The Corpus of Singapore English Messages (CoSEM)

Wilkinson Daniel Wong Gonzales[1] | Mie Hiramoto[2] | Jakob Leimgruber[3] | Jun Jie Lim[4]

[1]Department of Linguistics, University of Michigan, Ann Arbor, USA, [2]Department of English Language and Literature, National University of Singapore, Singapore, [3]English Department, University of Basel, Switzerland, [4]Linguistics Department, University of California, San Diego, USA

Correspondence

Wilkinson Daniel Wong Gonzales, Department of Linguistics, University of Michigan, 440 Lorch Hall 611 Tappan Street Ann Arbor MI 48109-1220.

Email: wdwg@umich.edu

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

Abstract

This article introduces the first version of the Corpus of Singapore English Messages (CoSEM), a 3.6-million-word monitor corpus of online text messages collected between 2016 and 2019, compiled and managed by a group of scholars who share an interest in Colloquial Singapore English (CSE)

research. The paper explains the motivations behind developing a new corpus for the investigation of CSE. It also documents the process of compiling and organizing CoSEM and describes the corpus's initial structure and composition. We further discuss the social variables used in tagging the data, as well as ethical challenges, advantages, and disadvantages unique to online message datasets. In addition, to demonstrate CoSEM's applications and highlight the importance of its sociolinguistic information tagging, we present preliminary analyses of two selected CSE features: (1) the Hokkien-derived expression *(bo)jio* and (2) sentence-final adverbs (*already*, *also*, *only*). As CoSEM is an ongoing project, we conclude the article with notes on future directions.

Keywords: Colloquial Singapore English; sociolinguistics; contact linguistics; finger speech; computer-mediated communication

1 | INTRODUCTION

Colloquial Singapore English (CSE) is a linguistic variety used primarily in multiracial and multilingual Singapore. It began to emerge in the colonial period (1819–1942), becoming a lingua franca among speakers of Singapore's major heritage languages: Hokkien, Cantonese, Malay, and Tamil. Research on CSE played an important role in the development of the world Englishes paradigm, having started at least a decade before the formation of the International Committee of the Study of World Englishes in 1988 (see Richards & Tay, 1977; Tongue, 1979; Platt & Weber, 1980; Platt et al., 1983; Kwan-Terry, 1989; Gupta, 1994; Alsagoff, 2010; Wong, 2014; Bao, 2015; Ziegeler, 2015). Extensive research has been conducted on the linguistic properties of CSE (for example, phonetics: Lim, 2009; Starr & Balasubramaniam, 2019; morphosyntax: Bao 2010b; Bao & Wee, 1999; semantics: Bao, 2009, Hiramoto & Sato 2012; pragmatics: Lim, 2007; Hiramoto, 2012; Leimgruber, 2016), and on the CSE speech community and how speakers use CSE in relation to other languages in the Singaporean language ecology (for example, Leimgruber, 2012, 2018; Siemund et al., 2014; Leimgruber et al., 2018; Hiramoto, 2019; Starr, forthcoming).

2

Although studies focusing on CSE's structural features tend not to delve into its internal variation, scholars, especially sociolinguists, generally agree that CSE varies across generations and racial groups, and has likely been changing constantly over the course of its history (see Wee, 2003; Leimgruber, 2014; Lim, 2015; Hiramoto, 2019). For example, new features and expressions are frequently introduced, and existing ones replaced, in CSE. Computer-mediated communication (CMC) has made the dynamic nature of CSE particularly salient – remarkably so in the netspeak of Singapore netizens. Noting this trend of CSE communication, some researchers have recently started to utilize CMC as a data source for linguistic investigations (Botha, 2018; Deuber et al., 2018). However, there is a lack of systematically compiled archives of CMC-based data for CSE research. We suggest that using CMC-based data is crucial for capturing the historical trajectories of CSE features and to advance research in the field. Therefore, as scholars who are committed to the study of CSE and its role in Singaporean society, we present in this paper the first version of the Corpus of Singapore English Messages (CoSEM), a monitor corpus whose primary objective is to further our understanding of the systematic and dynamic nature of CSE in a new format of communication, namely, text messaging. The major goal of this project is to provide scholars with a contemporary corpus of CSE data (see Leimgruber et al., 2021).

In what follows, we offer a preliminary report of the ongoing CoSEM project. Section 2 discusses CoSEM in relation to other corpora to cast light on the motivations for creating CoSEM. Section 3 details how the data are being collected, organized, and compiled. In section 4, we report the structure and composition of the corpus, including the distribution of the data with respect to social factors. The format of the social information tags is discussed in section 5. Sample sociolinguistic analyses on two CSE features – (*bo*)*jio* '(no) invite', and the clause-final adverbs *already*, *also*, and *only* – are presented in section 6 to demonstrate the applicability and utility of CoSEM. Section 7 concludes the paper with discussions of the corpus's advantages and disadvantages, the practical and ethical considerations of collecting and using online text message data, and possible future directions for the project.

2 | COSEM IN RELATION TO OTHER CORPORA

3

Using corpora to investigate CSE features is a widely applied practice. Scholars conducting CSE research have used existing corpus-based data such as the National Institute of Education Spoken Corpus of English in Asia (Low, 2015). Other corpora that contain CSE data (as part of a larger corpus) include the GloWbE and NoW corpora (Davies, 2013, 2016). The most well-known and accessible corpus for the CSE research community within a world Englishes paradigm, however, has been the International Corpus of English (ICE). ICE, which features data from both spoken and written sources, has made data-driven, quantitative analysis of English varieties possible, especially for correlative observations across different world English varieties. With the primary aim of collecting material for comparative studies of Englishes worldwide, the ICE project was initiated in 1988 by Sidney Greenbaum, who coordinated it until 1996; the current coordinator is Gerald Nelson (see Greenbaum, 1988; Greenbaum & Nelson, 1996; Nelson, 2012). This computerized database made 'reliable usage-based studies possible and practical' (Bao, 2010a: 1729) and allowed scholars of world Englishes to document the varying properties of Englishes around the world – documentation that was necessary to substantiate the world Englishes model of linguistic variation. Among the first components of the earliest version of ICE is the Singapore component (ICE-SIN), which has been used by many CSE studies, including some of those mentioned earlier. In particular, the portion of ICE-SIN known as Grammar of Spoken Singapore English Corpus (GSSEC), collected between 1998 and 1999 (see Lim, 2001; Lim & Foley, 2004), has been outstandingly useful as a source of colloquial speech style data.

ICE-SIN, however, has not been updated since its compilation; thus, the Singapore English data it makes available are from the 1990s. This fact was one of the motivations for the CoSEM project. As researchers, we strongly felt the need for a digital database that captures how present-day CSE-speakers use CSE. We also recognized the need for a database with *easily accessible* social metadata (i.e., social information transparent in each utterance); the ICE-SIN has similar metainformation, but this information cannot be immediately associated with the utterance, unlike CoSEM. CoSEM will not only complement GSSEC by enlarging the pool of data available on CSE, it will also provide an invaluable addition of contemporary CSE data in a medium that did not exist when ICE-SIN was compiled. The availability of corpus data from two different time frames, namely, 1998–1999 in the case of GSSEC and 2016–2019 in the case of CoSEM, will also make diachronic CSE comparisons possible.

4

Apart from the metadata structure and lack of updates, another drawback of the ICE-SIN has to do with the observer effect. At least in the spoken section, many of the participants in the ICE-SIN project were aware that they were being recorded and, as such, had the tendency to erase 'non-standard' English features, introducing bias. Evidence for this is found in the corpus itself (e.g., limited instances of *lah*, a quintessential CSE feature). CoSEM has the advantage of reducing if not eliminating this bias.

