Received: 22 March 2022 | Revised: 20 June 2022 | Accepted: 27 June 2022

DOI: 10.1111/weng.12597

ORIGINAL ARTICLE

WORLD ENGLISHES WILEY

Tagging Singapore English

Li Lin | **Kunmei Han** | **Jia Wen Hing** | **Luwen Cao** | **Vincent Ooi** | **Nick Huang** | **Zhiming Bao**

Department of English, Linguistics and Theatre Studies, National University of Singapore, Singapore

Correspondence
Zhiming Bao and Nick Huang, Department of English, Linguistics and Theatre Studies, National University of Singapore, Block AS5, 7 Arts Link, Singapore 117570.
Email: bao@nus.edu.sg and znhuang@nus.edu.sg

Funding information
DSO National Laboratories - Singapore ; Humanities and Social Sciences Research, National University of Singapore; Social Science Thematic Research Grant, Ministry of Education, Singapore

Abstract

It is well-known that Outer Circle English has undergone extensive contact-induced lexical and grammatical restructuring. Is it possible to use common NLP tools developed for Inner Circle English to process Outer Circle English texts? Here, we report our experience of using the Stanford PoS tagger to tag the Singaporean component of the International Corpus of English (ICE-SIN). We isolate two major contact-related causes of tagging errors: (1) lexical and grammatical loans directly borrowed from the local languages; and (2) English-origin words with new grammatical meanings acquired from the local languages. While the first type may be easy to overcome, the latter type is intractable, creating an extra layer of morphosyntactic complexity. We achieved comparable accuracy rates in the more formal registers, and a lower but still decent 88% in the informal register of private conversations. A tagged ICE-SIN allows us to investigate lexical and grammatical restructuring at unprecedented levels of detail.

1 | INTRODUCTION

Since Greenbaum (1988) first proposed it, the International Corpus of English (ICE) has served the World English community for well over a quarter of a century, providing valuable data for research on world Englishes, enriching not only the world English literature, but also the contact linguistics literature, and indeed the linguistics literature more generally. A large number of studies, informed by data from the ICE country corpora, have been published in edited volumes (Nelson, Wallis, & Aarts, 2002) and in various linguistics journals, including the generalist journals *Journal of Linguistics* and *Language*. Indeed, *World Englishes*, the flagship journal of the world Englishes community, has devoted no less than three special issues to the ICE project. It has grown to include 14 country corpora, from the Inner Circle countries of Britain, Canada, New Zealand and the US to the Outer Circle countries of Singapore, India, and the Philippines (http://ice-corpora.net). Some ICE corpora have been grammatically annotated. This is a remarkable feat, given the

fact that the teams compiling the country corpora needed to follow the same corpus design with limited resources. The project is still growing, albeit slowly, to include more countries or regions (Wales and California, for example) and to add more linguistically-relevant features to the existing corpora, such as phonological annotation, aligning sound recordings with transcriptions in the spoken subcorpora, and lexical and structural annotation of the data (Kirk, 2017; Gut & Fuchs, 2017; Kirk & Nelson, 2018). Indeed, the idea of annotating ICE corpora for parts of speech and grammatical structure started as early as the project itself (Fang & Nelson, 1994; Greenbaum & Nelson, 1996). It is, unfortunately, a labor- and resource-intensive enterprise. Despite recent advances in NLP technologies, modern taggers and parsers are still with errors, and their output needs to be checked by linguistically trained researchers. At the present, only a few ICE country corpora have been annotated. The Sri Lankan corpus, for example, has been tagged but not checked (Bernaisch, Mendi, & Mukherjee, 2019), and the Philippine corpus was prepared for tagging with the exploration software ICECUP (Wallis, 2012). According to the UCL Survey of English Usage, only ICE-GB has been tagged, parsed and checked by linguists, with the caveat that the checked corpus is not perfect. It is, needless to say, an improvement over the tagger- or parser-generated output. The annotated ICE-GB is distributed through the website of UCL Survey of English Usage, together with ICECUP.

British English, of course, is an Inner Circle variety. Common PoS taggers, such as CLAWS (Garside & Smith, 1997), which is used to tag the 100-million-word British National Corpus and ICE-GB, are designed to work with texts of Inner Circle Englishes. Tagging Outer Circle Englishes may present unique problems due to contact-induced changes these varieties have undergone. In this paper, we report our experience of tagging ICE-SIN, the Singaporean corpus of the International Corpus of English. Our experience has been generally positive. Modern PoS taggers, trained on Inner Circle standard materials, can tag Outer Circle English texts with lower but still decent accuracy. The relatively high accuracy provides some relief for vetting automatically-assigned PoS tags. The vetted texts can in turn be used as part of the gold standard for training the PoS tagger to work optimally on Outer Circle English materials. Many off-the-shelf and freely available NLP tools have functions for training taggers; prominent among them are the Stanford NLP Group's PoS tagger (Toutanova, Klein, Manning, & Singer, 2003), the more recent Stanza Python package (Qi, Zhang, Zhang, Bolton, & Manning, 2020), and the spaCy Python package (Honnibal, Montani, Van Landeghem, & Boyd, 2020).

Modern taggers are useful tools for quantitative and corpus linguistics, and indeed for theoretical linguistics as well. Tagging Singapore English for parts of speech gives us insight into its lexicon and morphosyntax. We report some of the subtle changes in the lexical distribution in Singapore English that can be revealed only with data from an annotated corpus.

2 | TAGGING ICE-SIN

Scholarly interest in Singapore English started in earnest in the 1970s and 1980s, with the publication of Tongue (1974), Platt (1975), Crewe (1977), Tay (1979, 1982), Platt and Weber (1980), and Ho (1986). Since then the literature on the variety has grown extensively, making Singapore English, we would like to venture, the most studied variety among Outer Circle Englishes. Some of the more recent monograph-length works on the variety include Chew (2013), Leimgruber (2013), Wong (2014), Bao (2015), Ziegeler (2015), Low and Pakir (2018), Wee (2018), Buschfeld (2020), Lee (2020), Teo (2020), and Li (2021), covering a wide spectrum of topics from sociolinguistics to corpus linguistics to formal linguistics. Works on the morphosyntax of Singapore English typically rely on overt lexical markers that manifest novel functions in aspect (*already*, *got*), quantification (*got*, *also*), voice (*kena*), and pragmatic overtone (*lah*, *meh*), drawing data from native-speaker intuition, field observations, or computer corpora. These sources provide complementary data, often with convergent analytical results. For example, Brown (1999) observes that *already* and *also* are largely clause-final in Singapore English, an observation that can be readily made by casual visitors to Singapore today. This is clearly borne out by corpus data. The usage patterns of *already* and *also* in ICE-SIN and ICE-GB corroborate Brown's (1999) observation; see Table 1, cited from Table 4 of Bao & Hong (2006).[1]

