Ethnic and gender variation in the use of Colloquial Singapore English discourse particles[1]

JAKOB R. E. LEIMGRUBER
University of Freiburg

JUN JIE LIM
University of California, San Diego

WILKINSON DANIEL WONG GONZALES
University of Michigan

and

MIE HIRAMOTO
National University of Singapore

Discourse particles are among the most commented-upon features of Colloquial Singapore English (CSE). Their use has been shown to vary depending on formality, context, gender and ethnicity, although results differ from one study to another. This study uses the *Corpus of Singapore English Messages* (CoSEM), a large-scale corpus of texts composed by Singaporeans and sent using electronic messaging services, to investigate gender and ethnic factors as predictors of particle use. The results suggest a strong gender effect as well as several particle-specific ethnic effects. More generally, our study underlines the special nature of the grammatical class of discourse particles in CSE, which is open to new additions as the sociolinguistic and pragmatic need for them develops.

Keywords: discourse particles, Singapore English, ethnicity, gender, variation

1 Introduction

Among the better-known features of Colloquial Singapore English (CSE, also known as Singlish) are its discourse particles: derived from various substrate languages, they are monosyllabic and typically restricted to the clause-final position (see e.g. Wee 2003; Starr forthcoming). These particles have been subjected to extensive scholarly attention focusing on their origins (e.g. Lim 2007), their syntax and pragmatics (e.g. Gupta

2006; Hiramoto 2012), and their ethnic and social distribution (e.g. Platt 1987; Khoo 2012; Smakman & Wagenaar 2013). In their role as discourse markers, they are excellent diagnostic features of informality and have achieved something of a stereotypical status, often drawn upon for the purpose of indexing national identity. This is perhaps best illustrated by the presence of such particles (particularly the hyper-stereotypical *lah*) in pop-culture artefacts, including internet memes, as well as on products such as t-shirts and other touristic souvenirs. While many of these particles have been described in a fair bit of detail, there remain uncertainties as to their exact behaviour in use and their variation across social factors. Ethnicity is one such factor which has attracted some attention (Platt 1987; Begum & Kandiah 1997; Gupta 2006; Leimgruber 2009; Smakman & Wagenaar 2013; Botha 2018), but the corpora involved were often small and not revealing enough for a proper variationist analysis. This is what we attempt to address in this article, using CoSEM, the *Corpus of Singapore English Messages*, a large corpus of text messages gathered from 2017 to 2019.

2   Background

The island of Singapore, located in Southeast Asia at the southern tip of the Malay Peninsula, is a small (approximately 720 square kilometres, resident population of around 5 million) multi-ethnic city-state. While settlements of regional importance have existed on the island for several centuries, the history of modern Singapore is typically taken to begin with the arrival of the British in 1819, at which point the island was populated almost entirely by the indigenous Malays, bar a few dozen Chinese (Turnbull 1989). The British invested heavily in their presence there, establishing a tariff-free port on the lucrative and strategically important entrance to the Malacca Straits, a key point on the trade routes linking China and Europe (Wong 2016). As a direct result of the now more supraregional relevance of the port city, migrants began arriving from various places, some brought in more forcefully than others (e.g. convicts out of British India), and some drawn by the colony's economic promise (see e.g. Leimgruber 2013a: 3–4). While the Malays were still in the majority in the first census in 1824, three years later immigrants from southern China had arrived in such numbers that they became the dominant ethnic group in the city (Leimgruber 2013a: 3). These Chinese migrants, however, came from regions within China that were 'perceived as culturally and linguistically distinctive' (Chew 2013: 47). The languages they spoke were typically Southern Min (Hokkien, Teochew, Hainanese) and Cantonese, languages that have low levels of mutual intelligibility, resulting in high internal heterogeneity – as shown, for instance, in the separate school systems run by the respective ethnolinguistic groups (see e.g. Starr & Hiramoto 2019: 5). Mandarin, the present-day official Chinese language of Singapore, was not present in any significant proportion in the early days of the colony. Even after independence, it took several decades for Mandarin to become relevant in the local ecology: it was used as a home language by a mere 0.1 per cent of the population in 1957 (Kuo 1980; see also figure 1), and only after the launch of the Speak Mandarin Campaign in 1979 did
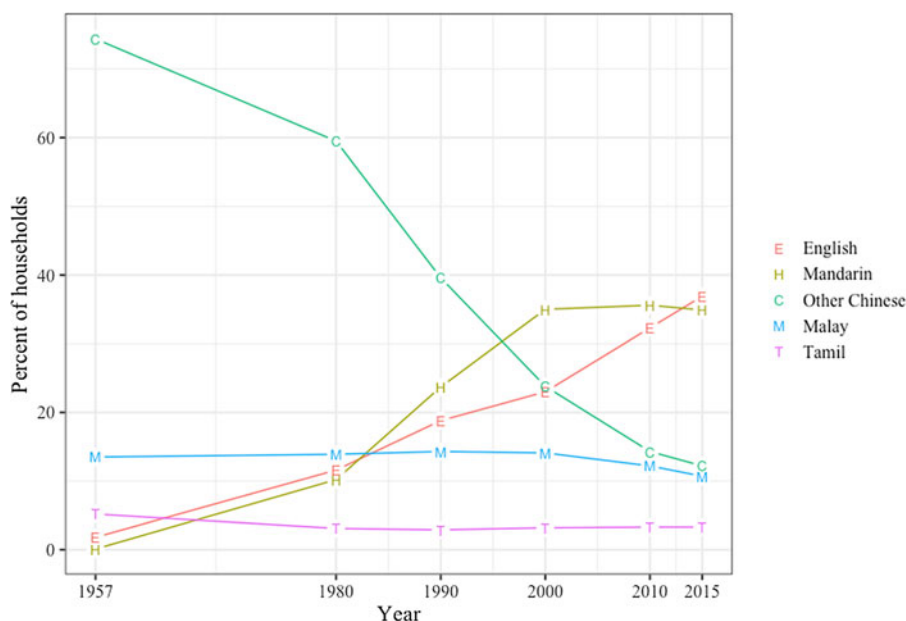
Figure 1. Changes in responses to the census item 'language predominantly used at home', 1957–
2015 (Leimgruber, Siemund & Terassa 2018)

Mandarin gradually establish itself as a serious competitor in Singaporean homes (see e.g. Bokhorst-Heng 1999; Teo 2005; Starr & Kapoor 2020), a language shift largely at the expense of the traditionally spoken varieties of Chinese.

