



**Cambridge
Elements**

World Englishes

Young Singaporean English

**Jakob R. E. Leimgruber,
Wilkinson Daniel Wong
Gonzales, Mohamed Hafiz,
Jun Jie Lim, Nick Huang
and Mie Hiramoto**

ISSN 2633-3309 (online)

Cambridge Elements

Elements in World Englishes

edited by

Edgar W. Schneider

University of Regensburg

YOUNG SINGAPOREAN ENGLISH

*Current Developments in CSE
Morphosyntax in a Digital World*

Jakob R. E. Leimgruber

University of Regensburg

Wilkinson Daniel Wong Gonzales

The Chinese University of Hong Kong

Mohamed Hafiz

National University of Singapore

Jun Jie Lim

California State University San Marcos

Nick Huang

National University of Singapore

Mie Hiramoto

National University of Singapore



CAMBRIDGE
UNIVERSITY PRESS



CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781009634700

DOI: [10.1017/9781009634717](https://doi.org/10.1017/9781009634717)

© Jakob R. E. Leimgruber, Wilkinson Daniel Wong Gonzales,
Mohamed Hafiz, Jun Jie Lim, Nick Huang and Mie Hiramoto 2026

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, with the exception of the Creative
Commons version the link for which is provided below, no reproduction of any part
may take place without the written permission of Cambridge University Press &
Assessment.

An online version of this work is published at doi.org/10.1017/9781009634717 under
a Creative Commons Open Access license CC-BY-NC 4.0 which permits re-use,
distribution and reproduction in any medium for non-commercial purposes providing
appropriate credit to the original work is given and any changes made are indicated.
To view a copy of this license visit <https://creativecommons.org/licenses/by-nc/4.0>

When citing this work, please include a reference to the DOI [10.1017/9781009634717](https://doi.org/10.1017/9781009634717)

First published 2026

A catalogue record for this publication is available from the British Library

*A Cataloging-in-Publication data record for this Element is available
from the Library of Congress*

ISBN 978-1-009-63470-0 Hardback

ISBN 978-1-009-63473-1 Paperback

ISSN 2633-3309 (online)

ISSN 2633-3295 (print)

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

For EU product safety concerns, contact us at Calle de José Abascal, 56, 1º, 28003
Madrid, Spain, or email eugpsr@cambridge.org

Young Singaporean English

Current Developments in CSE Morphosyntax in a Digital World

Elements in World Englishes

DOI: 10.1017/9781009634717

First published online: June 2026

Jakob R. E. Leimgruber
University of Regensburg

Wilkinson Daniel Wong Gonzales
The Chinese University of Hong Kong

Mohamed Hafiz
National University of Singapore

Jun Jie Lim
California State University San Marcos

Nick Huang
National University of Singapore

Mie Hiramoto
National University of Singapore

Author for correspondence: Jakob R. E. Leimgruber, jakob.leimgruber@ur.de

Abstract: The analysis of newly or recently emerged grammatical and lexical forms in Colloquial Singapore English (CSE) is the main objective of this Element. Using corpus, survey, and interview data from different age groups, we shed light on the spread of language change across generations and ethnicities. Existing descriptions of CSE as a high-contact L1 variety of English in the late stages of endonormative stabilisation do not fully capture Singapore's continued multilingual ecology: source languages remain in active use alongside CSE, enabling ongoing cross-linguistic influence. Innovative uses resulting from contact can be observed in apparent and real time. In this volume, we use a range of sources to look at recent changes in the lexicon and grammar of CSE, pointing to a dynamic variety that is difficult to fully capture with existing models of variation. This title is also available as Open Access on Cambridge Core.

Keywords: Singapore, Colloquial Singapore English, morphosyntax, corpus linguistics, computer-mediated communication

© Jakob R. E. Leimgruber, Wilkinson Daniel Wong Gonzales, Mohamed Hafiz, Jun Jie Lim, Nick Huang and Mie Hiramoto 2026

ISBNs: 9781009634700 (HB), 9781009634731 (PB), 9781009634717 (OC)

ISSNs: 2633-3309 (online), 2633-3295 (print)

Contents

1	Introduction	1
2	Corpus Approaches to CSE Change	6
3	Integrating Corpus and Survey Methods to Examine CSE Use and Change	13
4	Structural and Stylistic Variation in the Use of <i>(Bo)Jio</i>	31
5	Variation and Change in CSE Passive Marking	48
6	Conclusion	65
	Glossing Conventions	73
	References	74

1 Introduction

The aim of this Element is the analysis of newly or recently emerged grammatical and lexical forms in Colloquial Singapore English (CSE). The variety in question, occasionally subsumed (locally and globally) under the moniker ‘Singlish’, has been described as a high-contact L1 variety of English (Kortmann, Lunkenheimer, & Ehret 2020) and as being in the late stages of endonormative stabilisation (Schneider 2007). In Singapore, the source contact languages continue to be used alongside CSE, with widespread bi- and multilingualism, which makes for a dynamic setting with continuous cross-influence between the languages involved. Innovative uses resulting from contact can be observed in apparent and real time. In this volume, we use a range of sources to look at recent changes in the lexicon and grammar of CSE, pointing to a complex dynamic variety that is difficult to fully capture with existing models of variation based on founder effects (Mufwene 1996), substrate transfer (Mufwene 1990; Siegel 1999), and even more diachronic approaches (Schneider 2007). In this first section, we begin with a short introduction to Singapore and its linguistic ecology. This is followed in Section 2 by a presentation of CoSEM, the *Corpus of Singapore English Messages* (<https://github.com/wdwgonzales/CoSEM>), which is the dataset on which our analyses are primarily based. Section 3 reports the results of a sociolinguistic survey examining, among other things, the usage of three innovative features in CSE – namely, clause-initial *is it*, adverbial *confirm*, and the sentence-final particle (SFP) *sia* – across age, gender, and ethnicity. The findings complement previous corpus-based findings and highlight age as the most reliable predictor of self-reported usage, while gender and ethnicity show limited effects. Section 4 presents a close-up study of the lexical item *(bo)jio*, a relatively recent innovation in CSE of Hokkien origin and with the surface translation ‘NEG invite’. We show how the alternation between *(bo)jio* and its English counterpart is conditioned, as well as the extent to which it has been subjected to morphological integration into CSE. Section 5 considers grammatical change, with a case study of passive constructions (both CSE and standardised English counterparts), following up on Bao and Wee’s (1999) discussion of CSE *kena* and *give* passives. Unlike the Malay-influenced *kena* passive, the data suggest that the Chinese-influenced *give* passive has now fallen out of popular use, at least among younger CSE speakers. At the same time, we document the emergence of a novel *tio* passive, also derived from Chinese (more specifically, from Southern Min), in the same population.

1.1 Background: Singapore

The city-state of Singapore is characterised by a high degree of diversity, ethnically, culturally, and linguistically. This is in no small part due to its advantageous geographical location at the southern tip of the Malay Peninsula, just at the south-eastern entrance to the Straits of Malacca. Already in pre-colonial times, the island on which the current city stands enjoyed its strategic situation at the crossroads of several local, regional, and global maritime trade routes, attracting the attention of several of the empires of the region (Kwa et al. 2019; Baker 2020). The settlement at the mouth of the Singapore River served as a trading and political centre of varying size and importance in the centuries leading up to the British landing of 1819. It is under the rule of the East India Company and, later, of the Empire, that modern Singapore saw the beginning of an economic and social development that shaped its affluent present (see e.g., Wong 1978). The natural deep-sea harbour was expanded by the British into a major stop on the trade route from China, via India, onto Europe (eventually via the Suez Canal), and the tariff-free port city quickly developed into a prized destination for merchants and workers in search of a better life (Baker 2020: chap. 5).

The history of migration into Singapore has been described widely (see inter alia Turnbull 1989; Gupta 1994; Bao 2021; Leimgruber 2026). Apart from the indigenous Malay population that lived on the island and in the surrounding archipelago, Chinese traders settled in what is now Singapore well before the arrival of the British. The latter brought with them, early on, colonial subjects from India for various administrative and military duties. Famines and instability in southern China in the nineteenth and early twentieth century led to a large number of migrants primarily from Fujian, Guangdong, and Hainan leaving for Singapore; by 1836 the Chinese were the majority ethnic group (Starr & Hiramoto 2019: 5; Hafiz 2025: 3). These Chinese brought with them their languages, mostly Southern Min (Hokkien, Teochew, and Hainanese) and Cantonese. When the territory became a Crown Colony in 1867, the local education system in English and other languages was further consolidated. By the end of the nineteenth century, the population's ethnic distribution of ca. 75 per cent Chinese, 15 per cent Malays, 8 per cent Indians, and 2 per cent 'Others' (1911 census, see Leimgruber 2026: 234) had been achieved, which would not change until the present day by more than a few percentage points (Pak & Hiramoto 2023).

The British military defeat against the Japanese in 1942 ushered in a brutal three years of occupation, which had few linguistic consequences but did its part to stoke anti-colonial sentiment in the years following the war (Turnbull 1989: 209;

Matsuoka 2018; Huff & Huff 2020). The first Legislative Assembly was elected in 1955, and self-government was granted four years later. In 1963, Singapore became independent from Britain by merging with Peninsular Malaya, Sabah, and Sarawak into the Federation of Malaysia and becoming one of its fourteen states. This short-lived solution came to an end in 1965, when Singapore became independent from the Federation and became the independent nation-state it is today (Hafiz 2025: 3). Economic policies of global market integration and a focus on education, high technology, and high added-value services were combined with nation-building efforts of a social, cultural, and linguistic nature to ensure the survival of this small city-state.

1.2 Background: Linguistic Ecology

The migrations described in Section 1.1 resulted in a highly linguistically diverse population. At least four language families have been present in significant numbers since the early days of Singapore: Austronesian, Sino-Tibetan, Dravidian, and Indo-European. Austronesian is the autochthonous language family and includes Malay and its related varieties in the region, such as Javanese, Boyanese, and several others. These predate the colonial presence, and the Malay language is, in fact, legislated as the national language of the country, in addition to its co-official status (Hafiz 2025: 5). Malay – or more precisely, the variety now known as Bazaar Malay – was once the lingua franca of Singapore given its status as a trade port part of the larger Malay Archipelago (Bao & Aye 2010). Malay was said to be used for communication between various communities in Singapore (Aye 2005), in ‘non-formal areas of discourse in particular as the language of inter-ethnic communication’ (Tham 1996: 1). Singaporean languages from the Sinitic branch of Sino-Tibetan include the Southern Min varieties Hokkien, Teochew, and Hainanese, as well as several varieties of Yue (Cantonese). Mandarin, one of the co-official languages, and nowadays the second most widely used language after English, was not brought into Singapore by migrants, but was rather imposed by language policy on the Sinophone majority as an intra-ethnic lingua franca, thereby serving a unifying function as well as helping to reduce diversity to a manageable number of officially endorsed languages (see Wong & Tan 2017; Lim, Chen & Hiramoto 2021). Statistics from census data report a dramatic shift towards Mandarin as a dominant home language among the Chinese population, with the co-official language rising from close to 0 per cent in 1957 to over a third of households in 2010. Proficiency in non-Mandarin varieties of Chinese is, nowadays, rare among the under-50s: census data indicates that they are used as ‘most or second most frequent home language’

by just 2.4 per cent of that age group (Department of Statistics Singapore 2020). The third language family, Dravidian, includes the co-official language Tamil, Malayalam, and other southern South Asian languages. Additional South Asian languages belonging to the Indo-European family are also represented in Singapore, such as Hindi and Gujarati. Last (but certainly not least), the Indo-European family includes the co-official language English, which is now (according to 2020 census data) the most widely used home language.

The English language in Singapore has benefitted from strong institutional support, also (and perhaps even more so) in post-colonial times. It is one of the four co-official languages but has the additional non-statutory status of the country's 'working language'. In short, it is the language in which institutions of the state work and function, the medium of education in all (since 1987) state schools, the official language of the conscription-based armed forces, and the dominant language of the white-collar workplace. Most tellingly, despite legislation elevating English above the other official languages, all legislative texts are in English, as it is the language of judicial (and generally governmental) records. Proficiency in the language has long been a marker of social class, adding to its attraction in a culturally competitive educational landscape. The fact that English is itself not only a school subject but also the medium of instruction of all other subjects (excluding the 'mother tongue' subject) further puts pressure on Singaporeans to add it to their repertoires. Census data since the mid twentieth century show that there has been a steady increase in the percentage of Singaporeans who report predominantly using English as their home language, reaching 48 per cent of households in 2020. The rest use Mandarin (30%), Malay (9%), other Chinese varieties (9%), Tamil (3%), and other languages (1.4%) (Department of Statistics Singapore 2020).

Given the diversity mentioned earlier, it is not surprising that multilingualism is the norm rather than the exception in Singapore. This is true both in terms of dominant home language, for which 81 per cent responded using more than one language at home, and in terms of literacy, for which 74 per cent of the literate population reported literacy in two or more languages (Department of Statistics Singapore 2020). Most of this multilingualism is actually bilingualism in English and one of the other three official languages (65%), in line with governmental language policy, and confirmed by sociolinguistic studies (Leimgruber, Siemund, & Terassa 2018). Such a linguistic ecology, in which the languages involved in the contact situation during the foundational period (cf. Schneider 2007) are in continued, sustained contact with English, and taught, learnt, and spoken to a considerable degree side by side with English, results in a situation in which constant admixture and transfer are possible in both directions. As such, we have here a dynamic contact setting in which

contact-induced changes are continuously likely, meaning that today's CSE may be quite different from the one spoken a generation ago or so. This is a question that we will actively engage with in [Section 3](#).

1.3 Colloquial Singapore English

Our recurrent and main argument in this Element is that the vernacular form of English used in Singapore (often called 'Singlish' locally; we will use herein the term CSE) is constantly exposed to influence from the languages with which it has been in contact since its inception, thus leading to continuing innovation originating in this contact. Unlike other frameworks such as the third-wave variationist framework (Johnstone 2016; Ilbury 2020) that views CSE as one of the enregistered styles of English used in Singapore (Leimgruber 2012; Khoo 2023; Lee 2023; Lu 2025), from the perspective of world Englishes scholars, CSE is a contact form of English, with influences from Sinitic languages and Malay, along with Tamil to a lesser extent. The variety has long been subject to scholarly attention; the interested reader is referred to the book-length works by Ho and Platt (1993), Gupta (1994), Lim (2004), Deterding (2007), Leimgruber (2013), and Bao (2015) for a general overview; specific features of the variety have been described in a range of scientific articles over the past three decades. Among the more stereotypical features of CSE are its discourse particles (the most salient being *lah*), which have received some attention and which we also touch on in [Section 3](#).

The dynamic nature of the continued contact between especially Sinitic varieties (but also Malay and, to a lesser extent, Tamil) and English in this context, as well as widespread bilingualism, have allowed us to argue in previous work (e.g., Lim, Hiramoto, et al. 2024; Hafiz et al. 2025; but also Leimgruber 2016) that innovations in CSE can continuously be 'transferred in' from the putative substrate languages (e.g., *is it* discussed in Hiramoto et al. 2025). This view goes against traditional ideas of such substrate influence being restricted to early contact (e.g., Siegel 1999), for instance, by the so-called founder principle (Mufwene 2001). Therefore, the variety of CSE used by younger speakers, reflected in our corpus and described in this Element, is influenced not only by global youth language phenomena and Americanisation (see e.g., Tan 2012 on rhoticity; or Starr 2019 on the BATH–TRAP split), but also by locally grown cross-fertilisation between languages that are actively used in the community. This impact goes both ways, too, perhaps best conceptualised in terms of a complex dynamic system (on which see Schneider 2025). While Singaporean instantiations of Mandarin, Malay, and so on exist, our focus here

remains on the local form of English and the ways in which the complex linguistic environment has shaped it.

2 Corpus Approaches to CSE Change

The primary source of data used in this Element is the Corpus of Singapore English Messages (CoSEM). In this section, we present the corpus and show how it differs from other existing resources of CSE. A fuller description of the corpus can be found in Gonzales et al. (2023a).

2.1 Brief Description

Compiled between 2016 and 2022, CoSEM (version 5) currently stands at 6.6 million tokens. Previous publications using the corpus were based on earlier versions, hence the difference in word counts (e.g., 3.6 million in Gonzales et al. 2023a). For version 5, the data was further cleaned for duplicates, and punctuation was excluded from the token/word count. It was anonymised using the *scrubadub* package as well as customised RegEx scripts. As with any type of automated scrubbing, there is a chance of remnant utterances that are not properly scrubbed.

The corpus was compiled over several semesters spanning 2016 to 2022 at the National University of Singapore as part of an advanced sociolinguistics class (EL3211: Language in Contact) taught by one of the authors. The idea behind the class project was to give students the opportunity to experience data collection and curation, corpus preparation, and to critically engage with linguistic data from a variety with which they would be, for the most part, familiar. Students built their own dataset, consisting of WhatsApp group chats (later extended to include Telegram chats), anonymised the data, and made sure they obtained informed consent from all chat participants in line with university research ethics guidelines. The group chats included users in the students' wider social circle, that is, family members, classmates, friends, colleagues, and so on. As the corpus was being developed, we also received institutional support from Undergraduate Research Opportunities Programme (UROP) students and research assistants at the National University of Singapore, who contributed to the ongoing process of cleaning and refining the dataset.

The Corpus of Singapore English Messages exhibits a pronounced age-based distribution, with younger individuals dominating contributions, as shown in Table 1. Speakers aged 21–29 account for 70.7 per cent of lines and 69.3 per cent of words, followed by those aged 18–20 at 23.5 per cent for both metrics. Together, these groups comprise over 94 per cent of the dataset, highlighting a strong generational skew. By contrast, contributions from those

Table 1 Distribution of lines and words by age groups in CoSEM

Age category	Lines (n)	Lines (%)	Words (n)	Words (%)
18–20	325,312	23.5	1,551,172	23.5
21–29	979,303	70.7	4,577,378	69.3
30–39	15,575	1.1	78,717	1.2
40–49	23,323	1.7	108,873	1.7
50–59	38,018	2.8	267,850	4.1
60–69	1,916	0.1	14,046	0.2
Others	1,265	0.1	5,421	0.1
Total	1,384,712	100.0	6,603,457	100.0

Table 2 Distribution of lines and words by gender of user in CoSEM

Gender	Lines (n)	Lines (%)	Words (n)	Words (%)
Female	729,473	52.7	3,551,742	53.8
Male	655,184	47.3	3,051,502	46.2
Others	55	0.0	213	0.0
Total	1,384,712	100.0	6,603,457	100.0

over 30 are minimal. The 30–39 group contributes just 1.1 per cent of lines and 1.2 per cent of words, while the 40–49 group adds 1.7 per cent for both measures. The 50–59 demographic, though slightly more represented, remains small at 2.8 per cent of lines and 4.1 per cent of words. The 60–69 group is nearly absent, with only 0.1 per cent of lines and 0.2 per cent of words. The ‘Others’ category, likely representing speakers outside standard age brackets, accounts for a mere 0.1 per cent of lines and words.

This generational imbalance reflects broader trends in digital language data, where younger users dominate. Consequently, the corpus captures linguistic forms and interactional styles aligned with younger speakers, particularly those in their twenties, which is of interest in this Element.

Table 2 indicates that CoSEM exhibits a slight imbalance, with female users contributing a majority of the data. Women account for 52.7 per cent of lines and 53.8 per cent of words, while men contribute 47.3 per cent of lines and 46.2 per cent of words. This suggests that not only do female users produce slightly more lines, but they also tend to use more words per line on average. The ‘Others’ category, representing non-binary, unspecified, or alternative gender identities, is effectively absent, contributing just 55 lines and 213 words, which is statistically negligible within the dataset.

Table 3 Distribution of lines and words by chat gender make-up groups in CoSEM

Chat gender make-up	Lines (<i>n</i>)	Lines (%)	Words (<i>n</i>)	Words (%)
Exclusively female	438,014	31.6	1,996,941	30.2
Dominantly female	166,087	12.0	844,555	12.8
Half male/half female	283,434	20.5	1,373,316	20.8
Dominantly male	335,544	24.2	1,661,556	25.2
Exclusively male	161,633	11.7	727,089	11.0
Total	1,384,712	100.0	6,603,457	100.0

In terms of chat gender composition, notable differences in participation and linguistic output across different gender groupings can be found, as summarised in Table 3. Chats composed exclusively of female participants account for the largest share of the corpus, contributing 31.6 per cent of all lines and 30.2 per cent of all words. This suggests that all-female spaces are among the most linguistically active in the dataset. Mixed-gender settings also make up a significant portion of the corpus. Chats with an equal split between male and female participants contribute 20.5 per cent of lines and 20.8 per cent of words, while ‘Dominantly female’ chats, where women form a majority but not the totality, account for 12.0 per cent of lines and 12.8 per cent of words. On the male-dominated side, ‘Dominantly male’ chats represent 24.2 per cent of lines and 25.2 per cent of words, and ‘Exclusively male’ chats contribute the smallest share, with 11.7 per cent of lines and 11.0 per cent of words. These figures suggest that gender composition influences participation levels and word output, with female-majority spaces contributing more lines overall. However, the relatively balanced distribution between male- and female-dominated chats suggests that neither gender group is vastly underrepresented.

Table 4 shows the distribution of lines and words by nationality and racial groups in the corpus, which reveals a striking concentration of contributions from a single dominant group, with notable disparities among others. Singaporean Chinese users overwhelmingly account for the largest share of the corpus, contributing 77.4 per cent of all lines and 76.9 per cent of all words. This suggests that the linguistic patterns captured in the dataset will be heavily shaped by this demographic. The next largest contributors are Singaporean Indian users, who make up 10.0 per cent of lines and 10.0 per cent of words, followed by Singaporean Malays, who contribute 6.2 per cent of lines and 6.6 per cent of words. Singaporean Eurasians are a smaller but still notable presence, accounting for 0.9 per cent of lines and 0.9 per cent of words. Other ethnic groups, including Indonesian (0.7%), Indian users from India (0.6%), and

Table 4 Distribution of lines and words by racial/ethnic groups in CoSEM

Race/ethnicity	Lines (n)	Lines (%)	Words (n)	Words (%)
Australian	2,015	0.2	9,013	0.1
China Chinese	4,951	0.4	22,727	0.3
Spanish	16	0.0	102	0.0
French	20	0.0	114	0.0
British Pakistani	51	0.0	121	0.0
Hongkong Chinese	345	0.0	1,994	0.0
Indonesian Chinese	4,497	0.3	21,411	0.3
Indonesian	9,809	0.7	47,068	0.7
India Indian	7,997	0.6	31,205	0.5
Japanese	46	0.0	265	0.0
Korean	1,230	0.1	4,238	0.1
Myanmar Chinese	791	0.1	4,249	0.1
Malaysian Chinese	291	0.0	1,032	0.0
Malaysian Malay	519	0.0	4,452	0.1
Malaysian	2,119	0.2	10,706	0.2
Filipino	46	0.0	245	0.0
Saudi Arab	883	0.1	2,936	0.0
Singaporean Armenian	86	0.0	450	0.0
Singaporean Arab	570	0.0	2,780	0.0
Singaporean Bengali	90	0.0	687	0.0
Singaporean Chinese	1,071,534	77.4	5,075,449	76.9
Singaporean Chinese–Indian	2,538	0.2	10,881	0.2
Singaporean Chinese–Malay	294	0.0	2,001	0.0
Singaporean Eurasian	12,046	0.9	57,783	0.9
Singaporean Indian	138,724	10.0	660,927	10.0
Singaporean Malay	85,568	6.2	432,584	6.6
Singaporean (Mixed)	780	0.1	4,888	0.1
Singaporean Filipino	26	0.0	138	0.0
Singaporean Sinhalese	29,044	2.1	160,820	2.4
Singaporean	5,176	0.4	17,849	0.3
Thai	716	0.1	4,233	0.1
Taiwanese Chinese	815	0.1	3,634	0.1
USA/American	118	0.0	931	0.0
Vietnamese	890	0.1	5,212	0.1
Others	71	0.0	332	0.0
Total	1,384,712	100.0	6,603,457	100.0

Chinese users from China (0.4%) and Indonesia (0.3%), contribute modestly to the corpus. The remaining racial and ethnic categories, including Japanese, Filipino, Vietnamese, and various other Chinese subgroups, each account for less than 0.5 per cent of the total dataset. Many groups, such as those from the USA (0.01%) and Australia (0.2%), contribute minimally, suggesting that their linguistic representation in the corpus is negligible. These figures highlight the strongly localised nature of the dataset, with Singaporean users, especially Singaporean Chinese, being the dominant contributors.

With respect to the year, the corpus reveals clear temporal patterns, with a peak in contributions between 2015 and 2017, followed by a steady decline in more recent years. This is depicted in Table 5. The corpus sees a gradual increase in contributions from 2012 to 2014, with the number of lines rising from 13,877 (1.0%) in 2012 to 77,922 (5.6%) in 2014. A significant jump occurs in 2015 (15.2%) and 2016 (25.0%), which together account for over 40 per cent of the total corpus. The peak contribution year is 2016, with 346,545 lines and 24.0 per cent of the total words. Following this peak, the number of lines decreases, though 2017 (19.2%) and 2018 (12.9%) remain substantial contributors. A sharper decline follows in 2019 (6.5%) and 2020 (10.0%), suggesting reduced engagement or shifts in data collection. From 2021 onward, contributions drop dramatically, with only 1.51 per cent of lines in 2021 and virtually no data in 2022 (0.01%). This suggests that the corpus is primarily reflective of conversations occurring between 2015 and 2018, with more recent linguistic trends being underrepresented.