TABLE 1 *Already* and *also* in ICE-GB and ICE-SIN, in percent. The numbers do not add up to 100 due to rounding, and in the case of ICE-SIN, to the omission of tokens of *already/also* found in formulaic expressions (I, initial; M, medial; F, final)

	already						*also*					
	GB			SIN			GB			SIN		
	I	M	F	I	M	F	I	M	F	I	M	F
Private Dialogue	5	80	16	2	29	66	24	76	0	13	36	41
Public Dialogue	4	88	7	1	83	14	10	87	3	7	87	6
Monologue	3	93	3	4	81	16	8	91	1	6	93	1
Writing	0	95	5	5	92	4	5	95	0	6	94	0

As we can see, the preferred position for both *already* and *also* is clause-final in the PRIVATE DIALOGUE register of Singapore English. This is typical of Outer Circle Englishes, where contact-induced grammatical restructuring largely affects informal registers, sparing the more formal registers.

For lexical markers such as *already* and *also*, ICE-SIN, SCoRE (Hong, 2009), and other databases of Singapore English provide ready quantitative data that can be fruitfully analyzed with common concordance tools, such as Antconc (Anthony, 2017). However, there is no practical way to examine changes at a more morphosyntactically sophisticated level. A tagged and parsed corpus opens up new possibilities and frontiers to explore the linguistics of contact.

Tagging ICE-SIN is part of our effort to build a tagged and parsed treebank of Singapore English (Huang, Hing, Lin, & Bao, 2021), modeled on the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993; Taylor, Marcus, & Santorini, 2003). Like other ICE corpora (Wallis, 2012), ICE-SIN contains formatting materials that needed to be removed before we tag it. Since there is no off-the-shelf tagger of Singapore English, we chose the Stanford PoS tagger (Toutanova et al., 2003), available on the Stanford NLP website (https://nlp.stanford.edu/), for convenience and for compatibility with the conventions established for the Penn Treebank. The Stanford PoS tagger is trained on standard American English, and uses the Penn Treebank tagset, summarized below (Marcus et al., 1993):

(1) The Penn Tagset

a.	Nouns:	NN (*book*), NNS (*books*), NNP (*Time*), NNPS (*Times*)
b.	Verbs:	VB (*do*), VBD (*did*), VBG (*doing*), VBN (*done*), VBP (*do*), VBZ (*does*)
c.	Adjectives:	JJ (*good*), JJR (*better*), JJS (*best*)
d.	Adverbs:	RB (*slowly*), RBR (*more*), RBS (*most*)
e.	Pronouns:	PRP (*he*), PRP$ (*his*)
f.	Prepositions:	IN (*in*)
	Subordinators:	IN (*if*)
g.	Particles:	RP (give *up*)
h.	Modals:	MD (*must*)
i.	Infinitive marker:	TO (*to*)
j.	Wh-words:	WDT (*which*), WP (*who*), WP$ (*whose*), WRB (*how*)
k.	Others	CC (*or*), CD (*two*), DT (*the*), EX (*there*), PDT (*all*), POS (*'s*)

In addition, there are tags for symbols (SYM), foreign words (FW), and interjections (UH). Punctuations are their own tags. We add two more tags, SFP (sentence-final particle), to tag a class of particles unique to Singapore English, and GOT, to tag uniquely local uses of *got* (*Got once I first* 'There was one time I was first'). The Stanford PoS tagger uses IN to tag both prepositions (*in*) and subordinating conjunctions (*if*, *that*), and TO to tag *to* as a preposition and as the infinitive marker. Although the immediate context can disambiguate the two functions, we follow the Brown corpus,

which is part of the Penn Treebank, and separate the two *to*'s, using IN for the preposition *to* (*to school*), and TO for the infinitival *to* (*to do*). We keep IN for both prepositions and conjunctions.

Common English PoS taggers, such as CLAWS, spaCY (www.spacy.io) and the Stanford PoS tagger, are capable of tagging standard English texts with an estimated accuracy rate between 95% and 97% (Fang & Nelson, 1994; Leech, Garside, & Bryant, 1994; Manning, 2011). When we tag ICE-SIN with the Stanford PoS tagger, we achieve variable accuracy rates, depending on the registers of the ICE-SIN texts. Table 2 shows the overall accuracy rates in the four registers of Singapore English:

The Stanford PoS tagger performs well across all registers in Singapore English, even though it is trained on American English data. The accuracy rate in the more formal registers of Singapore English is 96%, similar to the rates reported for British and American English in the works cited earlier. Compared with the CLAWS' performance with British English (Fang & Nelson, 1994) and the Stanford tagger's performance with the Penn Treebank (Manning, 2011), the success rate declines by nearly 10 percentage point in the informal register of private conversations in ICE-SIN. This is entirely within our expectations. Some of the issues that cause tagging errors identified in Leech et al. (1994) and Manning (2011) are just as valid for Singapore English as they are for British or American English. Take for example consistency, when the tagger assigns different tags to the same word forms, seemingly randomly. Manning (2011) cited expressions like *the 1930s*, which the Penn Treebank tags as cardinal numbers (1930s_CD) or as plural nouns (1930s_NNS), by chance. Whereas *the 1930s* is clearly a plural noun, numerals such as *1930* are truly ambiguous, as a number or as a year. Such issues arise in Singapore English too, and we decided to follow the Brown corpus and keep the tag CD as assigned by the tagger. The same consistency issue arises from tagging words that belong to multiple grammatical categories, such as *that* (determiner, complementizer, adverb), *what* (determiner, wh-phrase), and *which* (determiner, complementizer, wh-phrase), as exemplified in (2):

(2)
a. I should think that_IN he would want to go. (s1a-093)
b. I told Bee Guan that_WDT Eileen is not going. (s1a-014)
c. What_WP best friend you're talking about? (s1a-018)
d. What_WDT took me so long to write my memoirs? (w2b-009)
e. Which_WDT part of USA are you going to? (s1a-026)
f. One thing which_WDT I don't like to do is to shop. (s1a-005)

In both (2a) and (2b), *that* is a complementizer but is assigned different tags. *What* is a wh-determiner in (2c) and a stand-alone wh-phrase in (2d), so the two tags should swap. The Stanford PoS tagger treats all tokens *which* as a wh-determiner (WDT), even when it introduces a relative clause (2f).