The Malay population of Singapore was relegated to minority status very soon after the foundation of the colony, at 42 per cent by the first census in 1824, and 13 per cent today (Department of Statistics 2016). The Malays came to Singapore from different parts of the Malay Archipelago, and spoke various Malayo-Polynesian languages including Malay, Boyanese and Javanese (Cavallaro & Ng 2020). Chew (2013: 38–43) points out that members of these various ethnic groups, like the Chinese immigrants, originally also considered themselves separate, with a pan-Malay identity emerging only gradually. Unlike Chinese languages among the Chinese, however, the Malay language is firmly entrenched in the Malay population, with 78 per cent reporting its use at home in a 2015 household survey (as opposed to 62 per cent of Chinese reporting using any Chinese variety). Language shift within the Malay community, where it exists, is solely in the direction of English (Cavallaro & Ng 2020); the Malay language, with its official and national status, however, has seen little change in use rates over the last sixty years' censuses (see figure 1).

The third-largest ethnic group with official recognition in Singapore is the Indian group. At 9 per cent, it is the smallest (Department of Statistics 2016) and it is also the most linguistically heterogeneous in that the languages lumped together under the Indian heading come from the Dravidian (Tamil, Telugu, Malayalam, etc.) and

Indo-Aryan (Punjabi, Gujarati, Hindi, etc.) language families. Tamil, the official language, is spoken by 38 per cent of Indian Singaporeans (Starr & Balasubramaniam 2019), and is the default 'mother tongue' taught to Indian pupils at school. Recent reforms to make other Indian languages (Gujarati, Hindi, Punjabi, Bengali and Urdu) available as options have proven most popular in the case of Hindi, which saw its intake double within six years (from 3,771 in 2011 to 7,199 in 2017; Jain & Wee 2019: 7–8). Indians are also the ethnic group with the highest percentage reporting English as a dominant home language (44 per cent in 2015).

English, in general, has had great success in terms of language shift to it from other languages. As can be gleaned from figure 1, while there is not much change (in terms of percentages) for Malay and Tamil, a massive decrease in non-Mandarin varieties of Chinese can be observed, particularly from the 1980s onwards, from 67 to almost 17 per cent. This loss is compensated for by a concurrent rise of Mandarin and English, languages which have seen their share increase dramatically in the same period. In the case of Mandarin, this is clearly the result of a concerted language policy to actively promote Mandarin and actively discourage the so-called 'dialects'.[2] The very visible Speak Mandarin Campaign, with its posters and school-based activities, is just one element of this policy; the structure of the education system, the regulation of language in the media (see e.g. Leimgruber 2013b: 245–6) and pronouncements by the political elite all follow the same goal of increasing Mandarin use and proficiency (Lim, Chen & Hiramoto 2021). English, meanwhile, has experienced similar success, due to other kinds of language policies. English enjoys the unofficial status of the country's 'working language'; it is the sole medium of instruction in schools, and the language of the armed forces, of the government and generally of any white-collar workplace. While language shift towards English is not the stated aim of policymakers (which is 'English-knowing bilingualism', to quote Pakir 1991), educational and socioeconomic pressures combine to make the acquisition of English the most viable goal and the one into which the most resources are invested. The consequence is visible in figure 1: the latest Household Survey from 2015 indicates that English has now overtaken Mandarin as the most common home language in Singapore.

Regardless of the reasons behind this shift, the fact is that English has been, over the last few decades, spoken by an increasing share of the population, either as a 'dominant home language' (as per the census measure), as a first language outside the home, or as an additional language used in a variety of settings and domains. These include, as mentioned, the education system, most workplaces and the armed forces (in which male Singaporeans compulsorily serve for two years). Such high degrees of use in a

---

[2] The term *dialect* to refer to different varieties of Chinese is problematic, as highlighted previously by many scholars such as De Francis (1998: 53–67), Ramsey (1989: 16–17, 28–9) and Mair (1991), among others. Speakers of Chinese, in Singapore and elsewhere, frequently refer to them as dialects, even though they are often mutually unintelligible (Cheng 1996), and some (such as Cantonese) even have their own (more or less) standardised writing system. Our use of 'dialect' (in quotation marks) is an attempt to balance local non-specialist (and language policy) use and linguistic realities.

wide range of settings and domains unsurprisingly led to the emergence of a local vernacular form of English, known as Colloquial Singapore English (CSE, or Singlish), which has spread to large parts of the population.

CSE has been widely described in the literature (see, inter alia, Platt 1975; Gupta 1994; Foley *et al.* 1998; Lim 2004; Low & Brown 2005; Deterding 2007; Leimgruber 2013a). Based on these studies, structural features such as copula-deletion, pro-drop and the *kena* passive construction, among others, can be explained by the variety's continuous contact with Sinitic and Malay languages. It features lexical admixture primarily from Hokkien and Malay; its grammatical structures come mostly from Chinese (or Sinitic) languages. In the phonology, CSE shares features with other Asian Englishes (consonant cluster reduction, final devoicing, non-contrastive vowel length, fewer instances of schwa, etc.). Among the more stereotypical features of CSE are the substrate-derived discourse particles under investigation here; they are widely used and recognised in Singapore. The substrate languages present in the country explained above certainly had a role to play with regard to these particles: Chinese languages feature many such discourse particles (Lim 2007), as does Malay (Yap *et al.* 2016), whereas Indian languages typically do not (but see Baskaran 1988). The existence of CSE has long been considered a problem by policymakers, mostly because it is seen as having a negative influence on Standard English proficiency. The standard form is, of course, the one taught in schools and the one encouraged by authorities as key to maintaining the country's performance in the globalised economy. CSE, on the other hand, is largely used in informal contexts, resulting in what Gupta (1994) identifies as a classic diglossia. In addition to its informal character, the vernacular has also taken on additional societal roles, as it is deployed in language use and conceptualised in terms of cultural orientation (Alsagoff 2010) or indexicality (Leimgruber 2012). The cultural and national indexing facilitated by CSE is also what affords it a certain amount of covert prestige (see the attitudes expressed in, e.g., Leimgruber 2014; Siemund, Schulz & Schweinberger 2014; Leimgruber, Siemund & Terassa 2018), to the extent that a select few features of the variety were even allowed by decision makers to appear in visual displays during the country's fiftieth anniversary of independence in 2015 (Hiramoto 2019: 457). It is fair to say, however, that the basic top-down policy orientation of minimising the use of CSE remains largely intact (Wee 2011).