Table 5 Distribution of lines and words by year of utterance in CoSEM

Year	Lines (<i>n</i>)	Lines (%)	Words (<i>n</i>)	Words (%)
2012	13,877	1.0	65,215	1.0
2013	42,792	3.1	186,043	2.8
2014	77,922	5.6	347,019	5.3
2015	210,354	15.2	939,696	14.2
2016	346,545	25.0	1,584,041	24.0
2017	265,292	19.2	1,229,895	18.6
2018	178,696	12.9	902,801	13.7
2019	90,446	6.5	545,972	8.3
2020	137,770	10.0	705,037	10.7
2021	20,887	1.5	97,227	1.5
2022	131	0.01	511	0.01
Total	1,384,712	100.0	6,603,457	100.0

The dataset presented in this section provides a comprehensive overview of CoSEM, highlighting its scope, structure, and demographic balance. Building upon this descriptive foundation, the [following section](#) offers a closer examination of how the corpus composition informs the patterns of linguistic usage observed within it. By understanding the demographic and temporal dimensions, we gain insight into how sociolinguistic factors shape digital communication practices among Singaporean users.

As the data shows, CoSEM is largely shaped by contributions from younger speakers, particularly those in their twenties, given that the data was primarily collected from university students. The gender distribution further highlights meaningful participation from both male and female speakers, with a slight female majority, suggesting dynamic gendered interactions within Singapore's digital discourse spaces. This imbalance reflects the general tendency in linguistics courses at the National University of Singapore, where female students typically outnumber their male counterparts, and aligns with the demographic composition of the course from which much of the data was collected. The breakdown by nationality and race demonstrates the corpus's strong local grounding, with Singaporean Chinese users making up the majority of participants, followed by Indian and Malay contributors. This composition underscores how local demographic realities and social networks are reflected in digital communication patterns.

2.2 CoSEM in Relation to Other Corpora

The Corpus of Singapore English Messages stands apart from other corpora of CSE in several ways. First, it is a corpus of written electronic communication, a type of conversational fingered speech (Faulkner & Culwin 2005), which consists, therefore, of language belonging to a separate medium from the purely spoken or written type found, for instance, in the International Corpus of English (ICE). Due to its largely informal style, the genre has been said to be sited halfway between spoken and written, though this qualification is by no means unchallenged (cf. Maynor 1994; Crystal 2006; Androutsopoulos 2006). Text messaging is semi-synchronous in nature, which sets it apart from spoken conversation as well as from written emails, for example. The earliest publicly available text messaging corpus with Singapore English content, as far as we know, is the NUS SMS Corpus (Chen & Kan 2012), although it is much smaller than CoSEM (~56,000 messages) and not focussed on CSE. Other previous research on CSE has also drawn on online discussion forums as a source of data (e.g., Deuber, Leimgruber, & Sand 2018; Ziegeler 2020; Ziegeler 2024), which are yet another genre characterised by less synchronous interaction than the text

messages in CoSEM. More recently, the YouTube Corpus of Singapore English Podcasts (YCSEP, see Coats et al. 2025) has expanded the landscape of CSE corpora by providing a large-scale, fully transcribed dataset of informal spoken Singapore English. Like CoSEM, YCSEP captures informal, spontaneous use of CSE across a wide range of speakers, offering valuable insights into contemporary CSE use. However, while YCSEP represents spoken interaction primarily through long-form podcast discourse, CoSEM captures short-form, everyday digital communication.

Second, CoSEM features a form of CSE more contemporary than is found in other recent corpora. ICE-SIN in particular, which is often used in cross-varietal analyses due to its consistent internal architecture in line with other national corpora of the same suite, is among the earliest ICE components collected and dates from the 1990s. As such, it is at least one generation older than CoSEM and does not reflect current linguistic practices or the societal changes that have taken place in the country since then, including a larger English-speaking base in the population and changing attitudes towards CSE. Third, CoSEM contains a decidedly *young* variety of CSE, as evidenced by the age distribution given in Table 1. While this is at least in part an artefact of the data collection procedure as well as the electronic medium used, it sets CoSEM somewhat apart from other corpora such as ICE-SIN and the Sgforums corpus (Deuber, Leimgruber, & Sand 2018). Furthermore, especially compared with ICE-SIN, CoSEM features much less acrolectal speech, given the low-impact data collection method removing many of the issues associated with the interviewer paradox. As such, the corpus is a timely addition to existing databases on CSE, allowing new insights into young and innovative uses of language in the Singapore context.

For these reasons, the CSE captured in CoSEM should be of interest to a wide range of linguistics researchers: from World Englishes researchers and sociolinguists to formal linguists and psycholinguists interested in CSE and/or nonstandard varieties of English. To ensure that the corpus itself can be used easily by the field, we are currently exploring ways to enrich CoSEM further with part-of-speech (POS) tagging and syntactic parses. Following the approach outlined in Lin et al. (2023) and Huang et al. (2025), we plan to apply statistical taggers and parsers to automatically annotate the corpus, followed by manual validation of tags and parses, especially for linguistic phenomena of interest like SFPs and constructions like topic-comment structures, ellipsis, and so on. The manual validation process will require a substantial investment of time and resources, but it will help ensure that the tags and parses are sufficiently accurate for research purposes.

In summary, the overview of CoSEM in this section establishes the foundation for subsequent analyses in this volume. The [next section](#) introduces a recently administered survey on how Singaporeans use CSE in social media, designed to supplement the corpus data. The remainder of this Element will explore the findings, highlighting both corpus-based observations and survey-driven perspectives on language use, identity, and sociocultural meanings.

3 Integrating Corpus and Survey Methods to Examine CSE Use and Change

3.1 Introduction

3.1.1 CoSEM Studies Thus Far

In the [previous section](#), we discussed CoSEM's design and logistics; in this section, we provide an overview of research conducted with the corpus to date. One of our first publications, Gonzales et al. (2023a), introduced CoSEM, situating it alongside existing corpora of Singapore English, detailing its construction and 3.6-million-word composition (v.1), and illustrating its analytical potential through preliminary case studies on two CSE features: the Hokkien-derived (*bo*)*jio* 'NEG invite' and clause-final adverbs (CFAs). In conjunction with the larger and more recent CoSEM (v. 5) described earlier, this work established the corpus as a key resource for tracing change in CSE in computer-mediated communication. Subsequent studies have used CoSEM to investigate SFPs – arguably the most well-known feature of CSE and an enduring area of interest in CSE research. Leimgruber et al. (2021) showed that gender and ethnicity shape the use of certain particles in CSE, reinforcing the view that SFPs are a particularly sensitive locus of sociolinguistic differentiation in CSE (see also Leimgruber 2009; Smakman & Wagenaar 2013; Botha 2018). Building on Khoo (2012), Hiramoto et al. (2022) traced the development of the novel SFP *sia* from its origins as a Malay swearword into a marker of youth identity and a 'coarse but cool' style, used across a variety of speakers and interactional contexts. Hafiz et al. (2025) extended this work, examining the distribution of *sia* and its variants (*sial*, *siak*, and *siol*) across demographic groups and showing that the use of *sia* and its variants has spread beyond young male CSE speakers and is increasingly used by younger CSE speakers.

Building on the pilot study in Gonzales et al. (2023a), J. J. Lim, Hiramoto et al. (2024) revisited CFAs in CSE, showing that *already* and *only* have increased in frequency over time while *also* has remained relatively stable. These results provide evidence for growing morphosyntactic endonormativity

(Schneider 2007) in CSE. The authors also demonstrated that younger and male speakers lead in the use of CFAs, reinforcing earlier findings that age and gender are highly relevant for understanding variation in CSE.

The Corpus of Singapore English Messages has also facilitated our investigation of innovative CSE features, such as adverbial *confirm* and the clause-initial question tag *is it*. Drawing on data from CoSEM and ICE-SIN, J. J. Lim, Hafiz et al. (2024) documented the emergence and formal properties of a novel adverb in CSE – *confirm* – used to express strong speaker certainty in an utterance. Strikingly, the authors also showed that adverbial *confirm* was used more frequently in CoSEM over other surface-semantically equivalent adverbials such as *definitely* and *for sure*. As the adverbial use of *confirm* has not been reported elsewhere in the CSE literature, the authors introduce it as an emerging CSE resource and use CoSEM text message data to describe its formal properties and distribution. Similarly, Gonzales et al. (2023b) first drew attention to the innovative use of *is it* in clause-initial position in CSE, documenting not only its formal properties but also showing that this construction is more likely to be used in rhetorical, playful contexts and by younger Singaporean speakers. Building on this work, Hiramoto et al. (2025) showed that the use of clause-initial *is it* in CSE has increased over time, and drawing on Siegel's (2000, 2008) substrate transfer theory, argued that it entered CSE due to contact with local Sinitic languages in Singapore.

In addition to the abovementioned features, CoSEM contains many other playful and innovative CSE forms that may indicate ongoing changes in the variety. Using data from CoSEM, several undergraduate honours theses have investigated features like *already* and *only* (Cheong 2016), *toh* 'fall, die' of Hokkien origin (Chee 2019), *dei*, a vocative directed particularly at Indian men (Kaur 2019), and the lesser-studied SFPs *bah* and *wor* (Choo 2024). Other promising avenues for further study have surfaced in exploratory student projects, such as the use of double SFPs (see Boo, Lee, & Tan 2023). Together, these studies underscore the richness of the CoSEM dataset and its potential to capture socially salient innovations that have not yet been systematically described in the literature. At the same time, our reliance on CoSEM necessarily constrains the kinds of claims we can make about CSE more broadly. As a text-message corpus, CoSEM privileges relatively young, digitally active speakers and offers only indirect access to speakers' metalinguistic awareness, social evaluations, and reported usage outside the messaging context. The digital communication landscape has also shifted considerably since we began building CoSEM, with many Singaporeans now using other messaging platforms beyond WhatsApp and Telegram and making it increasingly difficult to capture CSE usage in Computer-Mediated communication. This

may be why some of the innovative items identified in the corpus (e.g., *toh*, *dei*, and double SFPs) occur infrequently, making it challenging for fine-grained quantitative modelling by age, ethnicity, or gender. However, these features are recognisable and stylistically salient for many CSE speakers. To address these limitations and test whether these features are diffusing beyond the specific cohorts represented in CoSEM, it became crucial to supplement the corpus with an attitudinal and usage survey.

3.1.2 Motivations for the Survey

A central question in sociolinguistics research concerns the relationship between how speakers perceive their own language use and their actual linguistic behaviour. Classic work by Labov (1966, 1972, 1975) demonstrated that speakers' self-reported usage of certain linguistic features often diverges from observed usage, suggesting that metalinguistic intuitions may not always be reliable indicators of linguistic behaviour (see also Trudgill 1972, 1974a). Such mismatches may arise because speakers' perceptions are shaped by their attitudes towards certain linguistic features, as well as broader prescriptive norms about language. At the same time, other studies (e.g., Bailey, Wikle & Tillery 1997; Chambers 1998; Bailey & Tillery 1999; Tillery, Wikle & Bailey 2000; Benson et al. 2009; Dollinger 2012; Dollinger 2015; Kostadinova 2018) have demonstrated that speakers' perceptions can, under certain conditions and for certain linguistic features, reliably track actual patterns of language use. The supplementary survey to CoSEM allows us to probe how widely these forms are known, how often and in what contexts they are used (e.g., speech vs. texting), and what social meanings speakers attribute to them, including possible associations with particular ethnic groups, genders, or stance styles. In this way, corpus data and survey evidence can be brought into productive dialogue: the corpus provides ecologically valid records of usage *in situ*, while the survey offers a broader window onto the penetration, social evaluation, and perceived normativity of emerging CSE features among speakers outside CoSEM.

We administered an online survey containing questions about respondents' demographic info, CSE use (awareness, acceptability, frequency, etc.) (see Section 3.3 for details), and contact information for possible follow-up. The survey was distributed across our social networks in Singapore between 2024 and 2025 and was targeted at adults over eighteen who reported being native CSE speakers and were living in Singapore. While it was not possible to exclude CoSEM participants in the survey entirely, we tried our best to recruit participants who are outside of the CoSEM cohorts by encouraging the survey participants to distribute the survey to a wider range of people in terms of

age, race, occupation, education, and language background. This snowball method yielded 230 participants, of which 214 met our criteria and were retained for analysis.

In this section, we focus on comparing survey data on Singaporeans' self-reported usage of CSE linguistic features with actual usage patterns attested in CoSEM as our initial step of utilising the combination method. To do this, we highlight three salient CSE linguistic features mentioned earlier as our test samples: clause-initial *is it* (Gonzales et al. 2023b; Hiramoto et al. 2025), adverbial *confirm* (Lim, Hafiz, et al. 2024), and the Malay-derived SFP *sia* (Khoo 2012; Hiramoto et al. 2022; Hafiz et al. 2025). These linguistic tokens were selected due to the fact that they show frequencies in CoSEM substantial enough to support a meaningful comparison. We anticipate that insights gained from these high-frequency items will help guide future analyses of less frequent CSE features by revealing how survey-based perceptions may align with or diverge from corpus-based observations.

It is important to note that the survey respondents and corpus participants do not constitute the same group of individuals. Because both datasets were collected anonymously, it is not possible to determine any overlap between them, nor to directly compare a respondent's perception of their linguistic behaviour with their actual observed usage. However, this is not our objective. Rather, by examining whether self-reported usage by survey respondents aligns with or diverges from previous corpus-based findings, we aim to contribute to ongoing discussions about the extent to which self-reports can serve as reliable indicators of language use across social variables.

Section 3.2 provides a brief overview of the three linguistic features under consideration. Section 3.3 outlines the survey design and our hypotheses. Section 3.4 presents the results of our analysis, and Section 3.5 concludes with a discussion of how self-reported usage compares with observed patterns in CoSEM, and what this might reveal about language variation in CSE.

3.2 Overview of *Is It*, *Confirm*, and *sia*

3.2.1 Clause-Initial *Is It*

In addition to its canonical use in standardised English, where *is it* functions primarily as a question tag seeking agreement from the listener (1), CSE also contains an innovative clause-initial use of *is it*, which serves mainly as a rhetorical device conveying a playful or teasing affect (2). It is hypothesised that this clause-initial use arose through substrate influence from the clause-initial A-not-A question construction found in the Sinitic languages spoken in Singapore. See Gonzales et al. (2023b) and Hiramoto et al. (2025) for detailed

discussions of CSE clause-initial *is it*, including how it differs from the question-tag *is it* in standardised English.

- (1) it's not their problem **is it**
<ICE-SIN:S1A-062#89:1:B>
- (2) *wa is it you think i wuliao*
'Wow, are you saying that I am boring?'
<COSEM:18CF59-2829-20CHF-2017>

Gonzales et al. (2023b) find that younger speakers are more likely to use clause-initial *is it* than older ones. However, inadequate evidence of gender and ethnicity effects is found.

3.2.2 Adverbial Confirm

As in standardised English, the lexical item *confirm* can function as a verb in CSE (3). However, unlike in standardised English, CSE also features an adverbial use of *confirm*, which expresses strong speaker certainty, as illustrated in (4). J. J. Lim, Hafiz, et al. (2024) argue that CSE adverbial *confirm* represents a case of grammatical replication (Heine & Kuteva 2003; Heine & Kuteva 2005; Heine et al. 2021) from Mandarin *kending*, which similarly functions both as a verb meaning 'to confirm' and as an adverb meaning 'definitely'. See J. J. Lim, Hafiz, et al. (2024) for intricate evidence that supports treating CSE adverbial *confirm* as a strongly subjective speaker-oriented adverb.

- (3) *But let's wait for <name> to confirm the timing?*
'Let's wait for <name> to confirm the time, shall we?'
<COSEM:17CF24-111-23CHM-2016>
- (4) *Aiyah I will confirm be late*
'Damn, I will definitely be late.'
<COSEM:18CF47-1636-20CHF-2017>

While previous work has not reported any effect of age, gender, and ethnicity on the use of adverbial *confirm*, ongoing research in the team indicates that younger speakers, male speakers, and Chinese speakers are more likely to use adverbial *confirm*.

3.2.3 Sentence-Final Particle *sia*

A recent addition to CSE's well-known inventory of SFPs is *sia*, which contains four variants: *sia*, *sial*, *siak*, and *siol*. Previous research has suggested that *sia* functions as an intensifier signalling coarseness (Khoo 2012) and may also be

pragmatically employed to index youthfulness and coolness (Hiramoto et al. 2022). See Hafiz et al. (2025) for a detailed discussion of the pragmatic meaning differences between the four variants of SIA. An example of *sia* is shown in (5).

- (5) *How much honey he buy sia*
‘How much honey did he buy?!’
<COSEM:18CF42-1803-23MAF-2016>

Building on Khoo (2012) and Hiramoto et al. (2022), Hafiz et al. (2025) show that younger speakers, male speakers, and Malay speakers are more likely to use SIA.

3.2.4 Intermediate Summary

Table 6 summarises the findings regarding the relationship between the use of clause-initial *is it*, adverbial *confirm*, and SFP SIA, and the speaker’s age, gender, and ethnicity.

First, younger Singaporeans are found to be more likely to use all three linguistic features, reinforcing the idea that these variables represent recent innovations in CSE. However, limited evidence of the effect of ethnicity has been found on the use of clause-initial *is it* in CoSEM, while Chinese speakers are more likely to use adverbial *confirm*, and Malay speakers are more likely to use SFP SIA. This contrast suggests that adverbial *confirm* and SFP SIA may represent later developments in CSE, while clause-initial *is it* may represent an earlier development and is therefore broadly shared across Singapore’s major ethnic groups. The hypothesis that adverbial *confirm* and SFP SIA are newer innovations receives some support from gender patterns: inadequate gender effects are observed for clause-initial *is it*, while males are more likely to use adverbial *confirm* and SFP SIA. Previous research has similarly identified the male gender as a positive predictor of the use of morphosyntactic features in CSE (e.g., Leimgruber et al. 2021; Lim, Hiramoto, et al. 2024; Hafiz et al. 2025).

Table 6 Previous findings on the relationship between CSE features and social variables

	Age	Gender	Ethnicity
Clause-initial <i>is it</i>	Younger	(Not enough evidence of effect)	(Not enough evidence of effect)
Adverbial <i>confirm</i>	Younger	Male	Chinese
SFP SIA	Younger	Male	Malay

3.3 Methodology and Hypotheses

3.3.1 Structure of the Survey

We designed the survey using Google Forms, titled *Survey on Singlish use*. The aim of the survey was to investigate the contexts in which CSE is used among Singaporeans, how Singaporeans perceive the use of CSE across different age groups, and respondents' perceptions of their own use of specific CSE morpho-syntactic features. The survey was divided into eight sections, and respondents could only proceed to the [next section](#) after completing all mandatory fields in the current one.

[Section 1](#) introduced the survey and included its title, purpose, and the investigators involved in the research (i.e., the authors of this monograph). It also provided an overview of the survey questions and information about data privacy and contact details. Participation in the survey was entirely voluntary, and respondents did not receive any compensation for completing the survey.

[Section 2](#) collected demographic information from respondents, including their age (numerical value), gender (male, female, or option to self-identify), ethnicity/race (Chinese, Malay, Indian, Eurasian, or option to self-identify), nationality (Singaporean or option to self-identify), years of residency in Singapore (numerical value), and social media platforms used (TikTok, WhatsApp, Twitter, Line, Viber, WeChat, Facebook, Messenger, Instagram, Signal, Telegram, SMS, Reddit, and an option to specify others). Respondents could also optionally self-identify their generational membership (Gen Alpha, Gen Z/Zoomer, Gen Y/Millennial, Gen X, Baby Boomers, Silent Generation, Greatest Generation, or option to select 'No idea' or to specify others).

[Section 3](#) asked respondents to indicate which social media platforms they frequently use CSE on, with the same options available as presented in [Section 2](#). Respondents were then asked to describe, based on their experience, the distinctive ways in which children, teenagers, young adults (ages twenty-one to fifty), and older adults (over fifty) use CSE. Respondents were also invited to share examples and reflections on whether they perceive there to be any age-related differences in the use of Singlish on social media.

[Sections 4 to 7](#) focussed on respondents' self-reported frequency of use of specific morphosyntactic features in CSE. For each feature, respondents indicated whether they used it 'a lot', 'occasionally', 'very rarely', or 'never'. Because the survey covered a wide range of features, the questions were divided across three sections to reduce participant fatigue. [Section 4](#) included a mix of relatively new and older CSE features, as well as a question asking whether respondents believed *jio* or *bojio* entered CSE first, whether both emerged simultaneously, or whether they were unsure. [Section 5](#) concentrated on

relatively older CSE features. Sections 6 and 7 focussed primarily on SFP usage, with particular attention to double SFPs. To minimise misinterpretation, each feature was accompanied by an example sentence – some drawn from CoSEM – illustrating how the feature might be used in a CSE utterance. The features investigated in each section are as follows:

Section 4

- (1) *never jio* (e.g., *Why you never jio me to the party?* ‘Why didn’t you invite me to the party?’)¹
- (2) *confirm* + verb (e.g., *I confirm fail the test.* ‘I will definitely fail the test’.)
- (3) *is it* (e.g., *Izzit you think I’m crazy?* ‘Do you think I’m crazy?’)
- (4) *shag* (e.g., *I damn shag these few days sigh can’t help much.* ‘I’m damn tired these days, I can’t help it.’)
- (5) *spoil market* (e.g., *You do so much for what? Don’t spoil market leh.* ‘What are you doing so much for? Don’t make us look bad.’)
- (6) *chop* (e.g., *You got chop your card?* ‘Did you stamp your card?’)

Section 5

- (1) *kena* + verb (e.g., *I kena scold by the teacher yesterday.* ‘I was scolded by the teacher yesterday.’)
- (2) *shiok!* (e.g., *This drink damn shiok!* ‘This drink is very refreshing/tasty!’)
- (3) *rabak* (e.g., *He feeling sick. I saw him today really quite rabak.* ‘He is feeling sick. I saw him today and he looked really terrible.’)
- (4) *rabz* (e.g., *Yeah my cough also rabz af.* ‘Yeah, my cough is also very terrible.’)
- (5) *pon* (e.g., *I’m going to pon tomorrow’s training.* ‘I’m going to skip tomorrow’s training.’)
- (6) *cheem* (e.g., *His English very cheem leh.* ‘His English is very difficult to understand.’)
- (7) *sibeh* (e.g., *Her mee rebus sibeh good sia!* ‘Her mee rebus is very tasty!’)
- (8) *goondu* (e.g., *She bring the wrong book again. Damn goondu leh her.* ‘She brought the wrong book again. She’s really absent-minded.’)

Section 6

- (1) *sia* (e.g., *So hot sia?* ‘It’s very hot, isn’t it?’)
- (2) *toh* (e.g., *anyways I just ended my event and gonna toh till 9* ‘Anyway, my event just ended and I’m going to crash (i.e., rest) till 9.’)
- (3) *sia la* (e.g., *ya sia, don’t think we stupid siala* ‘Yeah, don’t think we are stupid.’)

¹ The standardised English translations are provided here for clarification; they were not given in the survey.

- (4) *nia* (e.g., *I providing beer nia leh* ‘I’m only providing beer.’)
- (5) *nia la* (e.g., *I think still count as 3hrs nia la* ‘I think it still counts as only 3 hours.’)
- (6) *liao sia* (e.g., *Can preorder liao sia* ‘We can [begin to] preorder.’)
- (7) *liao lor* (e.g., *Eh cannot buy drinks liao lor* ‘Eh, it is no longer possible to buy drinks.’)

Section 7

- (1) *le meh* (e.g., *Can open le meh?* ‘Can we [begin to] open it?’)
- (2) *le leh* (e.g., *Its very loud le leh* ‘It is already very loud.’)
- (3) *le hor* (e.g., *Now u know why fat le horh* ‘Now you know why [you’re] fat.’)
- (4) *la eh* (e.g., *Good luck la eh* ‘Good luck!’)
- (5) *la hor* (e.g., *Should be fine la hor* ‘It should be fine, right?’)
- (6) *la sia* (e.g., *Tell her she can leave la sia!* ‘Tell her she can leave!’)
- (7) *leh hor* (e.g., *John has been coming to work late. You’re worried that he might get into trouble if he’s late again. You say to John, ‘Don’t come late again leh hor’.* ‘Don’t come late again, please.’)
- (8) *leh sia* (e.g., *You are hiking with your friends. You are hungry but your friends want to continue. Finally, you say, ‘I’m hungry leh sia. Can we stop for lunch?’* ‘I’m hungry. Can we stop for lunch?’)
- (9) *what* (e.g., *I don’t know we need to bring IC. They also never tell us what.* ‘I didn’t know we had to bring our IC (identification card). They didn’t tell us!’)

Finally, Section 8 asked whether the respondents would be willing to be contacted for a follow-up interview, with the option to provide their contact information.

3.3.2 Survey Distribution

The survey was distributed between 2024 and 2025, beginning with the authors’ personal networks and subsequently extending to the networks of their friends and family members. The first responses were collected on 17 May 2024, and the final response was received on 13 March 2025.

