# Treebanks and World Englishes

## A Singapore English perspective

Nick Huang[1] Li Lin[2] Kunmei Han[1] Jia Wen Hing[1]
Luwen Cao[1] Vincent Ooi[1] and Zhiming Bao[1]
[1] National University of Singapore | [2] East China University of Political Science and Law

Treebanks (parsed corpora) play an important role in linguistic research, but creating high-quality parses can be very labor-intensive. This paper discusses the prospects of creating such parses in the context of New Englishes and what kinds of research insights parses can deliver. We present Singapore English as a case study. We suggest that despite the many contact-derived lexical and grammatical properties of Singapore English, it is quite feasible to apply an off-the-shelf American English parser to generate parses of Singapore English. In addition, we present an exploratory analysis of noun phrases in a Singapore English treebank, to illustrate the potential of parses and treebanks in research on World Englishes.

**Keywords:** treebank, corpus, World Englishes, syntactic parsing, variation, Singapore English, New Englishes, language contact

## 1. Introduction

Research on English varieties around the world has benefited greatly from corpus resources. For almost three decades, the *International Corpus of English* (ICE; Greenbaum 1988; Greenbaum and Nelson 1996; Kirk 2017; Kirk and Nelson 2018) has provided scholars with rich examples of the lexical and syntactic diversity in English varieties. However, it is worth noting that these corpora come with little or even no annotation. Of the 14 sub-corpora in ICE, only the ICE-GB (Great Britain) corpus has been annotated with both part of speech tags and syntactic parses (cf. Wallis and Nelson 2000; Nelson, Wallis, and Aarts 2002). The more recent Corpus of Global Web-based English (Davies and Fuchs 2015) and the even more recent Twitter Corpus of Philippine English (Gonzales 2023) are tagged for parts of speech but lack full syntactic parses.

https://doi.org/10.1075/eww.23069.hua | Published online: 21 March 2025
*English World-Wide* 46:1 (2025), pp. 93–121. ISSN 0172-8865 | E-ISSN 1569-9730

This lack of rich linguistic annotation, especially at the level of syntactic parsing, is unsurprising. The creation of parsed corpora (treebanks) is traditionally time-consuming and labor-intensive, requiring careful work by linguistically-trained researchers. The lack of treebanks in turn imposes major constraints on the type of research that can be conducted. Largely due to language contact, there are often substantial differences in lexicons and grammars of English varieties around the world, especially for the so-called "New Englishes" spoken in former British or American colonial possessions originally settled by non-native speakers of English (corresponding to what Kachru [1982, i.a.] calls "Outer Circle" varieties). While one can study the range of variation by using concordance software for individual words and collocations, it is far more challenging to do the same for grammatical constructions without access to parses.

In this article, we address these issues through a case study on Singapore English (henceforth "SgE"). We argue that creating syntactic parses for SgE material might be less difficult than expected if one leverages freely-available parser software developed for standard American English ("AmE"). Even though SgE can be linguistically very different from AmE, many distinctive aspects of SgE, such as the borrowing of content words or grammatical constructions from Chinese and Malay, actually have a limited impact on parser accuracy. Although parsers typically do not come with documentation explaining how they make parsing decisions, our analysis of parsing errors for these contact-induced features suggests that the relatively high quality of parses is not coincidental. Rather, it reflects the strategy used by parsers for handling novel words and constructions. We also describe some strategies for improving parser performance for use by other researchers. That said, we note that these parsers are not perfect: parsing accuracy is sensitive to other factors, such as speaker/writer/transcriber errors and non-standard orthography in the material to be parsed. Nonetheless, our experience suggests that parsers can make syntactic annotation and analysis easier, allowing researchers to pose and answer new research questions. Put differently, these parsers, even though developed for a specific register of a particular variety of English, can be helpful for the study of a wider range of English varieties.

This article is organized as follows. Section 2 provides some background on SgE and on how we evaluate the feasibility of using the Stanford Parser, originally developed for AmE, to parse SgE material. Section 3 discusses parser strategy and performance, highlighting which contact-induced linguistic features the parser struggles with, and which ones it can parse accurately. We organize our discussion around different types of features, namely, borrowings of content words, function words, and constructions, so that it will be helpful for readers who are curious about parser performance for equivalent features in the English varieties of interest to them. Section 4 presents an exploratory analysis of SgE noun phrases (NPs) to illustrate how parses and treebanks can be useful in World English research. Section 5 concludes.

## 2. Background

As mentioned above, we propose using off-the-shelf parsers to address a gap in treebanks for the study of World Englishes. Even though these parsers were originally developed for AmE, there is reason to give this option serious consideration. For instance, SgE, the variety of interest in this article, differs from standard AmE in lexicon, grammar, and even orthography. But there is still enough overlap, to the extent that an AmE reader unfamiliar with SgE can still draw some inferences about the SgE sentences in Examples (1)–(3), taken from informal conversation ("private dialogue") transcripts from the International Corpus of English–Singapore (ICE-SIN) subcorpus. It is not implausible that parsers could exploit the same overlap to generate relatively accurate syntactic analyses of these sentences.

(1) My office not many people will come...
('As for my office, not many people will come'; s1a-007)[1]

(2) ...I think can understand. ('...I think I can understand'; s1a-090)

(3) You got go underwater. ('You did go underwater'; s1a-085)

In this section, we provide more background on the contact-induced linguistic features of SgE and on one well-known AmE parser, the Stanford Parser (version 4.2; see Klein and Manning 2003 and subsequent work, e.g. Qi et al. 2020). We then describe how we apply the Stanford Parser to SgE material to evaluate the feasibility of our proposed approach.

### 2.1 Singapore English

As many researchers have noted, SgE has acquired a set of distinctive linguistic features due to intensive language contact between (British) English, varieties of Malay and Chinese, and, to a smaller extent, South Asian languages like Tamil (Platt 1975; Crewe 1977; Tay 1979; Tongue 1979; Gupta 1994; Leimgruber 2013; Bao 2015; Ziegeler 2015; Wee 2018; Teo 2020; i.a.). This reflects the settlement of modern Singapore by immigrants from various parts of Asia and the use of English as an official language of administration and education. For scope reasons, we

[1] For all examples, all "s1a-..." examples are from ICE-SIN private dialogue transcripts, "s1a-..." being the filename. Text in quotes are our paraphrases. In certain examples, we have bolded words of interest.

will set aside issues related to Singapore's language ecology and internal variation within SgE; interested readers should consult the references above. Instead, our focus here will be on the linguistic features that distinguish SgE from (standard) British English (henceforth "BrE") and AmE. These features, attributable to language contact, can be readily detected in informal SgE. To facilitate our discussion in Section 3, we further sort them into three broad categories.