These consistency issues are traceable to English morphosyntax. They are common to Inner Circle British or American English and to Outer Circle Singapore English, giving rise to the comparable accuracy rates in the formal registers we see in Table 2. The lower accuracy rate in the informal register of private conversations is, needless to say, due to the usual registral differences in grammar and lexical choice, and to the contact-induced changes that are characteristic of Singapore English, and of other Outer Circle Englishes.[2] We now proceed to document our tagging experience, focusing not on the NLP technologies that underpin the tagger, but on the linguistics of contact.

TABLE 2 Accuracy rates of tagging Singapore English, by register

PRIVATE DIALOGUE	88.9%
PUBLIC DIALOGUE	96.4%
MONOLOGUE	96.4%
WRITING	96.0%

3 | ISSUES IN TAGGING SINGAPORE ENGLISH CONVERSATIONS

Singapore English has borrowed words and phrases extensively from the languages that make up its linguistic ecology, mainly Chinese and Malay. We classify the extensive borrowings broadly into two types: lexical borrowings, and grammatical borrowings. Included in the latter type are English-derived words that have acquired novel grammatical meanings. We examine lexical borrowings first.

3.1 | Tagging lexical borrowings

We recognize three major types of lexical borrowing in Singapore English: words with currency beyond Singapore, words with currency within Singapore, and words which are clearly foreign and indistinguishable from code-switching. These are exemplified below:

(3)	a.	kiasu	Chinese	'selfish; afraid of losing out'
		kopitiam	Chinese/Malay	'coffee house'
		parang	Malay	'machete'
	b.	bochap	Chinese	'nonchalant, ignore'
		sian	Chinese	'bored'
		tahan	Malay	'endure, hold on'
	c.	gong	Chinese	'stupid'
		hock	Chinese	'fortunate, fortune'
		lang	Chinese	'people'

For Inner Circle varieties of English, the words in (3) are all foreign. There are differences among them, however subtle. *Kiasu*, *kopitiam* and *parang* have left the shores of Singapore, having found their way into the *Oxford English Dictionary*. Their use is still limited to Singapore and Malaysia, however. A quick search in the Factiva database of *The Times* of London yields seven tokens of *kiasu*, with the first appearing in a 1993 story about the kiasu Singaporeans, and since 1997, three tokens of *kopitiam* and 17 tokens of *parang*. These three words fare much better in Singapore's own *Straits Times*, where we find, since 1989, 1,926 tokens of *kiasu*, 1,510 tokens of *kopitiam*, and 451 tokens of *parang*. *Bochap*, *sian* and *tahan* may not be known internationally; in Singapore, they are common. *Sian* and *tahan* appeared in the *Straits Times* for the first time in 1989 (*so sian*; *I can only tahan (endure) up to 25 minutes*), and *bochap* in 1997 (*dogs are more bochap*). *Gong*, *hock* and *lang* are clearly foreign words, and appear mainly in code-switching environments. Since the words exemplified in (3a,b) are not foreign to Singaporeans, and the four Chinese-origin words have lost their tones, we tag them as regular words. The words in (3c), which make up the phrase (4h), retain their tones. We tag them as foreign words (FW).

These words are all found in ICE-SIN, as shown in (4) (the correct tag appears after the slash).

(4)	a.	Let's not be so kiasu_JJ	(s1b-029)
	b.	the Killiney kopitiam_NN	(s2b-031)
	c.	I think I feel quite sian_JJ	(s1a-057)
	d.	I just cannot tahan_VB	(s1a-084)
	e.	We are kiasu_NN/JJ	(s1b-029)
	f.	People can say I'm kiasu_NNP/JJ	(w2c-019)

g. Buay_NNP/MD tahan_NNP/VB (s1a-088)
cannot endure
'Cannot endure.'

h. Gong_NNP/FW lang gong_VBG/FW hock (s1a-083)
stupid person stupid fortune
'Stupid people have their own fortune.'

In (4g), we treat *buay*, which is of Hokkien origin, as a modal verb, and tag it accordingly.

Most English words are multi-categorial, and human speakers and computer taggers alike must rely on contextual morphosyntactic information to resolve the categorial uncertainty of words. Contextual information is the only clue for words like those in (3), which are most likely not listed in American or British corpora that English PoS taggers typically train on. The Stanford PoS tagger tags correctly if the contextual clue is sufficiently rich and unambiguous (4a–d), and makes wild guesses when it is lacking (4e–h), and inconsistent guesses to boot—*kiasu* is tagged as adjective (JJ), correctly, in (4a), as common noun (NN) in (4e), and as proper noun (NNP) in (4g), even though its local environment is the same (*be…kiasu*) in all three contexts. Incidentally, *gong* and *lang* in (4h) are regular English words, although they are not intended as such in ICE-SIN.

The non-English context, as expected, offers confusing clues. Manning (2011) lists seven types of error from tagging the Penn Treebank with the Stanford tagger,[3] which can be grouped into two broad categories: lexical errors caused by gaps in the training data and grammatical errors caused by difficult linguistics. These error types are attested in tagging Singapore English. The lexical loans in (4a–g) and the code-switching foreign words in (4h) further complicates tagging that relies on standard English dictionaries and morphosyntactic contexts.

3.2 | Tagging grammatical borrowings

Singapore English has undergone extensive contact-induced restructuring not only in the lexicon, but in grammar as well. There are two ways in which it is manifested: in foreign words which are directly borrowed with their foreign-sourced grammatical functions, and in English words which converge in grammatical function and usage with their foreign counterparts. These changes present different challenges to PoS tagging. We look at direct foreign borrowings first.

Singapore English has a productive system of sentence-final particles that express various subtle overtones of attitudes and emotions (Gupta, 1992; Lim, 2007). Many of the particles are direct borrowings from the local languages, mainly Cantonese, Hokkien or Malay. They are attested in ICE-SIN; three are exemplified in (5), along with the tag assigned by the Stanford PoS tagger.

(5) a. You got to take the word for it lah_FW (s1a-012)
Of course they threatened lah_NN (s1a-023)

b. Your niece came back already leh_JJ (s1a-088)
But today got some rice left leh_NNP (s1a-007)
'But today there is some rice left.'

c. So probably I'll cook for them lor_FW (s1a-007)
Anyway we can go church lor_NN (s1a-023)

d. Must top up meh_NN? (s1a-074)
'(I) must top up?'
(On visiting castles) All of them meh_VBP? (s1a-016)
'(Visit) all of them?'

According to Gupta (1992), *lah*, *leh* and *lor* are assertive particles with variable force of assertiveness, and *meh* expresses surprise and is used in interrogatives as a mild retort. They are tagged SFP. These words are obviously not in the Stanford PoS tagger's English training dictionary, and contextual information explains some of the tags—*threaten* and *top up* in (5a,d) require nominal objects, and *all of them* appears to be the subject that calls for a verb (5d).