3 | DATA AND METHODOLOGY

The current version of CoSEM comprises nearly 3.6 million words (around 900,000 lines) of online text messaging data. It is a compilation of what McWhorter (2013) calls 'finger speech' – also known as textspeak or texting language – collected between 2016 and 2019 from the messaging platform WhatsApp, at the National University of Singapore. Most of the data were collected as part of a project done by students enrolled in an advanced sociolinguistics class taught by one of the authors. The majority of the students were women between the ages of 18 and 22; most were also linguistics majors, while the rest were from different disciplines in the humanities and social sciences (psychology, language studies, economics, and so on), science, engineering, and business.

In the project, students were tasked to build a mini corpus consisting of their own existing chat logs from WhatsApp, which they would then use to conduct a linguistic analysis of a unique CSE feature of their interest. They were asked to sample at least 5000 words of text from at least three different group chats, all of which involve multiple speakers (for example, with classmates, with family, friends, and colleagues). We required this to ensure that their databases did not contain idiosyncrasies of chat-group-specific speech patterns. Overall, 500 group chats were sampled by the students for the first version of CoSEM.

The students were also told to collect data only from chats that had begun at least one week before the semester had started. This was to eliminate any undue influence (in the form of observer bias) that could emerge from chat participants knowing their utterances would be used for linguistic analysis. After collecting the data, the students were instructed in how to systematically clean, organize, and tag their data in a spreadsheet document (Excel, Google Sheets, Numbers). They were

5

given a set of guidelines. They were required to (1) include all timestamps of the utterances, (2) record demographic information of each speaker such as age, gender, and race, and (3) replace all media components such as gifs, stickers, embedded videos/audios, and so on, from the raw data with the placeholder *[media omitted]*. Emojis, on the other hand, were kept in original text format, as the spreadsheet software can handle emojis without decoding problems. URLs were kept as is. Students were required to obtain consent from every participant in the collected chat. In cases when participants did not agree to release (part of) their chat logs (or chat metadata), the student removed all instances of utterances from those individuals. After completing these steps, the students uploaded their spreadsheets to a class corpus folder. Prior to submission, they removed information related to personal identifiers from their contribution to ensure participant anonymity.

After the semester, the students were invited to donate their mini corpus to CoSEM. Data from students who agreed to release their files were compiled to form the anonymized and socially tagged CoSEM. All mini-corpus data were screened and double-checked by the authors to ensure the accuracy of data entry formats and the removal or anonymization of all personal identifiers, before being exported to the main CoSEM file. These screening and double-checking procedures were necessary in order to secure data compatibility within the corpus; however, they are extremely time-consuming, which is one of the reasons CoSEM is not yet publicly available. In addition, data collection is still ongoing. Our target is a corpus of at least five million words.

4 | COMPOSITION AND DISTRIBUTION

In this section, we detail the preliminary composition and distribution of the data processed thus far. Designed to be suitable for sociolinguistics analysis, CoSEM is tagged with key social information about the participants, namely, age, gender, race, and nationality, as well as metadata such as the year of collection and year of utterance.

6

4.1 Age

In the current database, the youngest participant is 18, and the oldest is 69. Table 1 shows the frequency distribution according to age group (categorical age). The majority of the contributors of CoSEM, as demonstrated in Table 1, were in their 20s; 70.20% of the words or 71.66% of the utterance lines are sourced from this age group. Those below 21 years of age also contributed a substantial portion of the data, amounting to 21.74% of the words or 23.05% of the lines in the corpus. The contribution of older speakers (8.06% of the total words) – those between 30 and 69 – is dwarfed by the contributions of speakers from 18 to 29 years of age (91.94 % of the total words).

TABLE 1 Word and line breakdown (raw and proportion, age group)[1]

Age group	Words	%words	Lines	%lines
18-20	782,122	21.74	206,293	23.05
21-29	2,525,037	70.20	641,247	71.66
30-39	30,535	0.85	7,396	0.83
40-49	60,813	1.69	12,511	1.40
50-59	194,311	5.40	26,894	3.01
60-69	4,343	0.12	538	0.06
Total	3,597,161	100	894,879	100

The corpus is dominated by data from young CSE speakers, because the students' group chats were generally with their same-age peers. The participants from the older-age groups are typically the students' family members and/or colleagues. There are no participants below the age of

7

18 because the students were told to collect data only from participants over 18 at the time of text messaging.

4.2 Gender

The corpus adopts a binary classification of gender following the conventions of traditional variationist sociolinguistics. The classification relies on participants' self-reporting. The breakdown shows a balance of data from males and females (Table 2) – 50% of words and 50.84% of lines come from female participants, while the rest come from males.

TABLE 2 Word and line breakdown (raw and proportion, gender)

Gender	Words	%words	Lines	%lines
Female	1,798,711	50.00	454,926	50.84
Male	1,798,450	50.00	439,953	49.16
Total	3,597,161	100	894,879	100

CoSEM's relatively balanced distribution of data with respect to gender is due to the fact the students were instructed to ensure that their mini corpora comprised gender-balanced data.

The current CoSEM data exclude a few utterances from individuals who identified as gender-fluid, non-binary, or transgender (at different stages of their lives). We decided to remove these speakers' data out of respect for their feelings of not wanting to be labeled as a binary category, and to avoid any possibility of inadvertently revealing sensitive gender identity information.

8

4.3 Gender composition

CoSEM comprises chat logs – a unit of analysis that transcends the word and the line. The corpus preserves chat information, principally that involving the gender make-up of each individual chat. Chats that do not have male participants, for example, are labeled 'female only'. Those that do not have female participants are labeled 'male only'. Chats that have both female and male participants are characterized as 'mixed', and this label is applied even in a chat among, for example, ten participants of whom only one is male. The absence of more gradient types of coding, which would allow a continuous variable of gender composition, is a limitation of CoSEM. The bulk of the corpus consists of 'mixed' chats, which constitute 74.53% of the words and 70.46% of the lines. Female-exclusive chats constitute 18.76% of words and 22.82% of lines of the entire corpus (Table 3).

TABLE 3 Word and line breakdown (raw and proportion, gender composition)

Composition	Words	%words	Lines	%lines
Female only	674,760	18.76	204,244	22.82
Male only	241,446	6.71	60,085	6.72
Mixed	2,680,955	74.53	630,550	70.46
Total	3,597,161	100	894,879	100

4.4 Year of utterance

9

Although the corpus was compiled between 2016 and 2019, it contains data that date back to 2012 (1.45% of words, 1.55% of lines). Most of the data, however, come from 2015 to 2017. The largest set of single-year data is from 2016 (37.28% of words, 37.68% of lines) (Table 4).

TABLE 4 Word and line breakdown (raw and proportion, year of utterance)

Year	Words	%words	Lines	%lines
2012	52,284	1.45	13,892	1.55
2013	147,943	4.11	42,783	4.78
2014	288,799	8.03	77,202	8.63
2015	800,529	22.25	208,563	23.31
2016	1,340,866	37.28	337,180	37.68
2017	930,408	25.87	207,429	23.18
2018	36,332	1.01	7,830	0.87
Total	3,597,161	100	894,879	100

4.5 Race and nationality

The definition of race adopted here is a categorical one – a label that groups individuals' self-reportings such as 'Chinese' or 'Indian'. Nationality, on the other hand, refers to the status of being part of a nation, typically indicated by eligibility to possess a passport from that nation. The race and nationality data tagged in CoSEM are self-reported by participants.