#### **2.1.1** Borrowing of content and function words

SgE has borrowed content words from Chinese, Malay, and South Asian languages; examples of loans include *kiasu* 'afraid of losing out' (from Southern Min Chinese), *shiok* 'enjoyable', or *tahan* 'endure' (both from Malay). In SgE, these are the same parts of speech and subject to the same rules as (near-)equivalents in standard BrE or AmE, e.g. *tahan* is a verb that takes animate subjects and can follow a modal auxiliary.

Function word borrowings include sentence-final particles, which express speakers' attitudes and/or emotions about the preceding material in the sentence (Gupta 1992; Lim 2007, i.a.). Example (4) shows *lah*, borrowed from Chinese and Malay (see Lim 2007, i.a. for details about *lah*'s functions and origins). Beyond sentence-final particles, another example is the adversative passive marker *kena*, from Malay (Bao and Wee 1999); see Example (5).

(4) But I look forward to go **lah**. (s1a-040)

(5) I **kena** shocked... ('I was shocked...'; s1a-096)

#### **2.1.2** English-origin words with novel uses

In SgE, some English words have acquired grammatical properties that are absent in BrE (or AmE), as shown in Examples (6)–(8). These are often calques from Chinese or Malay. They include several sentence-final particles: *also* (cf. Malay *juga* 'also, too'), *already*, which expresses a change of state (cf. Southern Min *liao* or Mandarin Chinese *le*; Bao 1995; i.a.), emphatic *one* (cf. e.g. Mandarin *de*; Bao 2009; note that *one* is also a numeral and noun, like in BrE and AmE). Another example is *got*, which can indicate existence and a perfective reading, like Southern Min *u* or Mandarin *you* (Lee, Ling and Nomoto 2009; Hiramoto and Sato 2012; Bao 2014, and references therein). That said, not all English words in this category have obvious Chinese or Malay analogues. One example is the sentence-final particle *what*, which is used to contradict a prior assertion. While the sentence-final use of *what* reflects Chinese and Malay influence, the exact connection *what* has with these languages is unclear (Lim 2007).

(6) ...Yen is tone deaf **also.** (s1a-066)

(7) I finish **already**. ('I have now finished [it].'; s1a-049)

(8) **Got** such thing **one**. ('There is indeed such a thing'; s1a-051)

(9) Then you can laminate **what**.
(Speaker argues that lamination is possible, contrary to what the listener is suggesting; s1a-061)

### 2.1.3 Borrowing of constructions distinguished by their syntax

More precisely, these are grammatical constructions distinguished not by the appearance of a particular word, but solely by word order or the omission of certain grammatical elements. This category includes topic-comment constructions, as in Example (10), which are used more freely in SgE due to Chinese and Malay influence (e.g. Ziegeler 2000; Bao 2001; Leuckert 2019). While these constructions are easily identified by their syntax – a topic followed by a comment clause – neither the topic nor the comment clause is marked morphosyntactically. Similar comments apply to bare conditional constructions (also found in Chinese; Bao and Lye 2005), illustrated in Example (11), in which both conditional and consequent clauses are unmarked.

(10) My office not many people will come…
('As for my office, not many people will come.'; s1a-007)

(11) You eat already you can die one… ('If you eat [them], you could die.'; s1a-007)

Additionally, SgE allows the omission of copulas, articles, and arguments (subject and objects), as shown in Examples (12)–(14).[2] This again reflects influence from Chinese and Malay, which allow copulas to be omitted more easily, lack definite articles, and allow null arguments (Ziegeler 2015; Lin 2022; Bao 2001 and references therein).

(12) My English __ very bad. ('My English is very bad'; omitted copula; s1a-031)

(13) But halfway through they changed __ music.
('… they changed the music'; omitted *the*; s1a-041)

[2] SgE also allows the omission of noun and verb inflections, e.g. there is no perfective or past tense morphology on *finish* in Example (7). This is also attributable to Chinese and Malay, which lack this kind of morphology. Additionally, the complementizer *that* (e.g. *I think __ it's raining*) can be omitted. In the rest of this article, our discussion of how omission affects parsing will focus on the omission of subjects and copulas. As far as we can tell, the other constructions do not pose as severe a parsing challenge, because there are similar AmE constructions without these elements. As a particularly straightforward example, the Stanford Parser can parse sentences without the complementizer *that*, because AmE also allows the omission of *that*.

(14) Actually one day __ must go to the beach…
(‘Actually one day [we] must go to the beach…’; omitted subject; s1a-030)

## 2.2 Parser software

This section briefly introduces freely available parsers developed for (standard) AmE. For scope reasons, we will not discuss the technical aspects or the statistical principles underlying these parsers. We do note, though, that these parsers are “trained” (created) by exposing them to a large corpus of syntactic trees that have been manually checked by linguistically-trained researchers. These parsers can produce high-quality parses for written standard AmE, with F-measure accuracy in the 80–90% range[3] (e.g. Klein and Manning 2003) when tested with material from the *Wall Street Journal* corpus of the Penn Treebank (Marcus, Santorini and Marcinkiewicz 1993; Taylor, Marcus and Santorini 2003).

Our study focuses on the Stanford Parser, which has been maintained since the early 2000s and is now part of a larger software suite (Qi et al. 2020). The standard installation of the parser comes with several AmE “grammars,” which were trained on, among other sources, the *Wall Street Journal* corpus. Consequently, it can produce tags and parses consistent with the tagset and phrase structure conventions of the Penn Treebank (Table 1). The parser can also parse texts that have been tagged for parts of speech; it also allows users to train their own parser using their own corpora. Both features offer ways to produce more accurate parses, as discussed in Section 3.3.

Other parsers potentially of interest include the Berkeley Neural Parser (Kitaev, Cao and Klein 2019), the now-unmaintained AllenNLP parser (Gardner et al. 2018), and the spaCy parser (Honnibal and Johnson 2015). The Berkeley and AllenNLP parsers are reported to be more accurate for written AmE than the Stanford Parser and allow training on user-supplied corpora. However, as far as we know, one limitation is that these parsers cannot process texts that have been tagged for parts of speech. As we noted above (also Section 3.3.1), more accurate parses can be produced if the parser has this feature and if users can provide such a parser with tagged texts.

[3] Briefly, to calculate the F-measure, the parse produced by the parser is compared with the correct parse that the parser should have produced. The F-measure reflects the percentage of constituents common to both parses.

