	172.90	218,011	2,479,299
Yah Lah BUT	372	288.32	347,736	3,836,627
YouGotWatch	314	63.40	83,807	842,654
Total	**1,312**	**620.29**	**757,072**	**8,380,826**

[6] See, e.g. https://plbinsights.com.
[7] See, e.g. https://podcast.feedspot.com/singapore_podcasts

Although the subcorpora vary in size, they collectively offer a rich mixture of speech styles, speaker identities, and communicative contexts. This diversity supports a broad range of sociolinguistic and discourse-analytic investigations into contemporary spoken English in Singapore.

3.2 Data harvesting and corpus creation

After identification of suitable podcast channels, yt-dlp was used to extract the audio, in WAV format, for all available videos in each channel, typically several hundred. Wav files, which ranged in size from approximately 30 MB to 2 GB per file, were stored in a cloud container. Individual .wav files were then sent to a GPU processing pipeline on servers hosted by Finland's Centre for Scientific Computing. Whisper's large-v3 model (Radford et al. 2023), as implemented in the Python libraries Faster-Whisper[8] and WhisperX (Bain et al. 2023), was used to generate ASR transcripts from the audio; speaker diarization (i.e. an indication of which speaker produced a particular utterance turn) was implemented with pyannote (Bredin 2023; Plaquet and Bredin 2023), via the WhisperX package. For each .wav file, an output file in Praat's TextGrid format was generated, with each automatically detected speaker in that video assigned to their own tier. A custom parser was then used to organize the transcript content in order of speaker turns in each podcast episode. Automatic annotation of part-of-speech tags or dependency relations can increase the utility of corpora, but only recently been implemented for SgE corpora (Lin 2022; Huang et al. 2025). For YCSEP, part-of-speech tags were appended using SpaCy's en_core_web_sm model (Honnibal et al. 2020). Some utterance turns were incorrectly transcribed by Whisper as repeated single-word utterances (e.g. "talk talk talk talk talk talk"). A regular expression was used to remove 141 of these hallucinations from the final dataset. The corpus is available in two versions: A static, tabular version, downloadable from the Harvard Dataverse, and a searchable online version.[9] The online version includes the playable audio segments.

4. Corpus analysis of SgE features

4.1 Features associated with SgE

As extensively shown in previous studies (e.g. Deterding 2007; Leimgruber 2013; Bao 2015; Ziegeler 2015; Basile 2024, among others), the variety of English spoken in Singapore is the result of intense contact with different languages. This contact dates back to at least 1819, when the East India Company officially established a trading post in Singapore (Wong 2016). Today, English has the status of official language in the island, alongside Mandarin Chinese, Malay, and Tamil. The wide-ranging influence of other Chinese languages – including Hokkien, Cantonese, and Southern Min varieties – as well as, to a lesser extent, Malay and Tamil, is evident in various morphosyntactic constructions typically found in this Asian variety. Bao points out (2010b: 1727)

[8] https://github.com/SYSTRAN/faster-whisper
[9] https://doi.org/10.7910/DVN/B7JRID; https://ycsep.corpora.li.

that some features transferred from these languages include (1) topic prominence, (2) clause-final particles, (3) the use of *got* to express possession or existence, and (4) the use of *already* as a marker of the perfective aspect:

(1) Good suggestion hah this one. (ICE-SIN:S1A-013)
'(A) good suggestion, this one is'.
(2) Be more professional lah, you know. (ICE-SIN:S1A-011)
'(You should) be more professional, you know'.
(3) Apple pie inside got wine or not? (ICE-SIN:S1A-006)
'Does the apple pie contain wine or not?'
(4) Actually we challenge him for two years already lah. (ICE-SIN:S1A-013)
'Actually, we have challenged him for two years'.

A large corpus such as YCSEP allows researchers to conduct further investigations on features of this type by exploiting recent spoken data across channels dealing with varied topics and reflecting different registers. Until now, most corpora of SgE containing recent data from the last decade – such as the *Corpus of Global Web-based English* (GloWbE) (Davies and Fuchs 2015), the *Corpus of Singapore English Messages* (CoSEM) (Gonzales et al. 2023), the *HardWareZone Corpus* (Basile 2024), and the *Salary Corpus* (Basile 2024) – have primarily relied on written sources, where contributors' identities often remain unknown. The audio files for the *International Corpus of English – Singapore* (ICE-SIN), collected in the 1990s, are not openly available. The *Singapore National Speech Corpus*, developed for natural language processing applications, is a large and systematically documented dataset. However, because its conversational speech samples were recorded either under laboratory conditions or via video conferencing, they may not fully reflect the face-to-face contexts and broader range of discourse topics typically observed in naturalistic speech communication. The creation of YCSEP thus marks a significant advancement, providing a rich resource for exploring features of Singaporean English synchronically while also enabling diachronic comparisons with transcripts of spoken data from the 1990s both on a phonological and on a grammatical level.

Discourse particles of non-English origin, such as *meh, lor, ah, sia, bah*, and *lah*, are a prominent feature of SgE. Particularly *lah* is a central feature of the variety, as argued by Leimgruber et al. (2020: 605), appearing in nearly all linguistic accounts, including early ones such as Richards and Tay (1977) and Kwan-Terry (1978). According to Lim (2007), *lah* is a particle that "draws attention to mood or attitude and appeals for accommodation; indicates solidarity, familiarity, [and] informality" (Leimgruber et al. 2020: 607). In order to illustrate the usefulness of YCSEP as a dataset for investigating typically oral features, we provide an overview of the relative frequency of the particle in recent corpora or datasets (Table 3).