3.3.3 Respondent Pool

We collected a total of 230 survey responses. Sixteen were excluded because the respondents were either non-Singaporean or Singaporean individuals who had spent the majority of their lives outside Singapore. Of the 214 remaining responses, 118 (55.1%) belonged to female respondents and 96 (44.9%) belonged to male respondents, creating a slight imbalance in gender distribution. We also

assigned the respondents to generational cohorts based on their age: 56 (26.2%) Zoomers (ages 12 to 27 in 2024; 13 to 28 in 2025), 83 (38.8%) Millennials (ages 28 to 43 in 2024; 29 to 44 in 2025), 59 (27.6%) Gen X respondents (ages 44 to 59 in 2024; 45 to 60 in 2025), and 15 Boomers (ages 60 to 78 in 2024; 61 to 79 in 2025). Only one respondent was 80 years of age – above 78 – in 2024, and we opted to group this individual with the Boomers due to the proximity of their age rather than creating a separate category. This brings the total number of Boomers to 16 (7.5%). In terms of ethnicity, 167 (78.0%) respondents were Chinese, 20 (9.3%) respondents were Indian, 13 (6.1%) respondents were Malay, and 14 (6.5%) respondents identified as Other. A limitation of the study is the uneven distribution across generational cohorts and ethnic groups in the respondent pool.

3.3.4 Predictions

As mentioned earlier, this study investigates whether CSE speakers' self-reported frequency of use of three features – clause-initial *is it*, adverbial *confirm*, and SFP *sia*² – aligns with previously reported and ongoing corpus-based findings. To recap, these findings have shown that younger speakers are more likely to use all three features, male speakers and Chinese speakers are more likely to use adverbial *confirm*, and male speakers and Malay speakers are more likely to use SFP *sia*. Based on these findings, we predict that younger CSE speakers – which we define as belonging to the Zoomer and Millennial cohorts – will report higher frequencies of use for all three features. We further expect that male speakers will report higher use for both adverbial *confirm* and *sia*, while Chinese speakers will report higher use of adverbial *confirm* and Malay speakers higher use of *sia*. By testing these predictions, we aim to assess whether self-reported usage matches corpus-based distributions, thereby contributing to broader discussions on the reliability of self-reports as indicators of language variation across social groups.

3.4 Results

For the purposes of running correlation tests, we recoded categorical variables in the survey into numeric values to calculate the strength and direction of associations between social variables and reported usage patterns. Gender was recoded as a binary variable (0 = female, 1 = male). Because several ethnic categories contained very small cell counts, including zeroes, making groupwise comparisons unreliable, ethnicity was collapsed into a binary contrast between Chinese

² In this survey we only queried one variant of *sia*. This is why the notation here differs from that used in Section 3.2, where a range of different spellings and variants of a larger variable *SIA* was discussed.

and non-Chinese respondents (0 = Non-Chinese, 1 = Chinese). Finally, self-reported frequency of use was recoded onto a four-point scale (0 = Never, 1 = Very rarely, 2 = Occasionally, 3 = A lot).

3.4.1 Self-Reported Frequency of Clause-Initial *Is It* Usage

A Spearman correlation test between continuous age and recoded frequency revealed a significant association between the age of the respondent and self-reported frequency of clause-initial *is it* usage ($n = 214$, $r = -0.276$, $p < 0.001$). In other words, as age increases, the less likely it is for individuals to report using clause-initial *is it*.

The data also display clear generational patterns. To recap, we grouped participants into generations based on their ages in 2024/2025: Zoomers (ages 12/13–27/28), Millennials (ages 28/29–43/44), Gen X (ages 44/45–59/60), and Boomers (ages 60/61–78/79). As seen in Figure 1, younger cohorts (Zoomers and Millennials) are more likely to report using clause-initial *is it* ‘A lot’, while older cohorts (Gen X and Boomers) are less likely to do so, creating a descending staircase pattern for ‘A lot’ across the age groups.

Figure 2 shows the self-reported frequency of clause-initial *is it* usage across gender. Inferential statistics (i.e., Mann–Whitney *U*-test) revealed a significant difference between females ($n = 118$) and males ($n = 96$), $U = 4789$, $p < 0.05$, indicating that females are more likely to report using clause-initial *is it*.

Finally, the same test was conducted to assess whether there was an effect of ethnicity on the participants’ self-reported frequency of clause-initial *is it* use. The test did not find any significant difference between Non-Chinese ($n = 47$)

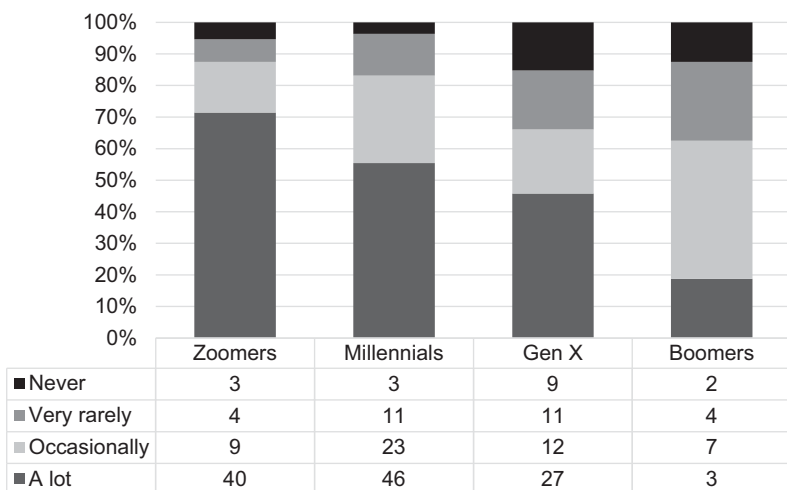


Figure 1 Self-reported frequency of clause-initial *is it* usage across generations

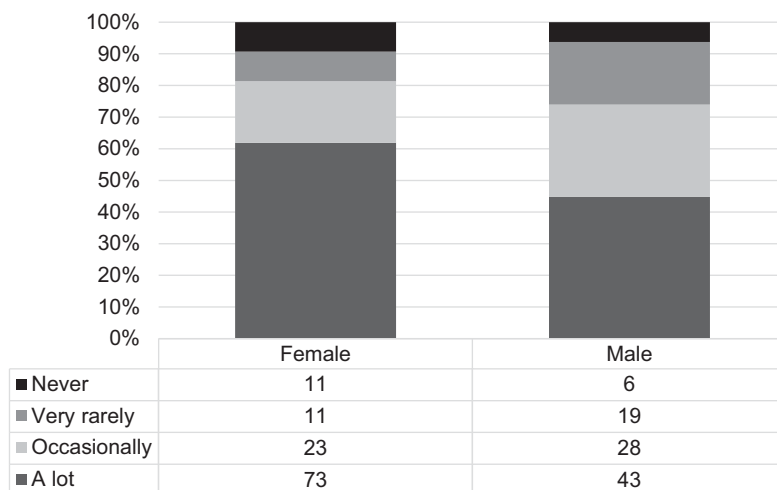


Figure 2 Self-reported frequency of clause-initial *is it* usage across genders

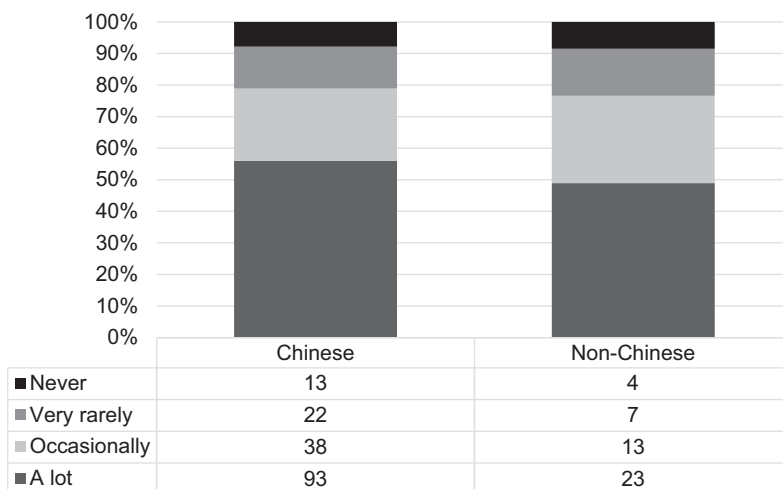


Figure 3 Self-reported frequency of clause-initial *is it* usage across ethnic groups

and Chinese ($n = 167$) respondents, $U = 3,684$, $p = 0.48$, despite the descriptive variation shown in Figure 3.

3.4.2 Self-Reported Frequency of Adverbial Confirm Usage

Turning next to adverbial *confirm*, a Spearman correlation test between continuous age and recoded frequency revealed a significant association between

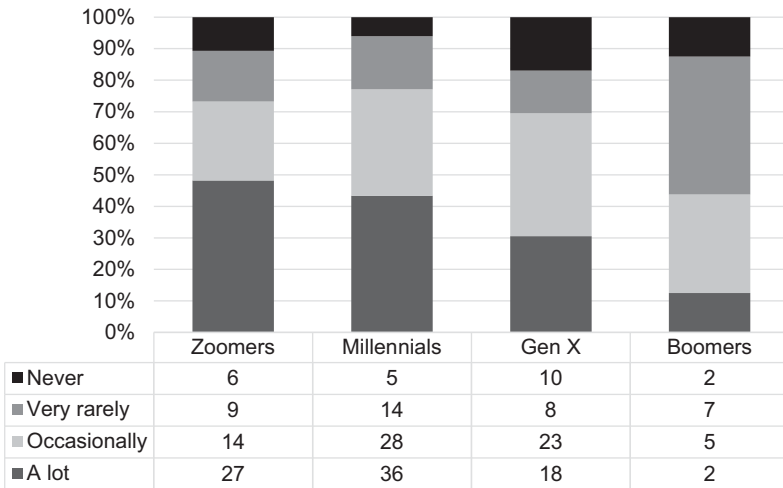


Figure 4 Self-reported frequency of adverbial *confirm* usage across generations

the age of the respondent and self-reported frequency of use ($n = 214$, $r = -0.18$, $p < 0.001$). This suggests that as age increases, the likelihood of reporting using adverbial *confirm* decreases.

The data also shows clear generational patterns, as shown in Figure 4. Younger speakers (Zoomers and Millennials) are more likely to report using adverbial *confirm* 'A lot', while older cohorts (Gen X and Boomers) are less likely to do so, creating a descending staircase pattern for 'A lot' across the age groups.

Figure 5 shows the self-reported frequency of adverbial *confirm* usage across gender. A Mann–Whitney U -test revealed no significant difference between females ($n = 118$) and males ($n = 96$), $U = 5,567.5$, $p = 0.82$.

Finally, the same test was conducted to assess whether there was an effect of ethnicity on the participants' self-reported frequency of adverbial *confirm* use. The test did not find any significant difference between Non-Chinese ($n = 47$) and Chinese ($n = 167$) respondents, $U = 3,429.5$, $p = 0.16$, despite the descriptive variation shown in Figure 6.

3.4.3 Self-Reported Frequency of Sentence-Final Particle *Sia*

Turning finally to the SFP *sia*, a Spearman correlation test between continuous age and recoded frequency revealed a significant association between the age of the respondent and self-reported frequency of use ($n = 214$, $r = -0.45$, $p < 0.001$). In other words, as age increases, the likelihood of reporting using *sia* decreases.

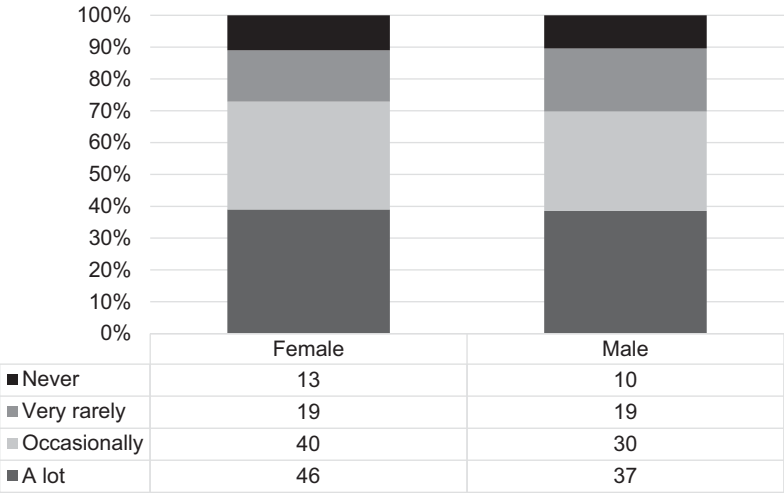


Figure 5 Self-reported frequency of adverbial *confirm* usage across genders

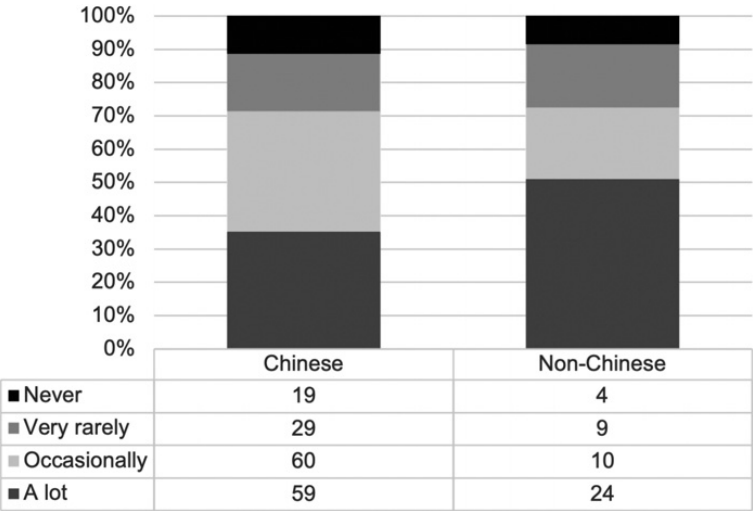


Figure 6 Self-reported frequency of adverbial *confirm* usage across ethnic groups

The data also shows clear generational patterns, as shown in Figure 7. Younger speakers (Zoomers and Millennials) are more likely to report using *sia* ‘A lot’, while older cohorts (Gen X and Boomers) are less likely to do so, creating a descending staircase pattern for ‘A lot’ across the age groups.

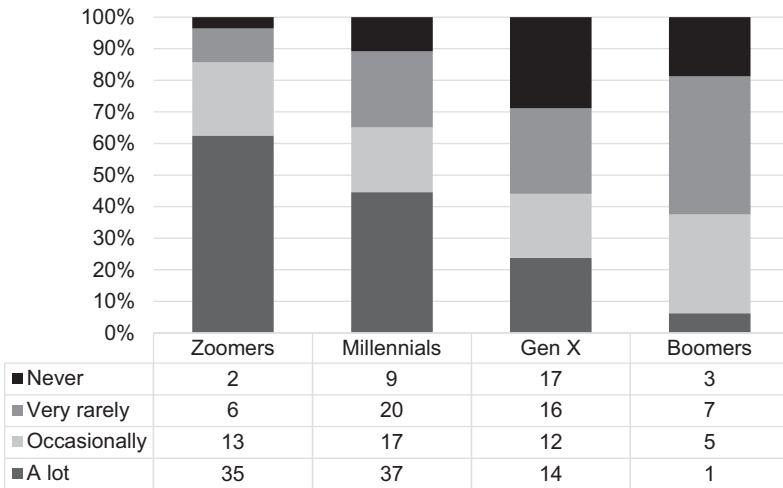


Figure 7 Self-reported frequency of *sia* usage across generations

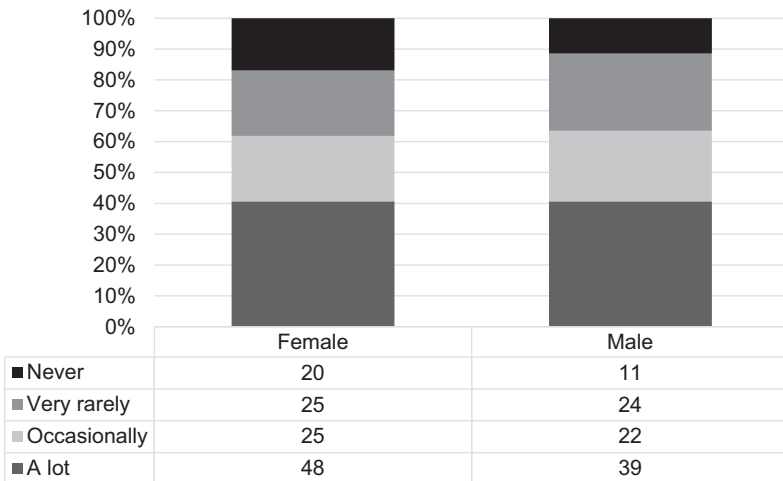


Figure 8 Self-reported frequency of *sia* usage across genders

Figure 8 shows the frequency of *sia* usage across genders. A Mann–Whitney *U*-test revealed no significant difference between females ($n = 118$) and males ($n = 96$), $U = 5,507$, $p = 0.71$.

Finally, the same test was conducted to assess whether there was an effect of ethnicity on the participants' self-reported frequency of use of *sia*. The test did not find any significant difference between Non-Chinese ($n = 47$) and Chinese ($n = 167$) respondents, $U = 3,356$, $p = 0.3$, despite the descriptive variation shown in **Figure 9**.

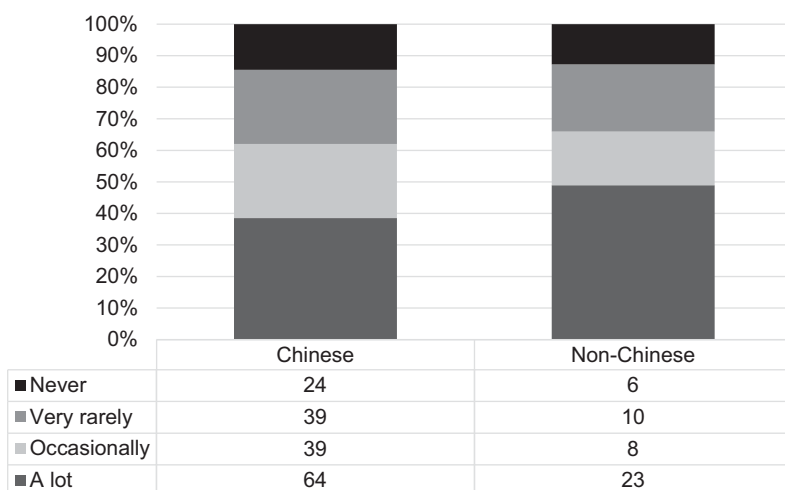


Figure 9 Self-reported frequency of *sia* usage across ethnic groups

3.4.4 Perceived Age-Related Differences in CSE Use on Social Media

Given that younger speakers self-report greater use of innovative CSE features while older speakers self-report less, we also wanted to find out whether respondents perceive age-related differences in CSE usage. One survey item asked participants whether they perceived any age-related differences in the use of CSE on social media, and to describe the differences, if any. Based on the response, we categorised it as ‘Yes’ (i.e., the participant perceived age-related differences in the use of CSE on social media) or ‘No/Unsure’ (i.e., the participant did not perceive any age-related differences). Four participants did not respond to this question and were excluded ($n = 210$). A Mann–Whitney U -test revealed no significant difference between continuous age and ‘Yes’ vs. ‘No/Unsure’ responses, $U = 4,585$, $p = 0.05$. Nonetheless, **Figure 10** indicates a clear generational trend: as their age increases, the participants’ responses are more likely to indicate ‘No/Unsure’.

Several respondents also highlighted differences in the digital domains where Singaporeans use CSE. As one noted: ‘Facebook generally has an older population, TikTok probably has the youngest. X/twitter is probably a mix of teens + 21–50 since [it’s] kinda old and new at the same time’. These observations suggest that platform choice may correlate with age-based variation in CSE usage – a promising avenue for further investigation.

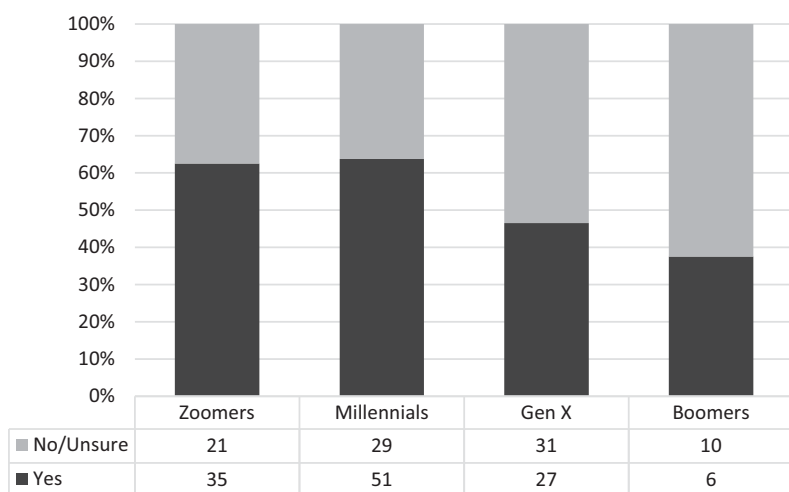


Figure 10 Perception of age-related differences in CSE usage on social media across generations

3.4.5 Intermediate Summary

Table 7 summarises the survey-based findings in this subsection and compares it to previous and ongoing corpus-based findings.

Across both the data from CoSEM and the survey, age consistently emerges as the most robust predictor of variation in the three CSE features examined herein. In every case, younger speakers show higher rates of use, regardless of whether the data comes from naturally occurring CMC or self-reported usage. In contrast, gender and ethnicity effects are largely confined to corpus data, as most of the self-reported data – apart from the gender effect on clause-initial *is it* – do not show significant differences.

3.5 Summary

This section has investigated the self-reported usage of three innovative CSE features – clause-initial *is it*, adverbial *confirm*, and the SFP *sia* – using survey data as an independent lens on contemporary CSE practices and as a complement to earlier corpus-based findings from CoSEM. The findings suggest that self-reported data may serve as reliable indicators of linguistic behaviour with respect to age, but not necessarily for gender or ethnicity. This position aligns with previous research arguing that self-reports are more accurate for non-phonetic linguistic features and/or features that are less stigmatised (e.g., Bailey, Wikle, & Tillery 1997; Benson et al. 2009; Dollinger 2015). As several scholars have noted, attitudes towards CSE have become increasingly

Table 7 Corpus-based vs. survey-based findings on the relationship between CSE features and social variables (bold text = corpus and survey results align)

	Age	Gender	Ethnicity
Clause-initial <i>is it</i>			
CoSEM	Younger	(Not enough evidence of effect)	(Not enough evidence of effect)
Survey	Younger	Female	(Not enough evidence of effect)
Adverbial <i>confirm</i>			
CoSEM	Younger	Male	Chinese
Survey	Younger	(Not enough evidence of effect)	(Not enough evidence of effect)
SFP <i>sia</i>			
CoSEM	Younger	Male	Malay
Survey	Younger	(Not enough evidence of effect)	(Not enough evidence of effect)

positive in recent years (see e.g., Leimgruber 2014; Lim 2015; Leimgruber, Siemund & Terasa 2018; Hiramoto 2019; Lee & Ahn 2021; Lim & Aman 2024), suggesting that stereotypically ‘Singlish’ features may carry less stigma among younger speakers. In this way, the survey results contribute to broader discussions about the utility of self-reported data in linguistics research.

Methodologically, the section demonstrates the value of putting corpus data and survey data into systematic dialogue. CoSEM offers ecologically valid records of actual usage in a specific CMC register, allowing us to trace the emergence and diffusion of new forms with considerable temporal and social resolution. The survey extends this perspective by tapping into how CSE users outside the corpus understand their own usage and that of others, and by providing explicit evidence of how features are evaluated, stereotyped, or linked to particular age groups and identities. Where corpus and survey findings converge, our confidence in the robustness of a pattern is strengthened; where they diverge, we gain insight into how linguistic practice and linguistic awareness may be decoupled.

Overall, the findings reinforce the idea that youth identity is closely tied to innovative linguistic developments in CSE. Younger CSE speakers’ recognition of age-based differences in CSE reflects an awareness of their role as linguistic trendsetters (cf. Labov 1972; Eckert 2000). This may explain why younger speakers both self-report in our survey and are observed in CoSEM to use clause-initial *is it*, adverbial *confirm*, and the SFP *sia* more frequently. This

converging finding strengthens our interpretation that these CSE features are part of an ongoing generational shift within CSE, with younger speakers driving the use of these innovative features. On the other hand, the divergence between our corpus-based and survey-based findings for gender and ethnicity effects suggests that these social-indexical associations, while observable in actual linguistic behaviour, may not be salient in speakers' metalinguistic consciousness. It may also indicate that speakers are less inclined to report gender- or ethnicity-based differences in the use of CSE features.

Finally, this section points towards several avenues for future work. On the empirical side, further survey-based and experimental studies could probe the social meanings of features such as *toh*, *dei*, double SFPs, and other low-frequency innovations that remain underattested in CoSEM. On the methodological side, expanding the respondent pool – especially among non-Chinese and older speakers – and incorporating follow-up interviews would allow a more fine-grained account of how CSE users talk about their own language. More broadly, the combined corpus-survey approach will offer a model for investigating other contact varieties in which digitally mediated data and speaker perceptions both play crucial roles in ongoing change. In future work, these survey-based insights can be brought into dialogue with CoSEM analyses to broaden our understanding of modern-day CSE across different communicative contexts. One such feature from contemporary CSE use found in CoSEM is *(bo)jio*, which we investigate in the [next section](#).