**Table 1.** Examples of Penn Treebank part of speech tags and constituent labels

| Part of speech tag | Description |
|---|---|
| NN | Noun, singular |
| NNS | Noun, plural |
| VBP | Verb, plural |
| VBD | Verb, past tense |
| JJ | Adjective |
| JJR | Adjective, comparative |
| RB | Adverb |
| RBR | Adverb, comparative |
| DT | Determiner (e.g. *the, a*) |

| Constituency label | Description |
|---|---|
| S | Non-interrogative clause ("sentence") |
| NP | Noun phrase |
| VP | Verb phrase |
| ADJP | Adjective phrase |
| ADVP | Adverb phrase |
| PP | Prepositional phrase |
| SBAR | Embedded clause marked with a complementizer *that, if*, or *wh*-phrase |
| SBARQ | *Wh*-interrogative clause |
| SQ | Yes-no interrogative clause |

### 2.3 Applying the Stanford Parser to SgE corpus material

While the Stanford Parser can parse AmE rather accurately, we are interested in how well it can parse SgE, or more specifically, SgE material containing contact-induced linguistic features that are absent in AmE. To address this question, we used the Stanford Parser to parse sentences from the ICE-SIN "private dialogue" transcripts, which are transcripts of informal conversations and so are a rich source of these features. We generated part of speech tags and phrase structure parses using one of the AmE grammars (the so-called "englishPCFG model") that comes with the Stanford Parser.

There are certainly limitations to using ICE-SIN data. For instance, there is variation in the distribution of these features within transcripts, as one might expect. A more serious limitation is the fact that the ICE-SIN data, being conversation transcripts, contain speaker errors, such as false starts, ellipses, and grammatical errors. Speaker errors themselves can affect the accuracy of the Stanford Parser, whose training material mostly consists of carefully-edited written material.

Despite these limitations, we believe ICE-SIN provides a good source of SgE material with which to test the Stanford Parser. There are certainly corpora of written colloquial SgE that we could have used, such as the Corpus of Singapore English Messages (Gonzales et al. 2023); we could also have analyzed social media posts. However, it is not obvious that these sources are superior to ICE-SIN. The same issues with variation and errors (including typographical errors) also occur in these sources. An additional complication is that these sources often feature non-standard English orthography, such as initialisms and other abbreviations. This poses a problem for the Stanford Parser, whose training material was prepared in standard AmE orthography. In contrast, orthography is not an issue for the ICE-SIN material, which, aside from borrowings, was transcribed according to standard British English (BrE) orthography.

In order to fairly represent the parser's performance with the contact-induced features of interest, our examples below were restricted to ICE-SIN material that we judge is free of speaker errors. This way, any parser error observed can be confidently attributed to the presence of these features, rather than the parser being misled by speaker errors.

To identify parser errors, we checked the parses produced by the Stanford Parser against our own parsing conventions and parses for ICE-SIN; as part of a SgE treebank project (Huang et al. 2022), we had created and vetted phrase structure parses for all 100 ICE-SIN private dialogue files. Like the Stanford Parser's output, our parsing conventions closely adhere to the Penn Treebank's, which are richly documented and cover a wide range of constructions, many of which are also present in SgE.

## 3. Analysis of Stanford Parser performance

To preview this section, we found that the parser can relatively accurately parse many of the SgE features of interest, despite the linguistic differences between SgE and AmE. Generally, only function words from Chinese or Malay and grammatical constructions like sentence-final particles and topic-comment constructions present complications. Even so, we have noticed that certain function words and grammatical constructions can sometimes be parsed without much trouble.

Our analysis of errors suggests that there are two factors that determine parser accuracy. The first factor is the parser's "strategy" for novel words. Although the Stanford Parser does not provide documentation on exactly how it makes parsing decisions, our observations lead us to conclude that the parser is likely making educated guesses for novel words, based on what other words or constituents that it can identify in the same utterance, on the assumption that the overall utterance conforms to patterns present in AmE. This strategy allows the parser to perform relatively well for content word borrowings, for example.

The second factor is, given a SgE construction that lacks an AmE counterpart, whether that construction can be reasonably approximated using another AmE construction, i.e. whether some AmE construction contains a very similar sequence of constituents and/or words. If such an AmE construction exists, the parser will still generate reasonably accurate parses, even though native speakers of AmE might find the SgE construction unidiomatic or even completely unacceptable. This is the case for *got* and null subject constructions, for example. However, if no such AmE construction exists (e.g. sentence-final particles), then inevitably the parser will generate inappropriate parses.

To better illustrate parser performance, we will group together SgE features based on the categories outlined in Section 2.1: borrowings of content words, function words, and constructions. We contrast parser errors with the correct parse (or tags) according to our conventions. We then explain why the parser might have made these errors with reference to the two factors.

### 3.1 Parsing content word borrowings

As mentioned previously, SgE often features content words of Chinese or Malay origin. Since they are almost never used in AmE, one might wonder whether the Stanford Parser will misanalyze them and the phrases in which they occur.

We observed some variability in how the parser assigns part of speech tags to these borrowings, similar to what Lin et al. (2023) observed. Tagging accuracy depends on the context in which the word appears: in Example (15), the parser correctly tags *sian* 'bored' as JJ (adjective), probably because of the immediately preceding adverb *very*. In contrast, the parser incorrectly tags *tahan* 'endure' in Example (16) as NN (singular noun), likely because it lacks useful context to determine that *tahan* is a verb – the preceding word is another borrowing, *buay* 'cannot' (from Chinese), which the parser incorrectly tags as an adjective (and not a modal auxiliary). Similarly, for the three Malay borrowings in Example (17), the parser incorrectly tags the adverb *tak* 'not' as NN and the modal auxiliary *boleh* 'can' as VBP (plural verb). In this case, however, *tahan* is correctly tagged as a verb.

(15) Actually I feel very sian_JJ. (s1a-057)

(16) Buay_JJ tahan_NN.
cannot endure
'[I] cannot endure [this].' (s1a-088)

(17) Tak_NN boleh_VBP tahan_VB them
not can endure.
'[I] cannot endure them.' (s1a-067)

As one might expect, tagging errors are correlated with parsing errors. For instance, *buay tahan* in Example (16), tagged as adjective and noun, is incorrectly parsed as a noun phrase (NP). The case of *Tak boleh tahan them* in Example (17) is more interesting, as it illustrates how the parser can still produce reasonably good parses despite tagging errors. As Example (18) shows, both *tahan* and *boleh* are correctly attached to verb phrases (VPs), since the parser analyzed both words as verbs. The only parsing error is with *tak*. *Tak*, tagged incorrectly as a noun, is parsed as a NP, as if it were the sentence's subject.