As Table 5 shows, the racial distribution in CoSEM is unbalanced, with data from individuals who identify as Chinese, namely, Chinese Singaporeans, forming the bulk of the corpus (78.81% of total words). The remainder of the corpus data come from individuals who identify as Malay, Indian,

10

Singaporean of other races, Singaporean mixed-race, and non-Singaporean (21.19% of total words). Almost all of the non-Singaporeans are from neighboring ASEAN regions such as Malaysia, Indonesia, Vietnam, and the Philippines; the few others are mainly exchange students attending the university. At this stage, some of the race and nationality tagging, which is inherited from the students' mini corpora, is not consistent. For instance, data from participants who are citizens of the People's Republic of China may be labeled PC, PR, or PRC. We will standardize the abbreviations in the next stage of the CoSEM project.

TABLE 5 Word and line breakdown (raw and proportion, race and nationality) (HK = Hong Konger, MX = Myanmar Chinese, TW = Taiwanese, FR = French, JP = Japanese, KR = Korean, PH = Filipino)

Nationality	Race	Abbreviation	Words	%words	Lines	%lines
Singaporean	Chinese	CH	2,824,988	78.53	713,539	79.74
	Eurasian	EU	8,533	0.24	1,857	0.21
	Indian	IN	362,918	10.09	95,906	10.72
	Malay	MA	308,679	8.58	61,399	6.86
	Chinese-Indian	CI	10,733	0.30	2,549	0.28
	Chinese-Malay	CM	96	0.00	29	0.00
Indian	Indian	II	18,351	0.51	4,589	0.51
Malaysian	Chinese	MC	339	0.01	107	0.01
	Malay	MM	4,467	0.12	520	0.06
	Unidentified	MS	1,002	0.03	303	0.03
Chinese	Chinese	PC, PR, PRC	9,811	0.27	2,846	0.32
	Indonesian	IC	15,253	0.42	3,378	0.38
Other	Chinese	HK, MX, TW	8,162	0.23	1,937	0.22

11

| | Other | FR, JP, KR, PH, and so on. | 23,829 | 0.66 | 5,920 | 0.66 |
| Total | | | 3,597,161 | 100 | 894879 | 100.00 |

The racial imbalance in the corpus can be accounted for by the general racial make-up of Singapore. People of Chinese ancestry form 74.35% of the citizens and permanent resident population of Singapore, while those of Malay ancestry account for 13.43%; of Indian ancestry, 9.00%; and Others, 3.20% (Singapore Department of Statistics, 2019). 'Others' refers to individuals whose racial identification lies outside of the three major racial groups of Singapore, including individuals who are 'Eurasian' or have mixed European and Asian ancestry. While a wide range of nationalities is represented in the corpus, Singaporeans dominate it, at 97.74% (Table 5). This group includes both citizens and non-citizen long-term residents of Singapore.

Although not directly relevant to the CoSEM project, some local language background is pertinent to the sample analyses in the following sections. Before the 1980s, Hokkien was a dominant language in Singapore, and it was at one time the lingua franca of Singapore's Chinese community (Starr & Hiramoto, 2018: 11). Today, Chinese Singaporeans' dominant home languages other than English are reported to be Mandarin (46.1%) or a Southern Chinese language like Cantonese, Hokkien, or Teochew (16.1%) (Singapore Department of Statistics, 2015); speakers of the latter languages are most likely to be senior citizens. Malay Singaporeans' most common home language other than English is Malay, whereas the most dominant home language other than English of those classified by the Singaporean government as Indian is Tamil (37.7%) or a non-Tamil Indian language like Hindi or Marathi (12%) (Singapore Department of Statistics, 2015).

12

From the 1960s, the state introduced a series of initiatives to implement a bilingual policy (Mother Tongue policy) in schools; prior to 1987, not all schools used English as a medium of instruction (Dixon, 2005: 25; Leimgruber 2013). This policy gradually became firmly set, and in 1980, Mandarin became a compulsory second language for Chinese students in English-medium schools. In 1983, most schools shifted their medium of instruction to English – a change that was completed by 1987 (see Starr & Hiramoto 2018). By default, children are now taught, in addition to English, a state-assigned Mother Tongue language associated with their racial group – Chinese children are assigned Mandarin, Indians Tamil, and Malays Malay (Ministry of Education, 2017). It is noteworthy, however, that most individuals' true heritage languages do not match their assigned Mother Tongue. This is perhaps illustrated most starkly with Chinese Singaporeans, whose heritage languages include Hokkien, Cantonese, Teochew and Hainanese but are required to study Mandarin as their Mother Tongue in school (see Lim, Chen & Hiramoto 2021 for discussion).

5 | FORMAT

CoSEM comes in two formats: .txt files primed for concordance software including AntConc, CasualConc (Figure 1), and spreadsheet files for spreadsheet software like Microsoft Excel, Numbers, Google Sheets (Figure 2). Both formats include the sociolinguistic variable tags (Figures 1 and 2).

```
<COSEM:17CF02-1484-24MAM-2016>    Mud

<COSEM:17CF02-1485-24CHM-2016>    Mud

<COSEM:17CF02-1486-22TWF-2016>    omg TY/F/CH/21 and hyhy played

<COSEM:17CF02-1487-21CHF-2016>    The photos r so nice!!

<COSEM:17CF02-1488-24CHM-2016>    hahahah

<COSEM:17CF02-1489-24CHM-2016>    i am sad

<COSEM:17CF02-1490-24CHM-2016>    i am not in it

<COSEM:17CF02-1491-22TWF-2016>    Sobs can we take when I play

<COSEM:17CF02-1492-24MAM-2016>    No

<COSEM:17CF02-1493-24CHM-2016>    can

<COSEM:17CF02-1494-24CHM-2016>    no problem

<COSEM:17CF02-1495-24CHM-2016>

<COSEM:17CF02-1496-21CHF-2016>    me too pls
```

FIGURE 1 **.txt format of CoSEM**

NUSID	Chat	FileID	Line#	Age	Race	Sex	Initials	UserID	Makeup	Year	Plac	<COSEM:FileID -Line#-UserID-	Excerpt
A0164269W	A	19MM01	1	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-1-19MAM-	Finally changed to iphone guise
A0164269W	A	19MAF01	2	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-2-20MAF-20	BAIK ji
A0164269W	A	19MM01	3	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-3-19MAM-	But i lost half of my contacts lmao
A0164269W	A	19MF01	4	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-4-20MAF-20	apparently they're not ours
A0164269W	A	19MF01	5	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-5-20MAF-20	hahaha
A0164269W	A	19MF01	6	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-6-20MAF-20	retrieve from sim ah
A0164269W	A	19MM01	7	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-7-19MAM-	Actually
A0164269W	A	19MM01	8	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-8-19MAM-	Including urd
A0164269W	A	19MF01	9	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-9-20MAF-20	how rude
A0164269W	A	19MF01	10	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-10-20MAF-2	ok
A0164269W	A	19MF01	11	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-11-20MAF-2	bye
A0164269W	A	19MF01	12	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-12-20MAF-2	don't talk to strangers
A0164269W	A	19MM01	13	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-13-19MAM	I managed to save like since j2 onwards
A0164269W	A	19MM01	14	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-14-19MAM	N i knew u guys since j1
A0164269W	A	19MM01	15	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-15-19MAM	So lmao
A0164269W	A	19MF01	16	20	MA	F		20MAF	Mixed	2016	[]	<COSEM:19MF01-16-20MAF-2	aku touched
A0164269W	A	19MM01	17	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-17-19MAM	Hehe
A0164269W	A	19MF01	18	19	MA	F		19MAF	Mixed	2016	[]	<COSEM:19MF01-18-19MAF-2	<emoji>
A0164269W	A	19MF01	19	19	MA	F		19MAF	Mixed	2016	[]	<COSEM:19MF01-19-19MAF-2	What colour
A0164269W	A	19MM01	20	19	MA	M		19MAM	Mixed	2016	[]	<COSEM:19MM01-20-19MAM	Space grey