4 Structural and Stylistic Variation in the Use of *(Bo)Jio*

The CSE construction *bojio* can be traced to the two Hokkien elements *bô* 'not' and *chio* 'invite'³ (Wong 2020: 104; Babcock 2022: 337). Together, they are used primarily as an interjection that can be translated, rather clumsily, as 'why did you not invite me'; its use is illustrated in example (6). Both Wong (2020) and Babcock (2022) give a good account of its linguistic relevance, which can be further sited within the concept of *kiasu* 'afraid of losing out'. Wong (2020: 103) surmises that the term *bojio* is rather recent (and 'only entered Singlish in the last five or six years' at the time of his publication). Our corpus analysis in Gonzales et al. (2023a) supports this view, with the item entirely absent from ICE-SIN, for instance.

- (6) A: Yall not gonna drink right?
 B: We drinking now bruh
 A: Wtf

³ These words in Hokkien are written in Peh-oe-ji, which is different from how users of Singlish write it in Singlish 'jio'. The Singlish form is typically pronounced [dʒɪɔ].

A: **Bojio**

‘You didn’t invite me!’

<COSEM:17CF12-31046-21SGINM00-2015>

Whereas the interjection *bojio* has by now achieved a certain degree of salience within the Singaporean speech community,⁴ the term, especially its *jio* ‘invite’ element, has also taken on a life of its own, with even inflectional and derivational morphology applied to it: bare verbal and nominal forms are illustrated in (7) and (8), whereas inflected (regular past tense and plural forms, as well as *-ing* forms) are seen in (9), (10), and (11). Rarer examples include derivational *jioness*.

- (7) *House heads went out for supper and never jio us!!!*
 ‘The house heads went out for supper and didn’t invite us!’
 <COSEM:21DC19-13118-23SGCHMEK-2020>
- (8) <name> *just ignored my jio*
 ‘DT/F/CH/20 just ignored my invitation.’
 <COSEM:17CF11-12621-20SGCHF00-2015>
- (9) *we should’ve jioed you for lunch*
 ‘We should’ve invited you for lunch.’
 <COSEM:19CM03-39813-22SGXXFNW-2018>
- (10) *SPEAKING OF jios . . . can I jio someone to help me collect something later at 7+
 + at the drop off point*
 ‘Speaking of invitations, can I invite someone to help me collect something later at 7+ at the drop off point?’
 <COSEM:18CF58-9614-19SGCHF00-2017>
- (11) *I specifically recall jioing everyone*
 ‘I specifically recall inviting everyone.’
 <COSEM:21DC31-7892-20SGMAM0A-2018>

The aim of this section is to take a comprehensive look at (*bo*)*jio* (henceforth labelled (*jio*) as a sociolinguistic variable) in contrast to (*not*) *invite*, its non-CSE counterpart. In the following, we explain our methods (Section 4.1), present the results (Section 4.2), and discuss the findings from our datasets (Sections 4.3 and 4.4).

⁴ *Bojio* is now visible in the linguistic landscape, used, for example, in advertising as well as in items of clothing or bags, capitalising on the mercantile value of Singlish much as the discourse particle *lah* does.

4.1 Methodology

4.1.1 Data Sources and Data Pre-Processing

The dataset for this study draws from two primary sources: CoSEM (Gonzales et al. 2023a) and the social media platform X (formerly Twitter).⁵ Data underwent preprocessing using both manual and automated tools to facilitate linguistic and sociolinguistic analysis. To examine the distribution and usage of *jio*, the corpus was queried using R, with regular expressions extracting all occurrences of *jio*, including related tokens such as *bojio* and *invite*. This process yielded 1,745 tokens, which were subsequently reviewed by two proficient speakers of CSE to ensure accuracy and exclude spurious matches (e.g., cases where *jio* appeared within unrelated strings like *emoji only*). While CoSEM provided a structured dataset of text messages, no comparable corpus existed for Twitter data at the time of this study. Moreover, the introduction of restrictive and costly API policies in 2023 precluded the construction of a specialised Twitter corpus akin to CoSEM. To approximate CoSEM's scope and relevance, we leveraged data scraping techniques. Fortunately, in late 2021, a research team member had already conducted preliminary data extraction using Gonzales' Twitter Corpus Suite (Gonzales 2024, 2025b), a Python-based framework. This yielded a dataset of 5,286 tokens of *jio* (including *bojio* and *invite*), distributed over 2010 to 2021, circumventing API-related constraints. As with CoSEM, two proficient speakers of CSE reviewed the extracted tokens to ensure accuracy. Unlike CoSEM, the Twitter dataset lacked sociolinguistic metadata, necessitating computational methods to infer demographic variables. Specifically, we adopted an approach informed by the social construction of age and gender, following Gonzales' (2024) methodology for demographic inference on Twitter. Using Wang et al.'s (2019) deep learning framework, we extracted age and gender information based on profile pictures, self-descriptions, and other textual markers. However, no existing computational tools achieved sufficient accuracy (>60%) for race and nationality inference, leading to their exclusion from the dataset. While this omission limits cross-corpus generalisability for these variables, it does not significantly affect the study's primary focus on age as a sociolinguistic variable.

The final dataset, integrating WhatsApp (CoSEM) and Twitter (*jio*) data, comprised 7,031 tokens. These were coded for key linguistic factors, including presence/absence of morphological affixation (e.g., *jio-ing*, *bo jio-ed*) and negation (e.g., *jio* vs. *bojio*, *invite* vs. *not invite*). To streamline annotation,

⁵ The data analysed here was obtained before the change in name, which is why we refer to it as 'Twitter' data.

Natural Language Processing (NLP) tools such as *nlk* and *spaCy* (Bird, Klein, & Loper 2009; Montani et al. 2023) automated the coding of English tokens, while non-English items underwent manual annotation by two proficient speakers of CSE.

Beyond linguistic features, sociolinguistic variables were computationally derived to assess their influence on *jio* usage. These included measures of sociolinguistic style, sentiment, intensity, and platform-based variation. Following Kiesling et al.'s (2018) discussion of sentiment as one operationalisable component of social and interactional meaning in computational work, sentiment was analysed using the *sentimentR* package (Rinker 2021), which assigns each text a polarity score representing the direction of affective evaluation: positive values indicate approval, enthusiasm, or endorsement, whereas negative values indicate disapproval, frustration, or complaint. In Kiesling et al.'s terms, this measure captures the affective dimension of stancetaking, that is, (positive, neutral, and negative) evaluation towards discourse entities, while abstracting away from the dialogic processes of alignment and investment that emerge across turns and participants. To operationalise this concept, we thus draw on polarity as a metric, which broadly captures what kind of stance is being taken. In contrast, sentiment intensity captures how strongly that stance is expressed, regardless of whether it is positive or negative. Intensity was operationalised as the absolute value of the polarity score, such that both highly positive and highly negative utterances receive high intensity values, while neutral or weakly evaluative utterances receive low values. For example, a strongly negative warning and a strongly positive invitation differ in polarity but may exhibit comparable intensity, reflecting similar levels of emotional investment. Distinguishing between polarity and intensity allows the analysis to test whether *jio* usage is more sensitive to evaluative direction (e.g., positive vs. negative stance) or to affective strength and engagement more generally.

Stylistic variation was examined along three dimensions. First, platform-based style was categorised based on dataset origin (text message vs. Twitter). Second, formality was inferred computationally following Grafmiller et al. (2018), incorporating markers such as mean word length, nouniness, passive constructions, and noun-verb ratio. Third, informational versus interpersonal tendencies were measured through linguistic indicators, including subordinating conjunction frequency, stranded prepositions, and mean sentence length. These stylistic features were distilled into two overarching dimensions: formality and informationality/interpersonality. Additionally, POS tagging was applied, as the syntactic environment of *jio* – particularly adjacent POS categories – was hypothesised to condition its variability.

Both datasets underwent systematic collection and preprocessing to enable cross-platform analysis of *jio* usage. Prior research has shown that linguistic variation is shaped by platform-specific constraints, including character limits, multimodal affordances (e.g., images and videos), and interactional norms (Davies & Fuchs 2015; Bohmann 2016). By incorporating text message and Twitter data, this study offers a comprehensive account of the sociolinguistic factors shaping the distribution and use of *jio*, *bojio*, and *invite* in Singapore English.

4.1.2 Data Analysis

Using a curated dataset, we examined variation in the use of *jio* through a combination of frequency distribution analysis, where we get the proportional distribution of variants and compare them by factors of interest. To enhance generalisability, Bayesian logistic regression analysis was also used, following contemporary approaches in linguistics (Vasishth et al. 2018; MacKenzie 2020). To assess the robustness of our model beyond the training data, we implemented an 80/20 training-test data split. The training set was fitted using a mixed-effects binomial model, appropriate for the binary syntactic variable under investigation – *jio* vs. *invite*. We estimated model parameters via Markov chain Monte Carlo sampling implemented in the *brms* package in R.

The model was trained on 5,625 observations, with 30,000 iterations per chain across four Markov chains. To mitigate the influence of early stage sampling bias, we applied a burn-in period of 15,000 iterations per chain and set a thinning parameter of 2. Weakly informative priors (normal distribution [0,5]) were assigned to both intercepts and slopes. Testing alternative priors (e.g., uniform, Cauchy [0,5]) revealed no substantial impact on the resulting posterior distributions. Model convergence was verified using $\hat{R}$ values and Effective Sample Size (ESS) estimates, in line with Vehtari et al. (2021: 683). Specifically, we ensured that $\hat{R}$ remained below 1.01 and ESS exceeded 400 (see ‘Results’ for specific values). We did not conduct a frequentist regression analysis, as Bayesian methods, despite their computational demands, provide more interpretable probabilistic inferences and allow for explicit statements regarding the absence of effects – something frequentist approaches do not permit (McElreath 2020; Levshina 2022).

The final model incorporated a range of predictors (Table 8), selected based on sociolinguistic research on CSE, previous studies, and ethnographic insights into variables influencing (*jio*) variation. Multi-level categorical variables (e.g., race) were coded using Weighted Helmert coding (Sonderegger 2023). Additionally, we specified random intercepts for the POS of adjacent words

Table 8 Factors tested in the model

Classification type	Predictor	Levels	Type of variable	Effect type
Linguistic	Affix present	Absent, present	Categorical	Main effect
	Negated context	Negated, non-negated	Categorical	Main effect
	L1 POS	adjective, noun, etc.	Categorical	Random effect
	L1 Lexeme	<i>you, not, please</i> , etc.	Categorical	Random effect
	R1 Lexeme	<i>me, the, that</i> , etc.	Categorical	Random effect
	R1 POS	adjective, noun, etc.	Categorical	Random effect
Socio-stylistic	Age	younger to older	Continuous	Main effect
	Gender	Male-presenting, female-presenting	Categorical	Main effect
	Race	Indian, Eurasian, Malay, Chinese	Categorical	Main effect
	Nationality	Singaporean, Non-Singaporean	Categorical	Main effect
	Year	2010 to 2022	Continuous	Main effect
	User	User 1, User 2, User 3, etc.	Categorical	Random effect
	Sentiment	Positive to negative	Continuous	Main effect
	Intensity	High to low intensity	Continuous	Main effect
	Style	WhatsApp, Twitter	Categorical	Main effect
	Style	Interpersonal to informational	Continuous	Main effect
	Style	Informal to formal	Continuous	Main effect

and for lexeme, though, following Levshina (2016: 253), only lexemes and users with more than five tokens were retained as individual factor levels. Those with fewer than five occurrences were collapsed into an ‘other’ category to ensure sufficient data representation.

Following the Bayesian posterior summary, we assessed the predictors influencing (*jio*) variation using the probability of direction (*pd*) measure. A predictor was considered to have a meaningful effect on the dependent variable (e.g., likelihood of using *jio*) if its median estimate was substantively distant from zero or if its credibility intervals excluded zero. The probability of direction (*pd*) – defined as the proportion of posterior draws that share the sign of the median – serves in this study as a probabilistic measure of effect existence

and certainty. A *pd* value approaching 1 suggests strong evidence for an effect in the indicated direction (positive or negative), whereas a *pd* near 0.5 implies that the effect may be negligible or non-existent (Makowski et al. 2019). For instance, if a predictor's median estimate in our model is -1.2 with a *pd* of 0.95, this indicates that 95 per cent of the posterior distribution falls below zero, while only 5 per cent of values exceed zero (Levshina 2016). Given that *jio* was coded as the reference category, this would imply a 95 per cent probability that the predictor in question reduces the likelihood of selecting the *jio* variant.

4.2 Results

We first present the results of the frequency distribution analysis to establish the presence of variation, which serves as a necessary justification for the use of regression while also providing readers with an overview of the variability under investigation. We then introduce the regression model, outlining its key features. Following this, we evaluate the model's predictive accuracy and other relevant metrics before proceeding to a detailed analysis of (jio) variation as conditioned by the factors of interest.

4.3 (Jio) Variation: Frequency Distribution Analysis

Before proceeding with the regression analysis, it is crucial to establish evidence of variation in the use of (jio) and present an overview of variation patterns involving (jio). The frequency analysis presented in Table 9 demonstrates that speakers exhibit a preference for the English-derived *invite* (59.59%) over the Hokkien-derived *jio* (40.41%). However, this distribution is not heavily skewed in favour of *invite*, as the difference is relatively modest – approximately 10 percentage points – suggesting that *jio* remains a salient and productive variant within the linguistic repertoire of speakers. A closer examination of variation by negation context reveals a systematic pattern: while *invite* is favoured in affirmative contexts, *jio* predominates in negated constructions.

Table 9 Distribution of variants of (jio)

	Negated	Non-negated	Total
<i>Invite</i>	274 (27.18%)	3,078 (66.67%)	3,352 (59.59%)
<i>Jio</i>	734 (72.82%)	1,539 (33.33%)	2,273 (40.41%)
	<i>bojio</i> : 466 (46.23%)		
	<i>never, no, not, don't</i> :		
	268 (26.59%)		
Total	1,008 (100%)	4,617 (100%)	5,625 (100%)

Crucially, the majority of tokens in negated contexts occur in the lexicalised expression *bojio*, a case of what Labov (1972) characterises as ‘stable variation’, in which speakers consistently alternate between variants in accordance with social or discursive constraints. This distribution suggests that the selection of *jio* in negation contexts is not arbitrary but reflects an embedded social constraint shaping its use.

4.4 (Jio) Variation: Regression Analysis

4.4.1 General Introduction

The Bayesian model helps clarify which factors make speakers more or less likely to use (jio) rather than its alternative. Table 10 reports the model’s posterior estimates, with the median values indicating the direction and typical size of each effect (positive values increase the likelihood of (jio), negative values decrease it). Importantly, all predictors show a high probability of direction ($pd > 0.95$), meaning that the model is highly confident about the direction of each effect.

A clear quantitative pattern emerges when comparing groups of predictors. Intralinguistic factors have larger effects on average (mean absolute median $n = 1.75$, $SD = 1.11$) than socio-stylistic factors (mean $n = 1.15$, $SD = 1.02$). This means that grammatical and structural conditions exert a stronger pull on speakers’ choices, while social and stylistic factors still matter but operate more subtly. In practical terms, structure sets the main limits on where (jio) is likely to appear, and social meaning fine-tunes usage within those limits.

4.4.2 Model Predictions and Accuracy

To check how reliable the model’s results are, we tested how well it predicts new data that were not used to fit the model. Specifically, we compared the model’s predictions to the actual choices speakers made in a held-out test dataset ($n = 1,406$). This comparison is summarised in a confusion matrix (Table 11), which simply shows how often the model correctly predicts each variant and where it makes mistakes. The model predicts speakers’ choices very well: it correctly identifies the observed variant in about 94 per cent of cases (Table 12). This performance is far better than a simple baseline strategy of always guessing the most frequent variant (*invite*), which would be correct only 59 per cent of the time. Statistical tests confirm that this improvement is not due to chance ($p(\text{Acc} > \text{baseline}) < 0.001$; McNemar’s $p < 0.001$), and the high Kappa value (0.87) indicates strong agreement between the model’s predictions and the observed data beyond what would be expected by chance. In practical terms, these results

Table 10 Posterior estimates from the Bayesian model showing how different predictors increase or decrease the likelihood of using *jio* rather than *invite*.

Predictors	Median	SD	89% CI (HDI)	<i>pd</i>	<i>BF</i>	$\hat{R}$	ESS
Fixed effects							
Intercept	506.13	68.0	398.95–615.41	1	11.62	1	22,668
Affix present (absent vs. present)	−0.58	0.22	−0.92 – −0.22	1	0.58	1	22,757
Negated context (negated vs. non-negated)	−1.31	0.3	−1.78 – −0.83	1	6.4	1	24,604
Sentiment	−3.36	0.4	−4.02 – −2.73	1	17.4	1	22,097
Intensity	0.25	0.09	0.11–0.41	1	−0.38	1	22,965
Age	−0.04	0.01	−0.06 – −0.03	1	1.7	1	22,469
Gender (male-presenting vs. female-presenting)	−0.37	0.15	−0.61 – −0.12	0.99	−0.62	1	22,454
Race (Indian vs. Eurasian)	2.45	1.34	0.41–4.67	0.98	0.55	1	27,181
Race (Malay vs. Eurasian, Indian)	1.43	0.83	0.15–2.79	0.97	−0.25	1	26,866
Race (Chinese vs. Indian, Malay, Eurasian)	2.09	0.5	1.31–2.89	1	7.33	1	26,910
Nationality (Singaporean vs. Non-Singaporean)	0.97	0.6	0.02–1.94	0.95	−0.81	1	24,801
Style (WhatsApp vs. Twitter)	−0.97	0.34	−1.49 – −0.43	1	1.07	1	26,648
Year	−0.25	0.03	−0.3 – −0.2	1	12.01	1	22,764
Style (Interpersonal-Informational)	0.88	0.1	0.73–1.05	1	17.57	1	23,296
Style (Informal-Formal)	−0.17	0.1	−0.33 – −0.02	0.98	−2.13	1	22,437

Table 10 (cont.)

Predictors	Median	SD	89% CI (HDI)	<i>pd</i>	<i>BF</i>	$\hat{R}$	ESS
Random effects (Intercepts only)							
User (SD intercept)	1.77	0.42	1.12–2.43	1	11.04	1	19,236
L1 POS (SD intercept)	2.74	1.1	1.33–4.42	1	10.29	1	19,481
L1 Lexeme (SD intercept)	3.34	0.4	2.72–3.96	1	29.89	1	17,557
R1 Lexeme (SD intercept)	1.85	0.21	1.53–2.19	1	29.7	1	18,772
R1 POS (SD intercept)	0.67	0.37	0.24–1.24	1	2.76	1	19,102

Note: Multi-level categorical predictors are coded using Weighted Helmert contrasts, which compare each level to the average of the remaining levels; reference levels are shown in **boldface**. For each predictor, the table reports the median estimate (the most typical effect size), SD (uncertainty around that estimate), the 89 per cent highest density interval (HDI) indicating the most credible range of values, *pd* (probability that the effect is positive or negative), *BF* (Bayes factor, indicating strength of evidence for an effect), $\hat{R}$ (a convergence diagnostic showing that the model has run properly), and ESS (effective sample size, indicating the reliability of the estimates).

Table 11 *Jio/invite* model accuracy – Confusion matrix
($n = 1,406$ unobserved tokens)⁶

		Actual (Reference)	
		Invite	Jio
Predicted	<i>Invite</i>	803	52
	<i>Jio</i>	35	516

Table 12 *Invite/jio* model evaluation

Measure	Value
n	1,406
Acc. (Bal. Acc.)	0.94 (0.93)
CI [Acc]	0.92–0.95
NIR	0.59
p [Acc > NIR]	<0.001
ρ	0.87
CI [ρ]	0.86–0.88
t	66.5
p [ρ]	<0.001

show that the model has learned systematic patterns in how speakers choose between *jio* and *invite*, instead of merely memorising the training data. Its strong performance on unseen data indicates that the predictors included in the model provide a robust account of variation and generalise well beyond the specific dataset analysed here, supporting the validity of the analysis.

4.4.3 Structural/Linguistic Constraints on (*Jio*)

The analysis reveals that affix presence exerts minimal influence on the selection between the Hokkien-derived variant *jio* and the English-derived *invite*, where the use of Hokkien-derived *jio* is slightly more preferred when (*jio*) is not affixed (Figure 11a). The lack of a stronger effect is somewhat unexpected, as morphological factors often play a strong role in constraining variation, particularly in contact settings where phonological or morphosyntactic integration can influence borrowing patterns (Szmrecsanyi 2006; Gonzales 2025a). However, in this case, the relative stability of *jio* use regardless of affix presence suggests that speakers may not perceive affixation as a meaningful constraint on variant selection.

⁶ Kappa = 0.87, sensitivity = 0.96, specificity = 0.91, pos pred value = 0.94, neg pred value = 0.94, prevalence = 0.59, detection rate = 0.57, detection prevalence = 0.61, balanced accuracy = 0.93, Bayes R^2 = 0.82

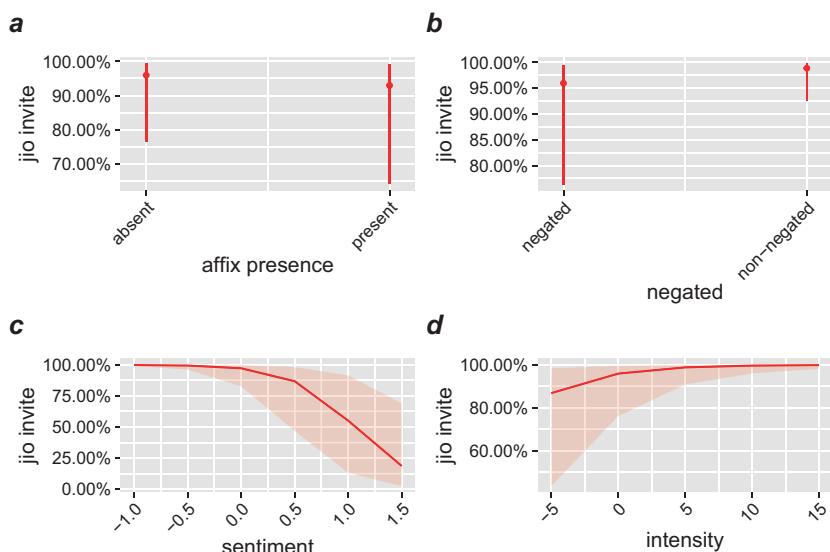


Figure 11 Marginal effects of structural factors on *jio* usage. The y-axis shows the model-estimated probability that *jio* is used rather than *invite*, with 100 per cent indicating a categorical preference for *jio*. Panels display posterior predicted probabilities as a function of (a) affix presence, (b) negation, (c) sentiment polarity, and (d) sentiment intensity, holding all other predictors constant. Points and lines represent predicted values, and error bars and shaded bands indicate 95 per cent credible intervals.

Negation shows a nuanced effect on *jio* usage once model-based predictions are considered. While raw frequencies indicate that *jio* is more common than *invite* in negated contexts overall (Table 9), the marginal effects in Figure 11b reveal that non-negated contexts exhibit a higher and more stable predicted likelihood of *jio* usage, as indicated by the tighter uncertainty interval. In contrast, negated contexts display greater variability in predicted probabilities, suggesting that negation does not uniformly favour *jio* across interactional settings. This divergence reflects the difference between aggregate distributional tendencies and conditional effects estimated by the model. The wider uncertainty associated with negated contexts likely arises from heterogeneity within the negation category, which subsumes both relatively fixed expressions such as *bojio* and more compositional constructions (e.g., *not jio* and *never jio*). These subtypes differ in their degree of routinisation and pragmatic function, resulting in less consistent model predictions. By contrast, non-negated *jio* usage appears more regular and contextually stable, yielding higher and more consistent predicted probabilities. These results collectively suggest that

negation does not categorically license *jio*, but rather introduces context-dependent variability, whereas non-negated environments support a more uniform preference for *jio* in the lexical alternation with *invite*.

The results for sentiment introduce a gradient social constraint that is implemented as a continuous structural predictor in the model, with *jio* being strongly associated with negative sentiment and *invite* emerging as the preferred variant in more positive contexts (Figure 11c). In this sense, sentiment occupies an intermediate position: it is social in its interpretation (indexing affective stance), but structural in its formal treatment as a scalar constraint on variant selection. This pattern is consistent with sociolinguistic work demonstrating that linguistic variants often carry affective or evaluative meanings, linking language choice to stance-taking and social positioning (Kiesling 2018; Pratt 2023). The association between *jio* and negative sentiment aligns with its frequent use in complaint-driven discourse, further reinforcing its pragmatic function in expressing exclusion or missed opportunities. However, the increasing variability in *jio* use at the more positive end of the sentiment scale suggests that speakers are not entirely constrained by sentiment alone, indicating that affective stance sets probabilistic tendencies rather than categorical limits, and leaving room for context-dependent variation.

Finally, the effect of intensity, while statistically significant, appears more subtle after an examination of median values. The trend suggests that as emotional intensity increases, speakers are more likely to use *jio* (Figure 11d), though the absolute-value effect (media $n = 0.25$) is weaker compared to negation (media $n = -1.31$) and sentiment (media $n = -3.36$). This suggests that speakers may prioritise conveying their attitudes or evaluations over merely reflecting the intensity of their emotions in their linguistic choices. Therefore, the current finding that affective stance exerts a stronger influence than intensity is consistent with previous research in variationist sociolinguistics. Another possible explanation is that intensity interacts with other factors, such as discourse framing or interlocutor expectations, which were not directly tested in the current model. This suggests avenues for future research that could explore whether *jio* is particularly sensitive to pragmatic or interactional constraints beyond those captured here.