The passive marker *kena*, derived from Malay, is attested six times in ICE-SIN; all are displayed in (6), with their assigned tags.[4]

(6) a. I feel like kena_FW sexual harassed. (s1a-031)
'I feel like being sexual harassed.'

b. I kena_VBP sexual harassed again, you know. (s1a-031)
'I was sexual harassed again, you know.'

c. She just said she kena_FW, right? (s1a-031)
'She just said she was (sexual harassed).'

d. There is guy…always kena_VB teased by this other guy. (s1a-079)
'There is a guy who…is always teased by this other guy.'

e. His tail like kena_NNP caught_VBD in the ratch hut. (s1a-052)
'His tail was caught in the ratch hut.'

f. I kena_VB shocked, you know. (s1a-096)
'I was shocked, you know.'

In Malay, *kena* 'strike' is a verb and carries adversative meaning as a passive marker, as exemplified below:

(7) a. Saya kena Covid.
I strike Covid
'I have Covid.'

b. Saya kena hantam.
I strike hit
'I got hit.'

Not surprisingly, all the tokens of the *kena* passive in (6) are adversative. We treat *kena* as a verb, and tag it accordingly. Since *kena* does not inflect (**John kenas/kena'ed covid*), we tag it as VBP if it is used as the main verb, even though the immediate context calls for VBD, which indicates that the action has already taken place at the time of utterance, as is the case in (6b,c,f). In the absence of English-style morphosyntactic context, the Stanford PoS tagger stumbles, in ways similar to the tagging of the sentence-final particles.

In addition to particles like *lah*, *leh*, *meh*, and the passive marker *kena*, which are directly borrowed from Chinese and Malay, Singapore English has also appropriated a number of foreign grammatical constructions which are marked by English words. There are two basic types: words which retain their original categorial status in English and words which do not. The former type includes aspectual markers of *already* and *ever*, which we have seen in Table 1, and the latter type includes sentence-final particles *what* and *one*, and the existential and aspectual marker *got*. These are exemplified in (8) and (9).

(8) a. *Already* (Bao, 1995, 2005; Ziegeler, 2021)
He go to New York *already*. ((perfective))
I cannot go inside *already*. (inchoative)

b. *Ever* (Ho & Wong, 2001)
The share *ever* hit forty dollars. (experiential)
'The share had hit forty dollars.'

(9) a. *What* (Gupta, 1992)
(on drawing on book) I never- I never ever draw *what*.
'I absolutely did not draw on your book.'

b. *One* (Bao, 2009)
I always use microwave *one*. (particle)
'I ALWAYS use microwave!'

c. *Got* (Lee, Ling, & Nomoto 2009; Bao, 2014)
Got people want to go. (existential)
'There are people (who) want to go.'
I *got* go Japan before. (perfective)
'I have been to Japan before.'

These forms, and their origins, have been studied extensively in the literature. They are all attested in ICE-SIN, and other databases of Singapore English. As aspectual markers in Singapore English, *already* and *ever* have lost their polarity status, as we can see in (8). Still, they are adverbs, and are tagged as such.

(10) a. So we can't reach him *already*_RB. (s1a-096)
Nowadays I switch to Mandarin *already*_RB. (s1a-007)

b. I *ever*_RB bought a jacket, a sweater. (s1a-057)
'I had bought a jacket, a sweater.'
(on dating) And he *ever*_RB ask me uh. (s1a-065)
'He had asked me.'

What, *one* and *got*, as exemplified in (9), pose serious challenges to PoS taggers based on English training dictionaries and morphosyntactic contexts. In Singapore English, on top of their English-derived lexical categories, these words acquired novel grammatical functions inconsistent with their English categorial status: *what* and *one* as sentence-final particles, and *got* as a base verb that marks existence and the perfective aspect.[5] Not surprisingly, the Stanford PoS tagger tags these words as regular English words, as shown below:

(11) a. People also sit what_WP/SFP. (s2b-043)
I should be able to see you what_WP/SFP. (s1a-094)

b. The adults no need to eat one_CD/SFP meh? (s1a-007)
'DON'T the adults need to eat?'
It's about making choices one_NN/SFP nuh. (s1a-025)
'It's about making choices!'

c. Cake inside got_VBD/GOT fruits. (s1a-006)
'As for the cakes, there are fruits inside.'
But today got_VBD/GOT some rice left leh. (s1a-007)
'But today there is some rice left.'
Sure got_VBD/GOT involve computer one. = (11b)
'(It) sure DID involve computers!'

The tags reflect the most common use of the words in English. *What* is tagged as either WP (*what did you say?*) or WDT (*what word did you say?*), to which we now add the tag SFP. *One* is a number, a noun, and now a sentence-final particle.[6] The novel uses of *got* and pronominal *one* are more subtle contact-induced grammatical changes that complicate English-based morphosyntax. They cause tagging challenges that can be attributed to difficult linguistics (Leech et al., 1994; Manning, 2011).

One and *got* pose a more serious problem. They inherit English-derived morphosyntactic frames, and have acquired a few more from local languages. In Singapore English, *one*'s pronominal function is extended to phrases, as shown in (12b):

(12) a. English-derived frames

A-*one*	the young *one*_CD/NN	(s1a-082)
N-*one*	the studio *one*_CD/NN	(s1a-097)
Pro-*one*	my *one*_NN	(w1b-002)

b. Chinese-derived frames

PP-*one*	From Thailand *one*_CD/NN	(s1a-080)
	'the one from Thailand'	
VP-*one*	Showing in Cathay *one*_CD/NN	(s1a-080)
	'the one (which is) showing in Cathay'	
S-*one*	You want *one*_CD/NN	(s1a-083)
	'the one that you want'	

Clearly, the forms in (12b) are ill-formed in English. The multiple functionality of *one* leads to multiple meanings, which require extended discursive contexts to disambiguate. Take for example *you want one*. It has three distinct meanings:

(13) You want one.

a.	*One* as NP object:	You want [$_{NP}$ one]	
b.	*One* as particle of emphasis:	[$_{S}$ You want one]	
		'You want (it)!'	
c.	*One* as pronominal:	[$_{NP}$ You want one]	= (12b)
		'the one that you want'	

One in (13a) requires no comment. With appropriate prosody, the utterance could be interpreted as stressing your wanting something, with *one* serving as the particle for emphasis, as in (13b). In (13c), *one* is analyzed as the pronominal, as indicated by the English gloss. This is the intended reading, given the extended context shown in (14):

(14) Three friends talking about Hawaii chocolates with macadamia nuts

A: I prefer Cadbury's macadamia.