Table 3: Relative frequency of *lah* in selected SgE corpora

Resource	**Frequency (pmw)**	**Source**
GSSEC	11,193	Smakman and Wagenaar (2013)
NSC spontaneous conversations	13,574	Boo et al. (2023)

(private corpus)	8,044	Botha and Bernaisch (2025)
GloWbE	19	english-corpora.org
YCSEP	1,395	

We find a frequency for *lah* of 11,692 (1,395 per million words) in YCSEP, a rate which is comparable, but somewhat lower than those reported in recent studies or corpora of spoken SgE: Botha and Bernaisch (2025) report a relative frequency of 8,044 pmw in a sample of transcribed conversational speech,[10] and Smakman and Wagenaar (2013) find 11,193 pmw in the GSSEC. Boo et al. (2023) find 68,149 *lah* tokens in the "spontaneous conversation component of the NSC", presumably Part 3 of the resource, which would correspond to a relative frequency of 13,574 pmw. In comparison, a corpus based on written data such as GloWbE, which consists of web-based English texts from the early 2010s, shows a dramatically lower frequency of *lah,* at just 18.76 instances per million words. While the reasons for the differences in frequency of *lah* in the transcript datasets may include differences in speech genre and formality as well as issues pertaining to ASR, the preliminary frequency data for this particle suggests that YCSEP may be a useful resource for analysing features specific to spoken interaction. A future version of YCSEP may utilize a speech-to-text model fine-tuned on conversational SgE speech. Experiments along these lines suggest that such a model may better capture some characteristic features of SgE, resulting in higher relative frequencies of *lah* and other discourse particles.

YCSEP is also a valuable tool for observing trends in grammatical change, such as the evolution of modal constructions, particularly when these are affected by contact phenomena with other languages. Since at least Leech (2003), it has been evident that traditional core modal auxiliaries in US and British English (e.g. *must* for the expression of necessity) have been declining in favor of alternative semi-modal constructions such as *have to* and *need to*, which may be perceived by speakers and hearers as less forceful or more neutral modals and therefore as more polite (cf. Mair and Leech 2020). More recently, Hansen (2018) and Basile (2023, 2024), have investigated whether this trend extends to SgE. Basile (2024: 194) demonstrated that *need to* is increasing in frequency in this variety, especially among younger generations. He attributed this change to the influence of Mandarin, where the construction *xūyào* exhibits semantic and syntactic parallels to English *need to* (cf. also Hansen 2018: 217). By analysing data spanning 30 years, Basile (2024) further argued that despite a slight decline in frequency, *must* remains productive in SgE and is grammaticalizing at a statistically significant rate ($p < 0.005$) toward epistemic functions. This result challenged Bao's (2010b) claim that *must* is almost exclusively limited to non-epistemic functions in this variety due to substrate influence.

Future research leveraging YCSEP could offer further insights into the behavior of these modal constructions. A preliminary analysis, as shown in Figure 1, highlights the frequency of *must*, *have to*, and *need to* across the six podcast channels of the corpus. Overall, the data confirm that the *need to* construction is particularly frequent across most podcast channels,

[10] The dataset, comprising ~100,000 words of transcribed conversational speech, does not seem to be publicly available.

despite their differences (cf. Section 3.1). This trend contrasts sharply with British English data, such as ICE-GB, where *need to* appears less than a third as often as *have to* in both public and private dialogues (236 vs. 722 occurrences per 1 million words, respectively; cf. Basile 2024: 147).

Further investigations are warranted to examine differences in frequencies between channels (i.e. lower use of *need to* in the channels *Randomly Relatable* and *Yah Lah BUT*), which may reflect different conversational styles of podcast participants. In addition, the distribution of functions (dynamic, deontic, and epistemic) for each of these modal constructions could be examined to confirm or challenge previous claims and findings. For example, it has been suggested that *must* in SgE is more commonly used to express dynamic modality – indicating general necessity rather than future obligation as in (5) – compared to British English, its historical ancestor (see Basile 2024). The YCSEP opens up new avenues for these types of explorations.

(5) [Talking about general steps to follow to create the perfect Tinder profile]
Speaker 1: Then you ***must*** choose the correct photo lah. You post with people uglier than you lah.
Speaker 0: No, no, no, no, no. You put photo with like handsome people lah. Then I will think like, oh, meaning this guy probably handsome in real life (YCSEP, *Randomly Relatable SG*, July 7th, 2021).