Collectively, these findings reinforce the idea that linguistic variation is shaped by structural constraints, with *jio* serving not merely as an interchangeable alternative to *invite* but as a variant whose use is systematically conditioned by negation, sentiment/affect, and to a lesser extent, intensity. While some results, such as the role of negation, were expected based on previous research, the minimal influence of affix presence and the relatively weak effect of intensity suggest that additional factors, such as discourse context and pragmatic function or even socio-stylistic factors, may be at play in shaping *jio* variation.

4.4.4 Socio-Stylistic Constraints on (*jio*)

Figure 12 presents the effects of key socio-demographic factors – age, gender, race, and year – on the variation between *jio* and *invite*. The results for age (Figure 12a) reveal a gradual decline in *jio* use as speaker age increases. Younger speakers exhibit the highest rates of *jio* usage, while older speakers show a clear shift towards *invite*. This pattern suggests that *jio* is more robustly maintained among younger individuals but becomes less frequent with age, aligning with research on age-based variation, where younger speakers tend to favour innovative or localised forms while older speakers conform more closely to dominant linguistic norms (Meyerhoff 2018).

From a sociolinguistic perspective, this age pattern can be interpreted in two theoretically distinct ways, depending on how community norms are assumed to behave over time. Under an age-grading account, individuals change their linguistic behaviour across the life course, but community-level norms remain stable: each new generation would be expected to use *jio* at similarly high rates

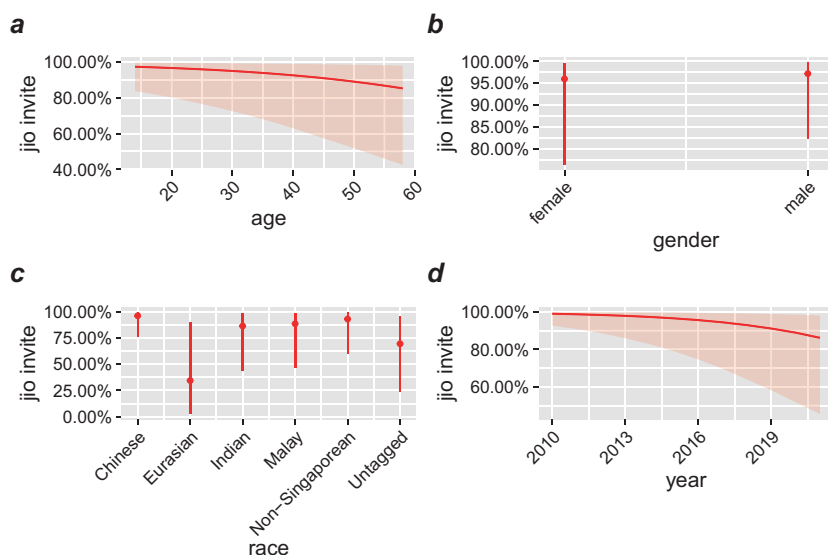


Figure 12 Marginal effects of socio-stylistic factors on *jio* usage. The y-axis shows the model-estimated probability that *jio* is used rather than *invite*, with 100 per cent indicating a categorical preference for *jio*. Panels display posterior predicted probabilities as a function of (a) age, (b) gender, (c) race (simplified categories), and (d) year of production, holding all other predictors constant. Lines and points represent predicted values, while error bars and shaded bands indicate 95 per cent credible intervals.

when young and to reduce usage as they age. Under a lifespan effect account, by contrast, individuals shift their usage in response to changing community norms, such that both individuals and the community distribution change over time. In this scenario, younger cohorts could use *jio* less than earlier cohorts did at the same age, signalling community-level change rather than stable age stratification. The present data cannot adjudicate conclusively between these interpretations, as doing so would require panel or real-time evidence, but the results nonetheless show that age is a strong and systematic predictor of *jio* variation, with important implications for its long-term vitality in speakers' repertoires.

If we examine the change in real time instead of apparent time (Figure 12d), we also find that *jio* usage shows a steady decline over time. While *jio* was more evenly distributed with *invite* in the early 2010s, its use has steadily decreased, particularly after 2017. This downwards trajectory suggests an overall shift towards *invite*, raising the possibility that *jio* is undergoing recession in favour of its English-derived alternative. This trend is further corroborated in Figure 13, which provides a year-by-year breakdown of *jio* and *invite* use. In earlier years, *jio* and *invite* were used at comparable rates, but by 2021, *jio* had nearly disappeared from the dataset. The accelerated decline of *jio* in recent years suggests that while it may have once been a stable variant, it is now becoming increasingly restricted in use. This pattern is in line with broader studies of lexical shift in contact varieties, where localised forms may recede due to ongoing globalisation and the increasing dominance of English.

Turning to gender (Figure 12b), the results show no meaningful difference in *jio* use between male and female speakers, as indicated by overlapping confidence intervals. This finding contrasts with much of the variationist literature, which frequently identifies gender-based differentiation, particularly when a variant carries strong social or indexical associations (Eckert 1989; Gonzales et al. 2026). The absence of a significant gender effect suggests that *jio* is not strongly linked to masculinity or femininity and may function as

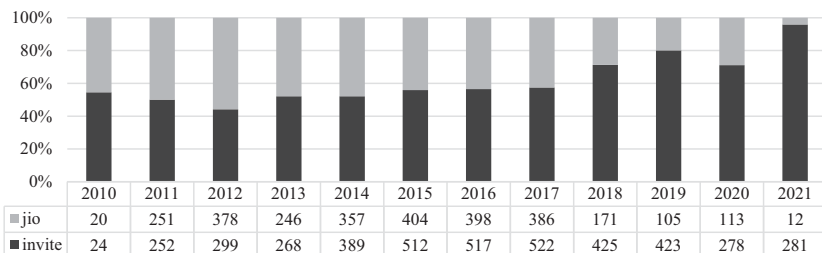


Figure 13 Change in (*jio*) over time

a broadly available linguistic resource for expressing other identities, which is consistent with research on gender and language contact (Meyerhoff & Birchfield 2019).

Ethnolinguistic background (Figure 12c) emerges as a more salient predictor, with *jio* use being highest among Chinese speakers and significantly lower among other ethnic groups. While Eurasian, Indian, and Malay speakers display a more balanced distribution between *jio* and *invite*, non-Singaporeans overwhelmingly favour *invite*. Although the presence of *jio* among non-Chinese speakers suggests some degree of diffusion, its lower frequency among these groups reinforces its continued association with Chinese linguistic heritage.

In addition to socio-demographic constraints, stylistic factors also play a significant role in shaping *jio* variation. Figure 14 examines the influence of style, revealing distinct patterns across the informational–interpersonal dimension, which is analytically independent from the formal–informal dimension. To clarify this distinction, two stylistic dimensions, formal–informal and informational–interpersonal, were derived separately based on the primary linguistic parameters identified by Grafmiller et al. (2018) for English, allowing these dimensions to vary independently rather than being treated as opposites.

In highly informational styles (Figure 14a), *jio* is used at higher rates, with usage increasing as the level of informational content rises. Importantly, this is not a contrast between formal and informal styles: texts can be highly informational while also being formal, and vice versa, as established in multidimensional models of register variation. This suggests that *jio* is preferred in more communicative settings, where speakers engage in high-information exchanges that prioritise efficient transmission of content rather than interpersonal or presentational concerns. This pattern aligns with Bell's (1984) audience design

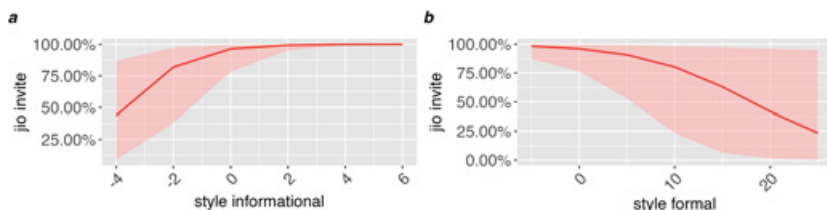


Figure 14 Marginal effects of sociolinguistic style on *jio* usage. The y-axis shows the model-estimated probability that *jio* is used rather than *invite*, with 100 per cent indicating a categorical preference for *jio*. Panels display posterior predicted probabilities across levels of (a) informational style and (b) formal style, holding all other predictors constant. Solid lines represent predicted values, and shaded bands indicate 95 per cent credible intervals.

framework, which posits that speakers modulate linguistic choice in response to communicative context.

By contrast, formal discourse (Figure 14b) exhibits the opposite trend, with *jio* usage declining as formality increases. At the highest levels of formality, *invite* is overwhelmingly preferred, suggesting that it is perceived as the more standard or socially neutral variant. This result is expected given long-standing findings that formal speech styles tend to favour standard or English-derived forms over localised variants (Labov 1972). The association between *jio* and informality suggests that it may be socially marked in a way that *invite* is not, further reinforcing the idea that *jio* is undergoing stylistic restriction.

4.5 Summary

The use of the innovative *jio* in CSE, in contrast with its counterpart *invite*, shows variation along structural and socio-stylistic axes. Here, ‘innovative’ refers to the relatively recent emergence of *jio* in the CSE repertoire, instead of its current direction of change. Model-based analyses qualify the apparent association between negation and *jio*: while raw distributions suggest a higher frequency of *jio* in negated environments, likely reflecting the early entrenchment of the borrowed form *bojio*, the marginal effects reveal that non-negated contexts in fact support a higher and more stable predicted likelihood of *jio* usage. Negated contexts, by contrast, show greater variability because, unlike *bojio*, these forms are not stored as fixed expressions but are constructed through ordinary negation, making their use more flexible and less predictable.

The sentiment and intensity measures are consistent with explicatory models of language contact, in which borrowed forms initially circulate in affectively marked niches before being redistributed. Apparent-time patterns showing higher *jio* use among younger speakers support the interpretation of *jio* as a relatively recent addition to the CSE vocabulary. At the same time, the real-time decline documented earlier suggests that this innovation seems reminiscent of an undergoing process of dialect focussing (Trudgill 1974b): instead of continuing to spread, *jio* is becoming increasingly restricted to particular ethnolinguistic groups, pragmatic functions, and interactional contexts. In this view, decline does not contradict innovation but reflects the narrowing and stabilisation of a previously diffuse variant as community norms consolidate.

The continued ethnolinguistic bias towards *jio* use among Chinese speakers, together with its decrease over real time, therefore suggests focussing rather than attrition, with *jio* persisting as a socially meaningful and locally indexed

resource even as its overall frequency drops. In the [next section](#), these issues are explored in the context of passive marking in CSE, for which several competing options are available.

5 Variation and Change in CSE Passive Marking

5.1 Introduction to Passives in CSE

This section focusses on passive constructions in CSE, which have received some scholarly attention in the literature (e.g., Bao & Wee 1999; Fong 2004; Lee, Ping, & Nomoto 2009; Bao 2011; Kim & Sato 2013). Bao and Wee (1999) and Fong (2004) argue that, in addition to two ‘standard’ English-derived passives containing *be* (12) and *get* (13), CSE speakers have innovated two passive(-like) constructions due to influence from locally prominent substrate languages: the *kena* passive (14), due to influence from Malay; and the *give* passive (15), due to influence from Sinitic languages such as Cantonese, Hokkien, and Mandarin.

- (12) *Be* passive
John **was** scolded by his boss.
(Fong 2004: 97)
- (13) *Get* passive
How many **got** arrested?
(Fong 2004: 97)
- (14) *Kena* passive
*The thief **kena** catch by the police*
‘The thief was caught by the police.’
(Bao & Wee 1999: 1)
- (15) *Give* passive
*John **give** his boss scold*
‘John was scolded by his boss.’
(Bao & Wee 1999: 5)

Bao and Wee (1999) note that the substrate-derived passive constructions (14)–(15) carry an adversative reading absent from standardised English passives (12)–(13). Thus, *kena* and *give* passives fulfil a distinct function from *be*- and *get*-passives in CSE and can be seen as competing constructions within its emerging grammar. They further propose that, in a language contact situation, between two competing variants of the same construction, speakers will favour the variant that ‘most closely resembles in structure’ the standard

form (Bao & Wee 1999: 9). Applying this to the context of Singapore English, this predicts that the *kena* passive will win out over the *give* passive, as the *kena* passive is more similar in word order to the standard *be-* and *get-* passives, sharing the structure PATIENT PASSIVE.MARKER *V-en by* AGENT.

In this section, we revisit Bao and Wee's (1999) prediction – proposed over a quarter of a century ago – by examining passive constructions in CoSEM. Our findings show that the *kena* passive is robustly attested in CoSEM, whereas the *give* passive is entirely absent. The robustness of the *kena* passive alongside the absence of the *give* passive in CoSEM lends support to Bao and Wee's prediction, suggesting that the inventory of passive constructions in CSE has stabilised. However, we also document the emergence of a novel passive construction that challenges this notion. This construction, which appears in a small set of CoSEM examples, involves the lexical item *tio*, derived from a passive marker in the Southern Min varieties of Hokkien and Teochew that we will transcribe as *tioh* (/tioʔ/ in Hokkien and Teochew, also /tieʔ/ in certain Teochew varieties).⁷ In the rest of this section, we will use *tioh* to refer to the Southern Min morpheme and *tio* to its CSE counterpart.

The emergence of the *tio* passive indicates that the passive system in CSE remains dynamic. To account for its emergence, we point to structural similarities between standardised English passives, *kena* passives, and the Southern Min *tioh* passive construction. An analysis of the uses of Southern Min *tioh* further reveals semantic similarities between English *get* and Malay *kena*. These facts lead us to suggest that the adoption of a variant might not only be dependent on structural similarities (as Bao and Wee suggest), but also on semantic similarities.

Since *kena* and *tio* but not *give* passives are attested in CoSEM, Section 5.2 begins with an overview of the use of *kena* and *tio* passives in CSE, comparing the CSE *kena* passive to how *kena* is used in Malay and the CSE *tio* passive to how its counterpart *tioh* is used in Singapore Southern Min varieties. Section 5.3 outlines the methods of our study, detailing how passive constructions were identified and extracted from CoSEM, how the sample was constructed, and how the dataset was coded for sentiment analysis. Section 5.4 presents and discusses the results of our study. Finally, Section 5.5 concludes by considering

⁷ We will transcribe Hokkien and Teochew data in this section using the International Phonetic Alphabet, instead of romanisation schemes like Peh-oe-ji for Hokkien and Peng'im for Teochew. This is because in Singapore, at least, the two varieties overlap substantially in phonology, vocabulary, and syntax. However, the romanisation schemes differ in how they represent key phonological features like aspiration and voicing and run the risk of confusing readers who are unfamiliar with these schemes.

factors that may have contributed to the emergence of the *tio* passive in CSE, and the relatively low frequency of *kena* and *tio* passives in our data.

5.2 Overview of *Kena* and *Tio* Passives

5.2.1 *Kena* Passives in CSE

The use of the Malay-derived *kena* passive in CSE was first reported by Bao and Wee (1999), who also asserted that it was more frequently used than the *give* passive. Similar to the standard *be* and *get* passives, in the CSE *kena* passive construction, the theme or patient Noun Phrase (NP) is promoted to subject, while the agent NP is optionally expressed as an oblique argument using a *by*-phrase. The *kena* passive in both Malay and CSE also usually involves an adversative reading. Examples (16) and (17) are taken from Bao and Wee (1999), demonstrating this documented use of the *kena* passive in CSE.

- (16) *The thief **kena** caught by the police*
 ‘The thief was caught by the police.’
 (Bao & Wee 1999: 2)

- (17) *The durian **kena** eaten already*
 ‘The durian got eaten already.’
 (Bao & Wee 1999: 2)

Examples (18), (19), (20), and (21) are taken from CoSEM, and they show how the *kena* passive is still attested in present-day CSE.

- (18) *Cried bc **kena** scolded by ms <name>*
 ‘He cried because he got scolded by Ms <name>.’
 <COSEM:21DC45-1514-19SGCHFST-2018>
- (19) *More and more ppl **kena** quarantined*
 ‘More and more people are being quarantined.’
 <COSEM:21DC09-9133-42SGCHFGH-2020>
- (20) *I **kena** cut from the team*
 ‘I was cut from the team.’
 <COSEM:19CF10-5176-23SGCHM0A-2018>
- (21) *Tmr my dinner **kena** cancelled so I can MAHJONG*
 ‘My dinner tomorrow got cancelled, so I can play mahjong.’
 <COSEM:18CF42-3081-23SGMAF00-2017>

5.2.2 Kena in Malay

Kena in Malay may function as a passive marker,⁸ as seen in (22)–(24). In the Malay *kena* passive construction, the agent can be omitted as shown in (22) and (23), and may also be expressed as an oblique argument using *oleh* as shown in (24).

- (22) *Aminah kena tampar*
 Aminah PASS slap
 ‘Aminah got slapped.’
 (Nomoto & Kartini 2012: 364)
- (23) *Dompét aku kena curi semalam*
 purse 1SG.POSS PASS steal yesterday
 ‘My purse got stolen yesterday.’
 (Nomoto & Kartini 2012: 364)
- (24) *Pencuri itu kena tangkap oleh polis semalam*
 thief that PASS catch by police yesterday
 ‘That thief was caught by the police yesterday.’
 (Nomoto & Kartini 2012: 367)

Note that *kena* possesses other functions in Malay. First, as shown in (25) and (26), *kena* can be used to express a deontic reading similar to ‘have to’ in English (see Nomoto and Kartini 2012 for more discussion on potential polysemy between these functions of *kena* in Malay).

- (25) *Nak kena bongkar kat rumah*
 will DEO ransack at house
 ‘(I) will have to ransack the house.’
 <COSEM:18MF03–19411–23SGMAF00–2017>
- (26) *saya kena bikin kuih*
 1SG DEO make snack
 ‘I have to make snacks.’
 <COSEM:18MF03–14392–23SGMAF00–2017>

Kena can also be used as a verb to mean ‘to strike or be struck by’ or ‘to receive or be subjected to’. As a verb, *kena* is followed by an NP object, as seen in (27) and (28).

- (27) *kereta-nya kena tiang eletrik*
 car-POSS hit pole electric
 ‘His car hit the electric pole.’
 (Dewan Bahasa dan Pustaka 2010: 735)

⁸ It should also be noted that the *kena* passive is not the only passive construction possible in Malay. Among other options, Malay also uses the prefix/circumfix *di-* ... *(-kan)* to passivise verbs, although this will not be covered as it is not important for the discussion of the *kena* passive in this section.

- (28) *kena penyakit demam kepialu*
 receive illness fever typhoid
 ‘(He) contracted typhoid fever.’
 (Dewan Bahasa dan Pustaka 2010: 735)

Lastly, *kena* also serves as an adjective that roughly means ‘suitable, right’, as seen in (29) and (30).

- (29) *Warna baju Aminah kena dengan warna selendang-nya*
 colour clothes Aminah suitable with colour shawl-POSS
 ‘The colour of Aminah’s clothes matches that of her shawl.’
 (Nomoto & Kartini 2012: 366)
- (30) *Dia menjadi berang kalau ada sesuatu yang tidak kena*
 3SG become furious if EXIST something that not right
di hati-nya
 at heart-POSS
 ‘She becomes furious if there is something that is not right in her heart.’
 (Nomoto & Kartini 2012: 366)

Of these functions, this section will primarily focus on the use of *kena* as an adversative passive marker (Bao & Wee 1999), which is well-attested in the CoSEM data set. However, it is pertinent to note that *kena* as a verb – meaning ‘to be subjected to’ – can also be found in CSE, as seen in the CoSEM examples (31) and (32). In these examples, CSE *kena* co-occurs with an NP object. This is in contrast with the passive marker *kena*, in which *kena* must precede the main verb.

- (31) *wah I kena sore throat*
 ‘Damn, I got a sore throat.’
 <COSEM:19SF01-47194-20SGSIFNA-2018>
- (32) *I hope u kana car accident*
 ‘I hope you get caught in a car accident.’
 <COSEM:18CF52-15075-23SGCHM00-2017>

It is also worth noting that the ‘have to’ or deontic function of *kena* can be found, albeit to a smaller extent, in some CSE utterances in CoSEM, as shown in (33) and (34).

- (33) *Yeah kena wait for 2nd round*
 ‘Yeah (you) have to wait for the second round.’
 <COSEM:18MF03-11905-24SGMAF00-2017>
- (34) *all kena scan IC*
 ‘(They) all had to scan their identity cards.’
 <COSEM:19MM02-206-42SGMAM00-2019>

5.2.3 The *Tio* Passive in CSE

The *tio* passive construction was not mentioned in Bao and Wee's (1999) study, and appears to be a relatively new feature of CSE. As such, it has not, to our knowledge, been studied in contemporary CSE research. Interestingly, the online dictionary Wiktionary does list *tio* as a lexical item of CSE, describing one of its functions as that of a passive marker (Wiktionary 2025). In this section, we present a more thorough description of the *tio* passive in CSE based on our findings from CoSEM.

In the corpus, *tio* is used similarly to *be* and *get* passives, as shown in examples (35) and (36). On the surface, *tio* is used in place of the standardised English passive markers, *be* and *get*.

- (35) *Can not say later i tio scolded again hahaha*
 'I can't say, otherwise I'll get scolded again.'
 <COSEM:17CF24-5711-26SGCHM00-2015>
- (36) *Wah today station games tio squeezed by guys*
 'Today at the station games, I got squeezed by guys.'
 <COSEM:17CF34-61588-21SGCHF00-2013>

Notably, example (36) shows the *tio* passive co-occurring with the optional *by*-phrase, *by guys*, which contains the agent argument. This is interesting since *tioh* passive constructions in Singapore Hokkien and Teochew, which we argue in what follows is the source of the *tio* passive construction in CSE, may not contain an oblique argument identifying the agent.

5.2.4 *Tioh* in Singapore Southern Min

There are three major varieties of Southern Min spoken in Singapore: Hokkien, Teochew, and Hainanese. Hokkien and Teochew can be reasonably identified as the source of CSE *tio*, as both varieties contain the phonologically similar passive marker /tioʔ/ (or /tieʔ/ in certain Teochew varieties). In contrast, Hainanese does not contain a phonologically similar passive marker,⁹ which suggests that Hainanese is unlikely to be the source of CSE *tio*. Demographically, Hokkien and Teochew speakers also greatly outnumber Hainanese speakers in Singapore (Department of Statistics Singapore 2020), which further supports the idea that Hainanese has had less of an influence on CSE than Hokkien and Teochew. For these reasons, we will base the following discussion on examples drawn from Singapore Hokkien and Singapore Teochew. As Hokkien and Teochew are

⁹ This is confirmed by our two Singapore Hainanese consultants, Yong Chuen Shin and Cheam Shu Kem.

closely related varieties, we have tried to pick examples that are intelligible in both varieties. We will henceforth refer to these two varieties as ‘Southern Min’ as a shorthand. The examples discussed here were obtained via consultation with two native speakers of Singapore Hokkien and two native speakers of Singapore Teochew.

In a *tioh* passive construction such as (37), the passive marker *tioh* precedes a transitive verb like ‘scold’, and the theme argument appears in the subject position, which precedes *tioh*. However, unlike the *kena* passive construction in Malay, the *tio* passive construction in CSE *tio*, or the *bei* passive construction in Mandarin, *tioh* passives do not allow the expression of an oblique agent. To express the agent in Singapore Southern Min, another passive construction using a marker derived from the verb ‘give’ must be used instead (cf. Bao and Wee’s (1999) discussion of Cantonese and Hokkien ‘give’ passive constructions).

- (37) *ua tio? me / p^ha?*
 1SG *tioh* scold beat
 ‘I got scolded/beaten.’

- (38) **ua tio? i me / p^ha?*
 1SG *tioh* 3SG scold beat
 Intended: ‘I got scolded/beaten by him/her.’

- (39) *ua puŋ / k^ho? i me / p^ha?*
 1SG give 3SG scold beat
 ‘I was scolded/beaten by him/her.’ (Teochew)

- (40) *ua ho i me / p^ha?*
 1SG give 3SG scold beat
 ‘I was scolded/beaten by him/her.’ (Hokkien)

For completeness, it is worth pointing out that there are dialectal differences in whether and how *tioh* is used as a passive marker in Southern Min more generally. Hing (2019), citing Zhou (2006: 683), reports that *tioh* passives are absent in mainland China varieties. However, *tioh* passives appear to exist in Taiwanese Southern Min (e.g., Lien 2001: 191–192; also cited in Ansaldo 2004); Hing (2019, 2023) also reports a *tioh* passive in Penang Hokkien. In both Taiwanese Min and Penang Hokkien, the agent may appear between *tioh* and the verb, unlike the case in Singapore Southern Min. In other words, the equivalent of example (38) is acceptable in these varieties.

In addition, like Malay *kena*, Southern Min *tioh* has other uses: as a deontic modal meaning ‘must’, a lexical verb that might be translated as ‘obtain, get, win’, an adjective meaning ‘correct’, and as an aspectual marker suffixing to

verbs, among others (see Lien 2001; Hing 2019 and references therein). For scope reasons, we will review in detail only the deontic and lexical verb uses, which, as we will see, might explain the emergence of the *tio* passive in CSE. In the deontic modal use, Singapore Southern Min *tioh* is often paired with another modal marker *ai* (in Hokkien, as shown in (41)) (Lien 2001; Chappell 2019; Hing 2019: 152) or *āĩ* (Teochew), which variously has desire ('want, like'), deontic ('must'), and/or future uses (we tentatively gloss this morpheme as deontic here). However, as (42) shows, *ai* / *āĩ* is not necessary for the modal interpretation.