FIGURE 2 **Spreadsheet format of CoSEM**

In the .txt format of the corpus, every line of utterance has been tagged with an identifier; (1) shows the format, while (2) provides an example. This format allows for easy identification of a line of utterance within the corpus, and for easy interpretation of relevant metadata. For example, in (2), the tag <17CF15-40341-20CHF-2016> shows that the utterance was collected in the year 2017 by

14

a Chinese female with identification number 15; the utterance is line 40341 in the corpus; and the line was produced by a 20-year-old Chinese Singaporean female in 2016.

(1)

<YearCollected-RaceOfCollector-GenderOfCollector-IdNumber-LineNumber-Age-Race-Gender-YearOfUtterance>

(2)   Are you gg study's

<17CF15-40341-20CHF-2016>

The .txt format of CoSEM is designed to make the corpus accessible to scholars who are accustomed to the 'standardized' corpus format used by widely recognized corpora such as ICE. Most traditional corpora can be characterized as a set of folders with numbered files consisting of text with individual lines, each tagged with a unique utterance tag, such as the one presented in (2). Researchers can simply load the corpus directory to any software that requires this text format, such as WordSmith or AntConc. If they want to investigate only a specific set of data as in 'Male only', they can choose to load specific folders. For instance, those interested in differences between Chinese and Indian CSE speakers could select the race version of the CoSEM .txt format and load the 'Chinese' and 'Indian' folder.

In the spreadsheet format of CoSEM, the data are organized into columns coded by the metadata and social information of participants. Thus, every utterance is preceded by information such as the speaker's age, race, gender, gender make-up of chat, year of utterance, and so on. This organization allows researchers to capitalize on spreadsheet tools like sorting and filtering, which can enable scholars to easily acquire the data they are interested in. For instance, a scholar interested in creating a sub-corpus of male CSE messaging could filter out lines that are not male speech. Again, this is possible because each individual line is tagged with background information. The column format of the spreadsheet is also useful for scholars who want to run statistical analyses,

15

as most statistical software requires coded data to be in 'long format', where each row corresponds to a single observation in a distinct category.

6 | COSEM IN ACTION: SAMPLE SOCIOLINGUISTIC ANALYSES

In this section, we demonstrate how CoSEM can be used for sociolinguistic analysis of CSE. We focus on the (continuous) sociolinguistic variables of age and year of utterance, showing how these relate to the use of one newly reported feature, the *(bo)jio* construction, as well as one well-discussed feature, sentence-final adverbs. We acknowledge that the statistical methods in our analyses may not be adequately sophisticated (e.g., use of multifactorial statistical models, see Gries 2018) despite the rich sociodemographic annotation of the data. However, the following analyses are only meant to be exploratory, in line with the scope of our current paper.

6.1 Age as a factor in the use of *(bo)jio*

Studies considering diachronic aspects of CSE features typically take an apparent time approach. Ziegeler (1995), for instance, used two main age groups of students (secondary school and university) in order to model the grammaticalization of counterfactual implicatures. Gries et al. (2018) used a corpus-based approach, comparing Singapore English data from various decades going back to the 1950s in order to account for change over time in the genitive alternation in Singapore English. In so doing, they took data from both ICE-GB and ICE-SIN and complemented it with a historical corpus of Singapore English to show that inferring diachronic developments from cross-varietal comparisons of synchronic corpora can be misleading. By contrast, Siemund and Li (2017) compared ICE-SIN with a corpus of oral history interviews conducted by the National Archives of Singapore. Their real-time study investigated the use of aspectual *already* and additive *also* as well as language attitudes from speakers born as far back as 1905. In our own corpus, the diachronic dimension can be probed by apparent time methods, based on the informants' age at the time of data collection. The present section exemplifies this approach with the construction *(bo)jio*.

16

The construction *bo + jio* derives from Hokkien (*bô* 'not' and *chio* 'to invite'), the variety of Southern Min that long served as a lingua franca among Chinese in Singapore. Hokkien remains an important source for many lexical items in contemporary CSE, even for speakers without a command of Hokkien. The term *bojio* has joined the general lexicon of CSE within the last decade or so. Its most basic use can be described as a jocular complaint about not having been invited, an interjection signaling willingness to join, or a request to join in on a given enterprise.

(3)   *A:*   *Study sleepover*

          A: *For the girls only*

          B: *Omg i dont mind!*

          C: **Bojio**

                'No invite!'

< 17CF11-15626-20CHF-2015>

through <17CF11-15629-20CHM-2015>

(4)   She will always not **jio** us

'She never invites us.'

<17CF15-18820-21CHF-2015>

(5)   Wanted to **jio** you to go running tmr

'[I] wanted to ask you to go running tomorrow.'

<17CF11-2961-20CHM-2015>

(6)   Please respond to the **jio**

'Please respond to the invitation.'

<17CF11-1057-20CHF-2015>

(7a)   Damn ;(but we **jio**ed u a month back for this

'Damn, but we invited you a month ago for this.'

<17CF11-21675-20CHF-2015>

(7b)   finally aware of the phoenix frisbee **jio**s

'[I'm] finally aware of the phoenix frisbee invitations.'

<17CF02-1105-22CHF-2016>

(8)   I think [he] **jio**ed Valerie out

'I think [he] asked Valerie out.'