3 Discourse particles

Given the sociolinguistic context established in the preceding section, we now turn to an explanation of the discourse particles, a central feature of CSE in most descriptions of the variety. The particle *lah*, for instance, appears in virtually all accounts of CSE, going back to the earliest works (e.g. Richards & Tay 1977; Kwan-Terry 1978; Bell & Ser 1983). While *lah* is undoubtedly of non-English origin, it is important to remember that the class of discourse markers in CSE also includes general English markers such as *you know*, *I mean*, *well* or *man* (see e.g. Gupta 1992; Choo 2016). In what follows, however, the focus will be on discourse particles that are of non-English origin and

fulfil the clause-final requirement. This includes the stereotypical *lah*, but a range of others too, illustrated in the following examples of online text messages from the CoSEM corpus (introduced in section 5):

(1) If you feel sad then stop spreading this bad fact about him **la**
'If you feel sad, then just stop spreading this bad fact about him.'
(COSEM:17CF15-35382-21CHF-2015)

(2) Wah u very free **ah**
'Wow, you are very free, aren't you?'
(COSEM:17CF31-4573-19CHF-2016)

(3) I think everyone plan timetable first **bah**
'I think it's best if everyone plans their timetable first.'
(COSEM:17CF07-13341-22CHM-2016)

(4) 30th got school **meh** it's a Sunday **Leh**
'Are you certain there is school on the 30th? It's a Sunday!'
(COSEM:17CF34-211-20CHF-2012)

(5) I find the black emojis dam funny **lor**
'I just find the black emojis damn funny.'
(COSEM:17CF07-8854-21CHF-2015)

(6) Sit where?? Still drizzling **sia** haha
'Where should we sit? It's still drizzling haha.'
(COSEM:17CM02-4764-21CHM-2016)

(7) Quite obvious **what** [3]
'It's quite obvious, isn't it?'
(COSEM:17IF03-3157-20INF-2014)

A primary objective of existing research discussing these particles is usually to describe their semantics and pragmatics. Table 1 gives an overview of some of the more commonly described particles of CSE and their definitions, as provided by various scholars. Some of these definitions are more straightforward than others, and sometimes scholars fail to agree on an exact definition. The particle *leh*, for example, has been reported as the equivalent of 'what about' in questions involving comparisons (Platt 1987), whereas for Wee (2004) it marks tentative suggestions or requests.

The origins of the particles are equally contested. Lim (2007) gives a convincing historical sociolinguistic account of the origins of a selection of particles, identifying two categories that can be distinguished diachronically. The first includes *lah*, *ah* and *wor*, of Malay and Hokkien origin, and are the older particles, whereas *lor*, *hor*, *leh*, *meh* and *mah*, the origins of which Lim traces to Cantonese, are more recent additions

---

[3] The particle *what* is generally considered to be an anglicised respelling of a CSE particle. Both Gupta (1994: 42), who spells it as *wo*, and Lim (2007) argue that the surface similarity with English is a mere coincidence; Lim (2007: 464) considers that the particle may have got its surface form from Cantonese, but fulfils functions more closely associated with those of Hokkien-origin *mah*. By contrast, Kuteva *et al.* (2018) do identify British English as the origin of CSE sentence-final *what*.

Table 1. *Major particles of Colloquial Singapore English and their definitions, in alphabetical order*

Particle	Definition
ah	signals continuation and keeps interlocutors in contact; softens command; marks a question expecting agreement … [or] requiring response (Lim 2007)
bah	marks uncertainty or non-commitment, has a hedging effect when giving advice (Leimgruber 2016)
hor	asserts and elicits support for a proposition (Gupta 1992)
lah	draws attention to mood or attitude and appeals for accommodation; indicates solidarity, familiarity, informality (Lim 2007)
leh	marks a question involving comparison; is equivalent to 'what about?' (Platt 1987); or marks a tentative suggestion or request (Wee 2004)
lor	indicates a sense of obviousness as well as resignation (Lim 2007)
mah	indicates obviousness (Lim 2007)
meh	marks a question involving scepticism (Lim 2007)
sia	serves as an intensifier or to mark coarseness (Khoo 2012); is a marker of casual speech denoting youth identity, coolness (Hiramoto, Lee & Choo 2017)
what/wor	indicates that information is obvious, contradicting something previously asserted (Lim 2007)

to CSE. This idea of discourse particles, which are normally treated as a fairly fixed and closed category, being in fact a much more fluid and dynamic category that new members can readily join, is underscored by more recent work describing the less-documented particles *bah* (Leimgruber 2016) and *sia* (Khoo 2012; Hiramoto, Lee & Choo 2017), which entered the CSE discourse particle category fairly recently. Whereas the Malay origin of *sia* is fairly uncontroversial, the purported Mandarin roots of *bah* point to a shift away from the traditional source languages of CSE particles towards Mandarin, a newcomer to the contact situation that has only gained ground as a home language in the last thirty years or so.

4   Social variation in discourse particles

Less attention has been given, to date, to social variation in the use of the discourse particles. In a corpus of young post-secondary students' English, Leimgruber (2009), for instance, found that ethnicity was a significant predictor of particle use in both formal and informal settings, except between Chinese and Indians in informal settings, where they behaved similarly. The general trends were that in formal settings, the Malays used particles most commonly, followed by the Chinese and then the Indians. In informal settings too, the Chinese and Indians used particles less frequently than the Malays. This corpus-based approach suggests that there are differences in the use of particles across ethnic groups, a suggestion later taken up by Smakman & Wagenaar (2013). Using data from the Singapore component of the *International Corpus of English* (ICE-SIN), they found three types of particles: those that did not exhibit variation by ethnicity (the particles *ah*, *lah* and *meh*); moderate ethnic indicators (such

as *lor*, which was used significantly more frequently by speakers of Chinese ethnicity); and absolute ethnic indicators (such as *hor* and *leh*, which were used exclusively by speakers of Chinese ethnicity).[4] Further, they contrasted particle use in inter- and intra-ethnic settings, and found that over all, there is more particle use in intra-ethnic settings. This result makes intuitive sense, seeing as such intra-ethnic settings would most likely be marked by lower levels of formality than inter-ethnic ones, in turn triggering features of informal speech such as substrate-derived discourse particles.

More recently, Botha (2018) compared particle use in spoken and written data, collecting, on the one hand, text messages, and on the other, recordings of informal face-to-face interaction – both, crucially, from the same group of informants. Among his findings is the fact that the particle *ah* occurs almost twice as often in speech as in writing. Further, he also considered a rather wide range of particles, including some that have not featured extensively in previous research (such as *orh*, *seh* and *nia*),[5] and found that less frequent particles are more common in written than in spoken data. We consider the inclusion of such low-frequency particles noteworthy, as they may be indicative of a natural ongoing language change towards the incorporation of an increasing number of different particles into CSE; we speculate that at least some of the less frequent particles are rather new additions to CSE. In terms of ethnicity, Botha notes that Indians use a smaller range of particles than the other two ethnicities, whereas the Chinese use a greater range of particles than the others. This finding is unsurprising because, as mentioned, Chinese varieties and Malay are known for the use of sentence-final discourse particles while Indian languages are not. Between-group preferences for certain particles were also reported, such as for the particle *orh*, which was used more often by Indians than Chinese but never by Malays, and the particle *eh*, which showed higher frequency among Malays than among Indians and Chinese.[6] Perhaps even more interesting, however, is Botha's account of the social networks in which his informants operate. He found more diversity in the particles used in 'second-order' zones of the networks, a fact explained as a result of the increased social distance between members, resulting in more need for overt identity construction and negotiation, a task to which discourse particles lend themselves particularly well. This, however, is in contrast to previous studies (Leimgruber 2009; Smakman & Wagenaar 2013), which suggest that informality, and therefore closeness within a network, triggers more particle use – the opposite of the explanation offered by Botha (2018).