C: Oh it's got macadamia and nuts. *You want one.*

B: Oh stuck to this piece ah so big ah.

A: This one you actually just chew it you know.

Without the extended discursive context, it is impossible to determine the precise meaning, and the part of speech, of *one*, for both the human interpreter and the computer tagger.

The same level of morphosyntactic ambiguity or vagueness can be seen with respect to *got*, which has undergone similar grammatical restructuring due to influence from Chinese. The word has been studied extensively (Lee, Ling , &

Nomoto, 2009; Hiramoto & Sato, 2012; Bao, 2014). Here, we focus on the two additional meanings of *got* that we have seen in (9c). These two uses of *got* have both been attested in ICE-SIN.

(15) *Got* V (perfective)

a. (about tape recorder) I just *got* purchase. (s1a-086)
'I just purchased (the tape recorder).'

b. You *got* go underwater. (s1a-085)
'You have been underwater.'

(16) *Got* N (existential)

a. That time lah, *got* two teachers start loving each other. (s1a-085)
'That time, two teachers started loving each other.'

b. Cake inside *got* fruits. (s1a-006)
'As for cake, there are fruits inside.'

c. Then June *got* Arts Fes. (s1a-025)
'Then there will be Arts Festival in June.'

Here, *got* is the base form and cannot be replaced with *get* (**I just get purchase/*Cake inside get fruits*).[7] In this respect, it behaves like the Chinese source *you* 'have,' or its Hokkien cognate *u*. As expected, the *got* N frame is often followed by a verb (*start*), and/or preceded by a locative or temporal expression (*cake*, *June*). Not surprisingly, the Stanford PoS tagger treats *got* in both frames as the past-tense form of *get* (VBD). However, while *got* V expresses the perfective, consistent with the tag VBD in terms of tense, the tense-related meanings of the *got* N frame in (16) are not derived from *got*. We use the label GOT to tag *got* as head of the two *you*-derived frames.

The existential meaning of *got* is extended to regular English sentences as well, as the data in (17c,d) show:

(17) a. He *got* his PhD last year. (s1a-019)

b. I tell you I just *got* my computer the other day. (s1a-061)

c. But Thursday night, I *got* church so I can't go. (s1a-051)
'But Thursday night, I have church so I can't go.'

d. You see we *got* four chairs here already what. (s1a-054)
'You see we already have four chairs here.'

All tokens of *got* in (17) are tagged VBD, which is appropriate for (17a,b), but strictly speaking, not for (17c,d), which do not express past events or states, as indicated by the glosses. Since the *got* tokens in (17) conform to English morphosyntax, we tag them VBD.

While all sentences in (17) contain the necessary context for arriving at the correct tense-related interpretation, there are many cases that require extended discursive context to disambiguate the precise meanings of *got*. One example follows.

(18) Let me know what prize you *got* lah. (s1a-014)

a. *Got* as past tense: Let me know what prize you received.

b. *Got* as existence: Let me know what prize you have.

Here, the two meanings share the same form, and require discursive context to disambiguate. As an existential marker, *got* does not carry any tense-related information. Indeed, within the larger discursive context of the utterance (18) shown below, (18b), not (18a), is the preferred reading:

(19) Two friends discuss dinner and dance party with lucky draw

A: So tomorrow let me know. Let me know what prize you *got* lah.

B: Aiyoh I don't know whether my I always not been so-called lucky.

Since the sentence *what prize you got* is grammatical in English, it could yield the past-tense reading if the two friends were talking about an event that has already taken place (*Let me know what prize you got last night*). In the exchange in (19), the event is in the future. We keep the tag VBD for such uses of *got*,[8] and reserve the tag GOT for tokens of *got* V and *got* N, as exemplified in (15) and (16). Such frames are ungrammatical in English, and yield readings unique to Singapore English.

The lexical borrowings and grammatical changes prove to be challenging to taggers that train on Inner Circle English materials and rely on English morphosyntax, aggravating the effect of difficult linguistics on tagging accuracy, especially in the informal register; see Table 2. Nevertheless, as our experience shows, the Stanford PoS tagger, and other English-based taggers, can be a useful tool to tag World English materials. The accuracy rate is lower, but manageably so.

4 | DOCUMENTING CONTACT-INDUCED CHANGES WITH TAGGED CORPORA

The lexical and grammatical restructuring that Singapore English has undergone has been well documented in the World English literature. Most of the studies rely on data from observations and from native-speaker judgment. The PoS tagged ICE-SIN not only allows us to substantiate the results of these studies with quantitative data, it also reveals some surprising facts about the lexical distribution between Singapore English and British English, and between informal and formal registers within Singapore English. A complete description of the lexical distribution of Singapore English is beyond the scope and aim of this paper. We now proceed to report some distributional patterns that can be corralled from the annotated ICE-SIN.

A total of 10 sentence-final particles are attested in ICE-SIN. They are shown in Table 3.

Not surprisingly, the vast majority of the 2,882 particle tokens are found in the informal register of PRIVATE DIALOGUE, with *lah* being the most frequently used particle, followed by *ah*.[9] Together, the particles account for 1.3% of the total word count in the register.

TABLE 3 All sentence-final particles attested in ICE-SIN

	PRIVATE	PUBLIC	MONO	WRITING
lah	1,600	68	48	6
ah	513	45	22	0
what	183	9	23	2
lor	145	0	3	0
one	73	8	3	2
leh	39	0	5	0
hor	36	1	5	0
hah	20	0	1	0
meh	15	1	1	0
mah	5	0	0	0
Total	2,629	132	111	10

TABLE 4 The incidence of use of *what* and *one* as sentence-final particle (SFP), and of *got* as aspectual and existential marker (GOT), in ICE-SIN

		PRIVATE	PUBLIC	MONO	WRITING
what	SFP	183	9	23	2
	total	1,818	1,067	976	808
	percent	10.1	0.8	2.4	0.2
one	SFP	73	8	3	2
	total	1,340	930	1,314	1,219
	percent	5.4	0.9	0.2	0.2
got	GOT	81	0	3	1
	total	606	178	307	182
	percent	13.4	0.0	1.0	0.5

TABLE 5 Counts of tagged words in type and token in the four registers of ICE-SIN and ICE-GB

	ICE-SIN		ICE-GB	
	Type	Token	Type	Token
PRIVATE DIALOGUE	11,186	197,559	11,034	185,507
	$\chi^2 = 12.69, p = .0004$			
PUBLIC DIALOGUE	11,254	172,412	11,351	166,304
	$\chi^2 = 10.52, p = .0012$			
MONOLOGUE	23,476	260,199	25,587	257,199
	$\chi^2 = 106.73, p<.0001$			
WRITING	65,668	428,246	67,233	421,233
	$\chi^2 = 46.07, p<.0001$			

The English words that have acquired the particle functions, *what* and *one*, are doing well in Singapore English, coming in at 3rd and 5th among the ten particles. Table 4 displays the incidence of use of *what* and *one*, and also *got*, reflecting the degree of contact-induced grammatical change in English-origin words that have acquired novel grammatical meanings in Singapore English.