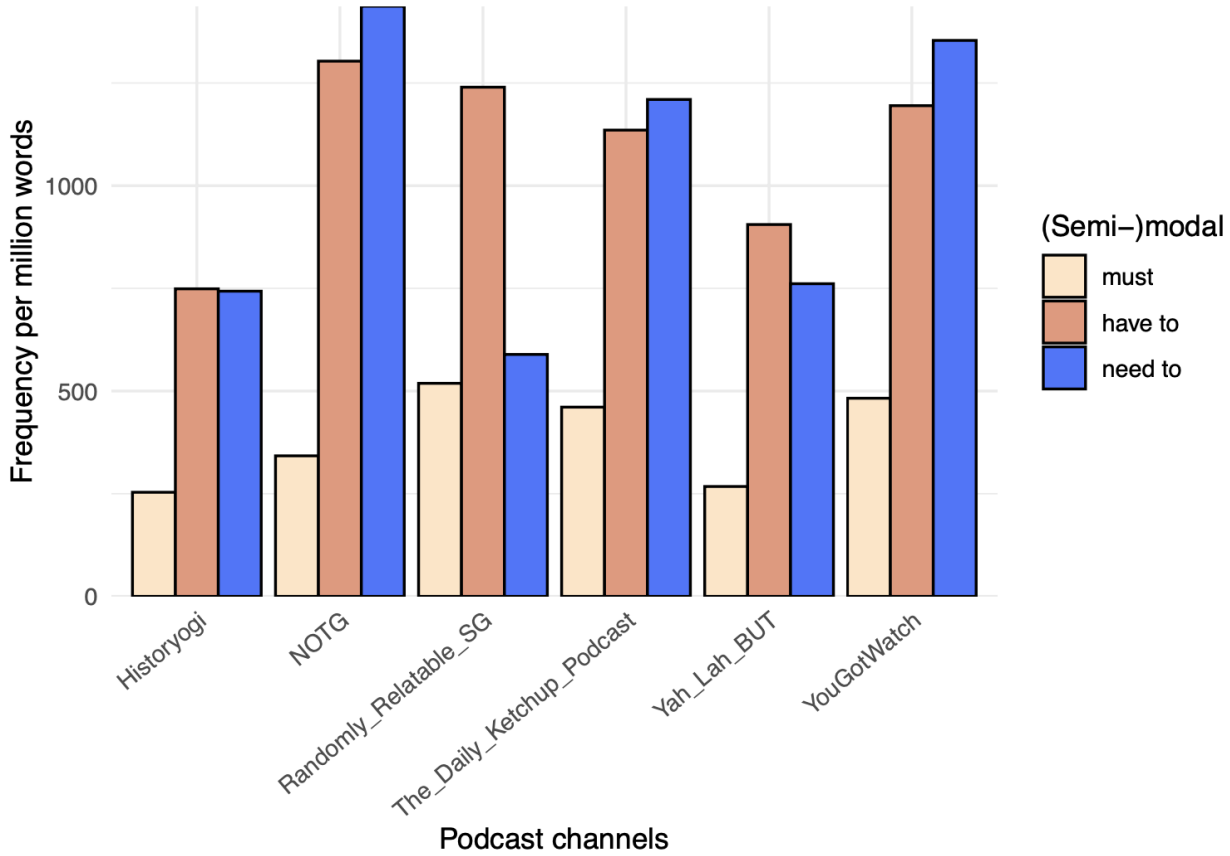

Figure 1. Distribution (per million words) of *must*, *have to*, and *need to* in the YCSEP podcast channels.

4.2 Other features

In addition to enabling the study of known features of SgE at scale in contemporary social media registers, YCSEP provides a unique opportunity to explore and observe previously unattested linguistic and grammatical features. To illustrate this, we conducted an exploratory study of the possible existence of double modals in SgE, which, if attested, may challenge both existing accounts of grammatical variation in that variety and existing typological accounts of double modals in English world-wide (Zullo et al. 2021).

Double modals are rare grammatical features that have long been considered restricted to a handful of British and American dialects of English, especially in the Deep South of the United States and in the Scottish Borders region (Coats and Morin 2024). However, recent studies utilizing a 'computational sociolinguistic' approach (Nguyen et al. 2016; Grieve et al. 2023) and relying on large corpora of geolocated social media data from Twitter and YouTube (Coats 2023b) have uncovered more widespread use of these constructions in the United States and the United Kingdom than previously considered (Coats 2023a, 2024; Morin and Grieve 2024; Morin et al. 2020; Morin et al. 2024). For example, Morin and Coats (2023) used a computational sociolinguistic method to attest double modals for the first time in two previously unaccounted for varieties of English: Australian and New Zealand English. Among their conclusions, the authors hypothesized that double modals may be a syntactic possibility for most speakers of English across the globe, and also raise a number of questions about their diachronic sources across different varieties, especially in the form of alternatives to the Scots-Irish double modal import hypothesis (Morin 2023). In particular, the exploration of double modals in a wider range of outer-circle varieties of English and varieties in English beyond the Atlantic, such as in the Pacific or Asia, may shed further light on these low-frequency constructions.

In what follows, we briefly report findings on possible double modals in YCSEP, attesting these features in SgE for the first time. We used regular expressions to search the corpus for all combinations of the modal forms *may*, *can*, *might*, *ought to*, *oughta*, *must*, *will* (and the contracted form *'ll*), *would*, *shall*, *should*, *could*, and *used to*. This yielded a total of ten search hits. Following manual inspection to remove false positives, conducted in line with the annotation protocol in Coats (2023, 2024), Morin and Coats (2023) and Coats and Morin (2024), we arrived at a final double modal dataset containing six double modal tokens, some of which are illustrated in Examples 6-8 below. The complete list contained *should ought to, could can, will might, would used to, would may, and will may*.

(6) you know, you **should ought** to serve. (Yah Lah BUT Podcast, 22 November 2022)
(7) and who knows maybe with someone that I really didnt like I **would may** not have done the same thing yeah (Yah lah BUT Podcast, 19 June 2023)

(8) he **would used to go** to a church to pray daily before (Daily Ketchup podcast, 25 May 2023)

The search results contain the metadata that permit the researcher to analyze the use of the double modal in context in the source video (see appendix) and to further adjudicate whether it appears to be a true, idiomatic double modal for the speaker (with no prosodic disfluencies and no visual cues of hesitation), a double modal innovation, or a borderline case of self-repair (Coats 2023).

While the dataset is smaller than the ones collected in the authors' previous studies of British, American, Australian, and New Zealand English, the relative frequency of 0.72 per million words for the feature in the Singapore data is comparable to that found for Australian and New Zealand English by Morin and Coats (2023); it is also corroborated by findings from a recent study of double modal frequencies in the blog subcorpus of GloWbE, where SgE shows 1.92 double modals per million words (Collins and Smith 2025). The YCSEP result contributes to our growing understanding of double modals in varieties of English, constituting further evidence that (i) double modals are valid, low frequency grammatical constructions of the English language, and that (ii) they are typologically more widespread than previously found.

Together, the case studies above, focusing on discourse particles and modal constructions, underscore the opportunities corpora like YCSEP offer for the study of features typical of spoken language and grammatical structures shaped by contact phenomena. This corpus not only facilitates the exploration of distinctive traits of SgE but also encourages broader discussions on the intricate dynamics of language contact and change.