(18) *Tak boleh tahan them.*
Stanford Parser: $[_{\text{S}}$ $[_{\text{NP}}$ tak] $[_{\text{VP}}$ boleh $[_{\text{VP}}$ tahan them]]]
Correct parse: $[_{\text{S}}$ $[_{\text{VP}}$ tak boleh $[_{\text{VP}}$ tahan them]]]
(Note: for presentation reasons, we will not show entire parse trees. Instead, we present parses in bracketed form to show which constituents words are attached to. We highlight only the brackets/constituents most relevant to the discussion.)

For the most part, however, there were few tagging or parsing problems for lexical borrowings like the problems in Examples (16) and (17). This is because in our data, these borrowings tend to occur in contexts like Example (15), where the part of speech and parse can be correctly inferred from the context. To the extent that the use of lexical borrowings in SgE is exemplified by tokens like Example (15) (and not Examples (16) and (17)), these lexical borrowings are not a major problem for the parser.

## 3.2 Parsing grammatical borrowings

### 3.2.1 English-origin words with novel uses

The Stanford Parser's accuracy for grammatical constructions is more uneven. Generally, the parser can accurately parse constructions if it contains a distinctive element that is derived from English and retains the same part of speech. Even though the word might be used in a novel way, this is not an issue for the parser, as long as structurally similar constructions exist in AmE. For illustration, consider sentence-final *also* and sentence-final *already* in Examples (19) and (20). Although these appear in positions that are less common in AmE (or BrE), the parser can still correctly tag them as adverbs and attach them to a VP. This is most likely because AmE allows many adverbs to attach to VPs and appear sentence-finally, as in Example (21), so sentence-final adverbs are common in the Stanford Parser's AmE training data.

(19) And Yen is tone deaf **also**. (s1a-066)

(20) I finish **already**. (s1a-049)

(21) I really want to see that happen **again**.
(*Wall Street Journal* subcorpus of Penn Treebank)

Similar remarks apply to *got* constructions (Lee, Ling, and Nomoto 2009; Hiramoto and Sato 2012; Bao 2014; i.a.). *Got* has acquired two novel uses under the influence of Chinese (via Southern Min *u*, Cantonese *yau*): it marks existence when paired with an NP, and a perfective reading when paired with a VP, as Examples (22) and (23) show, respectively. Both *got*s can describe situations in the present, even though *got* is ostensibly the past tense form of *get*. In order to differentiate them from standard English *got*, we follow Lin et al. (2023) in tagging them as GOT. In our parsing conventions, *got* appears with either a NP object (existence use) or a VP (perfective use) to form a larger VP.

(22) Where **got** time to make. ('Where is there time to make [this]?'; s1a-091)

(23) You **got** go underwater. ('You went underwater'; s1a-085)

Although these tagging and parsing conventions for *got* are new to the Stanford Parser, the parser still performs well. The typical problem encountered here was the parser tagging existential and perfective *got* as VBD, i.e. a past tense verb. Otherwise, the parser correctly parses *got* constructions; as Example (24) illustrates, perfective *got* combines with a VP to form a larger VP. The parsing accuracy is unsurprising. As Examples (25) and (26) show respectively, AmE *got* can also co-occur with an NP and with a VP, although in the VP case, *got* is a passive marker and not a perfective marker.

(24) *You got go underwater.*
Stanford Parser (also correct parse):
[$_{S}$ You [$_{VP}$ got [$_{VP}$ go underwater]]]

(25) They got [$_{NP}$ a small piece of the net profits]…
(*Wall Street Journal* subcorpus)

(26) The Giants got [$_{\mathrm{VP}}$ swamped in the second game]…
(*Wall Street Journal* subcorpus)

Conversely, the parser struggles when there is no obvious AmE analogue. Consider sentence-final *what* and *one*, shown in Examples (27) and (28). As sentence-final particles, *what* and *one* express the speaker's attitudes and/or emotions about the preceding material. Structurally, they ought to be attached high at the sentence level, directly to the S (= clause) node.

(27) Then you can laminate **what**.
(Speaker argues that lamination is possible, contrary to what the listener is suggesting; s1a-061)

(28) Got such thing **one**. ('There is indeed such a thing'; s1a-051)

However, because AmE lacks sentence-final particles, the parser consistently incorrectly tags *what* as WP (*wh*-word) and *one* as NN (noun) or CD (cardinal numeral) (see similar reports by Lin et al. 2023). These incorrect tags in turn lead to serious parsing errors. For instance, in Example (29), the parser incorrectly puts *what* inside a VP, as if *what* is the object of the verb *laminate*. Effectively, this sentence is mistakenly parsed as if it were an echo question (*You can laminate* what*?!*).

(29) *Then you can laminate what.* (s1a-061)
Stanford Parser:
[$_{\mathrm{S}}$ then you can [$_{\mathrm{VP}}$ laminate what]]
Correct parse:
[$_{\mathrm{S}}$ then you can [$_{\mathrm{VP}}$ laminate] [$_{\mathrm{SFP}}$ what]]

### 3.2.2 Function word borrowings

The same errors occur for other sentence-final particles with non-English origins, like *lor* or *lah*. Since these particles are new to the parser, it falls back on linguistic context to determine how to tag and parse them. However, linguistic context in this case can be misleading. In Example (30), *lor* is mistagged as a noun, likely because it follows *her*, which could be analyzed as a possessive. *Her lor* is then incorrectly parsed as a single constituent, namely, the NP object of the preposition *with*.

(30) *… we can play with her lor.* (s1a-091)
Stanford Parser:
[$_{\mathrm{S}}$ we can play [$_{\mathrm{PP}}$ with [$_{\mathrm{NP}}$ her lor]]]
Correct parse:
[$_{\mathrm{S}}$ we can play [$_{\mathrm{PP}}$ with her] [$_{\mathrm{SFP}}$ lor]]

In Example (31), the parser incorrectly treats the particle *lah*, appearing right after the verb *go*, as *go*'s object. Consequently, *lah* is tagged as a noun and is attached to the VP, instead of being attached to S, scoping over the rest of the sentence.

(31) *But I look forward to go lah.* (s1a-040)
Stanford Parser:
[$_{S}$ I look forward to [$_{VP}$ go [$_{NP}$ lah]]]
Correct parse:
[$_{S}$ I look forward to [$_{VP}$ go] [$_{SFP}$ lah]]

We next consider the adversative passive marker *kena*, which our conventions treat as a verb forming its own VP, like the auxiliary *be* in English. Although *kena* is yet another function word that is new to the parser, we find that the parser sometimes performs better with *kena* than with sentence-final particles. In Example (32), for instance, the parser correctly tags *kena* as a verb and attaches it to a VP. This is likely because the linguistic context for *kena* is more informative: *kena* appears immediately after a subject (*I*), and the parser (incorrectly) treats *shocked* as an adjective (not unreasonably, since passive participles can be used as verbs or adjectives). Since English sentences generally contain at least one verb, it seems that the parser preferred to analyse *kena* as the sentence's verb.