5   The CoSEM corpus

The corpus from which the data in this article are drawn is the *Corpus of Singapore English Messages* (CoSEM), compiled by a team at the National University of Singapore

---

[4] Note that the particles *lor*, *hor* and *leh*, which were found to be indicators for Chinese ethnicity, are clearly of Chinese origin (Lim 2007: 466), whereas this is less certain for *ah* and *lah*, for instance.

[5] We question the inclusion of *nia* (glossed as 'only' by Botha 2018: 276) among the CSE discourse particles, as it may in fact be fulfilling more complex grammatical functions (Lim *et al*. n.d.).

[6] Botha does not provide etymologies for *orh* and *eh*.

Table 2. *Number of words in the Corpus of Singapore English Messages (CoSEM), by gender and ethnicity, with totals*

	Chinese	Malay	Indian	Other	Total
Male	1,800,194	192,308	339,742	123	2,345,239
Female	1,861,267	194,827	234,098	12,995	2,290,315
Total	3,661,461	387,135	573,840	13,118	4,635,554

(Hiramoto *et al.* n.d.; Gonzales *et al*. n.d.). It consists of WhatsApp group chats provided by undergraduates between 2016 and 2019, although more cohorts are currently being recruited. After having obtained consent from all participants, the data were anonymised and tagged for ethnicity and gender, thereby adding precious sociolinguistic information typically absent from existing corpora of Singapore English. CoSEM currently stands at around 4.6 million words (including emojis), a number that is expected to keep growing as new cohorts are added. Table 2 shows the word counts by gender[7] and ethnicity in the dataset used here. Of note is the usual bias towards the Chinese component, which is a limitation inherent to the data collection process employed, but not unlike other corpora of CSE. Nevertheless, there are a substantial number of words in the Malay and Indian components, which should enhance the corpus's representativeness. In contrast, there is a fairly balanced distribution of words for male and female participants, which will allow for meaningful normalised frequencies.

6   Results

Table 3 gives an overview of ten frequent particles and their distribution by gender and ethnicity. For each ethnic group, the percentage of use among male and female participants is given. The last column shows that overall, in the entire dataset, 45,617 particles are used, mostly by men (58.3%). The actual particles show known trends: *ah* (14,216) and *lah* (13,616) are the high scorers, and *hor* (656) and *wor* (63) are much rarer. This is in line with previous work (e.g. Leimgruber 2009; Smakman & Wagenaar 2013; Botha 2018), and confirms the general pattern of occurrence of the various particles.

6.1   Ethnicity

Interesting trends become visible when taking ethnicity into account. A first overview is given in table 4, which shows individual and overall frequencies (in particles per thousand (‰) words of the respective subcorpus) of use for each particle. Several tendencies can be gleaned from this table. Overall, particle use frequencies differ from one ethnicity to the

---

[7] We are using the self-reported gender categories submitted by the CoSEM participants, in line with traditional variationist and dialectological research methods. We do recognise, however, that gender is not binary; to avoid misrepresentation, the data of the very small number of self-declared 'non-binary' participants are excluded from this study's dataset.

Table 3. *Selected particles in CoSEM, arranged alphabetically. For each ethnic group, the percentage of male and female uses of the particle is given, as well as the number of times the particle appears*

	Chinese			Malay			Indian			Overall		
	♀	♂	N	♀	♂	N	♀	♂	N	♀	♂	N
All	43.9	56.1	37,634	48.8	51.2	3,378	22.3	77.7	5,055	41.8	58.2	46,067
ah	41.0	59.0	10,770	56.8	43.2	1,560	16.6	83.4	1,868	39.5	60.5	14,198
bah	39.6	60.4	1,548	21.6	78.4	37	18.8	81.3	16	38.8	61.2	1,601
hor	47.8	52.1	627	40.0	60.0	15	0.0	100.0	12	46.6	53.4	654
lah	45.8	54.2	10,230	55.9	44.1	989	23.5	76.5	2,361	42.5	57.5	13,580
leh	52.0	48.0	4,700	28.8	71.2	163	52.4	47.6	126	51.1	48.9	4,989
lor	41.3	58.7	2,579	26.9	73.1	67	25.0	75.0	92	40.2	59.8	2,738
mah	50.9	49.1	1,297	8.8	91.2	68	57.4	42.6	61	49.1	50.9	1,426
meh	46.1	53.9	1,458	39.3	60.7	61	17.1	82.9	76	44.1	55.9	1,595
sia	37.7	62.3	4,363	23.9	76.1	418	27.4	72.6	442	35.5	64.5	5,223
wor	53.2	46.8	62	0.0	0.0	0	100.0	0.0	1	54.0	46.0	63

other, with the Chinese in the lead (10.28 particles per thousand words), followed by the Indians (8.81‰) and the Malays (8.73‰). The results of a linear regression implemented in the R environment (R Core Team 2013) and run on these frequencies show that the Chinese do not use significantly more or fewer particles than the rest (β = -0.008275, *SE* = 0.00434, *p* = 0.05654) and neither do the Malays (β = 0.004951, *SE* = 0.004139, *p* = 0.23158); the Indians, however, use significantly fewer particles compared to the

Table 4. *Selected particles in CoSEM, arranged alphabetically. For each ethnic group, the frequency per thousand words of the particle is given, as well as the number of times the particle appears*

	Chinese		Malay		Indian		Overall	
	Freq	N	Freq	N	Freq	N	Freq	N
All	10.28	37,634	8.73	3,378	8.81	5,055	9.97	46,067
ah	2.94	10,770	4.01	1,560	3.26	1,868	3.06	14,198
bah	0.42	1,548	0.10	37	0.03	16	0.35	1,601
hor	0.17	627	0.04	15	0.02	12	0.14	654
lah	2.79	10,230	2.55	989	4.11	2,361	2.93	13,580
leh	1.28	4,700	0.42	163	0.22	126	1.08	4,989
lor	0.70	2,579	0.17	67	0.16	92	0.59	2,738
mah	0.35	1,297	0.18	68	0.11	61	0.31	1,426
meh	0.40	1,458	0.16	61	0.13	76	0.34	1,595
sia	1.19	4,363	1.08	418	0.77	442	1.13	5,223
wor	0.02	62	0.00	0	0.00	1	0.01	63

others (β = -0.01081, SE = 0.004116, p < 0.001). This is in contrast to the findings reported by Leimgruber (2009), where the Malays were the ethnicity found to use particles most frequently.