5. Potential future use cases

5.1 Features associated with spoken language which are uncommon in writing

Spoken language is distinct in many ways from written language. A lack of large datasets of speech for many World Englishes has so far limited the extent to which research has been able to explore variation in speech (see Sec. 2). Not only is speech a distinct modality compared to writing, with its own linguistic characteristics, but informal and colloquial language, as well as linguistic innovation, is also more prevalent in speech compared to writing (Kirkpatrick 2007: 34–35, 172).

One notable feature of spoken SgE is reduplication, i.e. the repetition of a word that is semantically and pragmatically distinct from its single use and that is not due to performance issues such as hesitation, stuttering, or self-repair. Reduplication is a productive grammatical feature of Chinese languages, and its use in SgE has been proposed to represent a substratum-derived feature with limited productivity which can be employed to express meanings such as tentativeness, affection, intensity, continuity, or quantification (Bao 2010a: 806ff.). Bao found only a handful of reduplicatives in ICE-SIN data and noted that without access to the underlying audio, it can be difficult to disambiguate reduplicatives and instances of self-repair. Deterding (2007: 54–56) found no genuine instances of word reduplication in his academic interview data,

but reported a few cases from blogs. The examples he provided from his own data and from the literature include (i) cases involving nouns, such as *buddy buddy* (*I'm the kind who is buddy buddy person*) and *boy boy*, interpreted as expressing affection/intimacy, (ii) verbs, such as *wait wait wait* and *talk talk talk*, which are suggested to show continuity, as well as (iii) adjectives, such as *hot hot*.

The YCSEP data includes a large number of reduplicated lexical items: 20,866 instances, or 2,490 per million words. Table 4 shows the most common reduplications (i.e. 2 or more instances of the same word in sequence) for verbs, adjectives, nouns, adverbs, and discourse particles, determined on the basis of part-of-speech tagging.

Table 4: Most frequent reduplicated items in the corpus by word class and frequency

Verbs		**Modal verbs**		**Adjectives**		**Nouns**		**Adverbs**		**Discourse particles**	
wait	40	*can*	27	*correct*	132	*boom*	16	*so*	987	*yeah*	3066
go	23	*will*	11	*cool*	101	*something*	13	*very*	463	*like*	671
come	12	*cannot*	9	*many*	31	*thanks*	10	*just*	138	*no*	531
die	10	*might*	8	*interesting*	24	*nothing*	9	*not*	130	*okay*	290
talk	10	*would*	6	*big*	21	*people*	9	*really*	128	*yes*	126

In terms of grammatical category, the largest number of potential reduplications are discourse particles and adverbs such as *yeah*, *like*, *so*, *no*, or *very*. Many of these are likely to represent backchanneling or self-repair, although intensification is conceivable for some types. Among verbal, noun, and adjectival types that are repeated in utterance turns, several may represent genuine reduplications that serve to convey SgE-specific semantic and pragmatic values, including *wait wait*, *come come*, or *something something*. Some reduplications in SgE may also be more likely to occur utterance-initially, for example modal counter-assertions such as *can can* or *cannot cannot*. Reduplication may in some cases be part of formulaic language. For instance, out of 14 instances of *small small*, 8 occurred in the collocation *small small things* (and 2 more *small small little things*), as shown in Examples 9 and 10 below.

(9) [on the question 'Will moving out improve my relationship with my parents?']

I think especially like during COVID-19 now right that when I have to work from home right I have to see them every day then during meetings right they'll knock the door as if they cannot see that I'm in a meeting or I cannot hear that I'm in a meeting then they'll knock the door eh you lunch how ah what do you want to eat ah like all these **small small things** like it just builds up you know like (The Daily Ketchup Podcast, 28 March 2022)

(10) [on long-distance relationships]

yeah I got a tattoo of what he drew the first time we met and I realized I appreciate a lot of the **small small small little things** that I maybe would have looked past last time (The Hop Pod, 27 Sept. 2023)

This exploratory analysis of reduplication in YCSEP revealed evidence of its use in SgE. Although we have not examined the search hits in their original contexts in the audio data or the underlying YouTube videos in this proof-of-concept analysis, the structure of the corpus permits such an investigation, and manually checking search hits may be a useful way of distinguishing genuine reduplications from disfluencies. A more comprehensive study could not only consider a greater number of lemmata, but also analyze variables that may influence the likelihood of reduplication, such as word length, meaning, word class, or utterance position, as well as account for the frequency of reduplication in terms of likelihood or the percentage of reduplicated instances of each lemma. Multivariate approaches such as mixed effects logistic regression may be particularly useful in untangling the competing influence of various intra- and extralinguistic variables (see e.g. Fuchs 2017). Finally, a contrastive analysis of reduplication may also be instructive as to the extent to which reduplication is a feature that is more prevalent in SgE than in other varieties of English.

5.2 Phonetic features

With the availability of the audio data, YCSEP also lends itself to acoustic analyses. Such instrumental phonetic investigations could provide evidence for a comprehensive inquiry into the phonology of SgE. Relevant processes involving vowels include the monophthongization of the /eɪ/ and /əʊ/ vowels (Deterding 2000), the degree of vowel reduction (Gek and Deterding 2005) and possible vowel mergers (Deterding 2003) as well as a general, empirically grounded account of the vowel inventory. Consonants could also be of interest, among others rhotics (Kwek and Low 2021), dental fricatives (Moorthy and Deterding 2000) and the degree of aspiration of voiceless plosives. Morphophonological analyses at the cross-roads of phonology and syntax (e.g. Gut 2009) also benefit from the rich data available in a large speech corpus. Finally, suprasegmental investigations on intonation, stress (Low and Grabe 1999; Chong and German 2023) and speech rhythm (Low et al. 2000; Fuchs 2023) would round out the picture. All these variables could also be compared to other varieties of English, e.g. British English due to its role as historical input variety, and geographically close varieties such as Malaysian English, in order to trace similarities and differences (Tan and Low 2010).