(32) *I kena shocked…* ('I was shocked…'; s1a-096)
Stanford Parser: [$_{S}$ I [$_{VP}$ kena [$_{ADJP}$ shocked]]]
Correct parse: [$_{S}$ I [$_{VP}$ kena] [$_{VP}$ shocked]]]

In summary, sentence-final particles and the adversative passive *kena* potentially pose a problem because they are words that the parser has no prior exposure to. In such cases, the parser bases its analysis on linguistic context alone. For sentence-final particles, the linguistic context is unhelpful and there is no analogous AmE construction that the parser can exploit, which results in parsing errors. In contrast, *kena* can appear in a linguistic context that facilitates accurate tagging and parsing.

### 3.2.3 Constructions with only distinctive syntax

We next consider two types of constructions that are not distinguished by any particular word. The first is topic-comment constructions, including bare conditional constructions, which are distinguished by their structure. The second is constructions involving omission.

Our parsing conventions define the constructions in Example (33) as having a "topic-comment" structure. Although the syntax and pragmatics of these constructions vary, what they have in common is that the left edge of the sentence contains a constituent with a distinctive information structure property (underlined in the sentences), which we might term the topic. In our parsing conventions, both the topic and the clause serving as its comment are attached to a special constituent, TBAR. This contrasts with the Penn Treebank, which does not have TBAR and instead attaches the topic directly to the clause, as if the topic were part of the clause. In other words, we posit a phrase structure rule along the lines of TBAR → NP S, illustrated in Example (34a); NP is not directly attached to S. The same goes for bare conditional constructions. We follow Bao and Lye (2005), who argue that the conditional clause is a kind of topic. Consequently, we attach both conditional and consequent clauses to TBAR, illustrated in Example (34b), entailing another rule TBAR → S S.

(33) a. Left dislocation
*So this one got to look into it lah.* (‘So this one, we've got to look into it.’; s1a-045)
b. Fronting
*That one I arrange later lor.* (‘That one, I will arrange it later’; s1a-091)
c. Chinese-style topic-comment/ “double subject” construction; Li and Thompson 1976)
*My office not many people will come*... (‘As for my office, not many people will come.’; s1a-007)
d. Chinese-style “object preposing” (cf. Huang 2018 and references therein)
*I this book read already.* (‘I have read this book.’; not attested in ICE-SIN material but acceptable to native speakers)
e. Bare conditional
*You want to buy you go…* (‘If you want to buy, you go.’; s1a-007)

(34) a. [$_{\text{TBAR}}$ So [$_{\text{NP}}$ this one] [$_{\text{S}}$ got to look into it lah]]
cf. a Penn Treebank-style parse: [$_{\text{S}}$ So [$_{\text{NP}}$ this one] got to look into it lah]
b. [$_{\text{TBAR}}$ [$_{\text{S}}$ You want to buy] [$_{\text{S}}$ you go]]

The parsing challenge here is similar to that for sentence-final particles: there is no TBAR in the Stanford Parser's training data, so the parser has to resort to AmE constructions, which are inappropriate according to our parsing conventions. In Example (35), the parser analyzes the topic *this one* as if it were the subject, attaching it to S, instead to TBAR. As for bare conditionals, illustrated in Example (36), the parser tends to incorrectly analyze the main clause *you go* as part of a subordinate clause.

(35) *So this one got to look into it lah.*
Stanford Parser:
[$_{\text{S}}$ So [$_{\text{NP}}$ this one] [$_{\text{VP}}$ got to look into it lah]]
Correct parse:
[$_{\text{TBAR}}$ So [$_{\text{NP}}$ this one] [$_{\text{S}}$ [$_{\text{VP}}$ got to look into it lah]]]

(36) *You want to buy, you go...*
Stanford Parser:
[$_{\text{S}}$ You want [$_{\text{S}}$ to buy, you go]]
Correct parse:
[$_{\text{TBAR}}$ [$_{\text{S}}$ You want to buy], [$_{\text{S}}$ you go]]

We next discuss constructions involving the omission of subjects and copulas, which can be attributed to Chinese and Malay influence. Ideally, the parser should parse these constructions with exactly the same structure as their counterparts with overt subjects and copulas. However, we find that the parser does not always do so.

As mentioned before, omitted subjects ("null subjects") have a much wider distribution in SgE; specifically, subjects of finite and nonfinite clauses can be null in SgE, while as Example (37) illustrates, only the subjects of nonfinite clauses can be null in AmE. This difference, however, only occasionally poses a parsing problem. In Example (38), the finite clause *can understand* is correctly analyzed as a subjectless S, presumably by analogy to AmE nonfinite clauses, even though the sentence is ungrammatical in AmE. This again shows how the parser can generate an acceptable parse for a SgE construction, as long as there is a structurally similar AmE construction available.

(37) I really want [$_{\text{S}}$ __ [$_{\text{VP}}$ to see that happen again]].
(*Wall Street Journal* subcorpus)

(38) *So I think __ can understand*
('So I think I can understand.'; s1a-090)
Stanford parser (also correct parse):
So I think [$_{\text{S}}$ __ [$_{\text{VP}}$ can understand]]

That said, the parser can occasionally make mistakes with null subjects, especially when the complementizer *that* is also absent. In Example (39), omitting *that* and the subject obscures the fact that *was you* is a clause (an S attached to an SBAR, according to our parsing conventions). Coupled with the fact that *thought* could be analyzed as a noun, the parser incorrectly analyzes *was you* as just a VP, and *my mom thought* as an NP, the subject of *was you.*

(39) My mom thought __ was you ('My mom thought that it was you.'; s1a-066)
Stanford Parser:
[$_{\text{S}}$ [$_{\text{NP}}$ My mom thought] [$_{\text{VP}}$ was you]]
Correct parse:
[$_{\text{S}}$ [$_{\text{NP}}$ My mom] [$_{\text{VP}}$ thought [$_{\text{SBAR}}$ [$_{\text{S}}$ __ was you]]]]

Omitted copulas pose a more serious problem for the parser, because there is no analogous AmE construction. Superficially, omitted copula constructions are clauses that contain a subject (which itself can be omitted) and a predicate like a NP, ADJP (adjective phrase), or VP with a gerundival or participial main verb, but no copula verb. Setting aside the case of the VP predicate, this construction is inconsistent with the generalization that an English clause always consists of a subject followed by a verb. Example (40) illustrates this construction with an ADJP predicate: The embedded clause here is *now more expensive*, without a verb (or subject, in this case). The Stanford Parser therefore fails to correctly analyze this construction as an embedded clause (an SBAR and/or S); instead, it treats *now more expensive* as an ADJP and attaches that directly to *think*.