<17CF11-2970-20CHF-2015>

In (3), a mixed-gender group of ethnically Chinese 20-year-olds discusses examination revision, and female participant A proposes a girls-only 'study sleepover'. Female participant B responds enthusiastically. Participant C, male, attempts humor by interjecting *bojio*, which in this case functions as a request to be invited too. While the bimorphemic expression *bojio* has become common in CSE, notably as a playful expression used among friends and/or youths, bare *jio* is also found in the data. Often, as in (4), the verb remains negated in one way or another, which is reminiscent of its polarity in the original *bojio* sequence. Regardless of polarity, bare *jio* can appear both in its original verbal meaning 'to invite' (5) and in nominalized form (6). In both cases, *jio* can optionally undergo morphological marking, as exemplified in (7) and (8). Bare verbal *jio* can further be used in phrasal verb constructions such as (8).
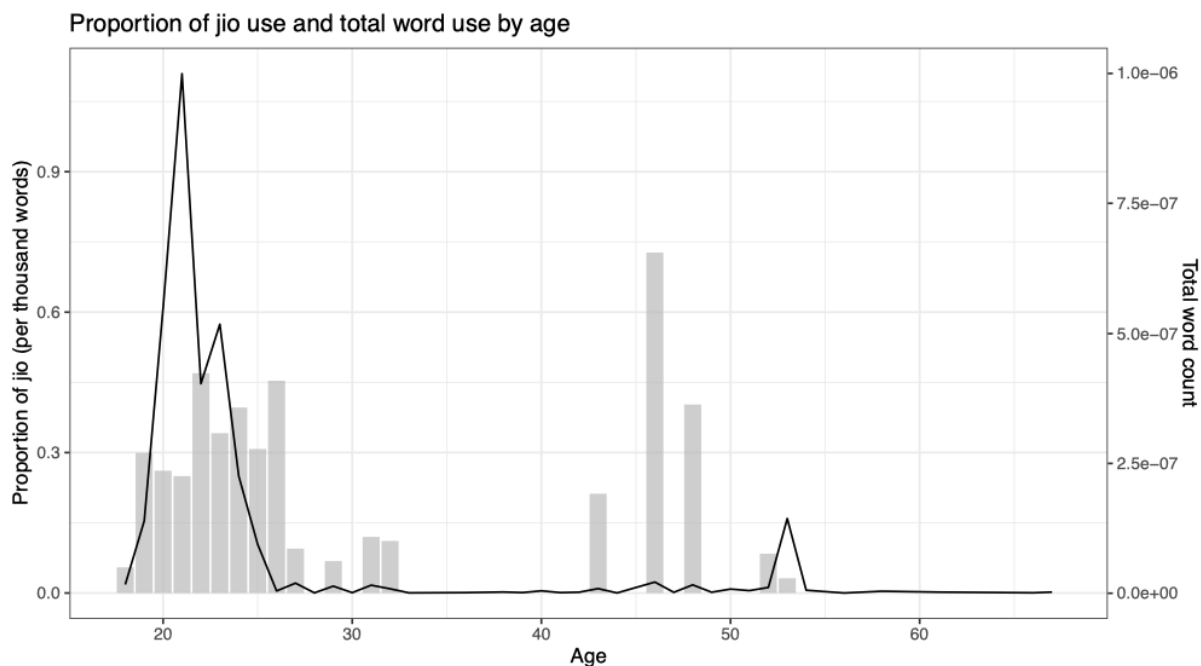
18

Proportion of jio use and total word use by age

FIGURE 3 **Proportion of (bo)jio tokens per 1,000 words by age (bars) in relation to total word count by age (line).**

Figure 3 shows that a certain amount of age-grading can be observed in the use of *(bo)jio*. Its occurrences cluster predominantly in the younger age group, although some participants in the 40–50 age bracket also use it at high rates. The latter are restricted to three individual high-frequency users whose overall contribution to the corpus is minimal. Interpreted through a diachronic apparent-time lens, the overall pattern suggests that *(bo)jio* is a comparatively recent innovation in CSE. More striking perhaps are the trends observed in Figure 4 (Absolute frequencies in Table 6): the nominal form of *jio* is clearly restricted to the under-25 age group, whereas users beyond that age favor the verbal form, suggesting that it is specifically the nominal *(bo)jio* that is an innovation of the younger informants in our sample. The interjection *bojio*, on the other hand, shows a fairly stable distribution.
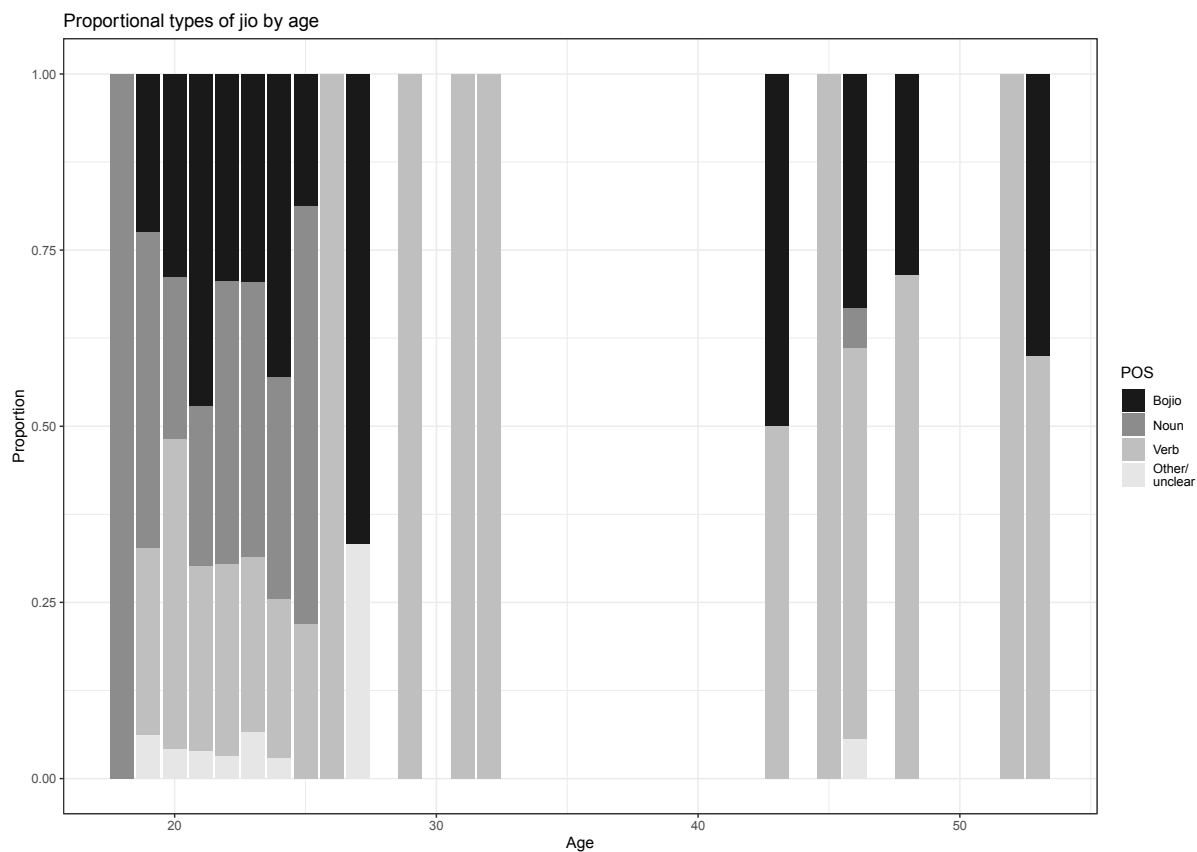
19

Proportional types of jio by age

FIGURE 4 **Types of (bo)jio, proportional, by age.**

TABLE 6 Absolute frequencies of *jio*

Age	Jio	Words	Frequency per thousand words	Age	Jio	Words	Frequency per thousand words
18	1	18235	0.055	43	2	9432	0.212
19	46	153616	0.299	44	0	100	0.000
20	159	608583	0.261	45	1	226	4.425
21	277	1109332	0.250	46	17	23377	0.727
22	210	447264	0.470	47	0	1444	0.000
23	196	574136	0.341	48	7	17385	0.403
24	99	249828	0.396	49	0	1673	0.000

20

25	32	104005	0.308		50	0	8381	0.000
26	2	4410	0.454		51	0	5103	0.000
27	2	21092	0.095		52	1	11912	0.084
28	0	291	0.000		53	5	159090	0.031
29	1	14679	0.068		54	0	5944	0.000
30	0	756	0.000		55	0	0	NA
31	2	16684	0.120		56	0	49	0.000
32	1	8994	0.111		57	0	0	NA
33	0	292	0.000		58	0	3832	0.000
34	0	0	NA		59	0	0	NA
35	0	0	NA		60	0	0	NA
36	0	824	0.000		61	0	1921	0.000
37	0	0	NA		62	0	0	NA
38	0	2112	0.000		63	0	0	NA
39	0	873	0.000		64	0	0	NA
40	0	4626	0.000		65	0	0	NA
41	0	822	0.000		66	0	428	0.000
42	0	1728	0.000		67	0	1994	2.000

The diachronic depth offered by the data in CoSEM is, of course, limited by the ages of our informants and does not permit generalizations into the past beyond 50 years. The imbalance towards young informants (as reported in Table 1) is another drawback. Notwithstanding these concerns, the mere presence, in our corpus, of age information on each informant is a significant improvement over existing corpora like ICE-SIN where these data are simply absent. Further, the availability in CoSEM of informal CMC data from age groups above 40 allows for some cautious investigation of age effects in language change.