Like *already* and *ever* that we have seen in Table 1 above, the locally-derived senses of the three words occur mainly in the informal register of PRIVATE DIALOGUE, and are negligible in the more formal registers. It is worth noting that the novel uses take up a large proportion: 10.1% for *what*, 5.4% for *one* and 13.4% for *got*.

Not all aspects of Singapore English grammar have undergone drastic contact-induced changes. The overall counts of tagged words in ICE-SIN and ICE-GB, in type and token, are similar, as shown in Table 5.

Since the tagset used in ICE-GB is not the same as the Penn tagset, the annotated data need to be adapted to make comparison meaningful. For content word classes this is straightforward: {n_com,sing} for NN and {n_com,plu} for NNS, for example. Function words is a problematic area. In ICE-GB, pronouns include words with the prefixes *some-*, *any-*, *no-*, and *every-* (*something*, *anyone*, *nobody*, *everything*). We treat them as nouns (NN). In ICE-SIN, we consider foreign words (FW) (268 type, 537 token) as content words, and sentence-final particles (SFP) and existential and aspectual *got* (GOT) as function words. Auxiliary verbs (*be*, *do*, *have*) are tagged as regular verbs, as is the practice of the Penn Treebank. Interjections (UH) and fragments (FRG) are excluded from our calculation.

TABLE 6 The distribution of nouns, verbs, adjectives and adverbs, and lexical density, in ICE-SIN and ICE-GB; in percent; df = 6

		PRIVATE	PUBLIC	MONO	WRITING	t	*p*
Nouns	SIN	16.1	22.1	26.7	30.2	0.59	.577
	GB	14.9	19.1	24.8	26.7		
Verbs	SIN	22.6	19.0	17.0	16.0	0.39	.711
	GB	22.3	18.0	15.5	15.4		
Adjectives	SIN	5.3	6.3	6.6	8.8	0.43	.682
	GB	4.7	5.4	7.0	8.0		
Adverbs	SIN	10.9	7.1	6.5	5.4	0.69	.518
	GB	8.4	6.6	5.9	5.3		
Lexical Density	SIN	55.1	54.6	56.8	60.4	2.42	.052
	GB	50.3	49.1	53.2	55.5		

By design ICE-SIN and ICE-GB have the same structure and size. It is not surprising that they have similar word counts in type and token, as we can see in Table 5. Compared with ICE-SIN, ICE-GB has a higher count in type, except PRIVATE DIALOGUE, and a lower count in token, resulting in a slight edge in lexical richness.[10] Lexically, there is no major difference between the two varieties of English despite the incorporation into the Singaporean variety of local words or meanings documented in Section 2. Singapore English remains a dialect of English after 200 years of intense language contact.

Not surprisingly, the distribution of the content word classes remains remarkably consistent as well, as shown in Table 6.

The *t*-test on the four content categories across the four registers is performed on SPSS v.27. The Shapiro–Wilk test shows normal distribution for all measures.

From Table 6, we can see that the distribution of content words in Singapore English parallels that in British English, and is consistent with the distribution in the Brown corpus (nouns 27.5%, verbs 16.3%, adjectives 7.9% and adverbs 5.4%), and with the findings reported in Biber et al. (1999) based on much larger corpora of American and British English. Nouns are most common in all registers except private conversations, where verbs dominate. In Singapore English conversations, we found very high frequencies for *so* and *then*. Indeed, *so* out-ranks function words *a* and *that*, and *then* out-ranks the negator *not* and the sentence-final particle *lah*. Discounting these two words, the adverb ratio in PRIVATE DIALOGUE is down to 8.8%, closer to the ICE-GB ratio of 8.3%. Overall, the usage rate is slightly higher across the four content word classes, and across the four registers. Although the differences are not statistically significant ($p < .05$), the *p* values indicate different degrees of variance between the two varieties: at $p < .577$ nouns exhibit greater variance than verbs at $p < .711$.

The differences in lexical density, at $p < .052$, are close to being statistically significant. Since lexical density measures the usage rate of content words against the total word count in the register, a higher lexical density figure implies a corresponding drop in the use of function words. It is likely that the drop is partially due to contact-induced changes in the grammar of Singapore English. Here, we will not include a full investigation of the distribution of function words in ICE-SIN and ICE-GB; see Lin (2022). Suffice it to say that the difference in lexical density between the two varieties of English is due to many factors. Here we mention two: modal verbs (ICE-SIN 18,818 v. ICE-GB 21,933) and numerals (ICE-SIN 14,780 v. ICE-GB 25,248). The lower frequencies of modals and numerals, both function word classes, contribute to the higher lexical density of ICE-SIN.

5 | CONCLUSION

The International Corpus of English project is now rounding off its third decade of serving the World English community. Although structural annotation of the data has been part of the original design, due to various reasons only ICE-GB has been fully annotated and checked to date. With recent advances in NLP technologies, modern English PoS taggers, among them CLAWS, spaCY and the Stanford PoS tagger, are able to achieve accuracy rates as high as 97% on British or American English texts, out-performing human annotators (Manning, 2011). Our experience of tagging Singapore English has been positive in demonstrating the feasibility of tagging Outer Circle English texts with readily available computer taggers trained on standard British or American English. The error rates are not significantly lower, even in tagging private conversations, where most of the foreign borrowings are attested. This is not only true of ICE-SIN, but also true of SCoRE, the corpus of classroom discourse in Singaporean schools (Hong, 2009). The Singapore English lexicon is basically English, with foreign lexical borrowings accounting for a very small portion, even in the informal register of private conversations.

Of the two main causes of tagging inaccuracy, unknown words and difficult linguistics, the former is easy to resolve. We expect lexical loans like *kiasu* and *lah* to behave like regular words if they are included in the dataset that the tagger trains on. Grammatical borrowings pose more serious challenges, creating an additional layer of structure or grammatical meaning unknown to English morphosyntax or outright ungrammatical. Tagging the particle uses of *what* and *one* and the existential and aspectual uses of *got* requires extensive linguistic context, including discursive context, for the human interpreter. It is inevitable that lexical gaps or linguistic context lower tagging accuracy. But our experience of tagging Singapore English gives us reason to be optimistic. Modern PoS taggers, trained on data from Inner Circle English, can be used as cost-effective tools to tag Outer Circle English. The accurate output of these taggers can further serve as an excellent starting point for the creation of gold-standard PoS tags for researchers who require even greater precision in lexical processing, facilitating what is otherwise a labor-intensive process. Such tools, therefore, open up new possibilities to explore the linguistics of contact.