(40) *I think now ___ more expensive.* ("...now it is more expensive"; s1a-011)
Stanford Parser:
... think [$_{\text{ADJP}}$ now more expensive]
Correct parse:
... think [$_{\text{SBAR}}$ [$_{\text{S}}$ now __ [$_{\text{ADJP}}$ more expensive]]]

### 3.3 Improving parsing accuracies

We next discuss two general strategies for improving parsing accuracies, which we have tested on the Stanford Parser but not on other off-the-shelf parsers (which might not have similar functionalities or be as easy to use). Note that both strategies require some degree of manual annotation, which might not always be feasible due to time and resource constraints.

#### 3.3.1 Providing the parser with part of speech tags

We saw above that parsing errors are correlated with tagging errors. One solution, therefore, is to first enrich the material with the correct parts of speech before parsing it. The Stanford Parser is designed so that it can incorporate user-supplied part of speech tags when making parsing decisions.

We found this strategy to be helpful when grammatical borrowings produce syntactically ambiguous sentences, which can cause the parser to select an inappropriate parse. For instance, consider Example (39), *My mom thought was you.* By explicitly tagging *thought* as a verb ("VBD") as in Example (41), we disambiguate the sentence and block the parser from analyzing *my mom thought* as an NP subject for *was you*. Consequently, the parser correctly analyzes *was you* as a subordinate clause.

(41) My_PRP$ mom_NN thought_VBD was_VBD you_PRP ._.
Stanford Parser (also correct parse):
[$_{S}$ [My mom] [thought [$_{SBAR}$ [$_{S}$ __ was you]]]]

Tagging is also helpful for lexical borrowings, provided they occur in contexts consistent with AmE phrase structure rules. Example (16), *Buay tahan*, presents a clear illustration. Untagged, it was incorrectly parsed as an NP. After tagging *buay tahan* as a modal auxiliary and verb, as in Example (42), the parser successfully parses it as a (subjectless) sentence, due to prior exposure to AmE sentences containing a modal auxiliary followed by a verb.

(42) Buay_MD tahan_VB ._.
('[I] cannot endure [this]')
Stanford Parser (also correct parse):
[$_{S}$ __ [$_{VP}$ Buay tahan]]

However, tagging is still of no help for grammatical borrowings that require novel part of speech tags or lack AmE analogues. In Example (43), *lah*, despite being correctly tagged as SFP, is still incorrectly analyzed as the NP object of *go*. This is because the SFP tag itself is new to the Stanford Parser: There is no such tag in its AmE training data.

(43) … I_PRP look_VBP forward_RB to_TO go_VB lah_SFP ._. (s1a-040)
Stanford Parser:
[$_{S}$ I look forward to [$_{VP}$ go [$_{NP}$ [$_{SFP}$ lah]]]]
Correct parse:
[$_{S}$ I look forward to [$_{VP}$ go] [$_{SFP}$ lah]]

As an estimate of how tagging improves parsing accuracy, we ran an analysis on all 100 private dialogue ICE-SIN files (s1a files), for which we had created and vetted parses. First, as a baseline, we simultaneously tagged and parsed sentences from these files using the Stanford Parser's AmE englishPCFG grammar. We then compared the parser's output with our manually-vetted parses using the software EVALB (Sekine and Collins 2013), which is commonly used for computing parser accuracy measures. Next, we repeated the same accuracy analysis, except this time we fed the parser tagged sentences. As Table 2 shows, doing so increases overall parsing accuracy (F-measure; paired sample t-test $t(99) = 26.5$, $p < 0.001$) and the percentage of sentences parsed without any errors (paired sample t-test $t(99) = 25.2$, $p < 0.001$). While these numbers are promising, we caution against expecting this approach to always produce similarly high accuracies for SgE (or other English varieties). This is because parsing accuracy is sensitive to other factors, such as speaker/writer errors, nonstandard orthography, and the availability of alternative parses (i.e. syntactic ambiguity; see Jurafsky and Martin 2023).

**Table 2.** Parsing accuracy for private dialogue ICE-SIN files

| | Mean (s.d.), without part of speech tags | Mean (s.d.), with part of speech tags | Difference |
|---|---|---|---|
| Overall parsing accuracy (F-measure) | 78.6 (3.9) | 83.0 (3.4) | +4.3 |
| % sentences without parsing errors | 41.8 (7.7) | 56.6 (7.7) | +14.8 |

### 3.3.2 Training the parser on hand-corrected parses

The only way for a parser to correctly analyze novel tags and grammatical constructions is to train it on parses already containing these tags and constructions, i.e. create a customized parser from a user-supplied treebank. Fortunately, some parsers, like the Stanford Parser, come with such a training function. Note, however, that this approach still does not guarantee perfect accuracy. Parses produced using this approach will still need to be manually reviewed and corrected.

This strategy is significantly more difficult than the strategy involving tagging (Section 3.3.1) because it requires first creating and vetting parses for training. The Penn Treebank team estimated that a human vetter with about three to four months' experience, working with machine-generated parses, can review and correct about 750–1,000 words per hour (Marcus, Santorini and Marcinkiewicz 1993). With SgE, however, we encountered a much lower rate of about 450 words per hour, possibly because our SgE material contains more production errors than the average Penn Treebank material, much of which has undergone careful editing. Moreover, for this strategy to work, one would need to assemble a substantial training corpus. In fact, our own experience suggests that the larger the training corpus, the greater the accuracy improvement. For this reason, we will not report estimates of improvements here, since that depends on one's training corpus.

Although this strategy is costly, we mention it because it is the only solution that eventually allows the parser to automatically generate parses conforming to one's own tags and parsing conventions (rather than preexisting ones, like the Penn Treebank's). This strategy is therefore ideal if one's material contains substantial amounts of borrowings that cannot be easily approximated using AmE phrase structure rules or part of speech categories.