While previous research has provided evidence from acoustic investigations into various aspects of the phonology of SgE, the data in YCSEP is much more comprehensive. With this wealth of data, and the help of forced phonemic alignment (MacKenzie and Turton 2020), the effects of coarticulation, lemma and variation between speakers can be studied.

5.3 Social variation in SgE

The new corpus also enables investigations of social variation in SgE pertaining to categories such as age, gender and ethnic group. While information on these variables is not part of the current version of the corpus, it can be determined and added, with reasonable effort and accuracy, from the audio and video data as well as information available online, where the participants are public figures or have an online presence. Information on the speakers' social characteristics could be crucial for many research questions. Age and gender are relevant to investigations into ongoing language change (e.g. Parviainen and Fuchs 2019), and ethnicity is particularly relevant to the multi-ethnic makeup of Singaporean society, not only due to the potential persistence of substrate language influences in SgE, but also in terms of how language identity can contribute to the maintenance of social boundaries between ethnic groups. Ethnicity is traditionally indicative in Singapore of a speaker's first language, but in complex ways, and increasingly less so as English becomes more widespread as a home language (see Sec. 1). Moreover, information on ethnicity, alongside age and gender, can be valuable in determining whether certain linguistic features are shared across different subgroups of the population or are more prominent within specific groups. For instance, in studies of modals in SgE or of discourse particles mentioned in Section 3.1, corpus analyses that lack this metadata limit researchers' ability to determine whether a particular feature is predominantly produced by one community – for example Chinese Singaporeans – or if it is also common among the Indian or Malay community. Evidence of the same feature being used by groups without shared substrate languages would challenge the hypothesis of cross-linguistic influence from substrate languages (see Teo 2019; Leimgruber et al. 2020; Morin and Basile 2022; Basile 2024) or could provide evidence of more recent panethnic convergence (Kalaivanan et al. 2021). Manual or semi-automatic annotation of demographic metadata will thus make it possible to gather evidence for the investigation of these kinds of research questions (cf. Botha 2018).

6. Caveats, summary and outlook

6.1 Limitations and caveats

While YCSEP should prove to be a valuable resource for the investigation of a range of research questions, several limitations and caveats should be noted. First, while large language models such as Whisper have resulted in remarkable increases in ASR quality, they are not infallible, and transcripts may contain errors. The provenance of Whisper's 680k hours of training data has not been disclosed in detail – the model almost certainly includes Singaporean English speech from resources such as Common Voice (Ardila et al. 2020), but may not include the kind of informal conversational speech that comprises much of the YCSEP content, and which includes frequent interruptions, overlaps, and self-repairs, in addition to out-of-vocabulary items that may not be in the model's training data. Whisper's base models (including large-v3 and the Faster-Whisper/WhisperX pipeline) have been described as reflecting a design philosophy that is focused on showing "clarity of intent" rather than the exact transcription of verbatim speech, including false starts, stuttering, and discourse particles (Lea et al. 2023). To

that extent, the model may also remove some discourse particles of SgE that could be of interest to analysts.

The issue of the applicability and suitability of LLMs for corpus-based research in dialectology and sociolinguistics is only beginning to be addressed (Grieve et al. 2025), but appropriate domain adaptation via fine-tuning with carefully curated data promises to be a viable approach. Recent efforts by a team associated with Singapore's Institute for Infocomm Research have resulted in the creation of language models for Singapore English, including fine-tuned ASR models for Singapore English speech (Wang et al. 2025).[11] A future step for YCSEP will be to fine-tune an ASR model on conversational SgE data from the NSC, which will likely improve the quality of the transcripts.

Similarly, YCSEP's automatically created turn boundaries are not always accurate, particularly for conversational speech that features multiple speakers and short speech turns: a speech segment with several speakers may be transcribed as a single turn by a single speaker, the automatically induced speaker identity labels (i.e. SPEAKER_00, etc.) may not be accurate, or some speech may not be transcribed. As is the case for ASR, diarization algorithms can be fine-tuned to increase accuracy, and this may be an additional future prospect for this resource. In general, longer passages by single speakers are accurately diarized, but analysts interested in interactional phenomena in speech are advised to check the segments via YouTube links.

As noted above, unlike some manually created resources, YCSEP does not include information such as speaker age, sex/gender, residence, residence history, educational attainment, occupation, or other personal characteristics. Some of these categories can be annotated with automatic, semi-automatic, or manual methods, but the resource is not suitable "out-of-the box" for the investigation of some research questions formulated along traditional sociolinguistic lines. Nevertheless, we expect YCSEP, as a corpus of situated language use, to prove useful for the investigation of a wide range of questions pertaining to SgE usage.

6.2 Summary and outlook

This article has described the creation of a new corpus of transcripts and audio, the YouTube Corpus of Singapore English Podcasts. The corpus content was compiled via automated methods from some of Singapore's most popular podcasts, utilizing a pipeline incorporating yt-dlp and WhisperX. YCSEP promises to advance the empirical study of Singapore English by providing a large-scale, naturalistic spoken corpus that captures a broad range of contemporary SgE linguistic features. In particular, YCSEP allows researchers to investigate phonetic, grammatical, and discourse-level phenomena that have previously been challenging to analyze due to the scarcity of spontaneous spoken data in most existing corpora of the variety. Preliminary analyses confirm that YCSEP not only confirms well-documented features of SgE, such as the use of characteristic discourse particles and substrate-influenced morphosyntactic structures, but also reveals emerging linguistic patterns that may be indicative of ongoing language change.

A key future development for YCSEP will be to improve transcription accuracy through domain-adapted ASR models. While the WhisperX-based pipeline approach used for the initial

[11] https://huggingface.co/MERaLiON

corpus creation has provides generally high-quality transcriptions, model fine-tuning on conversational Singapore English speech data could yield even better results; here, custom fine-tuned models such as those created by Infocomm Research's MERaLiON team or others could be employed. Similarly, improving speaker diarization accuracy may enhance the corpus' usability for interactional sociolinguistic research.