21

6.2 Increasing stabilization of sentence-final adverbs over time

The adverbs *already*, *also*, and *only* tend to occur clause-finally in CSE (Bao & Hong, 2006; Parviainen, 2012; Hiramoto, 2015; Cheong, 2016), as seen in (9) to (11). In keeping with previous literature, we continue to refer to them as sentence-final adverbs (SFAs), while describing them as clause-final.

(9)   Oh okay. Im at fifth floor **alrdy**

        'Oh okay. I'm already at the fifth floor.'

<18MF02-5714-23INF-2013>

(10)   I have alot of things to pass to you **also** haha

        'I also have a lot of things to pass to you haha.'

<17CF34-10659-21CHF-2012>

(11)   I ate one **only**

        'I only ate one.'

<18CF55-44567-50CHF-2017>

The data comprise 100 sentences containing *already* from each year from 2013 to 2018, 100 sentences containing *also* from each year from 2012 to 2018, and 100 sentences containing *only* from each year from 2012 to 2018, all randomly selected. The data were pruned to remove singletons, nominal modifiers, set phrases, and idiomatic usages, before each adverb in each sentence was manually coded for clause-finality.

22

Comparisons with the spoken component of the British National Corpus (BNC) (Table 7) and ICE-SIN (Table 8) show that the while SFAs are not impossible in 'inner-circle' varieties such as British English, the high frequency of use of SFAs is a distinctive feature of CSE, and that their use is increasing and stabilizing within the variety itself.

TABLE 7 Number of sentence-final adverbs in CoSEM and BNC

Adverb		CoSEM	BNC	Chi-Square p-Value
Already	SF	453 (77.17%)	75 (15.3%)	$p < 0.01$
	NSF	134 (22.83%)	414 (84.7%)	
Also	SF	250 (43.86%)	8 (0.02%)	$p < 0.01$
	NSF	320 (56.14%)	378 (99.98%)	
Only	SF	120 (22.18%)	4 (0.01%)	$p < 0.01$
	NSF	421 (77.82%)	313 (99.9%)	
Total		1698	1192	

In Table 7, random sampling of *already*, *also*, and *only* from CoSEM and BNC reveals that there is a greater tendency (p-value < 0.01) for all three adverbs to occur sentence-finally in CoSEM than in BNC, thus corroborating earlier claims that SFAs are a prominent feature of CSE.

TABLE 8 Number of sentence-final adverbs in CoSEM and ICE-SIN

Adverb		CoSEM	ICE-SIN	Chi-Square p-Value
Already	SF	453 (77.17%)	199 (68.4%)	$p < 0.01$

	NSF	134 (22.83%)	92 (31.6%)	
Also	SF	250 (43.86%)	156 (44.8%)	$p > 0.05$
	NSF	320 (56.14%)	192 (55.2%)	
Only	SF	120 (22.18%)	39 (15.7%)	$p < 0.01$
	NSF	421 (77.82%)	259 (84.3%)	
Total		1698	937	

As Table 8 shows, sentence-final *already* ($p < 0.01$) and *only* ($p < 0.01$) appear significantly more frequently in CoSEM than in ICE-SIN, suggesting a shift over the two-decade period between the compilation of ICE-SIN the 1990s) and the compilation of CoSEM. It is to be noted that we are aware that ICE-SIN and CoSEM are not directly comparable due to the differences in data type. Hence, the analysis here is suggestive but by no means conclusive. Meanwhile, there is no significant difference in the frequency of sentence-final *also* in CoSEM and in ICE-SIN.

One explanation for the observations in Table 8 comes from the fact that sentence-final counterparts for *already* and *only*, but not *also*, can be found in the substrate varieties of CSE such as Hokkien, Cantonese, and Malay (see Hiramoto, 2015 for discussion). Some of these features have also been fully borrowed into CSE as sentence-final particles – for example, Mandarin *le* (12) and Hokkien *liao* 'already' (13) – thus reinforcing the use of their English counterparts in clause-final position in CSE.

(12)   I'm walking there **le** paiseh!!

'I've already started walking over, sorry!'

<19CF06-452-22CHFGC-2018>

(13)   Damn long never see u **liao**

'I haven't seen you in a very long time.'

24

<19CM12-1511-25CHM-2017>

However, given that Mother Tongue of Chinese Singaporeans is now Mandarin Chinese by default, the increasing frequency and continuing stabilization of clause-final *already* and *only* is more likely attributable to continuing influence from speakers' knowledge of Mandarin Chinese and CSE, rather than of Cantonese and/or Hokkien, as has been claimed for 1990s CSE using data from ICE-SIN (Hiramoto, 2015). Another explanation might be that such adverbs in sentence-final position in CSE are becoming specialized. For example, Cheong (2016) and Erlewine (2018) noted that while standardized English *already* has an 'earlier than expected' meaning, CSE sentence-final *already* introduces a presupposition that the prejacent proposition did not hold at a prior time, which gives rise to its completive and inchoative/inceptive meanings, similar to Mandarin verbal *le* and clause-final *le* respectively (Bao, 2005). In recent studies, Teo (2019) also notes that *already* functions more frequently as an inchoative marker than completive marker uses, and Ziegeler (2020) reports that *already* is rapidly turning to be restricted in its functional scope in today's CSE. As SFAs continue to specialize in function, we expect their use to continue to increase and stabilize within CSE.

As these sample analyses indicate, CoSEM is useful in showing innovation in the use of *bo(jio)* in CSE among the younger CSE-speaking group, and in demonstrating how the use of SFAs in CSE has increased and stabilized over time, which in turn provides further evidence that other languages in the linguistic ecology of Singapore continue to influence the grammar of CSE. With the availability of timestamps and demographic data offered by CoSEM, more sophisticated diachronic analyses – such as looking at rates of change of a linguistic variable over time and across different social groups – promise to be fruitful future research directions.

7 | CONCLUSION

There remain certain limitations and challenges in using CoSEM. For one, the data were collected from students in undergraduate linguistics classes. As a result, a degree of bias is inevitable in the demographic composition of our sample. While a fair balance was achieved in terms of gender, it is likely that general participants in the corpus come from the upper socioeconomic classes and

25

educational levels in the country. While it could be argued that group chats have the potential to include members of diverse social strata, that likelihood should not be overestimated; network theory reminds us to be cautious of holding such expectations. In a similar vein, while CoSEM includes data from all three major ethnic groups, the distribution is skewed towards the majority Chinese population (84% of words in CoSEM vs. 74% in the resident population) whereas Malays are massively underrepresented (3% vs. 13%) and Indians are slightly overrepresented (12% vs. 9%). We recently reported our findings on sentence final adverb uses by speakers' gender and ethnicity; we also observed an effect of age differences in one of the particles (Leimgruber et al. 2021). While there are notable variations in speakers' use of CSE, we do admit that there is a bias in our monitor corpus. It is worth noting, however, that other corpora of Singapore English also often exhibit a Chinese bias unless they have been explicitly designed to control for race.