At the level of individual particles, however, there are notable ethnic effects, as shown in table 4. *Lah*, for instance, seems to be used much more often by the Indians (4.11‰) than by the Chinese (2.79‰) and Malays (2.55‰), a difference that is highly significant (β = 0.365898, SE = 0.084608, p < 0.0001). Closer inspection of this difference reveals that just over half of these Indian occurrences of *lah* come from one and the same chat: a mixed-ethnicity conversation in which seven Chinese and four Indians interact, but in which the Indians produce more text (190,258 words vs 167,250). Within this chat, the *lah* particle is used more frequently by the Indians (6.24‰) than the Chinese (5.70‰), although this difference is not significant (β = -9.18662, SE = 6.55987, p = 0.1614). When individual chat files are taken into account in the general model for *lah*, none of them stands out as having a significant effect on *lah* use, including this one (β = 11.71, SE = 1455, p = 0.993579). Therefore, we can discount any distortion brought about by this one chat file, and maintain that the significant effect observed in the overall corpus holds: Indians are significantly more likely to use *lah* than the other ethnicities.

Regression was used to query any statistically significant differences that might appear in the various particles. Its results are given in table 5. It appears that Chinese ethnicity has a strong effect on particle use: *bah*, *hor*, *leh*, *lor*, *meh* and *wor* are all used significantly more often by Chinese participants. Conversely, *ah* and *lah* are significantly less likely to be used by participants of Chinese ethnicity. The case of *lah*, discussed above, suggests a strong Indian effect. In the case of *sia*, with its postulated origin in the Malay language, the model shows no significant effect for Malay ethnicity (β = -0.115, SE = 0.4731, p = 0.80790); contrariwise, it is again Chinese ethnicity that shows a significant positive association with the particle (β = 1.0272, SE = 0.38286, p < 0.01).

Table 5. *General linear model querying the probability of ethnicity having a significant effect on particle use*

Particle	Base ethnicity	β	Std error	p
ah	Chinese	-0.5404	0.1386	p < 0.001
bah	Chinese	3.26949	0.47919	p < 0.001
hor	Chinese	3.72696	0.58003	p < 0.001
lah	Indian	0.365898	0.084608	p < 0.001
leh	Chinese	1.067	0.288	p < 0.001
lor	Chinese	1.035244	0.313706	p < 0.001
mah	Chinese	0.491192	0.339912	p = 0.14844
meh	Chinese	1.98368	0.41426	p < 0.001
sia	Malay	-0.115	0.4731	p = 0.8079
	Chinese	1.0272	0.38286	p < 0.01
wor	Chinese	2.71486	1.00765	p < 0.01

In the case of *ah*, by far the most frequent particle, the Chinese participants in the corpus show significantly lower rates of use compared to the other ethnicities (β = -0.5404, *SE* = 0.1386, *p* < 0.001), confirming the findings of Botha (2018) but partly refuting Leimgruber (2009) and Smakman & Wagenaar (2013), whose data suggest higher frequencies of *ah* use among Malays.[8] Finally, while *mah* does exhibit a higher frequency of use among Chinese, the effect is not statistically significant (β = 0.491192, *SE* = 0.339912, *p* = 0.14844). This is an interesting finding, considering previous research which found *mah* a marker of Chinese ethnicity (Platt 1987: 395; Leimgruber 2009: 185; Botha 2018: 271). We consider the larger (and more recent) corpus data used in this study to be a more accurate depiction of the ethnic distribution of particles in CSE, suggesting that *mah*, whatever its origin and erstwhile use might have been, has now crossed ethnic boundaries to no longer be directly associated solely with the Chinese group.

6.2   *Gender*

Several complex patterns can be observed with regard to gender (see table 6 and figure 2). The overall behaviour, which stands at 11.50 particles per thousand words among men and 8.40‰ among women, turns out to be highly significant (β = 0.03556, *SE* = 0.001718, *p* = 0.001). The same holds true for most individual particles, even where the difference in frequency is negligible: *mah*, for instance, is used at a comparable rate of 0.3061‰ (women, *N* = 701) and 0.3104‰ (men, *N* = 728), but this small difference

Table 6. *Male and female frequencies per thousand words for each particle; general linear model querying the probability of gender having a significant effect on particle use*

Particle	Freq. F	Freq. M	β	Std error	*p*
ah	2.45	3.67	1.033	0.1139	*p* < 0.001
bah	0.27	0.42	1.52697	0.40301	*p* < 0.001
hor	0.13	0.15	1.65744	0.54662	*p* < 0.01
lah	2.53	3.34	0.685410	0.072778	*p* < 0.001
leh	1.12	1.04	0.3552	0.1707	*p* < 0.05
lor	0.48	0.70	1.574353	0.208952	*p* < 0.001
mah	0.31	0.31	1.229510	0.237490	*p* < 0.001
meh	0.31	0.38	2.04475	0.33009	*p* < 0.001
sia	0.81	1.45	-0.5999	0.4347	*p* = 0.16756
wor	0.01	0.01	1.21649	1.23733	*p* = 0.32553

[8] Leimgruber (2009: 185) reports comparable *ah* usage rates among Chinese (5.73‰) and Indians (5.17‰), but not Malays (12.51‰). The different genres of text used in that study might in part explain this discrepancy. Smakman & Wagenaar (2013) calculated individual particle percentages out of the overall particle use, rendering the two measures not directly comparable.
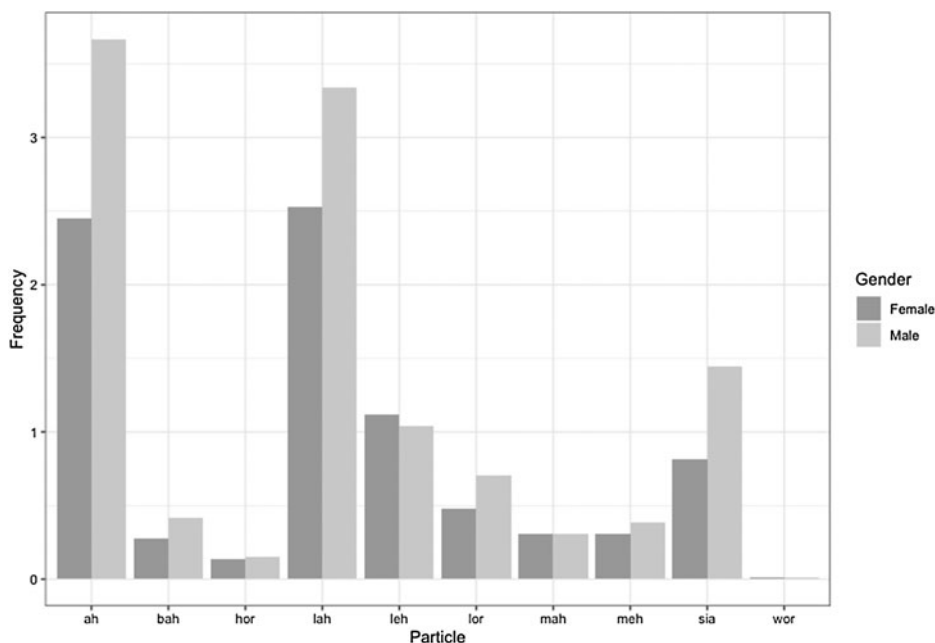
Figure 2. Particles per thousand words, by gender, arranged alphabetically