Beyond linguistic analysis, YCSEP offers opportunities for interdisciplinary research, particularly in fields such as multimodal discourse analysis. The availability of structured, timestamped transcripts linked to publicly accessible podcast videos enables multimodal studies of spoken discourse, including gesture, prosody, and interactional dynamics. By addressing the methodological and empirical challenges in corpus-based research on spoken Singapore English, YCSEP provides a foundation for future studies on language variation, change, and contact in postcolonial English varieties.

Funding information
This project has received support from the European Union – NextGenerationEU instrument and funding from the Research Council of Finland under grant number 358727.

Author addresses
Steven Coats, English, Faculty of Humanities, University of Oulu, 90014 University of Oulu, Finland, steven.coats@oulu.fi (address for correspondence)

Carmelo Alessandro Basile, Sorbonne Nouvelle University, Département du Monde Anglophone 8 Av. de Saint-Mandé 75012 Paris, France, alessandro.basile@sorbonne-nouvelle.fr

Cameron Morin, Université Paris Cité, 8 Place Paul Ricoeur 75013 Paris, France, cameron.morin@u-paris.fr

Robert Fuchs, Institut für Anglistik, Amerikanistik und Keltologie, Universität Bonn, Rabinstraße 8, 53111 Bonn. rfuchs@uni-bonn.de

7. References

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. "Common Voice: A Massively-Multilingual Speech Corpus". *arXiv*:1912.06670 [cs.CL]. https://doi.org/10.48550/arXiv.1912.06670

Bao, Zhiming. 2010a. "A Usage-based Approach to Substratum Transfer: The Case of Four Unproductive Features in Singapore English". *Language* 86: 792–820.

Bao, Zhiming. 2010b. "*Must* in Singapore English". *Lingua* 120: 1727–1737. https://doi.org/10.1016/j.lingua.2010.01.001

Bao, Zhiming. 2015. *The Making of Vernacular Singapore English: System, Transfer and Filter*. Cambridge: Cambridge University Press.

Bain, Max, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. “WhisperX: Time-Accurate Speech Transcription of Long-Form Audio”. In *Proceedings of Interspeech 2023*, 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

Basile, Carmelo Alessandro. 2023. “Necessity Modal Development in Singapore English: An Investigation of Substratist and Contact-Grammaticalisation Approaches”. *English World-Wide* 44: 276–302. https://doi.org/10.1075/eww.22019.bas

Basile, Carmelo Alessandro. 2024. *Modality in Contact: Necessity and Obligation in New Englishes*. Berlin: Mouton de Gruyter.

Boo, Ashley, Junwen Lee, and Ying-Ying Tan. 2023. “Particle Stacking in Singlish – New Data from the National Speech Corpus”. *Lingua* 287. https://doi.org/10.1016/j.lingua.2023.103513

Botha, Werner. 2018. “A Social Network Approach to Particles in Singapore English”. *World Englishes* 37: 261–281. https://doi.org/10.1111/weng.12250

Botha, Werner, and Tobias Bernaisch. 2025. “Social Network Effects on Particle Variation among Singapore Students”. *World Englishes* 44: 144–165. https://doi.org/10.1111/weng.12688

Bredin, Hervé. 2023. “Pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark and Recipe”. In *Proceedings of Interspeech 2023*, 1983–1987. https://doi.org/10.21437/Interspeech.2023-105

Chen, Wenda, Ying-Ying Tan, Eng Siong Chng, and Haizhou Li. 2010. “The Development of a Singapore English Call Resource”. In *Proceedings of Oriental COCOSDA 2010*.

Chong, Adam J., and James S. German. 2023. “Prominence and Intonation in Singapore English”. *Journal of Phonetics* 98: 101240. https://doi.org/10.1016/j.wocn.2023.101240

Coats, Steven. 2023a. “Double Modals in Contemporary British and Irish Speech”. *English Language and Linguistics* 27: 693–718.

Coats, Steven. 2023b**.** “Dialect Corpora from YouTube”. In Beatrix Busse, Nina Dumrukcic, and Ingo Kleiber (eds.), *Language and Linguistics in a Complex World*. Berlin: de Gruyter, 79–102.

Coats, Steven. 2024. “Naturalistic Double Modals in North America”. *American Speech* 99, 47–77.

Coats, Steven, and Cameron Morin. 2024. “Double Modals beyond the Atlantic: New Evidence from Computational Sociolinguistics”. *English Today*: 1–6. https://doi.org/10.1017/S0266078424000191

Collins, Peter, and Adam Smith. 2025. “The Double Modal Construction in English World Wide.” *World Englishes* 00: 1–19. https://doi.org/10.1111/weng.12735

Davies, Mark, and Robert Fuchs. 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: 1–28. https://doi.org/10.1075/eww.36.1.01dav

Davies, Mark. 2016–. *Corpus of News on the Web (NOW)*. Available online at https://www.english-corpora.org/now/

Deterding, David. 2000. “Measurements of the /eɪ/ and /əʊ/ Vowels of Young English Speakers in Singapore”. In David Deterding, Ee Ling Low, and Adam Brown, eds., *The English Language in Singapore: Research on Pronunciation*. Singapore: Singapore Association for Applied Linguistics, 93–99.

Deterding, David. 2003. “An Instrumental Study of the Monophthong Vowels of Singapore English”. *English World-Wide* 24: 1–16.

Deterding, David. 2007. *Singapore English*. Edinburgh: Edinburgh University Press.

Deterding, David, and Ee Ling Low. 2001. “The NIE corpus of spoken Singapore English (NIECSSE)”. *SAAL Quarterly* 56: 2–5.

Dunn, Jonathan. 2024. *Computational Construction Grammar: A Usage-Based Approach*. Cambridge: CUP.