## 3.4 Interim summary

The preceding sections reviewed contact-derived linguistic features of SgE. Sections 3.1 and 3.2 showed that some of these features are easier for the Stanford Parser than others. We argued that this outcome is not accidental but reflects the parser's strategy: when presented with unfamiliar words and constructions, it tries to analyze them using AmE grammatical rules and constructions. This works well, but only to the extent that there are clear AmE structural analogues. Section 3.3 discussed strategies for improving parser accuracy. Although these options require first investing resources into manual annotation, in our experience, they can deliver meaningful improvements in accuracy.

We believe that these results, even though they are based on SgE, are relevant to research on New English varieties more broadly. Like SgE, other New Englishes also differ from (standard) AmE in having content words, function words, and grammatical constructions borrowed from non-English languages. To the extent that an AmE parser can handle these linguistic features in SgE without much difficulty, the same is likely true for equivalent features in other New Englishes.

These findings have further implications for research on these varieties, as they could help researchers allocate their time and resources more strategically when creating parses. For instance, in our SgE treebank project, we used the two strategies described above – tagging all sentences and training the Stanford Parser – to quickly generate an initial set of parses that were relatively accurate. This then let us focus on checking, for instance, grammatical constructions that lack obvious AmE analogues, such as topic-comment constructions, instead of distributing our efforts uniformly across linguistic features, such as content word borrowings.

## 4. An exploratory analysis of noun phrases in SgE

Having shown that creating high-quality parses for a New English can be less resource-intensive than expected, we next give an example of how parses can be used with an exploratory analysis of SgE noun phrases (NPs). To manage reader expectations, we should clarify that this is not a comprehensive study of NPs. We see this article as primarily a study of parsing in the World English context. NPs are just one out of many phenomena for illustrating the value parses can bring.

However, NPs make an interesting case study, because they show substantial internal diversity that is best studied using parses. NPs can be realized as sequences of determiners, nouns, and other modifiers, or more simply as pronouns or demonstratives. Because of language contact with Chinese, Malay, and even South Asian languages like Tamil, SgE is pro-drop, allowing subject NPs to be null (omitted) even in finite clauses (Bao 2001; Leimgruber 2013; Sato 2016; Lee 2022, and references therein). Language contact also has led to SgE allowing "bare" NPs, without determiners, possessive pronouns, and quantifiers. Such bare NPs are acceptable even if the head noun is a singular count noun, as in Example (44); note that equivalent constructions are ungrammatical in BrE and AmE.

(44) Probably you'll see __ sunset… (s1a-001) ('…see the sunset')

### 4.1 Data and analysis

In this case study we investigate the diversity of SgE NPs by posing four questions:

1. In structural terms, what are the most common types of NPs?
2. Where do NPs appear: what structures are they attached to?
3. What is the rate of null subjects?
4. How often do singular nouns appear "bare"?

To answer the first two questions, we used the Python Natural Language Toolkit (NLTK; Bird, Klein and Loper 2009) to write a script to analyze NPs (excluding omitted NPs) in the parses we created for ICE-SIN private dialogues. In addition to identifying the constituents within an NP, we identified what constituent the NP is attached to: a clause, a VP, the topic-comment constituent TBAR, etc.

To estimate the rate of null subjects in our data, we count how many subordinate clauses have NP subjects (i.e. overt NPs attached to the clause) or lack overt subjects altogether. We exclude main clauses because main clauses can be open to alternative analyses. For example, consider the hypothetical example *can understand.* On its own, it is difficult to decide whether this is a main clause without an overt subject or just a verb phrase (VP) fragment. In contrast, in a sentence like *I think can understand*, it is clear that *can understand* is a subjectless subordinate clause of *think.*

Since SgE allows null subjects in finite clauses (unlike BrE or AmE), we also track whether clauses are finite. This is easily done with parses. For example, according to the Penn Treebank conventions we follow, finite clauses have main VPs headed by a modal auxiliary (tagged as MD) or a tensed verb (VBZ, VBP, or VBD), while nonfinite clauses typically contain the marker *to.*

To estimate the rate of bare singular NPs, we re-use the same NP data set, this time counting NPs containing a NN (singular noun). We then check whether that NP also contains a determiner (DT), possessive pronoun (PRP$) or a quantifier phrase (QP) or cardinal numeral (CD). We further run the same analysis on the Penn Treebank's Switchboard Corpus, to obtain a baseline rate of bare singular NPs in AmE. There are certainly other treebanks we could have used (e.g. ICE-GB or other Penn Treebank corpora), but the Switchboard corpus offers two advantages. First, it consists of informal AmE conversations, which is a good match for our ICE-SIN conversation transcripts. Second, as noted above, our own parsing conventions closely adhere to the Penn Treebank's, so it is straightforward to run the same analysis on the Switchboard data.

## 4.2 Results

As Table 3 shows, the most frequent type of overt NPs in our ICE-SIN material are pronouns, followed by determiner–noun sequences and singular nouns. Despite the diversity, the most frequent types are short, usually between one to three words long.

**Table 3.** The ten most frequent types of overt NPs in ICE-SIN private dialogue (s1a; SgE) files

| **Type (Penn Treebank part of speech tag / constituent label)** | **Absolute frequency** | **% of overt NPs** |
|---|---|---|
| Pronoun (PRP) | 28,449 | 46.3 |
| Determiner–noun (DT NN) | 4,358 | 7.1 |
| Singular noun (NN) | 2,981 | 4.9 |
| Determiner (DT) | 2,542 | 4.1 |
| NP–prepositional phrase (NP PP) | 1,997 | 3.3 |
| Proper name (NNP) | 1,979 | 3.2 |
| *Wh*-word (WP) | 1,698 | 2.8 |
| Possessive–noun (PRP$ NN) | 1,195 | 1.9 |
| Determiner–adjective–noun (DT JJ NN) | 1,004 | 1.6 |
| Plural noun (NNS) | 967 | 1.6 |
| **Total of top 10** | **47,170 (out of 61,390 total)** | **76.8** |

Table 4 shows that almost half of all overt NPs are attached to non-interrogative clauses (S). Further inspection shows that these are mostly subjects, with exceptions like temporal NPs *today* or *last week*, which typically have an adverb-like function. About 21% are attached to VPs, i.e. objects, and a slightly smaller number are part of a PP. Notably, about 0.6% of NPs are topics, attaching to a TBAR constituent. This number might seem low, given proposals that SgE is a topic-prominent language (in the sense of Li and Thompson 1976) due to influence from Chinese (Bao 2001; Bao and Lye 2005; Sato 2016; Leuckert 2019; Lee 2022; and references therein, among many others). In hindsight, however, this is not unexpected. Subjects often function as the topic of a sentence without necessitating a topic-comment construction. Moreover, once a topic-comment structure successfully establishes a topic, speakers presumably have less of a need to use these constructions again shortly after.