The presence of media other than text in the chats also presents some challenges, above all for qualitative discourse analyses. Images, videos, gifs, animated stickers, and other such non-text material – while an integral part of the messaging experience – were removed during the processing of the dataset. While optimal computer-based input and analysis requires the removal of these non-textual elements, the frequent references to them in the text are thus rendered less transparent. Furthermore, the written, instantaneous, and computer-mediated nature of the data in CoSEM creates a new set of considerations and challenges that researchers of netspeak will be familiar with (see King, 2009). These include the use of orthographies for reasons of economy or stylistic expression, and countless instances of (unintended) auto-corrected language use and typographical errors. These are nontrivial considerations, whether in preparing the data for processing using concordance software or interpreting the data itself. An example of a typographical error that could potentially result in misinterpretation is given in (14a), which the speaker immediately corrects with a following sentence (14b) with the use of an asterisk (*).

(14)   a)   *Bring cable meh*

              'Are you saying I have to bring a cable?'

        b)   *leh

              'Please bring a cable.'

<18CF48-5401-23MAM-2017>

through <18CF48-5402-23MAM-2017>

It is not unreasonable to think that many such errors exist in the corpus, and that many of them are uncorrected, unlike in (14). It is important to keep these confounds in mind, and where acceptability or felicity judgments are integral to the analysis, to consult native CSE speakers to prevent misinterpreting the data or using it erroneously. In this respect, we envision CoSEM and other corpora like it as complementary to data collected through more traditional linguistics research methods such as fieldwork and elicitation, and vice versa.

Despite these disadvantages, the advantages of CoSEM are many, particularly in comparison to other available corpora of CSE. First, it is a relatively large corpus, larger, for instance, than ICE-SIN, which comprises one million words. It is almost as large as the classroom component of the Singapore Corpus of Research in Education (SCoRE), with five million words. Second, CoSEM is a worthwhile new resource for highly informal registers of Singapore English. Even though most of the participants who provided the data in CoSEM come from educated segments of the population, the language in the corpus is decidedly less acrolectal than that found, for instance, in ICE-SIN. The informality of the text messaging context has had a clear effect on the resulting language use, which will allow scholars to investigate traditionally low-frequency forms (for example, discourse particles or other features). A third advantage of CoSEM is the presence of demographic data for all participants in the corpus. The often-lamented absence of clear ethnic or gender information in, for instance, ICE-SIN, is addressed here by having an informative identifier code for every participant. The recency of the data, spanning the years 2012 to 2018, is a further benefit of CoSEM, in that it will allow diachronic comparisons with earlier corpora, such as ICE-SIN. Another strength of CoSEM is that can be used for both 'sequential' and 'non-sequential' analyses. Scholars doing conversation analysis, for example, may find the CoSEM useful as the order of the chat messages are preserved and numbered in the corpus. Likewise, sociolinguists interested in 'non-sequential' variables such as gender, may also benefit from CoSEM because the spreadsheet format allows for filtering of data based on specific social variables. It is also worth pointing out the pedagogical value of the CoSEM compilation procedure, which was carried out as part of a student-led sociolinguistics class. As students provided their own data, they played an active role in discussing and implementing ethical

27

standards, such as in anonymizing the data, and in providing background social information about the participants. In so doing, the students were able to hone their skills in linguistic fieldwork and data collection.

One final issue to be addressed concerns the ethical and practical considerations of compiling CoSEM. As the nature and features of modern communication continue to change with the emergence of newer technologies, existing best practice guidelines (see Wynne, 2005) need to be adapted and adjusted to reflect such changes, and to allow for newer possibilities in corpora creation (Diemer et al., 2016). Several issues, some of them unexpected, that arose in the process of compiling CoSEM involved deciding between the kinds of data (text, emojis, images, and so on) to go into the final mark-up and standardizing the way the data and metadata is presented in the corpus, as the presentation of the original chat logs differed according to participants' phone models, phone operating systems, and versions of WhatsApp, among other factors. Anonymization also proved to be a key challenge, as information that could risk enabling personal identification but that is not usually found in corpora, such as contact numbers and bank account details, was frequently communicated in the WhatsApp chats that constitute CoSEM. To this end, we also hope that this paper will serve as an introduction and guide to the issues and challenges involved in the future creation of similar CMC-based corpora. In sum, this paper has provided an overview of CoSEM, reporting how the data were collected, organized, and compiled. It also demonstrated the efficacy and applicability of CoSEM in sociolinguistic analyses. As of the point of writing, CoSEM has not yet been released to the public; however, we plan to make it publicly available once data collection and screening of the data to ensure anonymity and conformity to ethical standards are completed. We hope that once CoSEM is available, it will inspire scholars and individuals interested in CSE not only to investigate the language itself but also to begin to pay serious attention to how CSE is changing and evolving in the context of the instantaneous communication enabled by existing and emerging technologies. Beyond that, this corpus should help us reveal the complexities of CSE, its dynamics with other languages in its ecology, and its interactions with social factors in a medium that is robustly used in the modern world yet relatively unexplored in relation to (socio)linguistics and allied fields.

NOTE

1. We performed a breakdown of both words and lines in case scholars would be interested in word density information (for example, words per line).

REFERENCES

Alsagoff, L. (2010). English in Singapore: Culture, capital and identity in linguistic variation. *World Englishes, 29*(3), 336–348.

Bao, Z. (2005). The aspectual system of Singapore English and the systemic substratist explanation. *Journal of Linguistics*, *41*, 237–267.

Bao, Z. (2009). *One* in Singapore English. *Studies in Language*, *33*, 338–365

Bao, Z. (2010a). *Must* in Singapore English. *Lingua*, *120*, 1727–1737.

Bao, Z. (2010b). A usage-based approach to substratum transfer: The case of four unproductive features in Singapore English. *Language*, *86*, 792–820

Bao, Z. (2015). *The making of vernacular Singapore English: System, transfer, and filter.* Cambridge: Cambridge University Press.

Bao, Z. & Hong, H. (2006). Diglossia and register variation in Singapore English. *World Englishes*, *25*, 105–114.

Bao, Z. & Wee, L. (1999). The passive in Singapore English. *World Englishes, 18*, 1–11.

Botha, W. (2018). A social network approach to particles in Singapore English. *World Englishes*, *37*, 261–281.

Cheong, P. S. E. (2016). *Sentence-final already and only in Singapore English* [Unpublished BA Honors thesis]. National University of Singapore.

29

Davies, M. (2013). *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE)*. Retrieved from https://www.english-corpora.org/glowbe/.

Davies, M. (2016). *Corpus of News on the Web (NOW): 10 billion words from 20 countries, updated every day*. Retrieved from https://www.english-corpora.org/now/.

Deuber, D., Leimgruber, J. R. E., & Sand, A. (2018). Singaporean internet chit chat compared to informal spoken language: Linguistic variation and indexicality in a language contact situation. *Journal of Pidgin and Creole Languages, 33*, 48–90.

Diemer, S., Brunner, M. L., & Schmidt, S. (2016). Compiling computer-mediated spoken language corpora: Key issues and recommendations. *International Journal of Corpus Linguistics*, *21*, 348–371.

Dixon, L. Q. (2005). Bilingual education policy in Singapore: An analysis of its sociohistorical roots and current academic outcomes. *International Journal of Bilingual Education and Bilingualism*, *8*, 25–47.

Erlewine, M. Y. (2018). A syntactic universal in a contact language: The story of Singlish *already* [Unpublished manuscript].

Greenbaum, S. (1988). A proposal for an international computerized corpus of English. *World Englishes, 7*, 315.