ACKNOWLEDGEMENTS

The work is partially supported by research grants from the Humanities and Social Science Research, National University of Singapore, DSO National Laboratories, and the Social Science Research Council, Singapore.

We thank one anonymous reviewer for his generous help with the statistical test results and their interpretation.

ORCID

Zhiming Bao https://orcid.org/0000-0002-7859-2654

NOTES

[1] Each ICE country corpus follows the same design, with 500 2,000-word texts in 32 categories (Greenbaum & Nelson, 1996). Bao and Hong (2006) groups these text categories into four broad registers, PRIVATE DIALOGUE (100 texts), PUBLIC DIALOGUE (80 texts), MONOLOGUE (120 texts) and WRITING (200 texts). The PRIVATE DIALOGUE contains spontaneous conversational data, and represents colloquial Singapore English, or more commonly, Singlish.

[2] We did an exhaustive count of tagging errors in two files, s1a-001 and w2a-001, as representative samples of informal conversations and formal writings. We counted 221 errors in s1a-001 (11%) and 63 errors in w2a-001 (3%). The most common types involve sentence-final particles (SFP), adverbs (RB) and interjections (UH):

	Count	Percent
SFP	27	12.2
RB	29	13.1
UH	68	30.8

These are exemplified below (s1a-001):

a. So_IN/RB you have to make your list lah_NN/SFP.
b. Aiyah_NNP/UH so_IN/RB you decided not to quit already ha_NN/SFP.

Of the 29 adverb errors, 25 are *so* tagged as subordinator (IN). Clearly, SFPs are due to contact, and the other two types are characteristic of conversations.

[3] Manning (2011) lists a total of seven types, shown in the table below:

1.	Lexical gap	4.5%
2.	Unknown word	4.5%
3.	Could plausibly get right	16.0%
4.	Difficult linguistics	19.5%
5.	Underspecified/unclear	12.0%
6.	Inconsistent/no standard	28.0%
7.	Gold standard wrong	15.5%

The first two types is a dictionary issue, and types 3, 4 and 5 are related to English morphosyntax. The remaining two types are due to standards being vague or not clearly specified. We label the first type lexical, and the second grammatical. In our vetting process, we keep a word's assigned tag if it is plausible for the word, and correct the tag if it is obviously wrong given the immediate context, as illustrated by the two sentences below:

a.	Can't see sunset_JJ/NN ah_NN/SFP	(s1a-001)
b.	You just slot_NN/VBP the whole thing in	(s1a-001)

Obviously, the tagger gets the morphosyntax of *sunset ah* wrong.

[4] Singapore English has a Chinese-derived passive as well, expressed by the English verb *give*, as shown below (Bao & Wee, 1999):

John *give* his boss scold.
'John was scolded by his boss.'

This passive is calqued on the Chinese passive expressed by the verb *gei* 'give.' We will not discuss this passive here. It is worth adding that unlike the *kena* passive, it is not attested in ICE-SIN, nor in SCoRE, the corpus of classroom discourse in Singaporean schools (Hong, 2009).

[5] *Got* has acquired more novel functions than these two; see Lee et al. (2009) and Bao (2014). Here we only discuss *got* as marker of existential and perfective meanings. The remaining novel uses of *got* pose the same type of challenge.

[6] We checked how *what* and *one* are tagged in the Brown corpus, which is part of the Penn Treebank. While the tagging of *what* is consistent, this is not the case for *one*. In the Brown corpus, *one* is tagged differently even though the context is the same:

As number:	One_CD wants a little more
As noun:	One_NN has to start early
As pronominal:	One_PRP must act

We will not attempt to draw a fine line between pronominal *one* and nominal *one*, and tag it as noun (NN), and *ones* as plural noun (NNS). *One* is tagged numeral (CD) only when it is used as such unambiguously (*one book*).

[7] The grammaticality judgment of Singapore English sentences with *get* N and *get* V is based on native-speaker intuition. The forms *get* V and *get* N are not attested in ICE-SIN, nor in the much larger database of classroom discourse SCoRE (Hong, 2009).

[8] In the SCoRE corpus of classroom discourse (Hong, 2009), utterances such as *I got math tomorrow* are common. Tagging *got* as VBD even though it is the base form is consistent with our decision to tag words as they are used. It has long been noted that inflectional categories are not marked in Singapore English as consistently as they are in native varieties

(Platt & Weber, 1980). A few specimens follow:

a. She *earn* a lot of money (s1a-006)
b. Nanyang Academy of Fine Arts was *establish* in 1938 (s2a-035)
c. One of these *day*, I'll be speaking in broken English (s1a-011)

The Stanford PoS tagger tags the uninflected verbs and nouns as they are, VBP and NN. We keep these tags in the vetting process.

[9] The spelling of the particles is not fully conventionalized. *Lah* is also spelled as *lar* and *la*, *lor* as *loh*, and so on. The *ah* count includes tokens of *nah* and its variant *nuh*, both of which usually follow words ending in the nasal (*Can nah*; *Can nuh*), as noted in Deuber, Leimgruber, and Sand (2018). The token counts include variable spellings. Many tokens of *nuh*, *ha*, etc. are indications of pauses or hesitations. We rely on the context to arrive at the most reasonable assessment. Two particles, *bah* and *liao*, have been reported (Leimgruber, 2016; Loo, 2016), but they are not attested in ICE-SIN.

[10] The chi-square tests indicate that the differences are statistically significant ($p < .05$). As is well-known, chi-square tests are sensitive to sample size. Given the large word counts of the four registers between ICE-SIN and ICE-GB, small differences inevitably lead to large χ^2 values, with infinitesimally small p values. Caution is needed in interpreting such p values for statistical significance (Brezina & Meyerhoff, 2014; Wallis, 2013).

We thank one anonymous reviewer for his generous help with the statistical test results and their interpretation.