- (41) *ua tio? ai k^hun.*
 1SG DEO DEO sleep
 'I must sleep.' (Hokkien)

- (42) *tio?!*
 DEO
 '[S/he] must!' (in response to a question like *i tio? ai k^hun?* 'Must s/he sleep?')

Second, *tioh* can be used as a lexical verb, with a meaning along the lines of 'obtain', 'get', and 'win'. This is a generally well-documented usage of *tioh* in Southern Min, even outside of Singapore (e.g., Lien 2001; Ansaldo 2004; Hing 2019). In Singapore Southern Min, the verbal use of *tioh* is prominently associated with winning games of chance (43) and contracting diseases (44). For thoroughness, it should be pointed out that the same verbal use can also be found in CoSEM.

- (43) *tio? bepio*
 win lottery
 'to win [a prize] in the lottery'
- (44) *i bo tio? COVID.*
 3SG NEG.PERF get COVID
 'S/he did not get (catch) COVID.'

5.2.5 Summary

This section has compared the use of *kena* in CSE to the use of *kena* in Malay, and the use of *tio* in CSE to the use of *tioh* in Singapore Southern Min (Hokkien and Teochew). While Malay *kena* can be used as a passive marker, a marker of deontic modality, and a lexical item to mean 'to match' or 'suitable', CSE *kena* may only be used as a passive marker. While *tioh* in Singapore Southern Min can be used as a passive marker, a marker of deontic modality, and a lexical verb

meaning ‘to obtain, get, win’, only the passive marker and lexical verb uses are found with CSE *tio*.

5.3 Passives in CoSEM

Next, we investigated the distribution of CSE passive constructions in CoSEM. A central goal of this investigation is to understand how *kena* and *tio* passives are used relative to other passive constructions in CSE, namely, *be*, *get*, and *give* passives. In particular, we sought to test Bao and Wee’s prediction that *kena* passives will ‘outcompete’ *give* passives in CSE.

Beyond this, we wanted to use the CoSEM data to address two other questions. The first is a high-level demographic question: who uses which passives, and how frequently? One might expect that the *kena* passive might have become more widespread among younger speakers if Bao and Wee’s predictions are correct. On the other hand, the newer *tio* passive, being of Chinese origin, might be more popular with younger Chinese speakers. The second concerns the sentiment-bearing properties of CSE passive constructions – specifically, the extent to which they occur with predicates expressing positive, neutral, or negative sentiment. This latter question addresses the popular claim that *kena* predominantly conveys an adversative meaning. We have also encountered claims online (e.g., Wiktionary 2025) that *tio* can be used positively, unlike *kena*. Since this claim has not been empirically verified and no prior description of the *tio* passive exists, we take this opportunity to evaluate it as well.

5.3.1 Identifying Passive Constructions in the Corpus

We first wrote a Python script to extract all lines containing *kena*, *tio*, *give*, *be*, and *get*, and their various inflected forms (e.g., *was* and *were*) and alternative orthographic forms (e.g., *kana* for *kena*). Each message was then slightly preprocessed to remove emojis and standardise characters (e.g., quotation marks), which are known to affect the accuracy of statistical parsing. The preprocessed messages were then parsed with a version of the Stanford Parser (Klein & Manning 2003 and subsequent work) customised for CSE – this was done by training the Stanford Parser on CSE parses manually validated by Bao, Huang, and colleagues in another CSE corpus project (see Lin et al. 2023; Huang et al. 2025). The parser parsed and generated phrase structure trees for all preprocessed lines except those that contain forty words or more, as these tend to create memory issues and cause the parser to crash. We note that lines of this length or greater are very rare, making up 0.3 per cent of all lines.

We then wrote another Python script to check each generated syntax tree. This was particularly important for the *be* and *get* tokens, since not all instances of *be* and *get* appeared with passives; *be*, of course, can appear with the present participle (a hypothetical standard English example: *They will be leaving*). The Python script extracted only tokens of *be* and *get* that co-occur with a VP headed by what the parser tagged as a past participle.

Within this initial sample, *be* and *get* were by far the most frequent markers. Their extremely high frequencies made a full annotation of them infeasible. As such, we extracted a 10 per cent random sample of the *be* and *get* parses. A linguistics-trained research assistant checked these samples and identified tokens that were genuine verbal passives. This was necessary because the parser is not perfectly accurate and might have made mistakes with tagging or parsing. A second reason is that the parser is also not always able to differentiate verbal passives from adjectival passives, as in *tired* in *They are/got tired*, which we consider to be adjectives rather than verbs. We excluded tokens that were clearly adjectival, where we can felicitously insert a degree adverb like ‘very’ immediately before the word (e.g., *They are/got very tired*).

The Python script also extracted all tokens of *give* co-occurring with a VP or clause as potential *give* passive tokens. Manual inspection, however, showed these to be parser errors; we did not find any tokens of genuine *give* passives.

Because the set of *kena* and *tio* tokens extracted was relatively small (*kena*: 673 tokens; *tio*: 94 tokens), we did not do any random sampling. However, we note that *kena* and *tio*, not unlike *be*, *get*, and *give*, also have non-passive uses, in which case they do not occur with verbs. As mentioned earlier, CSE *kena* can be a verb, as in *I kena COVID* ‘I got COVID’. To exclude these non-passive uses, the research assistant reviewed the full set of tokens, sorting them based on whether they appeared with verbs, NPs, intransitive, or other frames.

Table 13 lists the count of tokens for *be*, *get*, *kena*, and *tio* passives with unambiguous verb frames in our final sample. With this, we ran a multinomial regression with age, gender, and race as predictors, and choice of passive marker as the dependent variable, setting *be*-passive as the reference ‘standard’ category. For thoroughness, we created a reweighted data set that synthetically increased the *be* and *get* tokens ten times, so that the number of *be* and *get* tokens in this reweighted sample more accurately represents the actual distribution of passive markers. For visualisation purposes, we used the reweighted sample to generate all figures. We also re-ran the regression model on this reweighted sample and obtained qualitatively similar results.

Table 13 Distribution of passive markers in the sample

Passive Marker	Tokens in sample (10% for <i>be</i> and <i>get</i>)	Tokens in reweighted sample
<i>be</i>	415 (47.6%)	4,150 (70.6%)
<i>get</i>	141 (16.2%)	1,410 (24.0%)
<i>kena</i>	282 (32.3%)	282 (4.8%)
<i>tio</i>	34 (3.9%)	34 (0.0%)
Total	872 (100%)	5,876 (100%)

5.3.2 Coding for Sentiment

We coded this dataset for sentiment by referring to only the action or state described by the verb and ignoring all other contextual information. For example, we would consider all tokens of *paid* as positive, even in contexts like *getting paid peanuts* or *didn't get paid*, which most might consider as describing negative outcomes (note that these are hypothetical examples to illustrate the point). We do so because we judge that *kena* can be used for a verb with negative sentiment, even when the broader context has a positive sentiment, as illustrated by *disqualified* in the hypothetical sentence *My favourite team didn't kena disqualified* 'My favourite team didn't get disqualified'.

All tokens identified were coded as 'negative', 'positive', or 'neutral' by the same linguistics-trained student research assistant and one of the authors (either Mohamed Hafiz or Nick Huang); some examples for the *be*-passive are shown in (45). We chose to manually annotate for sentiment here, instead of using software (cf. Section 3 on *jio*). This is because we were interested in the sentiment expressed through the verb itself, and a number of verbs in our sample are of Malay or Chinese origin (e.g., *tekan* 'to treat poorly, to bully', from Malay; *buah* 'to scrape', from Southern Min) or are English words that have acquired novel uses (e.g., *arrow* 'to nominate [someone] for an undesirable task'). To ensure accurate annotation, we decided against using software, which is usually targeted at standard English varieties. In order to keep the process of manual annotation straightforward, we decided on a three-way categorical scheme, instead of a finer-grained one. Results reported in Section 5.4 are based on the research assistant's annotations, but we note that there was a relatively high degree of inter-annotator agreement (Cohen's kappa of .77).

- (45) (a) *Pls slowly do carefully and check if we miss out anything or if it can be improved.* (Positive because of *improve*)
 ‘Please do [the survey] slowly and carefully and check if we missed out anything or if it can be improved.’
 <COSEM:17CM05–8172–23SGCHM00–2016>
- (b) *i am being abused now* (Negative because of *abuse*)
 <COSEM:17CF12–65931–21SGCHM00–2016>
- (c) *I hope the stuff im writing can be absorbed in* (Neutral because of *absorb*)
 <COSEM:17CF34–54907–21SGCHF00–2015>

5.4 Results and Discussion

5.4.1 Relative Frequencies of Passive Constructions

As shown in Table 13, no *give* passives were found in CoSEM, suggesting that the construction might have fallen out of use for younger CSE speakers. This supports Bao and Wee’s (1999) prediction. It is also not unexpected, given Bao and Wee’s original observation that the *give* passive was not as robust compared to the *kena* passive at the time of their study. *Tio* passives are the least frequent in CoSEM, which supports our intuition that the *tio* passive has only recently emerged in CSE.

Table 14 presents the estimates for predictors influencing the likelihood of using a particular passive construction relative to others, with the *be*-passive serving as the reference ‘standard’ category. We discuss each predictor in turn.

Age. Table 14 indicates that there is strong evidence that as age increases, the likelihood of using *get*, *kena*, or *tio* decreases. First, this pattern aligns with trends reported for American English, where the *get* passive has been increasing at the expense of the *be* passive (Sneller & Fisher 2015). This parallel may provide support for the influence of American English on the English of younger Singaporeans, although the data cannot establish a causal link. Second, the results indicate that the substrate-influenced passives are relatively more popular among younger speakers than older ones. For the *kena* passive, the finding that younger speakers continue to use it aligns with Bao and Wee’s (1999) prediction that the *kena* passive will remain robust in CSE. The *tio* passive shows a similar trend, which is expected given its relatively recent emergence, likely introduced by younger speakers.

The distribution of passive constructions across age groups, as shown in Figure 15, also reveals some generational patterns. The *be* passive is the most frequently used across all age groups, with older speakers notably displaying the highest usage, followed by younger speakers, and next, middle-aged speakers. The *get* passive shows a gradual decline in usage with age. The *kena* passive is most frequent among middle-aged speakers but is considerably lower among younger and older speakers. Finally, the *tio* passive appears only

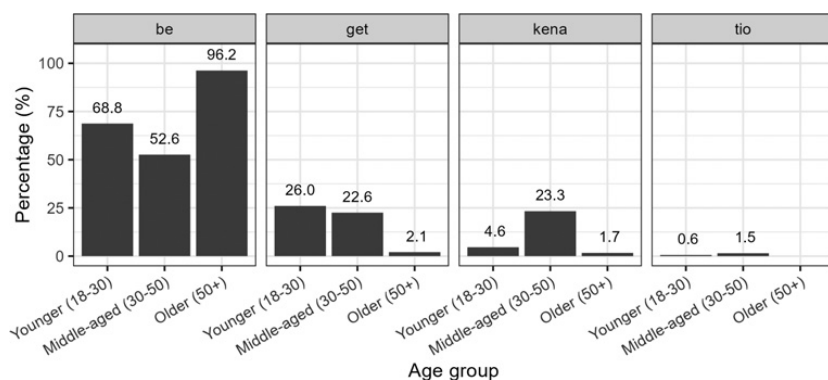


Figure 15 Percentage of a given passive marker relative to all passives used by an age group

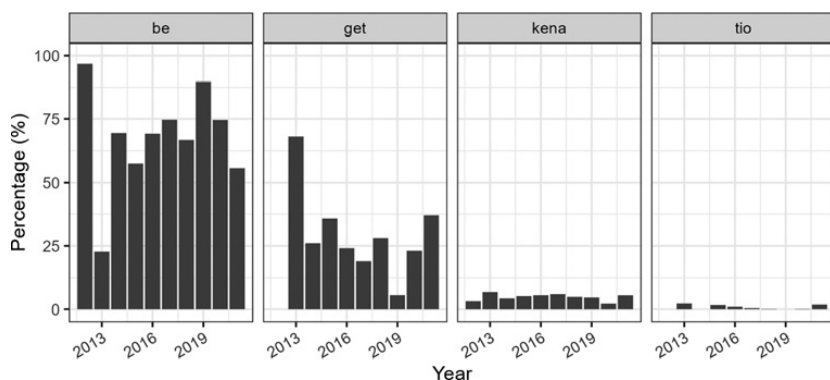


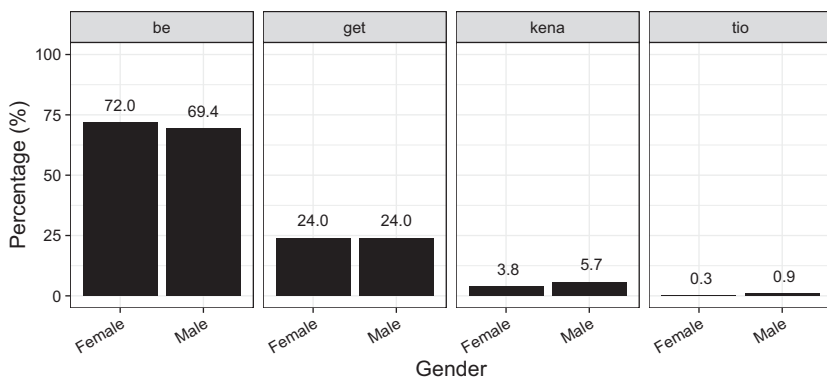
Figure 16 Percentage of a given passive marker relative to all passives in messages from a particular year

marginally in the data, with extremely low frequencies found among younger and middle-aged speakers. The patterns suggest that the use of standardised English *be* and *get* passives dominate the use of the substrate-derived *kena* and *tio* passives across different generations of CSE speakers.

Although our Bayesian results indicate that *kena* and *tio* have declined over time relative to *be*, it is not the case that *be* has been growing steadily at the expense of *kena* or *tio*. Rather, as Figure 16 shows, there is much fluctuation in the use of *be*. In contrast, *kena* is used at a low, but comparatively stable, rate. We take this relative stability to support Bao and Wee's (1999) prediction that the *kena* passive will continue to be used in CSE, which is especially notable given that their hypothesis was proposed over two decades ago. As for *tio*, Figure 16 shows that it is attested as early as 2013, but occurs even less

Table 14 Estimates of selected predictors from Bayesian models

Predictor (reference levels in bold)	<i>Get</i>		<i>Kena</i>		<i>Tio</i>	
	Median	<i>pd</i>	Median	<i>pd</i>	Median	<i>pd</i>
Intercept	360.47	99%	508.45	99%	801.28	99%
Age	−0.11	100%	−0.03	87%	−0.21	100%
Gender (male vs. female)	0.1	53%	−0.58	69%	−1.96	82%
Race (Indian vs. Other/Eurasian)	1.56	74%	−0.08	51%	0.47	57%
Race (Malay vs. Indian/Eurasian/Other)	−1.14	70%	−0.79	67%	0.68	60%
Race (Chinese vs. non-Chinese)	−1.24	77%	0.92	73%	−1.44	71%
Year	−0.18	99%	−0.25	99%	−0.4	99%

**Figure 17** Percentage of a given passive marker relative to all passives used by a particular gender

frequently. While *tio* remains far less established than *kena*, our current data do not yet allow us to determine whether its usage will increase or decline.

Gender. Table 14 provides limited evidence for an effect of gender on the likelihood to use *kena* or *tio* ($pd < 0.9$). This is unsurprising, as these passives are relatively ‘neutral’ linguistic features compared to other features in CSE. For example, the SFP *sia* in CSE has a vulgar origin and may thus show gender-based distributional differences (Hiramoto et al. 2022; Hafiz et al. 2025: 202). The substrate-influenced passives *kena* and *tio* lack such connotations, which may explain their relatively balanced usage across genders, as shown in Figure 17.

Table 15 Distribution of sentiment, by passive marker

Marker	Negative	Neutral	Positive	Total
<i>be</i>	46 (11.1%)	354 (85.3%)	15 (3.6%)	415 (100%)
<i>get</i>	50 (35.5%)	85 (60.3%)	6 (4.3%)	141 (100%)
<i>kena</i>	170 (60.3%)	111 (39.4%)	1 (0.0%)	282 (100%)
<i>tio</i>	17 (50.0%)	17 (50.0%)	0 (0.0%)	34 (100%)

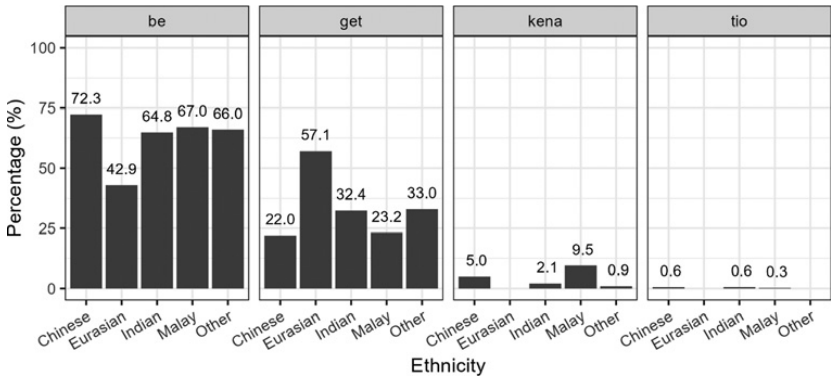


Figure 18 Percentage of a given passive marker relative to all passives used by members of an ethnic group

Race. Similarly, Table 14 shows that there is limited evidence of a racial effect on the use of *kena* ($pd < 0.9$). Figure 18 shows that Malay speakers use *kena* more frequently than the other ethnic groups, which is unsurprising given *kena*’s Malay origin. Notably, Chinese and Indian speakers also use *kena*, indicating that it has become a shared feature across ethnic groups. For *tio*, the data is too sparse to draw firm conclusions, though it is unsurprising that Chinese participants use it most frequently. Interestingly, Indian participants use *tio* about as often as Chinese participants, while Malay participants use it least – perhaps because they are more familiar with *kena*, the Malay-derived passive in CSE.

5.4.2 Distribution of Sentiment

Table 15 shows the distribution of sentiment for each passive marker in our data set; this is represented visually in Figure 19. These patterns let us answer, given a passive marker, what kind of verb (bearing positive, neutral, or negative sentiment) is most strongly associated with it. Within *kena* passives, the most

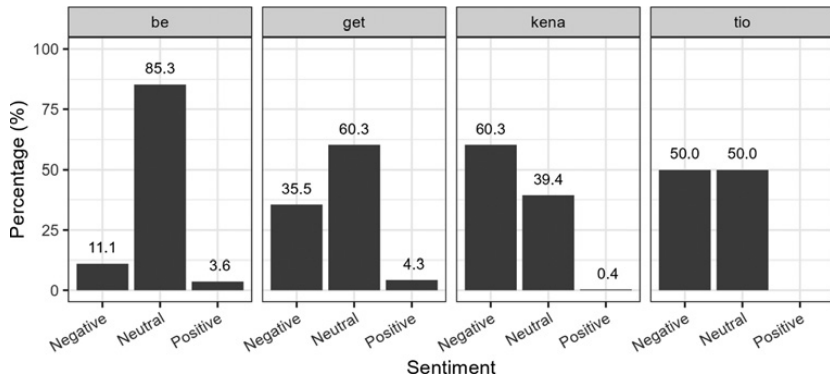


Figure 19 Distribution of sentiment, by passive marker.

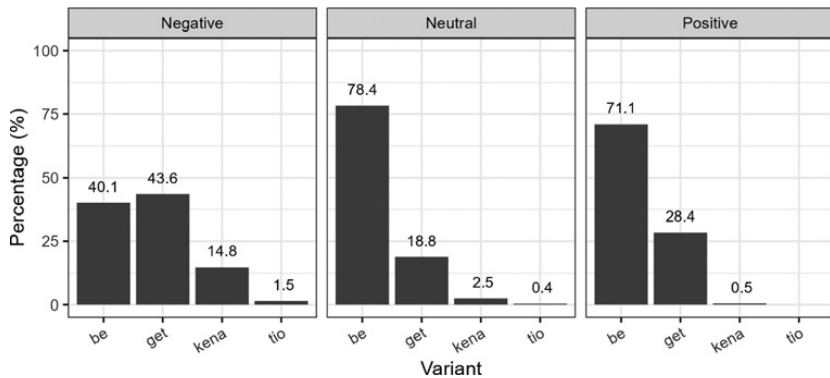


Figure 20 Distribution of passive markers used to express a particular sentiment

common verbs have a negative sentiment, compared with *be* and *get* passives (Fisher's exact test for *be*: $p < 0.01$, for *get*: $p < 0.01$). *Tio* also has a stronger negative bias compared with *be* (Fisher's exact test $p < 0.01$), but no more so than with *get* ($p = 0.22$). We found no evidence of a clear difference between *kena* and *tio* ($p = 0.35$).

We may also flip the question around: given a verb with a positive (or neutral, or negative) sentiment, which passive marker is likely to be used? To answer this question, we analysed the reweighted data set, which better represents the high frequency of *be* and *get* passives (Table 16 and Figure 20). Given the general preference for *be* and *get* passives (Table 13), it is unsurprising that at the sentiment level, *be* and *get* are also the most common options, with *kena* and *tio* trailing in distant third and fourth positions respectively (for each sentiment, the proportions of *be* and *get* both exceed that of *kena*, all $ps < 0.001$; the same

Table 16 Distribution of passive markers by sentiment

Verb sentiment	<i>Be</i> (reweighted)	<i>Get</i> (reweighted)	<i>Kena</i>	<i>Tio</i>	Total
Positive	150 (3.6%)	60 (4.3%)	1 (0.0%)	0 (0.0%)	211 (3.6%)
Neutral	3,540 (85.3%)	850 (60.3%)	111 (39.4%)	17 (50.0%)	4,518 (76.9%)
Negative	460 (11.1%)	500 (35.5%)	170 (35.5%)	17 (50.0%)	1,147 (19.5%)
Total	4,150 (100%)	1,410 (100%)	282 (100%)	34 (100%)	5,876 (100%)

for *tio* which is even less frequent than *kena*). In other words, although both *kena* and *tio* have a distinctive adversative flavour, they do not have a monopoly over negative or even neutral sentiment verbs. One plausible reason is that in these situations, speakers might still avoid *kena* and *tio* for register and/or sociolinguistic reasons; for instance, speakers might avoid *tio* in conversations with other speakers who (they suspect) do not use it (see Section 5.4.1).

5.5 Summary

We have documented the use of the *tio* passive construction among younger CSE speakers. To conclude, we discuss factors that may have contributed to the emergence of the *tio* passive in CSE. First, we note that Singapore Southern Min (Hokkien and Teochew) *tioh* passives are structurally similar to standard English passives: both constructions share the structure PATIENT PASSIVE.MARKER V and neither allows the overt expression of the agent of the verb before the verb. This similarity recalls Bao and Wee's (1999) hypothesis that speakers will favour variants that are structurally similar to the dominant (standard) form.

However, we propose that non-structural similarities may also have facilitated the emergence of the *tio* passive in CSE. As we noted earlier, Southern Min *tioh* has other functions, including a verbal meaning of 'obtain' or 'win', and also a deontic modal use like English *must* or *have to*. In these respects, Southern Min *tioh* resembles English *get*, which also carries the meaning of 'obtain' and has a deontic modal use (e.g., *I've got to go home*). Therefore, we propose a generalised version of Bao and Wee's hypothesis: in contact situations where there are competing options of realising a given construction (e.g., several passive markers), speakers may favour forms that are not only structurally but also semantically similar to the standard variant. This proposal aligns with and extends Baptista's (2020) observation that morphosyntactic and semantic features are more likely to be incorporated into a contact variety

if they preexist and are shared across some of the input varieties in the linguistic ecology.

A further factor may involve similarities between Southern Min *tioh* passives with CSE/Malay *kena* passives. Both types of passives share a similar structure and an adversative interpretation, and Malay *kena* additionally exhibits deontic modal uses. In an earlier discussion of CSE passives, Ansaldo (2004: 139) suggests that the influence of Malay *kena* on CSE might have been ‘reinforced’ by Southern Min *tioh* passives, which are structurally similar. Here, we suggest that the structural and semantic similarities between Malay *kena* and Southern Min *tioh* may have encouraged speakers to use *tioh* and *kena* interchangeably in CSE, thereby contributing to the emergence of the *tio* passive in CSE (see also discussion by Hing (2019, 2023) on how Malay *kena* might have influenced *tioh* in Penang Hokkien).

Overall, both *kena* and *tio* are used relatively infrequently in our dataset compared to the *be* and *get* passives. We believe that this does not undermine our claim about the emergence of the *tio* passive or the robustness of the *kena* passive in CSE. Rather, there are independent factors that constrain their use. Both *kena* and *tio* are more informal and may thus be avoided in more formal contexts. For *tio*, its low frequency may also reflect its recent emergence and association with specific demographic groups (Section 5.4.1). Moreover, *kena* and *tio* passives tend to convey an adversative reading, which further limits their use to contexts with a negative or neutral sentiment (Section 5.4.2). As a result, there is a relatively small subset of situations when speakers use *kena* and *tio*: in informal contexts when the main verb has a negative or (to a smaller extent) neutral sentiment. Finally, even within these contexts, *kena* and *tio* must compete not only with each other but also with the *be* and *get* passive constructions. These factors collectively drive down the use of *kena* and *tio*, making them seem less productive than *be* and *get* passives.