Fuchs, Robert. 2017. “Do Women Use More Intensifiers than Men? Recent Change in the Sociolinguistics of Intensifiers in British English”. *International Journal of Corpus Linguistics* 22: 345–374.

Fuchs, Robert. 2023. “A Synthesis of Research on Speech Rhythm in Native, Learner and Second Language Varieties of English – Introduction to the Volume”. In Robert Fuchs, ed., *Speech Rhythm in Learner and Second Language Varieties of English*. Singapore: Springer, 1–14.

Gek, Heng Mui, and David Deterding. 2005. “Reduced Vowels in Conversational Singapore English”. In David Deterding, Ee Ling Low, and Adam Brown, eds., *English in Singapore: Phonetic Research on a Corpus*. Singapore: McGraw-Hill Education, 54–63.

Gonzales, Wilkinson D. W., Jakob Leimgruber, Mie Hiramoto, and Junjie Lim. 2023. “The Corpus of Singapore English Messages (CoSEM)”. *World Englishes* 42: 371–388. https://doi.org/10.1111/weng.12534

Grieve, Jack, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. “The Sociolinguistic Foundations of Language Modeling”. *Frontiers in Artificial Intelligence* 7. https://doi.org/10.3389/frai.2024.1472411

Grieve, Jack, Dirk Hovy, David Jurgens, Tyler Kendall, Dong Nguyen, James Stanford, and Meghan Sumner. 2023. *Computational Sociolinguistics.* Lausanne: Frontiers Media SA.

Gut, Ulrike. 2009. “Past Tense Marking in Singapore English Verbs”. *English World-Wide* 30: 262–277.

Hansen, Beke. 2018. *Corpus Linguistics and Sociolinguistics: A Study of Variation and Change in the Modal System of World Englishes*. Leiden: Brill.

Hoffmann, Thomas. 2021. *The Cognitive Foundation of Post-colonial Englishes.* Cambridge: Cambridge University Press.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

Huang, Nick, Li Lin, Kunmei Han, Jia Wen Hing, Luwen Cao, Vincent Ooi, and Zhiming Bao. 2025. “Treebanks and World Englishes: a Singapore English Perspective”. *English World-Wide* 46: 93–121. https://doi.org/10.1075/eww.23069.hua

Kalaivanan, Kastoori, Firqin Sumartono, and Ying-Ying Tan. 2021. “The Homogenization of Ethnic Differences in Singapore English? A Consonantal Production Study”. *Language and Speech* 64: 123–140.

Kirkpatrick, Andy. 2007. *World Englishes: Implications for International Communication and English Language Teaching*. Cambridge: Cambridge University Press.

Koh, Jia Xin, Aqilah Mislan, Kevin Khoo, Brian Ang, Wilson Ang, Charmaine Ng, and Ying-Ying Tan. 2019. “Building the Singapore English National Speech Corpus”. In *Proceedings of Interspeech 2019*, 321–325. http://doi.org/10.21437/Interspeech.2019-1525

Kortmann, Bernd. 2019. “Global Variation in the Anglophone World”, in Bas Aarts, Jill Bowie and Gergana Popova, eds., *Oxford Handbook of English Grammar.* Oxford: Oxford University Press, 630–653.

Kwan-Terry, A. 1978. “The Meaning and the Source of the ‘la’ and the ‘what’ Particles in Singapore English”. *RELC Journal* 9: 22–36. https://doi.org/10.1177/003368827800900202

Kwek, Geraldine, and Ee-Ling Low. 2021. “Emergent Features of Young Singaporean Speech: An Investigatory Study of the Labiodental /r/ in Singapore English”. Asian Englishes 23: 116–136.

Leech, Geoffrey. 2003. “Modality on the Move: The English Modal Auxiliaries 1961–1992”. In Roberta Facchinetti, Manfred Krug, and Frank R. Palmer, eds., *Modality in Contemporary English*. Berlin: Mouton de Gruyter, 223–240.

Lea, Colin, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P. Bigham, and Leah Findlater. 2023. “From User Perceptions to Technical Improvement: Enabling People who Stutter to Better Use Speech Recognition”. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16.

Leimgruber, Jakob R. 2013. *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press.

Leimgruber, Jakob R., Jun Lie Lim, Wilkinson Gonzales, and Mie Hiramoto. 2020. “Ethnic and Gender Variation in the Use of Colloquial Singapore English Discourse Particles”. *English Language and Linguistics* 25: 601–620. https://doi.org/10.1017/S1360674320000453

Li, Lijun, Eliane Lorenz, and Peter Siemund. 2022. “The Ages of Pragmatic Particles in Colloquial Singapore English: A Corpus Study Based on Oral History Interviews”. *English World-Wide* 44: 91–117. https://doi.org/10.1075/eww.21016.li

Lim, Lisa. 2001. *Towards a Reference Grammar of Singapore English*. Final Research Report. Singapore: National University of Singapore.

Lim, Lisa. 2007. “Mergers and Acquisitions: On the Ages and Origins of Singapore English Particles”. *World Englishes* 26: 446–473. https://doi.org/10.1111/j.1467-971X.2007.00522.x

Lim, Lisa, and Joseph Foley. 2004. “English in Singapore and Singapore English: Background and Methodology”. In Lisa Lim, ed., *Singapore English: A Grammatical Description*. Amsterdam: John Benjamins, 1–18.

Lin, Li, Kunmei Han, Jia Wen Hing, Luwen Cao, Vincent Ooi, Nick Huang, and Zhiming Bao. 2023. “Tagging Singapore English”. *World Englishes* 42: 624–641. https://doi.org/10.1111/weng.12597

Low, Ee Ling, and Esther Grabe. 1999. “A Contrastive Study of Prosody and Lexical Stress Placement in Singapore English and British English”. *Language and Speech* 42: 39–56.