of 27 particles is nonetheless significant ($p < 0.001$). The opposite effect is seen with *leh*, which women use more often, though at a slightly lower level of significance ($p < 0.05$). Figure 2 suggests that *sia* shows a strong preference among men in frequency, but this difference is not significant ($p = 0.16756$).

Sia, we may add, is not always included in the basic inventory of CSE particles. Khoo (2012) provided one of the earlier treatises on *sia*; she identifies it as having its origin in the Malay swearword *sial*. It has, however, taken on meanings beyond simple swearing to include at least intensification in the form of coarseness and coolness as well as solidarity marking, in which case it indexes a casual stance, or a youth identity (Hiramoto *et al*. 2017). To illustrate these uses, (8) and (9) are examples of intensification, and (10) and (11) of solidarity marking.

(8) Wtf they damn mean **sia**
     'WTF, they are damn mean!'
     (COSEM:17CF34-7510-21CHF-2012)
(9) Actually im very clueless **sia**
     'Actually, I'm very clueless.'
     (COSEM:17CM02-2502-21CHM-2016)
(10) I need this in my life **sia**
     'I need this in my life!'
     (COSEM:17CF12-33682-21CHM-2015)

(11) And it's like less than a month alr **Sia**
  'And it's like less than a month already.'
  (COSEM:17CF25-5413-20MAF-2016)

The CoSEM data show that in the case of *sia*, ethnicity and gender interact strongly. Of all combinations, Malay men (1.65‰) use it most, whereas Malay women (0.51‰) use it least. Malay men, therefore, are significantly more likely to use the particle than any other ethnicity/gender combination ($\beta = 0.6102$, $SE = 0.121$, $p < 0.001$). The difference in frequencies between genders does not appear to be significant in the case of the Chinese (men: 1.51‰ vs women: 0.88‰) and the Indians (men: 0.94‰ vs women: 0.52‰). Our interpretation of these results is that the original Malay-language swearword meaning of *sia* is still latently present in the Malay population, thereby preventing Malay women from using the particle with the same frequency as their male counterparts. Meanwhile, for the Chinese and the Indians, such considerations would not be of consequence, given that proficiency in the Malay language (and thus knowledge of *sial*'s full connotations) is limited to certain lexical items or phraseological constructions among members of the Chinese and Indian communities.

7 Discussion

The findings presented in this article are summarised in table 7. There are several points of interest. Firstly, Chinese ethnicity as well as male gender are positive predictors for most particles, with varying degrees of statistical significance.

Secondly, we note the possible emergence of new ethnic markers, alongside those of Smakman & Wagenaar (2013): *bah* is used almost exclusively by Chinese speakers of CSE, with Malay men quite a distance behind. This is interesting, for at least two reasons: firstly, because *bah* is an unusual particle in that its origins lie in Mandarin (Leimgruber 2016) and not in Hokkien or Cantonese, like the bulk of the other discourse particles. As discussed above, Mandarin only became relevant in the local

Table 7. *Significant predictors for each particle*

Particle	Chinese	Malay	Indian	Male	Female
ah	✓			✓	
bah	✓			✓	
hor	✓			✓	
lah			✓	✓	
leh	✓			✓	
lor	✓			✓	
mah				✓	
meh	✓			✓	
sia		✓		✓	
wor	✓				

linguistic ecology after the 1980s. It is now a widely used home language (dominant in one-third of households in 2015), a change in status that undoubtedly has effects on language attitudes towards the variety – attitudes which are important in considerations of language shift and change (see e.g. Thomason 2001). Secondly, the behaviour of *bah* is somewhat surprising insofar as a particle with a putative Mandarin origin, while still significantly more common among the Chinese, is making its way into non-Chinese groups. Perhaps the close contact of men from different ethnic backgrounds during the compulsory two-year-long National Service has a role to play: the armed forces generally have been found to be fertile ground for linguistic transfer of several kinds (see e.g. Berthele & Wittlin 2013; Akande 2016).

Thirdly, some particles stand out as departing from the dominant Chinese and male effect. *Lah* shows a strong Indian bias. *Mah* is the only particle to not show any significant ethnic influence, whereas in the case of *sia*, the combination of Malay ethnicity and male gender predicts its use. It is safe to say, therefore, that gender and ethnicity are predictors of particle use.

8 Conclusion

The findings in this article complement the claims made by Botha (2018: 271) about the ethnic distribution of particles in respect to overall behaviour, but our findings depart from Botha's in the case of some particles; it is perhaps interesting to note that *ah* was found to be led by Indian users in Botha's study, whereas here it is the Chinese. Differences in methodology and corpus size may well explain these discrepancies. It is worth adding that ethnicity and gender alone do not account for the whole picture of variation in particle use, or in CSE in general: our model also accounted for age, on which we are not focusing in this article for lack of space. It is an important factor too, with, for example, *sia* being used almost exclusively by the younger generation (i.e. people under 30 years of age in CoSEM), an observation that we plan to test in future analyses of the corpus. Further, it appears that discourse particle usage may be an additional piece of evidence in support of Schneider's (2007) claim that Singapore is on its way into the differentiation stage of his dynamic model, in which subnational groups begin to use language differently in order to mark distinct identities. This seems to be the case here, with ethnic groups exhibiting different types of language use. This phenomenon has also been attested in the realm of phonetics; for instance, a study by Starr & Balasubramaniam (2019) found a tapped and trilled variant of /r/ to be used by Tamil Singaporeans as an index of Indianness, when otherwise the approximant [ɹ] would be used. Careful analysis of any sociolinguistic variable in the Singapore context will undoubtedly lead to the concurrent finding of instances of homogenisation and heterogenisation, as highlighted by Buschfeld (2020). The two are certainly not mutually exclusive, and are indicative of a speech community that is stable enough to have a common core, with social subvarieties emerging, as suggested by Schneider's (2007) placing of Singapore in the early stages of the differentiation phase of his model (see also Gonzales & Hiramoto 2020). A combination of such features, whether above

or below the level of consciousness, certainly suggests the presence of ethnicity-related subvarieties of CSE.