6 Conclusion

This Element has examined CSE as it continues to evolve among younger generations of speakers across digital domains, drawing on complementary sources of evidence such as corpora and an online survey of how Singaporeans perceive their own language use. Across the sections, we have observed a linguistic ecology shaped by intense, ongoing multilingual contact, youth identity, and the evolving landscape of digital communication. Altogether, these findings highlight the continued dynamism of CSE as a contact variety and illustrate how corpus-based and survey findings contribute to our understanding of ongoing change in CSE. In this concluding section, we summarise the main

findings in Sections 3 to 5, pull together several thematic insights from these sections, identify broader implications for research on contact-induced change and CSE, and outline promising avenues for further research.

In Section 3, we reported on the online survey we conducted with native CSE-speaking participants to explore both the contexts in which CSE is used and perceptions of its use. Our focus was on three innovations in CSE that were the subject of recent investigations – clause-initial *is it*, adverbial *confirm*, and SFP *sia* – to see whether the patterns in CoSEM align with CSE speakers' self-reported use. Our findings showed that younger speakers consistently report higher usage of these features than older speakers, reinforcing their role as linguistic trendsetters (e.g., Labov 1972; Eckert 2000). Age thus emerged as a reliable measure of comparability between self-reported use and actual use of morphosyntactic features in CSE, while gender and ethnicity appear less reliable.

Section 4 examined the Hokkien-influenced CSE construction *bojio* 'NEG invite', originally used as an interjection to mean literally '[you] didn't invite [me]' but pragmatically as a complaint: 'Why didn't you invite me?'. The CoSEM data, supplemented by corpus evidence from X (formerly Twitter), showed that while Singaporeans more commonly use *invite* overall, distributional patterns (frequency only results) indicate that *jio* clusters in negated constructions, whereas model-based predictions reveal a higher and more stable likelihood of *jio* in non-negated contexts. In addition, *jio* is strongly associated with negative sentiment and is likely to be used more by younger and Chinese speakers. The results indicate that while *jio* clustering in negated constructions reflects its origins in complaint-related usage, its model-predicted stability in non-negated contexts and its association with negative sentiment, younger speakers, and Chinese users point to a broader pattern of affective and social conditioning beyond structural factors alone. Stylistically, we found that *jio* is preferred in more informational and informal settings, highlighting its growing presence in CSE since its reported entry around 2015–2016 (Wong 2020: 103).

Section 5 investigated substrate-influenced passive constructions in CSE. We found that use of the Malay-influenced *kena* passive remains robust in CSE, confirming Bao and Wee's (1999) prediction, while the *give* passive appears to have fallen out of favour. We also documented the innovative *tio* passive in CSE – an emerging construction due to influence from Southern Min. Both *kena* and *tio* passives map closely onto standardised English *be* and *get* passives and may optionally indicate an oblique argument using a *by*-phrase. Our analysis showed that younger speakers are more likely to use *kena* and *tio* passives than older speakers. There is limited evidence that gender and ethnicity have an effect on the use of the *kena* passive. We also showed that *kena* and *tio* passives are strongly associated with negative sentiment. Finally, we suggested that the

infrequency of *tio* passives in CoSEM likely reflects their restriction to colloquial, adversative contexts and competition from the more established *kena* passive.

6.1 Dynamic Contact and Continual Substrate Influence in CSE

A key thread running through this Element is the ongoing role of substrate influence in shaping morphosyntactic development in CSE. Traditional models of language contact, particularly those rooted in creole formation, typically assume that substrate influence is most potent during the early stages of contact (e.g., Mufwene 1996; Lefebvre 1998; Siegel 1999; Siegel 2000; Mufwene 2001). However, the empirical evidence presented in recent research on CSE and in this Element challenges this assumption in the context of Singapore.

The linguistic ecology of Singapore remains one in which Sinitic varieties (including Mandarin), Malay, Tamil, and English continue to be used side-by-side, creating conditions that favour ongoing substrate influence. This is reflected in recent research and in this Element's discussion of innovative morphosyntactic features of CSE: clause-initial *is it*, adverbial *confirm*, SFP *sia*, the verb *(bo)jio*, and the passive marker *tio*. Accordingly, the development of CSE cannot be understood solely in diachronic, founder-driven terms, but as a dynamic process shaped by youth identity and sustained multilingual contact in Singapore.

While globalising influences, such as Americanisation (e.g., Tan 2012; Starr 2019; Lincoln et al. 2024; Venkatachalam & Starr 2024), certainly play a role in the ongoing development of CSE, the evidence here supports a more complex picture in which new features also arise from continual influence from locally present languages in Singapore's linguistic ecology. From this perspective, CSE is best understood as a complex dynamic system (cf. Schneider 2025) in which local and global forces interact to shape ongoing developments.

6.2 The Role of Corpora and Surveys in CSE Research

A second major contribution of this Element concerns the methodological merits of corpus-based analysis complemented with insights from survey responses. In previous work, CoSEM has not only allowed us to identify notable age-, gender-, and ethnicity-based variation in the use of CSE morphosyntactic features (e.g., Leimgruber et al. 2021; Hiramoto et al. 2022), but also to examine patterns shaped by morphosyntactic factors, pragmatics, and competition among grammatically similar variants (e.g., Gonzales et al. 2023b; Lim, Hafiz, et al. 2024; Hafiz et al. 2025). Building on this tradition, Sections 4 and 5 leverage data from CoSEM to explore how structural, social, stylistic, and discourse-pragmatic

factors shape the use of *jio* and passive constructions in CSE. Section 5 also demonstrates how corpus data can reveal features that have fallen out of use (e.g., the *give* passive) as well as features that are newly emerging in CSE (e.g., the *tio* passive).

Additionally, Section 3 brought previous findings from CoSEM into dialogue with self-reported and perception data gathered through an online survey. While CoSEM data allows us to establish actual usage patterns of CSE in naturally occurring contexts, the survey data provides insights into how speakers themselves conceptualise both their own and others' use of CSE, much like previous research has done in Singapore and elsewhere (Leimgruber 2014; Leimgruber, Siemund, & Terassa 2018; Leimgruber 2019; Leimgruber et al. 2022). By pairing these two perspectives, we were able to contribute a novel methodological angle on the relationship between observed and self-reported data in CSE. Specifically, our findings suggest that age, though not gender or ethnicity, emerges as a reliable predictor of alignment between self-reports and actual usage patterns.

This methodological pairing underscores the value of combining corpus-based and survey-based approaches to understanding ongoing developments in CSE. Corpus data capture ecologically valid, naturally occurring linguistic practices of Singaporeans in informal digital contexts, enabling us to observe and predict how CSE features are used across individuals and across time. On the other hand, survey data reveal speakers' awareness, attitudes, and metalinguistic intuitions. All in all, these complementary sources of evidence provide a more holistic understanding of contemporary CSE than either method could achieve alone.

6.3 Towards an Understanding of Sentiment in CSE

Another key contribution of this Element lies in its ability to empirically verify the evaluative tendencies borne by two innovative CSE features – verbal *jio* and the *tio* passive – which, up until now, have largely been discussed through ethnographic observation and informal description (see e.g., Gonzales et al. 2023a on *(bo)jio*). Importantly, however, the two features pattern differently with respect to evaluation, and should not be treated as instantiating a single sentiment-negation profile. For *jio*, existing descriptions have often suggested an association with unfavourable or problematic situations, but also note occasional positive or neutral uses, leaving its evaluative profile somewhat underspecified. By contrast, *tio* has received far less focussed attention in the literature, and it would be overstated to claim that prior accounts consistently portray it as tied to 'negative contexts' in a general sense.

Our analysis in Section 4 shows that *jio*'s relationship with negation is stratified: although raw frequency patterns indicate clustering in negated constructions, model-based predictions reveal a higher and more stable likelihood of *jio* in non-negated contexts, with negated environments exhibiting greater variability. Importantly, the relevant notion of sentiment for *jio* operates at the level of the utterance as a whole: across both negated and non-negated environments, *jio* is strongly associated with negative overall message-level sentiment, indicating that speakers tend to deploy *jio* when adopting a broadly negative evaluative stance towards an event or situation.

Section 5 demonstrates that the *tio* passive displays a different evaluative alignment. Rather than tracking sentence-level sentiment, *tio* is sensitive to the evaluative polarity of the main predicate itself. That is, *tio* preferentially occurs with verbs denoting adverse, undesirable, or negatively valued outcomes, even when the speaker's overall stance towards the situation is positive. For example, a Liverpool supporter could plausibly say 'Very good, Man U *tio* disqualified', where the global message is positive for the speaker, but the predicate *disqualified* encodes a negatively evaluated event. In this respect, *tio* patterns less like a discourse-level stance marker and more like a construction whose semantics encode adversative or negatively valued eventualities, comparable to related passives such as *kena* or English *get* passives.

Viewed in conjunction, these patterns suggest that it is not negation as a formal grammatical property that unifies *jio* and *tio*, but evaluativity operating at different structural levels. For *jio*, sentiment is a property of the utterance-level stance; for *tio*, it is anchored in the lexical semantics of the main verb. Negation is therefore best understood as one possible, but non-determinative, resource through which negative stances may be expressed in the case of *jio*.

More broadly, these results underscore the utility of corpus data in clarifying the evaluative profiles of emerging CSE features, especially in cases where speaker intuitions and informal descriptions remain partial, underspecified, or internally inconsistent. Instead of presupposing that such features necessarily carry stable, explicit 'social meanings', our findings show that corpus evidence is particularly well-suited to isolating where evaluative tendencies reside in the grammar – whether at the level of discourse stance, predicate semantics, or constructional meaning. In the case of *jio*, this opens the door to understanding how evaluative stance may feed into broader identity-related practices; in the case of *tio*, the contribution is more modest but no less important, refining our understanding of how adversative meanings are grammatically encoded in CSE without requiring strong claims about social indexing.

Finally, these findings can be situated within a complex dynamic systems perspective on contact-induced change (Schneider 2025), in which emerging

CSE features are understood to stabilise through recurrent usage patterns rather than entering the system with fixed meanings. For *jio*, repeated association with negatively evaluated utterances appears to yield a relatively stable discourse-level evaluative tendency that may become available for stance-taking and social interpretation. For *tio*, by contrast, stabilisation occurs at a lower level of abstraction, with negative evaluativity functioning primarily as a semantic constraint on predicate selection rather than as a discourse- or identity-linked stance marker. Negative evaluativity thus acts as an ‘attractor’ in the evolving CSE system, but one that operates at different structural levels for different constructions, underscoring that not all emergent features develop comparable social meanings.

6.4 Future Directions

By examining the text messages of predominantly young Singaporeans, this Element has shown that CSE is evolving in ways that are entirely in line with what we know about language change more generally. As a number of variationist scholars have argued (e.g., Labov 1972; Eckert 2000), youth styles are often at the forefront of linguistic innovation, and our analyses of *is it*, adverbial *confirm*, *sia*, *(bo)jio*, and the emerging *tio* passive support this: younger CSE speakers are consistently the ones leading in the use, extension, and refunctionalisation of new morphosyntactic and lexical resources. At the same time, our findings underline the continued value of older corpora of Singapore English such as ICE-SIN. These earlier datasets remain indispensable for establishing historical baselines and tracing change over the longer term. What CoSEM adds is a more recent, youth-centric snapshot of CSE in digital interaction, allowing us to identify current trends and incipient changes that are not visible in 1990s data. Bringing legacy corpora such as ICE-SIN into dialogue with newer resources like CoSEM, therefore, promises a richer account of how CSE has shifted across roughly a quarter of a century, both structurally and socially, and how it might continue to evolve (see e.g., Schneider 2025 for the use of *finish* as a completive marker in CSE). We also hope that the case studies in Sections 3–5 have made this point tangible. Each section has used concrete empirical evidence to illustrate how innovations circulate through the system: from the diffusion of youth-indexical items (*is it*, *confirm*, and *sia*), to the morphosyntactic integration of *jio* and its competition with English *invite*, to the reorganisation of the passive system with the decline of older contact forms and the emergence of the *tio* passive. Altogether, these examples show not only that CSE is changing, but that change is unevenly distributed across age groups, interactional contexts, and linguistic subsystems.

Looking ahead, one key direction concerns broadening the empirical base beyond CoSEM. As discussed in [Section 3](#), the corpus is heavily skewed towards younger, digitally active speakers, and to particular social networks anchored around university students. This profile has been extremely productive for identifying innovative features such as *toh*, *dei*, and double SFPs, but it also limits the generalisations we can make about CSE in the wider community. Some of the future work will benefit from collecting additional data from speakers outside the CoSEM cohorts – especially older Singaporeans, non-tertiary-educated speakers, and a wider range of occupational and ethnic backgrounds – to test whether the innovative forms described here have diffused more broadly or remain confined to particular youth-centric niches. A second avenue involves extending our combined corpus–survey framework. The present survey provided an initial sense of how CSE users outside CoSEM report their own usage patterns and evaluate specific forms. Some of the future studies could develop this approach by incorporating follow-up interviews, experiments (e.g., perception or acceptability-judgement tasks), and platform-specific analyses (e.g., TikTok, Instagram, and Discord) that capture multimodal and cross-platform variation in greater detail. This would allow us to refine our understanding of when self-reports align with observed usage and when they diverge, and how changing media ecologies shape the visibility and circulation of new CSE features. Third, the features investigated so far represent only a subset of the innovations attested in CoSEM. The corpus contains many other promising candidates for future work, including lesser-studied particles, multi-particle sequences, and further instances of grammatical replication or semantic shift. Systematic analyses of items such as *toh*, *dei*, double SFPs, and additional emergent forms will help clarify how contact-induced change, indexicality, and youth style interact in shaping contemporary CSE. Comparing these developments with related processes in other high-contact Englishes and urban youth varieties could also contribute to refining models of contact and variation that were largely developed on the basis of earlier, less digitally mediated data. Finally, we envisage future collaborative work that builds directly on CoSEM as an open resource. The corpus has been made publicly available to researchers (<https://github.com/wdwgonzales/CoSEM>), and we hope that colleagues will use it for a wide range of purposes – testing theoretical claims about contact and change, developing NLP tools sensitive to CSE, exploring stance and affect in digital communication, or designing teaching materials that better reflect the linguistic realities of Singaporean youth. As CoSEM is enriched with additional annotation and potentially extended with new data, it can serve as a long-term monitoring tool for tracking how CSE continues to develop in a rapidly changing sociolinguistic landscape.

Returning to the claim in the beginning of this Element that CSE is difficult to fully capture with existing models of variation, our findings allow us to be more precise about where these models fall short and how they might be refined. Traditional, stage-based accounts of contact varieties and postcolonial Englishes, such as those in, for example, Mufwene (1990), Siegel (1999), and, to some extent, Schneider (2007), typically treat substrate influence as largely completed once endonormative stabilisation has been reached, and pay relatively little attention to how contact-driven innovations are taken up, reallocated, and reinterpreted by subsequent generations. By tracing the trajectories of features such as clause-initial *is it*, adverbial *confirm*, *sia*, *(bo)jio*, and the *tio* passive across apparent time, platforms, and interactional contexts, we show that contact effects in CSE are ongoing, multi-sourced, and tightly bound up with youth style and affective stance-taking in our case studies. This combination of corpus, survey, and sentiment-based evidence suggests the need for models that foreground continuous, bi/multi-directional contact and socially embedded speaker agency, rather than linear founder-driven scenarios alone. In this sense, the CSE documented here offers a useful test case for rethinking how we conceptualise morphosyntactic change in high-contact L1 Englishes more generally.

In sum, the main contribution of this Element lies not only in documenting a set of specific innovations but in demonstrating how combining corpus data, survey findings, and targeted case studies can shed light on youth-led change in a contact variety that remains in motion. Our hope is that the work presented in Sections 3–5 will encourage further research that both deepens and widens the empirical coverage of CSE, while also contributing to broader theoretical discussions about language variation and change in the digital age.

Glossing Conventions

Listed here are the glossing abbreviations used in this Element, which follow the Leipzig glossing conventions (Bickel, Comrie, & Haspelmath 2008).

1SG	First person singular
3SG	Third person singular
DEO	Deontic
EXIST	Existential marker
PASS	Passive marker
POSS	Possessive marker
NEG	Negation
PERF	Perfective aspect

References

- Androutsopoulos, Jannis. 2006. Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics* 10(4), 419–438. <https://doi.org/10.1111/j.1467-9841.2006.00286.x>.
- Ansaldo, Umberto. 2004. The evolution of Singapore English: Finding the matrix. In Lisa Lim (ed.), *Varieties of English Around the World* G33, 127–150. Amsterdam: Benjamins. <https://doi.org/10.1075/veaw.g33.08ans>.
- Aye, Khin Khin. 2005. *Bazaar Malay: History, Grammar and Contact*. Singapore: National University of Singapore PhD thesis.
- Babcock, Joshua. 2022. Postracial policing, ‘mother tongue’ sourcing, and images of Singlish standard. *Journal of Linguistic Anthropology* 32(2), 326–344. <https://doi.org/10.1111/jola.12354>.
- Bailey, Guy & Jan Tillery. 1999. The Rutledge effect: The impact of interviewers on survey results in linguistics. *American Speech* 74(4), 389–402.
- Bailey, Guy, Thomas Wike & Jan Tillery. 1997. The effects of methods on results in dialectology. *English World-Wide. A Journal of Varieties of English* 18(1), 35–63. <https://doi.org/10.1075/eww.18.1.03bai>.
- Baker, Jim. 2020. *Crossroads: A popular history of Malaysia and Singapore*. 4th ed. Singapore: Marshall Cavendish.
- Bao, Zhiming. 2011. Convergence-to-substratum and the passives in Singapore English. In Claire Lefebvre (ed.), *Typological Studies in Language*, 95, 253–270. Amsterdam: Benjamins. <https://doi.org/10.1075/tsl.95.15zhi>.
- Bao, Zhiming. 2015. *The Making of Vernacular Singapore English: System, Transfer and Filter* (Cambridge Approaches to Language Contact). Cambridge: Cambridge University Press.
- Bao, Zhiming. 2021. The origins of Singapore English. In Peter Siemund & Jakob R. E. Leimgruber (eds.), *Multilingual Global Cities: Singapore, Hong Kong, Dubai*, 19–37. Singapore: Routledge.
- Bao, Zhiming & Khin Khin Aye. 2010. Bazaar Malay topics. *Journal of Pidgin and Creole Languages* 25(1), 155–171. <https://doi.org/10.1075/jpcl.25.1.06bao>.
- Bao, Zhiming & Lionel Wee. 1999. The passive in Singapore English. *World Englishes* 18(1), 1–11. <https://doi.org/10.1111/1467-971X.00117>.
- Baptista, Marlyse. 2020. Competition, selection, and the role of congruence in creole genesis and development. *Language* 96(1), 160–199.
- Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2), 145–204. <https://doi.org/10.1017/S004740450001037X>.

- Benson, Phil, Alice Chik, Xuesong Gao, Jing Huang & Wenfeng Wang. 2009. Qualitative research in language teaching and learning journals, 1997–2006. *The Modern Language Journal* 93(1), 79–90. <https://doi.org/10.1111/j.1540-4781.2009.00829.x>.
- Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Max Planck Institute for Evolutionary Anthropology – Department of Linguistics. www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf.
- Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural Language Processing with Python*. Cambridge: O'Reilly.
- Bohmann, Axel. 2016. Language change because Twitter? Factors motivating innovative uses of because across the English-speaking Twittersphere. In Lauren Squires (ed.), *English in computer-mediated communication*, 149–178. Berlin: Gruyter. <https://doi.org/10.1515/9783110490817-008>.
- Boo, Ashley, Junwen Lee & Ying-Ying Tan. 2023. Particle stacking in Singlish – New data from the National Speech Corpus. *Lingua* 287, 103513. <https://doi.org/10.1016/j.lingua.2023.103513>.
- Botha, Werner. 2018. A social network approach to particles in Singapore English: A social network approach to particles in Singapore English. *World Englishes* 37(2), 261–281. <https://doi.org/10.1111/weng.12250>.
- Chambers, Jack K. 1998. Social embedding of changes in progress. *Journal of English Linguistics* 26(1), 5–36. <https://doi.org/10.1177/007542429802600102>.
- Chappell, Hilary. 2019. Southern Min. In Alice Vittrant & Justin Watkins (eds.), *The Mainland Southeast Asia Linguistic Area*, 176–233. Berlin: Gruyter. <https://doi.org/10.1515/9783110401981-005>.
- Chee, Hannah Naomi. 2019. *Toh: Language Contact, Identity Construction and the Proliferation of CMC Amongst Singaporean Youth*. Singapore: National University of Singapore BA Honours thesis.
- Chen, Tao & Min-Yen Kan. 2012. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*. 47, 299–335. <https://doi.org/10.1007/s10579-012-9197-9>.
- Cheong, Phoebe Si En. 2016. *Sentence-Final Already and Only in Singapore English*. Singapore: National University of Singapore BA Honours thesis.
- Choo, Athena. 2024. *A Case Study of Bah and Wor: A Newer Variety of Discourse Particles in Singlish*. Singapore: National University of Singapore BA Honours thesis.
- Coats, Steven, Carmelo Alessandro Basile, Cameron Morin & Robert Fuchs. 2025. The YouTube corpus of Singapore English podcasts. *English World-Wide* 46(3), 274–298. <https://doi.org/10.1075/eww.25018.coa>.

- Crystal, David. 2006. *Language and the Internet*. 2nd ed. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511487002>.
- Davies, Mark & 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), 1–28. <https://doi.org/10.1075/eww.36.1.01dav>.
- Department of Statistics Singapore. 2020. Census of population 2020. Statistical release 1: Demographic characteristics, education, language and religion. www.singstat.gov.sg/publications/reference/cop2020/cop2020-sr1. (30 October 2024).
- Deterding, David. 2007. *Singapore English*. Edinburgh: Edinburgh University Press.
- Deuber, Dagmar, Jakob R. E. Leimgruber & Andrea Sand. 2018. Singaporean internet chat compared to informal spoken language: Linguistic variation and indexicality in a language contact situation. *Journal of Pidgin and Creole Languages* 33(1), 48–91. <https://doi.org/10.1075/jpcl.00002.deu>.
- Dewan Bahasa dan Pustaka. 2010. *Kamus Dewan*. 4th ed. Dewan Bahasa dan Pustaka.
- Dollinger, Stefan. 2012. The western Canada-US border as a linguistic boundary: The roles of L1 and L2 speakers. *World Englishes* 31(4), 519–533. <https://doi.org/10.1111/j.1467-971X.2012.01778.x>.
- Dollinger, Stefan. 2015. *The Written Questionnaire in Social Dialectology: History, Theory, Practice* (IMPACT: Studies in Language, Culture and Society). Vol. 40. Amsterdam: Benjamins. <https://doi.org/10.1075/impact.40>.
- Eckert, Penelope. 1989. The whole woman: Sex and gender differences in variation. *Language Variation and Change* 1(3), 245–267. <https://doi.org/10.1017/S095439450000017X>.
- Eckert, Penelope. 2000. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High* (Language in Society 27). Malden, MA: Blackwell.
- Faulkner, Xristine & Fintan Culwin. 2005. When fingers do the talking: A study of text messaging. *Interacting with Computers* 17(2), 167–185. <https://doi.org/10.1016/j.intcom.2004.11.002>.
- Fong, Vivienne. 2004. The verbal cluster. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 75–104. Amsterdam: Benjamins. <https://doi.org/10.1075/veaw.g33.06fon>.
- Gonzales, Wilkinson Daniel Wong. 2024. When to (not) split the infinitive: Factors governing patterns of syntactic variation in Twitter-style Philippine

- English. *English Language and Linguistics* 28(2), 305–339. <https://doi.org/10.1017/S1360674323000631>.
- Gonzales, Wilkinson Daniel Wong. 2025a. From tweets to trends: Analyzing sociolinguistic variation and change using the Twitter Corpus of English in Hong Kong (TCOEHK). *Asian Englishes* 27(1), 1–24. <https://doi.org/10.1080/13488678.2023.2251771>.
- Gonzales, Wilkinson Daniel Wong. 2025b. Sociolinguistic analysis with missing metadata? Leveraging linguistic and semiotic resources through deep learning to investigate English variation and change on Twitter. *Applied Linguistics* 46(3), 411–434. <https://doi.org/10.1093/applin/amad086>.
- Gonzales, Wilkinson Daniel Wong, Pui Yu Ivy Chan, Xiaohan Harry Zhang, Chui Yin Judy Ng & Chung Yan Ching Karina (2026). *Gender and Uptalk in Hong Kong English*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009634083www.cambridge.org/core/elements/abs/gender-and-uptalk-in-hong-kong-english/C08F6A2E02F23489D9F7E698B92065CA>.
- Gonzales, Wilkinson Daniel Wong, Mie Hiramoto, Jakob R. E. Leimgruber & Jun Jie Lim. 2023a. The Corpus of Singapore English Messages (CoSEM). *World Englishes* 42(2), 371–388. <https://doi.org/10.1111/weng.12534>.
- Gonzales, Wilkinson Daniel Wong, Mie Hiramoto, Jakob R. E. Leimgruber & Jun Jie Lim. 2023b. *Is it in Colloquial Singapore English: What variation can tell us about its conventions and development*. *English Today* 39(4), 243–256. <https://doi.org/10.1017/S0266078422000141>.
- Grafmiller, Jason, Benedikt Szmrecsanyi, Melanie Röthlisberger & Benedikt Heller. 2018. General introduction: A comparative perspective on probabilistic variation in grammar. *Glossa* 3(1), 94. <https://doi.org/10.5334/gjgl.690>.
- Gupta, Anthea Fraser. 1994. *The Step-Tongue: Children's English in Singapore*. Clevedon: Multilingual Matters.
- Hafiz, Mohamed. 2025. *A Sketch Grammar of Singapore Colloquial Malay*. Singapore: National University of Singapore, MA thesis.
- Hafiz, Mohamed, Mie Hiramoto, Jakob R. E. Leimgruber, Wilkinson Daniel Wong Gonzales & Jun Jie Lim. 2025. Sociolinguistic variation in Colloquial Singapore English *sia*. *World Englishes* 44(1–2), 218–236. <https://doi.org/10.1111/weng.12700>.
- Heine, Bernd, Gunther Kaltenböck, Tania Kuteva & Haiping Long. 2021. *The Rise of Discourse Markers*. 1st ed. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108982856>.
- Heine, Bernd & Tania Kuteva. 2003. On contact-induced grammaticalization. *Studies in Language* 27(3), 529–572. <https://doi.org/10.1075/sl.27.3.04hei>.