REFERENCES

Anthony, L. (2017). *AntConc* 3.5.0. Tokyo: Waseda University.
Bao, Z. (1995). *Already* in Singapore English. *World Englishes*, *14*(2), 181–188.
Bao, Z. (2005). The aspectual system of Singapore English and the systemic substratist explanation. *Journal of Linguistics*, *41*(2), 237–267.
Bao, Z. (2009). *One* in Singapore English. *Studies in Language*, *33*(2), 338–365.
Bao, Z. (2014). *Got* in Singapore English. In E. Green & C. F. Meyer (Eds.), *The variability of current world Englishes* (pp.147–165). Berlin: De Gruyter Mouton.
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*(1), 105–114.
Bao, Z., & Wee, L. (1999). The passive in Singapore English. *World Englishes*, *18*(1), 1–11.
Bernaisch, T., Mendi, D., & Mukherjee, J. (2019). *Manual to the International Corpus of English – Sri Lanka*. Ms. Justus Liebig University.
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
Brezina, V., & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, *19*(1), 1–28.
Brown, A. (1999). *Singapore English in a nutshell: An alphabetical description of its features*. Singapore: Federal Publications.
Buschfeld, S. (2020). *Children's English in Singapore: Acquisition, properties, and use*. London: Routledge.
Chew, P. G.-L. (2013). *A sociolinguistic history of early identities in Singapore: from colonialism to nationalism*. New York: Palgrave Macmillan.
Crewe, W. J. (1977). *Singapore English and Standard English: Exercises in awareness*. Singapore: Eastern Universities Press.
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*(1), 48–91.
Fang, A. C., & Nelson, G. (1994). Tagging the Survey Corpus: A LOB to ICE experiment using AUTASYS. *Literary and Linguistic Computing*, *9*(3), 189–194.
Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp .102–121). London: Longman.
Greenbaum, S. (1988). A proposal for an international computerized corpus of English. *World Englishes*, *7*(3), 315.
Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) project. *World Englishes*, *15*(1), 3–15.
Gupta, A. F. (1992). The pragmatic particles of Singapore colloquial English. *Journal of Pragmatics*, *18*(1), 31–57.
Gut, U., & Fuchs, R. (2017). Exploring speaker fluency with phonologically annotated ICE corpora. *World Englishes*, *36*(3), 387–403.
Hiramoto, M., & Sato, Y. (2012). *Got*-interrogatives and answers in Colloquial Singapore English: Aktionsart and stativity. *World Englishes*, *31*(2), 186–195.

Ho, M. L. (1986). *The verb phrase in Singapore English* [Doctoral dissertation, Monash University].
Ho, M. L., & Wong, I. F. H. (2001). The use of *ever* in Singaporean English. *World Englishes*, *30*(1), 79–87.
Hong, H. (2009). *A corpus-based study of educational discourse: The SCoRE approach*. [Doctoral dissertation, National University of Singapore].
Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.
Huang, N., Hing, J. W., Lin, L., & Bao, Z. (2021). *A tagged and annotated corpus of* Singapore English. MS. National University of Singapore.
Kirk, J. (2017). Developments in the spoken component of ICE corpora. *World Englishes*, *36*(3), 371–386.
Kirk, J., & Nelson, G. (2018). The International Corpus of English project: A progress report. *World Englishes*, *37*(4), 697–716.
Lee, K. M. (2020). *Exploring changes in English news writing in Singapore: A diachronic corpus-based study*. [Doctoral dissertation, National University of Singapore].
Lee, N. H., Ling, A. P., & Nomoto, H. (2009). Colloquial Singapore English *got*: functions and substratal influences. *World Englishes*, *28*(3), 293–318.
Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th international conference of computational linguistics* (pp. 622–628). Kyoto, Japan.
Leimgruber, J. R. (2013). *Singapore English: Structure, variation and usage*. Cambridge: Cambridge University Press.
Leimgruber, J. R. (2016). *Bah* in Singapore English. *World Englishes*, *35*(1), 78–97.
Li, L. (2021). *Language contact: A historical sociolinguistic reconstruction of Colloquial Singapore English in relation to its Chinese substrates*. [Doctoral dissertation, University of Hamburg].
Lim, L. (2007). Mergers and acquisitions: On the ages and origins of Singapore English particles. *World Englishes*, *27*(4), 446–473.
Lin, L. (2022). *A corpus-based grammar of Singapore English: Description and change*. [Doctoral dissertation, National University of Singapore].
Loo, J. (2016). *The grammar of already and liao in Singapore English*. [Doctoral dissertation, National University of Singapore].
Low, E. L., & Pakir, A. (Eds.) (2018). *World Englishes: Rethinking paradigms*. London: Routledge.
Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In A.F. Gelbukh (Ed.), *Computational linguistics and intelligent text processing*. CICLing 2011. Lecture Notes in Computer Science (vol. 6608, pp. 171–189). Berlin: Springer.
Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.
Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: Working with the British component of the international corpus of English*. Amsterdam: John Benjamins.
Platt, J. (1975). The Singapore English speech continuum and its basilect 'Singlish' as a 'creoloid'. *Anthropological Linguistics*, *17*(7), 363–374.
Platt, J. T., & Weber, H. (1980). *English in Singapore and Malaysia: Status, features, functions*. Oxford: Oxford University Press.
Qi, P., Zhang, Y. H., Zhang, Y., Bolton, J., & Manning, C. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Association for Computational Linguistics (ACL) System Demonstrations*, 101–108.
Tay, M. W. J. (1979). The uses, users and features of English in Singapore. In J. C. Richards (Ed.), *New varieties of Englishes* (pp. 91–111). SEAMEO Regional Language Centre, Singapore.
Tay, M. W. J. (1982). The phonology of educated Singapore English. *English Worldwide*, *3*(2), 135–145.
Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: An overview. In A. Abeilé (Ed.), *Treebanks* (pp. 5–22). Text, Speech and Language Technology, vol. 20. Dordrecht: Springer.
Teo, M. C. (2020). *Crosslinguistic influence in Singapore English: Linguistic and social aspects*. London: Routledge.
Tongue, R. K. (1974). *The English of Singapore and Malaysia*. Singapore: Eastern University Press.
Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of HLT-NAACL 2003* (pp.252–259).
Wallis, S. (2012). *Tagging ICE Philippines and other ICE corpora. Ms*. Survey of English, University College London.
Wallis, S. (2013). Binomial confidence intervals and continency tests: Mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, *20*, 178–208.
Wee, L. (2018). *The Singlish controversy: Language, culture and identity in a globalizing world*. Cambridge: Cambridge University Press.
Wong, J. O. (2014). *The culture of Singapore English*. Cambridge: Cambridge University Press.
Ziegeler, D. (2015). *Converging grammars: Constructions in Singapore English*. Berlin: De Gruyter Mouton.

Ziegeler, D. (2021). Convergence in contact grammaticalization in Singapore English: the case of *already*. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, 37 | 2021.

How to cite this article: Lin, L., Han, K., Hing, J. W., Cao, L., Ooi, V., Huang, N., & Bao, Z. (2023). Tagging Singapore English. *World Englishes*, *42*, 624–641. https://doi.org/10.1111/weng.12597