Discourse particles have been shown to play important roles in language contact situations (Torres 2002, 2006; Matras 2009: 137–44; Smith-Christmas 2016). Considering, as we do, discourse particles (as well as languages) to be fluid, we propose that they developed in CSE early on as a kind of lubricant required in conversations between speakers of different, mutually unintelligible languages. In other words, discourse particles were useful, pragmatically, for speakers to get their message across. At the same time, discourse particles naturally became part of the feature pool of CSE through the support of substrate languages in which the corresponding particles are well established. It is not surprising, then, that new discourse particles are still being incorporated to this day into CSE by speakers based on their conversational needs.

Authors' addresses:

English Department
University of Freiburg
Rempartstr. 15
79085 Freiburg
Germany
jakob.leimgruber@anglistik.uni-freiburg.de

Department of Linguistics
University of California, San Diego
9500 Gilman Drive #0108
La Jolla, CA 92093-0108
USA
jjlim@ucsd.edu

Department of Linguistics
University of Michigan
Lorch Hall
611 Tappan Street
Ann Arbor, MI 48109-1220
USA
wdwg@umich.edu

Department of English Language and Literature
National University of Singapore
Block AS5
7 Arts Link, FASS
Singapore 117570
Singapore
ellmh@nus.edu.sg

References

Akande, Akinmade. 2016. Multilingual practices in Nigerian army barracks. *African Identities* 14(1), 38–58.

Alsagoff, Lubna. 2010. English in Singapore: Culture, capital and identity in linguistic variation. *World Englishes* 29(3), 336–48.

Baskaran, Lohanayahi. 1988. Aspect of Malaysian English syntax. PhD thesis, University College London.

Begum, Rizwana & Thiru Kandiah. 1997. Misrecognitions of variability in new varieties of English: Tamil minority usage in Singapore English. In Edgar W. Schneider (ed.), *Varieties of English around the world*, vol. G19, 189–205. Amsterdam: John Benjamins.

Bell, Roger T. & Larry Peng Quee Ser. 1983. 'To-day la?' 'Tomorrow lah!' The LA particle in Singapore English. *RELC Journal* 14(2), 1–18.

Berthele, Raphael & Gabriele Wittlin. 2013. Receptive multilingualism in the Swiss Army. *International Journal of Multilingualism* 10(2), 181–95.

Bokhorst-Heng, Wendy. 1999. Singapore's Speak Mandarin Campaign: Language ideological debates in the imagining of the nation. In Jan Blommaert (ed.), *Language ideological debates*, 235–65. Berlin: Mouton de Gruyter.

Botha, Werner. 2018. A social network approach to particles in Singapore English. *World Englishes* 37(2), 261–81.

Buschfeld, Sarah. 2020. *Children's English in Singapore: Acquisition, properties, and use*. Abingdon: Routledge.

Cavallaro, Francesco & Bee Chin Ng. 2020. The language 'cleave': Unravelling the multicultural and multilingual weave of Singapore. In Peter Siemund & Jakob R. E. Leimgruber (eds.), *Multilingual global cities: Singapore, Hong Kong, Dubai* (Multilingual Asia). Singapore: Routledge.

Cheng, Chun-Chuan. 1996. Quantifying dialect mutual intelligibility. In James Cheng-Teh Huang & Audrey Yen-Hui Li (eds.), *New horizons in Chinese linguistics*, 269–72. Dordrecht: Kluwer Academic Publishers.

Chew, Phyllis Ghim Lian. 2013. *Sociolinguistic history of early identities in Singapore: From colonialism to nationalism*. Basingstoke: Palgrave Macmillan.

Choo, Jessica. 2016. One people, one nation, one Singlish? A study on ethnic differences in the use of Singlish particles. BA honours thesis, National University of Singapore.

De Francis, John. 1998. *The Chinese language: Fact and fantasy.* Honolulu: University of Hawai'i Press.

Department of Statistics. 2016. *General household survey 2015*. www.singstat.gov.sg/publications/ghs/ghs2015 (accessed 23 May 2020).

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

Foley, Joseph A., Thiru Kandiah, Zhiming Bao, Anthea Fraser Gupta, Lubna Alsagoff, Chee Lick Ho, Lionel H. A. Wee, Ismail L. Talib & Wendy Bokhorst-Heng (eds.). 1998. *English in new cultural contexts: Reflections from Singapore*. Singapore: Singapore Institute of Management/Oxford University Press.

Gonzales, Wilkinson Daniel Wong, Mie Hiramoto, Jakob R. E. Leimgruber & Jun Jie Lim. n.d. *Corpus of Singapore English Messaging* (CoSEM): Description and explorations. MS, National University of Singapore.

Gonzales, Wilkinson Daniel Wong & Mie Hiramoto. 2020. Two Englishes diverged in the Philippines? A substratist account of Manila Chinese English. *Journal of Pidgin and Creole Languages* 35(1), 125–59.

Gupta, Anthea Fraser. 1992. The pragmatic particles of Singapore colloquial English. *Journal of Pragmatics* 18(1), 31–57.

Gupta, Anthea Fraser. 1994. *The step-tongue: Children's English in Singapore*. Clevedon: Multilingual Matters.

Gupta, Anthea Fraser. 2006. Epistemic modalities and the discourse particles of Singapore. In Kerstin Fischer (ed.), *Approaches to discourse particles*, 243–63. Leiden: Brill.

Hiramoto, Mie. 2012. Pragmatics of the sentence-final uses of *can* in Colloquial Singapore English. *Journal of Pragmatics* 44(6–7), 890–906.

Hiramoto, Mie. 2019. Colloquial Singapore English in advertisements. *World Englishes* 38(3), 450–62.

Hiramoto, Mie, Tong King Lee & Jessica Choo. 2017. The sentence-final particle *sia* in a corpus of Colloquial Singapore English text message data. Paper presented at Methods 16, National Institute for Japanese Language and Linguistics, Tokyo.

Hiramoto, Mie, Wilkinson Daniel Wong Gonzales, Jun Jie Lim & Jakob R. E. Leimgruber. n.d. *Corpus of Singapore English Messages* (CoSEM). Unpublished dataset, National University of Singapore.

Jain, Ritu & Lionel Wee. 2019. Diversity management and the presumptive universality of categories: The case of the Indians in Singapore. *Current Issues in Language Planning* 20(1), 16–32.