- Heine, Bernd & Tania Kuteva. 2005. *Language Contact and Grammatical Change* (Cambridge Approaches to Language Contact). Cambridge: Cambridge University Press.
- Hing, Jia Wen. 2019. *A Descriptive Grammar of Penang Hokkien*. Singapore: National University of Singapore PhD thesis.
- Hing, Jia Wen. 2023. *Pún* and *tiòh* in Penang Hokkien. *Journal of Pidgin and Creole Languages* 38(1), 151–169. <https://doi.org/10.1075/jpcl.00108.hin>.
- Hiramoto, Mie. 2019. Colloquial Singapore English in advertisements. *World Englishes* 38(3), 450–462. <https://doi.org/10.1111/weng.12422>.
- Hiramoto, Mie, Wilkinson Daniel Wong Gonzales, Jakob R. E. Leimgruber, Jun Jie Lim, & Jessica Choo. 2022. From Malay to Colloquial Singapore English: A case study of sentence-final particle *sia*. In Aloysius Ngefacs, Hans-Georg Wolf & Thomas Hoffmann (eds.), *World Englishes and Creole Languages Today – Vol. 1. The Schneiderian Thinking and Beyond: In Honour of Edgar W. Schneider* (LINCOM Studies in English Linguistics 24), 117–130. Munich: Lincom.
- Hiramoto, Mie, Wilkinson Daniel Wong Gonzales, Jakob R. E. Leimgruber, Jun Jie Lim, & Mohamed Hafiz. 2025. Functional expansion and syntactic change in Colloquial Singapore English *is it*. In Sofia Rüdiger, Sven Leukert & Jakob R. E. Leimgruber (eds.), *World Englishes and social media: Platforms, variation, and meta-discourse*, 15–40. London: Bloomsbury. <https://doi.org/10.5040/9781350421448.ch-2>.
- Ho, Mian-Lian & John T Platt. 1993. *Dynamics of a Contact Continuum*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198248286.001.0001>.
- Huang, Nick, Li Lin, Kunmei Han, et al. 2025. Treebanks and World Englishes: A Singapore English perspective. *English World-Wide* 46(1), 93–121. <https://doi.org/10.1075/eww.23069.hua>.
- Huff, Gregg & Gillian Huff. 2020. The Second World War Japanese occupation of Singapore. *Journal of Southeast Asian Studies* 51(1–2), 243–270. <https://doi.org/10.1017/S002246342000017X>.
- Ilbury, Christian. 2020. ‘Sassy queens’: Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics* 24(2), 245–264. <https://doi.org/10.1111/josl.12366>.
- Johnstone, Barbara. 2016. Enregisterment: How linguistic items become linked with ways of speaking. *Language and Linguistics Compass* 10(11), 632–643. <https://doi.org/10.1111/lnc3.12210>.
- Kaur, Ershen. 2019. *Dei: The Sociolinguistic Investigation of an Indian Ethnic Marker in English*. Singapore: National University of Singapore BA Honours thesis.

- Khoo, Velda. 2012. *The sia particle in Colloquial Singapore English*. Singapore: Nanyang Technological University BA Honours thesis.
- Khoo, Velda. 2023. Voicing singlish from the ‘middle’: Indexical hybridities of class, race, language, and Singaporeanness. *Journal of Linguistic Anthropology* 33(2), 202–222. <https://doi.org/10.1111/jola.12403>.
- Kiesling, Scott F. 2018. Masculine stances and the linguistics of affect: On masculine ease. *NORMA* 13(3–4), 191–212. <https://doi.org/10.1080/18902138.2018.1431756>.
- Kiesling, Scott F., Umashanthi Pavalanathan, Jim Fitzpatrick, Xiaochuang Han & Jacob Eisenstein. 2018. Interactional stancetaking in online forums. *Computational Linguistics* 44(4), 683–718. https://doi.org/10.1162/coli_a_00334.
- Kim, Chonghyuck & Yosuke Sato. 2013. More on *Kena*-passives in Singapore English. *World Englishes* 32(3), 297–307. <https://doi.org/10.1111/weng.12032>.
- Klein, Dan & Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – ACL ’03* 1, 423–430. Sapporo, Japan: Association for Computational Linguistics. <https://doi.org/10.3115/1075096.1075150>.
- Kortmann, Bernd, Kerstin Lunkenheimer & Katharina Ehret. 2020. ewave: The Electronic World Atlas of Varieties of English. Zenodo. <https://doi.org/10.5281/ZENODO.3712132>.
- Kostadinova, Viktorija. 2018. Attitudes to usage vs. actual language use: The case of *literally* in American English: American English speakers know how and why they use *literally*. *English Today* 34(4), 29–38. <https://doi.org/10.1017/S0266078418000366>.
- Kwa, Chong Guan, Derek Heng, Peter Borschberg & Tai Yong Tan. 2019. *Seven Hundred Years*. Singapore: Marshall Cavendish.
- Labov, William. 1966. *The Social Stratification of English in New York City* (Urban Language Series). Washington, DC: Center for Applied Linguistics.
- Labov, William. 1972. *Sociolinguistic Patterns* (Conduct and Communication Series). Philadelphia: University of Pennsylvania Press.
- Labov, William. 1975. *What is a Linguistic Fact?* Berlin: Gruyter. <https://doi.org/10.1515/9783110882353>.
- Lee, Elsie Shu Ying & Hyejeong Ahn. 2021. Overseas Singaporean attitudes towards Singlish. *Asian Englishes* 23(3), 264–279. <https://doi.org/10.1080/13488678.2020.1795783>.
- Lee, Nala Huiying, Ling Ai Ping & Hiroki Nomoto. 2009. Colloquial Singapore English *got*: Functions and substratal influences. *World Englishes* 28(3), 293–318. <https://doi.org/10.1111/j.1467-971X.2009.01594.x>.

- Lee, Tong King. 2023. *Spiaking* Singlish: The politics of ludic English in Singapore. *Language in Society* 52(2), 241–261. <https://doi.org/10.1017/S0047404521001019>.
- Lefebvre, Claire. 1998. *Creole Genesis and the Acquisition of Grammar: The Case of Haitian Creole* (Cambridge Studies in Linguistics 88). Cambridge: Cambridge University Press.
- Leimgruber, Jakob R. E. 2009. *Modelling variation in Singapore English*. Oxford: University of Oxford DPhil thesis.
- Leimgruber, Jakob R. E. 2012. Singapore English: An indexical approach. *World Englishes* 31(1), 1–14. <https://doi.org/10.1111/j.1467-971X.2011.01743.x>.
- Leimgruber, Jakob R. E. 2013. *Singapore English: Structure, Variation, and Usage* (Studies in English Language). Cambridge: Cambridge University Press.
- Leimgruber, Jakob R. E. 2014. Singlish as defined by young educated Chinese Singaporeans. *International Journal of the Sociology of Language* (230), 45–63. <https://doi.org/10.1515/ijsl-2014-0026>.
- Leimgruber, Jakob R. E. 2016. *Bah* in Singapore English. *World Englishes* 35(1), 78–97. <https://doi.org/10.1111/weng.12175>.
- Leimgruber, Jakob R. E. 2019. *Language Planning and Policy in Quebec: A Comparative Perspective with Views from Wales and Singapore* (Language in Performance). Tübingen: Narr.
- Leimgruber, Jakob R. E. 2026. English in Singapore. In Raymond Hickey & Kate Burridge (eds.), *New Cambridge History of the English Language*. 229–252. Cambridge: Cambridge University Press.
- Leimgruber, Jakob R. E., Ahmad Al-Issa, Eliane Lorenz & Peter Siemund. 2022. Managing and investing in hybrid identities in the globalized United Arab Emirates. *Journal of Language, Identity & Education* 23(6), 955–972. <https://doi.org/10.1080/15348458.2022.2070849>.
- Leimgruber, Jakob R. E., Jun Jie Lim, Wilkinson Daniel Wong Gonzales & Mie Hiramoto. 2021. Ethnic and gender variation in the use of Colloquial Singapore English discourse particles. *English Language and Linguistics* 25(3), 601–620. <https://doi.org/10.1017/S1360674320000453>.
- Leimgruber, Jakob R. E., Peter Siemund & Laura Terassa. 2018. Singaporean students' language repertoires and attitudes revisited. *World Englishes* 37(2), 282–306. <https://doi.org/10.1111/weng.12292>.
- Levshina, Natalia. 2016. Control, causation and Google counts. *Nederlandse Taalkunde* 21(2), 253–263. <https://doi.org/10.5117/NEDTAA2016.2.LEVS>.
- Levshina, Natalia. 2022. Comparing Bayesian and frequentist models of language variation: The case of *help* + (to-)Infinitive. In *Data and Methods in*

- Corpus Linguistics: Comparative Approaches*, 224–258. Cambridge: Cambridge University Press.
- Lien, Chunfa. 2001. The semantic extension of *tioh8* 著 in Taiwanese Southern Min: An interactive approach. *Language and Linguistics* 2(2), 173–202.
- Lim, Joshua & Norhaida Aman. 2024. English language preservice teachers' attitudes towards Singlish and their beliefs about the utility of contrastive pedagogies in the language classroom. *Asian Englishes* 26(3), 666–681. <https://doi.org/10.1080/13488678.2024.2399406>.
- Lim, Jun Jie, Spencer C. Chen & Mie Hiramoto. 2021. 'You don't ask me to speak Mandarin, okay?': Ideologies of language and race among Chinese Singaporeans. *Language & Communication* 76, 100–110. <https://doi.org/10.1016/j.langcom.2020.10.003>.
- Lim, Jun Jie, Mohamed Hafiz, Wilkinson Daniel Wong Gonzales, Jakob R. E. Leimgruber & Mie Hiramoto. 2024. Adverbial *confirm* in colloquial Singapore English: Insights from a text message corpus. *Asian Englishes* 26(3), 648–665. <https://doi.org/10.1080/13488678.2024.2398853>.
- Lim, Jun Jie, Mie Hiramoto, Jakob R. E. Leimgruber & Wilkinson Daniel Wong Gonzales. 2024. Clause-final adverbs in Colloquial Singapore English revisited. *Journal of English Linguistics* 52(4), 321–345. <https://doi.org/10.1177/00754242241278838>.
- Lim, Lisa Li Su. 2004. *Singapore English: A Grammatical Description*. Amsterdam: Benjamins.
- Lim, Lisa Li Su. 2015. Coming of age, coming full circle: The (re)positioning of (Singapore) English and multilingualism in Singapore at 50. *Asian Englishes* 17(3), 261–270. <https://doi.org/10.1080/13488678.2015.1090772>.
- Lin, Li, Kunmei Han, Jia Wen Hing, et al. 2023. Tagging Singapore English. *World Englishes* 42(4), 624–641. <https://doi.org/10.1111/weng.12597>.
- Lincoln, Wesley Mark, Rebecca Lurie Starr, Víctor Fernández-Mallat & Jennifer Nycz. 2024. T-flapping in Singapore English: Americanization, innovation, or both? In *Dialect Contact: From Speaker to Community-Based Perspectives*, 161–182. Washington, DC: Georgetown University Press.
- Lu, Luke. 2025. Who's being elitist? A debate about the enregisterment of Singlish. *Language in Society* 54(3), 545–567. <https://doi.org/10.1017/S0047404524000162>.
- MacKenzie, Laurel. 2020. Comparing constraints on contraction using Bayesian regression modeling. *Frontiers in Artificial Intelligence* 3, 58. <https://doi.org/10.3389/frai.2020.00058>.
- Makowski, Dominique, Mattan S. Ben-Shachar, S. H. Annabel Chen & Daniel Lüdtke. 2019. Indices of effect existence and significance in the

- Bayesian framework. *Frontiers in Psychology* 10, 2767. <https://doi.org/10.3389/fpsyg.2019.02767>.
- Matsuoka, Masakazu. 2018. Media and cultural policy and Japanese language education in Japanese-occupied Singapore, 1942–1945. In Kayoko Hashimoto (ed.), *Japanese Language and Soft Power in Asia*, 83–102. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-5086-2_5.
- Maynor, Natalie. 1994. The language of electronic mail: Written speech? *The Publication of the American Dialect Society* 78(1), 48–54. <https://doi.org/10.1215/-78-1-48>.
- McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. London: Chapman and Hall. <https://doi.org/10.1201/9780429029608>.
- Meyerhoff, Miriam. 2018. *Introducing Sociolinguistics*. 3rd ed. Abingdon: Routledge. <https://doi.org/10.4324/9780429507922>.
- Meyerhoff, Miriam & Alexandra Birchfield. 2019. Gender and language contact: How gender is/isn't marked in language contact. In Jeroen Darquennes, Joseph C. Salmons & Wim Vandenbussche (eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)*, 246–256. Berlin: Gruyter. <https://doi.org/10.1515/9783110435351-021>.
- Montani, Ines, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem & Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements. Zenodo. <https://doi.org/10.5281/ZENODO.1212303>.
- Mufwene, Salikoko S. 1990. Transfer and the substrate hypothesis in creolistics. *Studies in Second Language Acquisition* 12(1), 1–23. <https://doi.org/10.1017/S0272263100008718>.
- Mufwene, Salikoko S. 1996. The founder principle in creole genesis. *Diachronica* 13(1), 83–134. <https://doi.org/10.1075/dia.13.1.05muf>.
- Mufwene, Salikoko S. 2001. *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.
- Nomoto, Hiroki & Abd.Wahab Kartini. 2012. Kena adversative passives in Malay, funny control, and covert voice alternation. *Oceanic Linguistics* 51(2), 360–386. <https://doi.org/10.1353/ol.2012.0017>.
- Pak, Vincent & Mie Hiramoto. 2023. Sticky raciolinguistics. *Signs and Society* 11(1), 45–67. <https://doi.org/10.1086/722622>.
- Pratt, Teresa. 2023. Affect in sociolinguistic style. *Language in Society* 52(1), 1–26. <https://doi.org/10.1017/S0047404521000774>.
- Rinker, W. Tyler. 2021. {Package sentimentr}. *Calculate Text Polarity Sentiment*. New York.

- Schneider, Edgar W. 2007. *Postcolonial English: Varieties Around the World* (Cambridge Approaches to Language Contact). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511618901>.
- Schneider, Edgar W. 2025. *World Englishes as Components of a Complex Dynamic System*. 1st ed. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009289504>.
- Siegel, Jeff. 1999. Transfer constraints and substrate influence in Melanesian Pidgin. *Journal of Pidgin and Creole Languages* 14(1), 1–44. <https://doi.org/10.1075/jpcl.14.1.02sie>.
- Siegel, Jeff. 2000. Substrate influence in Hawai'i Creole English. *Language in Society* 29(2), 197–236. <https://doi.org/10.1017/S0047404500002025>.
- Siegel, Jeff. 2008. In praise of the cafeteria principle: Language mixing in Hawai'i Creole. In Susanne Maria Michaelis (ed.), *Creole Language Library*, vol. 33, 59–82. Amsterdam: Benjamins. <https://doi.org/10.1075/cll.33.06sie>.
- Smakman, Dick & Stephanie Wagenaar. 2013. Discourse particles in Colloquial Singapore English: Discourse particles in Colloquial Singapore English. *World Englishes* 32(3), 308–324. <https://doi.org/10.1111/weng.12033>.
- Sneller, Betsy & Sabriya Fisher. 2015. When GET got noticed: The emerging salience of GET-passives. *U. Penn Working Papers in Linguistics* 21(1), 1–9.
- Sonderegger, Morgan. 2023. *Regression Modeling for Linguistic Data*. Cambridge, MA: The MIT Press.
- Starr, Rebecca Lurie. 2019. Cross-dialectal awareness and use of the BATH–TRAP distinction in Singapore: Investigating the effects of overseas travel and media consumption. *Journal of English Linguistics* 47(1), 55–88. <https://doi.org/10.1177/0075424218819740>.
- Starr, Rebecca Lurie & Mie Hiramoto. 2019. Inclusion, exclusion, and racial identity in Singapore's language education system. *International Journal of Applied Linguistics* 29(3), 341–355. <https://doi.org/10.1111/ijal.12242>.
- Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin: Gruyter. <https://doi.org/10.1515/9783110197808>.
- Tan, Ying-Ying. 2012. To *r* or not to *r*: Social correlates of /ɹ/ in Singapore English. *International Journal of the Sociology of Language* 2012(218), 1–24. <https://doi.org/10.1515/ijsl-2012-0057>.
- Tham, Seong Chee. 1996. *Multilingualism in Singapore: Two Decades of Development (Census of Population, 1990)*. Singapore: Department of Statistics.

- Tillery, Jan, Tom Wikle & Guy Bailey. 2000. The nationalization of a southernism. *Journal of English Linguistics* 28(3), 280–294. <https://doi.org/10.1177/00754240022005045>.
- Trudgill, Peter. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1(2), 179–195. <https://doi.org/10.1017/S0047404500000488>.
- Trudgill, Peter. 1974a. *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.
- Trudgill, Peter. 1974b. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3(2), 215–246. <https://doi.org/10.1017/S0047404500004358>.
- Turnbull, Constance M. 1989. *A history of Singapore, 1819–1988*. 2nd ed. Singapore: Oxford University Press.
- Varishth, Shravan, Bruno Nicenboim, Mary E. Beckman, Fangfang Li & Eun Jong Kong. 2018. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 71, 147–161. <https://doi.org/10.1016/j.wocn.2018.07.008>.
- Vehtari, Aki. 2021. On reparameterisation invariant Bayesian point estimates and credible regions. arXiv. <https://doi.org/10.48550/ARXIV.2109.10843>.
- Venkatachalam, Alamelu & Rebecca Lurie Starr. 2024. Variation in the acceptability of singular *they* in Singapore English. *U. Penn Working Papers in Linguistics* 30(2), 143–152.
- Wang, Yuxuan, Wanxiang Che, Jiang Guo, Yijia Liu & Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5720–5726. Hong Kong, China: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1575>.
- Wiktionary. 2025. tio. *Wiktionary, the Free Dictionary*. <https://en.wiktionary.org/w/index.php?title=tio&oldid=88480915#English>. (14 December 2025).
- Wong, Jock O. 2020. The Singlish interjection *bojio*. In Helen Bromhead & Zhengdao Ye (eds.), *Meaning, Life and Culture: In Conversation with Anna Wierzbicka*, 99–114. ANU Press. <https://doi.org/10.2307/j.ctv1d5nm0d.11>.
- Wong, Kevin Zi-Hao & Ying-Ying Tan. 2017. Being Chinese in a global context: Linguistic constructions of Chinese ethnicity. *Global Chinese* 3(1), 1–23. <https://doi.org/10.1515/glochi-2017-0001>.
- Wong, Lin Ken. 1978. Singapore: Its growth as an entrepot port, 1819–1941. *Journal of Southeast Asian Studies* 9(1), 50–84. <https://doi.org/10.1017/S002246340000953X>.

- Zhou, Changji. 2006. *Minnan fangyan da cidian* [= *Minnan dialect dictionary*]. Fuzhou: Fujian Renmin Chubanshe.
- Ziegeler, Debra. 2020. Changes in the functions of *already* in Singapore English: A grammaticalisation approach. *Journal of Pidgin and Creole Languages* 35(2), 293–331. <https://doi.org/10.1075/jpcl.00062.zie>.
- Ziegeler, Debra. 2024. *The Influence of the Lexifier: Beyond Grammaticalisation in Singapore English*. Berlin: Gruyter. <https://doi.org/10.1515/9783110785258>.

Acknowledgements

This Element and the larger research project on which it is based would not have been possible without the generous support of many individuals and institutions over the years. We gratefully acknowledge funding from the Singapore Ministry of Education (MOE) Academic Research Fund Tier 1, *Singapore Colloquial Singapore English in Social Media and Other Forms of Messaging* (R-103–000–167–115, 2019–2022), awarded to Mie Hiramoto as Principal Investigator. We also thank the National University of Singapore for its continued support, including the UROP scheme, which enabled sustained student involvement in the project. Nick Huang and Mohamed Hafiz are also supported in part by the MOE, Singapore, under the Social Science and Humanities Research Fellowship (SSRC2024-SSHR-002, 2026–2031), awarded to Nick Huang as Principal Investigator.

Our heartfelt thanks go to the research assistants who worked on various stages of the CoSEM project, in alphabetical order of their family names: Latané Bullock, Athena Choo, Choo Xue Ming Jessica, Sydelle D Souza, Aura Valentine Eden, Nadine Ng Hui Ning, Pak Yongkang Vincent, Luqman Aqil Bin Rozman, and Yeo Rei-Chi Lauren. We also thank Sara Blieninger and Nele Heaslip for their help in finalising this manuscript.

We are equally indebted to the UROP students, whose energy and curiosity helped shape many of the ideas and analyses in this Element, again in alphabetical order of their family names: Boey Yui Keat Ryan-Ashleigh, Chang Wei Xing, Athena Choo, Chloe Kwek Wing Yen, Lavonne Yew Yu Xuan, Lin Taohai, Mak Kai Feng Darren, Magdalene Ong Wenli, Megan Eleanor Wee Jia Xin, Pek Wee Kit Terence, Reno Sam Wei Jie, Ten Ting Kai, Francisca Ann Vincent, and Yeo Ming Wen, Megan Maria.

We would also like to thank our colleagues Susan Ang, Michael Yoshitaka Erlewine, Leslie Lee, Nala Lee, Edgar W. Schneider, and Rebecca Lurie Starr for their insightful comments, encouragement, and collegial support at various stages of this work. We are especially grateful to Edgar, in his role as Elements series editor, for his support and guidance throughout the development of this volume. Our deep gratitude extends to all survey participants and distributors, whose time and generosity made the empirical aspects of this Element possible.

There are far more people than we can properly name here – students, colleagues, friends, and family members who have offered feedback, ideas, and moral support over the years. Any omissions are entirely our own, and we hope that everyone who has contributed, in ways large and small, knows how sincerely grateful we are.

World Englishes

Edgar W. Schneider

University of Regensburg

Edgar W. Schneider is Professor Emeritus of English Linguistics at the University of Regensburg, Germany. His many books include *Postcolonial English* (Cambridge, 2007), *English around the World*, 2e (Cambridge, 2020) and *The Cambridge Handbook of World Englishes* (Cambridge, 2020).

Editorial Board

Alexandra D'Arcy, *University of Victoria*

Kate Burridge, *Monash University*

Paul Kerswill, *University of York*

Christian Mair, *University of Freiburg*

Christiane Meierkord, *Ruhr University*

Raj Mesthrie, *University of Cape Town*

Miriam Meyerhoff, *University of Oxford*

Daniel Schreier, *University of Zurich*

Devyani Sharma, *University of Oxford*

Sali Tagliamonte, *University of Toronto*

Bertus van Rooy, *University of Amsterdam*

Lionel Wee, *National University of Singapore*

About the Series

Over the last centuries, the English language has spread all over the globe due to a multitude of factors including colonization and globalization. In investigating these phenomena, the vibrant linguistic sub-discipline of "World Englishes" has grown substantially, developing appropriate theoretical frameworks and considering applied issues. This Elements series will cover all the topics of the discipline in an accessible fashion and will be supplemented by on-line material.

World Englishes

Elements in the Series

Uniformity and Variability in the Indian English Accent
Caroline R. Wiltshire

Posthumanist World Englishes
Lionel Wee

*The Cognitive Foundation of Post-colonial Englishes: Construction Grammar as the
Cognitive Theory for the Dynamic Model*
Thomas Hoffmann

Inheritance and Innovation in the Evolution of Rural African American English
Guy Bailey, Patricia Cukor-Avila and Juan Salinas

*Indian Englishes in the Twenty-First Century: Unity and Diversity in Lexicon
and Morphosyntax*
Sven Leuckert, Claudia Lange, Tobias Bernaisch and Asya Yurchenko

*Language Ideologies and Identities on Facebook and TikTok: A Southern
Caribbean Perspective*
Guyanne Wilson

Transnational Korean Englishes
Sofia Rüdiger and Alex Baratta

Multiscriptal English in Transliterated Linguistic Landscapes
Chonglong Gu

World Englishes as Components of a Complex Dynamic System
Edgar W. Schneider

*Optimizing Decision Trees for the Analysis of World Englishes and Sociolinguistic
Data*
Sarah Buschfeld and Claus Weihs

Philippine Englishes-in-Motion: The Case of Filipino Migrants in Japan
Julius C. Martinez

*Young Singaporean English: Current Developments in CSE
Morphosyntax in a Digital World*

Jakob R. E. Leimgruber, Wilkinson Daniel Wong Gonzales,
Mohamed Hafiz, Jun Jie Lim, Nick Huang and Mie Hiramoto

A full series listing is available at: www.cambridge.org/EIWE