Greenbaum, S., and Nelson, G. (1996). The International Corpus of English (ICE) Project. *World Englishes, 15*, 3–5.

Gries. S. T. (2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies, 1*, 276-308.

Gries, S. T., Bernaisch, T., & Heller, B. (2018). A corpus-linguistic account of the history of the genitive alternation in Singapore English. In S. C. Dehors (Ed.), *Modeling world Englishes: Assessing the interplay of emancipation and globalization of ESL Varieties* (pp. 245–280). Amsterdam: Benjamins.

Gupta, A. F. (1994). *The step-tongue: Children's English in Singapore.* Clevedon: Multilingual Matters.

Hiramoto, M. (2012). Pragmatics of the sentence-final uses of *can* in Colloquial Singapore English. *Journal of Pragmatics*, *44*, 890–906.

Hiramoto, M. (2015). Sentence-final adverbs in Singapore English and Hong Kong English. *World Englishes*, *34*, 636–653.

Hiramoto, M. (2019). Colloquial Singapore English in advertisements. *World Englishes 38*, 450-462.

Hiramoto, M. & Sato, Y. (2012). *Got*-interrogatives and answers in Colloquial Singapore English. *World Englishes*, *31*, 198–207.

King, B. (2009). Building and analysing corpora of computer-mediate communication. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 301–320). London: Bloomsbury Publishing.

Kwan-Terry, A. (1989). The specification of stage by a child learning English and Cantonese simultaneously. In H. W. Dechert & M. Raupach (Eds.), *A study of acquisitional processes, interlingual processes* (pp. 33–48). Tübingen: Gunter Narr Verlag.

Leimgruber, J. R. E. (2012). Singapore English: An indexical approach. *World Englishes 31*, 1–14.

Leimgruber, J. R. E. (2013). *Singapore English: Structure, variation, and usage*. Cambridge: Cambridge University Press.

Leimgruber, J. R. E. (2014). Singlish as defined by young educated Chinese Singaporeans. *International Journal of the Sociology of Language*, *230*, 45–63.

Leimgruber, J. R. E. (2016). *Bah* in Singapore English. *World Englishes*, *35*, 78–97.

Leimgruber, J. R. E. (2018). Itineracy immobilised: The linguistic landscape of a Singaporean hawker centre. *Linguistic Landscape*, *4*, 178–199.

Leimgruber, J. R. E., Lim, J. J., Gonzales, W. D. W. & Hiramoto, M. (2021). Ethnic and gender variation in the use of Colloquial Singapore English discourse particles. *English Language and Linguistics*. doi:10.1017/S1360674320000453

Leimgruber, J. R. E., Siemund, P. & Terassa, L. (2018). Singaporean students' language repertoires and attitudes revisited. *World Englishes*, *37*, 282–306.

31

Lim, J. J., Chen, S. C. & Hiramoto, M. (2021). "You don't ask me to speak Mandarin, okay?" Ideologies of language and race among Chinese Singaporeans. *Language & Communication*, *76*, 100–110.

Lim, L. (2001). *Towards a reference grammar of Singapore English* [Final Research Report]. National University of Singapore.

Lim, L. (2007). Mergers and acquisitions: On the ages and origins of Singapore English particles. *World Englishes*, *26*, 446–473.

Lim, L. (2009). Revisiting English prosody: (Some) New Englishes as tone languages? *English World-Wide*, *30*, 97–118.

Lim, L. (2015). Coming of age, coming full circle: The (re) positioning of (Singapore) English and multilingualism in Singapore at 50. *Asian Englishes*, *17*, 261–270.

Lim, L, & Foley, J. (2004). English in Singapore and Singapore English: Background and methodology. In L. Lim (Ed.), *Singapore English: A grammatical description* (pp. 1–18). Amsterdam: John Benjamins.

Low, E. L. (2015). The NIE Spoken Corpus of English in Asia (NIESCEA). Singapore: National Institute of Education, Nanyang Technological University.

McWhorter, J. (2013, February). *Txtng is killing language. JK!!!* [Video]. TED Conferences. Retrieved from https://www.ted.com/talks/john_mcwhorter_txtng_is_killing_language_jk/discussion#t-1876

Ministry of Education. (2017). General information on studying in Singapore. Retrieved from http://www.moe.gov.sg/admissions/returning - singaporeans/general - information - on - studying-in-singapore

Nelson, G. (2012). International Corpus of English. Retrieved from http://ice-corpora.net/ice/index.htm

Parviainen, H. (2012). Focus particles in Indian English and other varieties. *World Englishes*, *31*, 226–247.

32

Platt, J., & Weber, H. (1980). *English in Singapore and Malaysia.* Oxford; Oxford University Press.

Platt, J., Weber, H., & Ho, M. L. (1983). *Singapore and Malaysia.* Amsterdam: John Benjamins.

Richards, J. C. & Tay, M. W. J. (1977). The *la* particle in Singapore English. In W. Crewe (Ed.), *The English language in Singapore* (pp. 141–156). Singapore: Eastern Universities Press.

Siemund, P. & Li, L. (2017). Towards a diachronic reconstruction of Colloquial Singapore English. In D. Ziegeler & Bao Z. (Eds.) *Negation and contact: With special focus on Singapore English* (pp. 11–32). Amsterdam: Benjamins.

Siemund, P., Schulz, M. E., & Schweinberger, M. (2014). Studying the linguistic ecology of Singapore: A comparison of college and university students. *World Englishes, 33*, 340–362.

Singapore Department of Statistics. (2015). *Singapore residents by age group, ethnic group and sex, end June, annual.* Retrieved from https://www.singstat.gov.sg/

Singapore Department of Statistics. (2019). *Singapore residents by age group, ethnic group and sex, end June, annual.* Retrieved from https://www.singstat.gov.sg/

Starr, R. L. (forthcoming). Changing language, changing character types. In L. Hall-Lew, E. Moore, & R. J. Podesva (Eds.), *Social meaning and variation: Theorizing the third wave*. Cambridge: Cambridge University Press.

Starr, R. L. & Balasubramaniam, B. (2019). Variation and change in English /r/ among Tamil Indian Singaporeans. *World Englishes*, *38*, 630–643.

Starr, R. L. & Hiramoto, M. (2018). Inclusion, exclusion, and racial identity in Singapore's language education system. *International Journal of Applied Linguistics*, *29*, 341–355.

Teo, M. C. (2019). The role of parallel constructions in imposition. A syntactic study of already in Colloquial Singapore English. *Journal of Pidgin and Creole Languages*, 34(2): 347-377.

Tongue, R. K. (1979). *The English of Singapore and Malaysia* (2nd ed.). Singapore: Eastern Universities Press.

Wee, L. (2003). *The birth of a particle: Know in Colloquial Singapore English*. *World Englishes*, *22*, 5–13.

33

Wong, J. O. (2014). *The culture of Singapore English.* Cambridge: Cambridge University Press.

Wynne, M. (Ed.). (2005). *Developing linguistic corpora: A guide to good practice*. Oxbow Books. Retrieved from http://users.ox.ac.uk/~martinw/dlc/index.htm

Ziegeler, D. (1995). Diachronic factors in the grammaticalization of counterfactual implicatures in Singaporean English. *Language Sciences*, *17*, 305–328.

Ziegeler, D. (2015). *Converging grammars: Constructions in Singapore English.* Berlin: Mouton de Gruyter.

Ziegeler, D. (2020). Changes in the functions of already in Singapore English: A grammaticalization approach. *Journal of Pidgin and Creole Languages, 35*, 293–331.