Khoo, Velda. 2012. Conversation analytic approach to the *sia* particle in Singapore Colloquial English. Final-year project, Nanyang Technological University.

Kuo, Eddie C. Y. 1980. Singapore: Unity in diversity. In Evangelos A. Afendras & Eddie C. Y. Kuo (eds.), *Language and society in Singapore*, 39–62. Singapore: National University of Singapore Press.

Kuteva, Tania, Seongha Rhee, Debra Ziegeler & Jessica Sabban. 2018. On sentence-final 'what' in Singlish: Are you the Queen of England, or what? *Journal of Language Contact* 11(1), 32–70.

Kwan-Terry, A. 1978. The meaning and the source of the 'la' and the 'what' particles in Singapore English. *RELC Journal* 9(2), 22–36.

Leimgruber, Jakob R. E. 2009. Modelling variation in Singapore English. DPhil thesis, University of Oxford.

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

Leimgruber, Jakob R. E. 2013a. *Singapore English: Structure, variation, and usage*. Cambridge: Cambridge University Press.

Leimgruber, Jakob R. E. 2013b. The management of multilingualism in a city-state: Language policy in Singapore. In Peter Siemund, Ingrid Gogolin, Monika Edith Schulz & Julia Davydova (eds.), *Hamburg studies on linguistic diversity*, vol. 1, 227–56. Amsterdam: John Benjamins.

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

Leimgruber, Jakob R. E. 2016. *Bah* in Singapore English. *World Englishes* 35(1), 78–97.

Leimgruber, Jakob R. E., Peter Siemund & Laura Terassa. 2018. Singaporean students' language repertoires and attitudes revisited. *World Englishes* 37(2), 282–306.

Lim, Jun Jie, Spencer Chen & Mie Hiramoto. 2021. 'You don't ask me to speak Mandarin, okay?' Ideologies of language and race among Chinese Singaporeans. *Language & Communication* 76, 100–10.

Lim, Jun Jie, Mie Hiramoto, Jessica Choo, Wilkinson Daniel Wong Gonzales & Jakob R. E. Leimgruber. n.d. Sentence-final adverbs in Colloquial Singapore English revisited: Increasing frequency and stabilization in a WhatsApp corpus. MS, National University of Singapore.

Lim, Lisa. 2004. *Singapore English: A grammatical description*. Amsterdam: John Benjamins.

Lim, Lisa. 2007. Mergers and acquisitions: On the ages and origins of Singapore English particles. *World Englishes* 26(4), 446–73.

Low, Ee Ling & Adam Brown. 2005. *English in Singapore: An introduction*. Singapore: McGraw-Hill.

Mair, Victor H. 1991. What is a Chinese 'dialect/topolect'? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers* 29, 1–31.

Matras, Yaron. 2009. *Language contact*. Cambridge: Cambridge University Press.

Pakir, Anne. 1991. The range and depth of English-knowing bilinguals in Singapore. *World Englishes* 10(2), 167–79.

Platt, John T. 1975. The Singapore English speech continuum and its basilect 'Singlish' as a 'creoloid'. *Anthropological Linguistics* 17(7), 363–74.

Platt, John T. 1987. Communicative functions of particles in Singapore English. In Ross Steele & Terry Threadgold (eds.), *Language topics: Essays in honour of Michael Halliday*, vol. 1, 391–401. Amsterdam: John Benjamins.

R Core Team. 2013. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. www.R-project.org

Ramsey, Samuel Robert. 1989. *The languages of China*. Princeton, NJ: Princeton University Press.

Richards, Jack C. & Mary W. J. Tay. 1977. The *la* particle in Singapore English. In William Crewe (ed.), *The English language in Singapore*, 145–56. Singapore: Eastern Universities Press.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

Siemund, Peter, Monika Edith Schulz & Martin Schweinberger. 2014. Studying the linguistic ecology of Singapore: A comparison of college and university students. *World Englishes* 33(3), 340–62.

Smakman, Dick & Stephanie Wagenaar. 2013. Discourse particles in Colloquial Singapore English. *World Englishes* 32(3), 308–24.

Smith-Christmas, Cassie. 2016. Regression on the fused lect continuum? Discourse markers in Scottish Gaelic–English speech. *Journal of Pragmatics* 94, 64–75.

Starr, Rebecca Lurie. Forthcoming. Changing language, changing character types. In Lauren Hall-Lew, Emma Moore & Robert J. Podesva (eds.), *Social meaning and variation: Theorizing the Third Wave*. Cambridge: Cambridge University Press.

Starr, Rebecca Lurie & Brinda Balasubramaniam. 2019. Variation and change in English /r/ among Tamil Indian Singaporeans. *World Englishes* 38(4), 630–43.

Starr, Rebecca Lurie & Mie Hiramoto. 2019. Inclusion, exclusion, and racial identity in Singapore's language education system. *International Journal of Applied Linguistics* 29(3), 341–55.

Starr, Rebecca Lurie & Shrutika Kapoor. 2020. 'Our graduates will have the edge': Linguistic entrepreneurship and the discourse of Mandarin enrichment centers in Singapore. *Multilingua*. https://doi.org/10.1515/multi-2020-0033

Teo, Peter. 2005. Mandarinising Singapore: A critical analysis of slogans in Singapore's 'Speak Mandarin' campaign. *Critical Discourse Studies* 2(2), 121–42.

Thomason, Sarah G. 2001. Speakers' attitudes in language change, contact-language genesis and language preservation. *Sociolinguistic Studies* 2(2), 13–26.

Torres, Lourdes. 2002. Bilingual discourse markers in Puerto Rican Spanish. *Language in Society* 31(1), 65–83.

Torres, Lourdes. 2006. Bilingual discourse markers in Indigenous languages. *International Journal of Bilingual Education and Bilingualism* 9(5), 615–24.

Turnbull, Constance M. 1989. *A history of Singapore, 1819–1988*, 2nd edn. Singapore: Oxford University Press.

Wee, Lionel. 2003. The birth of a particle: *Know* in Colloquial Singapore English. *World Englishes* 22(1), 5–13.

Wee, Lionel. 2004. Reduplication and discourse particles. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 105–26. Amsterdam: John Benjamins.

Wee, Lionel. 2011. *Language without rights*. Oxford: Oxford University Press.

Wong, John D. 2016. *Global trade in the nineteenth century: The house of Houqua and the Canton system*. Cambridge: Cambridge University Press.

Yap, Ngee Thai, Mei Yuit Chan, Bee Eng Wong & Li Chia Tay. 2016. Discourse particles in Malaysian English: What do they mean? *Bijdragen tot de taal-, land- en volkenkunde / Journal of the Humanities and Social Sciences of Southeast Asia* 172(4), 479–509.