Low, Ee Ling, Esther Grabe, and Francis Nolan. 2000. “Quantitative Characterizations of Speech Rhythm: Syllable-timing in Singapore English”. *Language and Speech* 43: 377–401.

MacKenzie, Laurel, and Danielle Turton. 2020. “Assessing the Accuracy of Existing Forced Alignment Software on Varieties of British English”. *Linguistics Vanguard* 6(s1): 20180061.

Mair, Christian, and Geoffrey Leech. 2020. “Current Changes in English Syntax”. In Bas Aarts, April McMahon, and Lars Hinrichs, eds., *The Handbook of English Linguistics* (2nd ed.). Malden: Wiley-Blackwell, 249–276.

Moorthy, Shanti Marion, and David Deterding. 2000. “Three or Tree? Dental Fricatives in the Speech of Educated Singaporeans”. In Adam Brown, David Deterding, and Ee Ling Low, eds., The English Language in Singapore: Research on Pronunciation. Singapore: Singapore Association for Applied Linguistics, 76–83

Morin, Cameron. 2023. “Social meaning in Construction Grammar: Double Modals in Dialects of English”. PhD dissertation, Université Paris-Cité.

Morin, Cameron, and Carmelo Alessandro Basile. 2022. “Elicitation and Experimentation: Implications for English Sociolinguistics”. *Anglophonia* 34: 1–25. https://doi.org/10.4000/anglophonia.5184

Morin, Cameron, and Steven Coats. 2023. “Double Modals in Australian and New Zealand English”. *World Englishes* 00: 1–24. https://doi.org/10.1111/weng.12639

Morin, Cameron, and Jack Grieve. 2024. “The Semantics, Sociolinguistics, and Origins of Double Modals in American English: New Insights from Social Media”. *PLOS One* 19: E0295799.

Morin, Cameron, Guillaume Desagulier, and Jack Grieve. 2020. “Dialect syntax in Construction Grammar: Theoretical Benefits of a Constructionist Approach to Double Modals in English”. *Belgian Journal of Linguistics* 34: 252–262.

Morin, Cameron, Guillaume Desagulier, and Jack Grieve. 2024. “A Social turn for Construction Grammar: Double Modals on British Twitter”. *English Language and Linguistics* 28: 275–303.

Nelson, Gerald. 2002. *International Corpus of English: The Singapore Corpus*. User Manual.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. “Computational Sociolinguistics: A Survey”. *Computational Linguistics* 42: 537–593.

Pantos, Roger, and William May (2017). “HTTP Live Streaming”. *RFC* 8216. https://www.rfc-editor.org/rfc/rfc8216

Parviainen, Hanna, and Robert Fuchs. 2019. “‘I Don’t Get Time Only’: An Apparent-time IInvestigation of Clause-final Focus Particles in Asian Englishes”. *Asian Englishes* 21: 285–304.

Plaquet, Alexis, and Hervé Bredin. 2023. “Powerset Multi-class Cross Entropy Loss for Neural Speaker Diarization”. In *Proceedings of Interspeech 2023*, 3222–3226. https://doi.org/10.21437/Interspeech.2023-205

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. “Robust Speech Recognition via Large-Scale Weak Supervision”. In *Proceedings of the 40th International Conference on Machine Learning* (= *Proceedings of Machine Learning Research* 202: 28492–28518). https://proceedings.mlr.press/v202/radford23a.html

Richards, Jack C., and Mary W. J. Tay. 1977. “The La particle in Singapore English”. In William Crewe, ed., *The English Language in Singapore*. Singapore: Eastern Universities Press, 145–156.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.

Smakman, Dick, and Stephanie Wagenaar. 2013. “Discourse Particles in Colloquial Singapore English”. *World Englishes* 32: 308–324. https://doi.org/10.1111/weng.12033

Sodagar, Iraj. 2011. “The MPEG-DASH Standard for Multimedia Streaming over the Internet”. *IEEE Multimedia* 18: 62–67. https://doi.org/10.1109/MMUL.2011.71

Tan, Ying-Ying. 2014. “English as a ‘Mother Tongue’ in Singapore”. *World Englishes* 33: 319–339.

Tan, Rachel Siew Kuang, and Ee-Ling Low. 2010. “How Different are the Monophthongs of Malay Speakers of Malaysian and Singapore English?”. *English World-Wide* 31: 162–189.

Teo, Ming Chew. 2019. “The Role of Parallel Constructions in Imposition: A Synchronic Study of *already* in Colloquial Singapore English”. *Journal of Pidgin and Creole Languages* 34: 347–377. https://doi.org/10.1075/jpcl.00042.teo

Wang, Bin, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F. Chen, and AiTi Aw. 2025. “Advancing Singlish Understanding: Bridging the Gap with Datasets and Multimodal Models”. *arXiv*:2501.01034 [cs.CL]. https://doi.org/10.48550/arXiv.2501.01034

Wee, Lionel. 2004. “Singapore English: Morphology and Syntax”. In Bernd Kortmann, Kate Burridge, Rajend Mesthrie, Edgar W. Schneider, and Clive Upton eds., *A Handbook of Varieties of English: A Multimedia Reference Tool,* vol. 1. Berlin: Mouton de Gruyter, 1058–1073.

Wong, John D. 2016. *Global Trade in the Nineteenth Century: The House of Houqua and the Canton System*. Cambridge: Cambridge University Press.

Ziegeler, Debra. 2014. “Replica Grammaticalisation as Recapitulation: The Other Side of Contact”. *Diachronica* 31: 106–141.

Ziegeler, Debra. 2015. *Converging Grammars: Constructions in Singapore English* (Language Contact and Bilingualism 11). Berlin: Mouton de Gruyter.

Zullo, David, Simone Pfenninger, and Daniel Schreier. “A Pan-Atlantic ‘Multiple Modal Belt’?”. *American Speech* 96: 7–44.