**Table 4.** The ten most frequent constituents to which overt NPs are attached in ICE-SIN private dialogue (s1a; SgE) files

| **Type (Penn Treebank constituent label)** | **Absolute frequency** | **% of overt NPs** |
|---|---|---|
| Non-interrogative clause (S) | 28,219 | 46.0 |
| Verb phrase (VP) | 12,835 | 20.9 |
| Prepositional phrase (PP) | 10,480 | 17.1 |
| None, i.e. NP fragments | 2,757 | 4.5 |
| Yes-no question (SQ) | 2,576 | 4.2 |
| Fragment (FRAG) | 1,619 | 2.6 |
| Embedded clause with *that, for*, or *wh*-phrase (SBAR) | 1,146 | 1.9 |
| *Wh*-question (SBARQ) | 623 | 1.0 |
| Topic-comment construction (TBAR) | 383 | 0.6 |
| "Unlike coordination phrase" (UCP; e.g. a NP conjoined with a clause) | 239 | 0.4 |
| **Total of top 10** | **60,877** | **99.2** |

Table 5 shows that the finiteness of a subordinate clause affects what kind of subjects it has. In nonfinite clauses, most NP subjects are null (90.2%), followed by pronouns (5.8%). In contrast, in finite clauses, the most common NP subjects are pronouns (77.7%), followed by null subjects (5.0%). The percentage of null subjects might seem low, but again, it is not unexpected. First, null subjects are merely optional, even in prototypical pro-drop languages. Furthermore, null subjects are only possible if there are antecedents in the context. It is quite possible that antecedents for null subjects are not always available in our data. Since null subjects are present in Chinese and Malay, their relative productivity in our ICE-SIN materials (second-most common subject type in finite subordinate clauses) provides corpus-based quantitative evidence for the influence of Chinese and Malay on the grammar of SgE.

Table 6 compares singular NPs in the ICE-SIN and Switchboard data. Overall, the distribution of singular NPs is similar for both varieties. Nevertheless, there are differences that support existing characterizations of SgE. Notably, determiners are less common in ICE-SIN than in Switchboard (49.8% vs. 55.1%). Correspondingly, bare singular NPs are more common in ICE-SIN than in Switchboard (36.1% vs. 32.5%). Both observations are consistent with reports that SgE tends to omit determiners and allow bare (singular) NPs.

**Table 5.** The five most frequent types of NP subjects in nonfinite and finite subordinate clauses in ICE-SIN private dialogue (s1a; SgE) files

| **Nonfinite clauses** | | |
|---|---|---|
| **Type (Penn Treebank part of speech tags)** | **Absolute frequency** | **% of NPs** |
| Null | 3,502 | 90.2 |
| Pronoun (PRP) | 227 | 5.8 |
| Determiner–Singular noun (DT NN) | 31 | 0.8 |
| Singular noun (NN) | 19 | 0.5 |
| Possessive–Singular noun (PRP$ NN) | 11 | 0.3 |
| **Total of top 5** | **3,790 (out of 3,881 total)** | **97.7** |

| **Finite clauses** | | |
|---|---|---|
| **Type (Penn Treebank part of speech tags)** | **Absolute frequency** | **% of NPs** |
| Pronoun (PRP) | 6,961 | 77.7 |
| Null | 448 | 5.0 |
| *Wh*-phrase (WHNP) | 416 | 4.6 |
| Determiner–Singular noun (DT NN) | 177 | 2.0 |
| Determiner (DT) | 122 | 1.4 |
| **Total** | **8,124 (out of 8,960 total)** | **90.7** |

**Table 6.** Singular NPs in SgE (ICE-SIN private dialogue) and AmE (Switchboard)

| | **SgE (ICE-SIN private dialogues)** | | **AmE (Switchboard)** | |
|---|---|---|---|---|
| **Type of singular NP** | **Absolute frequency** | **%** | **Absolute frequency** | **%** |
| Not bare: determiner present | 6,898 | 49.8 | 23,960 | 55.1 |
| Not bare: possessive pronoun present | 1,562 | 11.3 | 4,649 | 10.7 |
| Not bare: quantifier present | 394 | 2.8 | 746 | 1.7 |
| Bare: Does not co-occur with determiners, possessive pronouns, or numerals | 5,001 | 36.1 | 14,121 | 32.5 |
| **Total** | **13,855** | **100.0** | **43,476** | **100.0** |

It is worth pointing out that this analysis has limitations. Our analysis considers only singular NPs but not plural ones, where determiners can also be omitted more freely in SgE (see Lin 2022 for additional discussion). We are further assuming that the distribution of mass nouns and count nouns is comparable in both data sets. For instance, an alternative explanation of why bare singular NPs are more common in ICE-SIN is simply because ICE-SIN happens to contain more mass nouns like *furniture*, which are grammatical and acceptable when appearing in a bare NP. This strikes us as unlikely, but confirming this assumption will require further study. Still, this analysis provides yet another illustration of how parses can enrich our understanding of syntactic differences between English varieties.

### 4.3 Future research directions using parses

The analyses of SgE NPs presented above are exploratory but they raise new research questions of their own. First, one could ask whether these findings can be replicated for other SgE corpora. Second, while we have attributed the presence of null subjects in SgE to influence from e.g. Malay and Chinese, it remains to be seen whether null subjects appear at a similar rate in comparable Malay and/or Chinese corpora. If they do, this would provide even stronger corpus-based evidence for the influence of these languages. Additionally, one might wonder whether within ICE-SIN transcripts the rates of null arguments and bare singular NPs are correlated with each other, since they are both said to be the result of language contact.

For scope reasons, we will set aside these questions for now. But we note that these are all questions about SgE and language contact that arise in response to this exploratory analysis of NPs. Although it is certainly possible to address these questions without parses, it should be evident that the availability of parses makes a large-scale quantitative analysis much easier. We hope that similar analyses, made possible with high-quality treebanks, will enable new areas of research.

## 5. Conclusion

Treebanks can play an important role in research on varieties of English, especially for research questions that require quantitative analyses of constructions and morphosyntactic features. Unfortunately, for cost reasons, few such resources are available currently. While there are off-the-shelf parsers available, they are almost always intended for varieties like standard AmE.

This paper aimed to confirm the value of treebanks as a research resource and to discuss how researchers can close this resource gap, using SgE as a case study. We found that an AmE parser can deliver surprisingly satisfactory results for SgE, despite SgE's contact-induced lexical and grammatical features. For content word borrowings, the parser can usually analyze the overall structure correctly by using the broader linguistic context the word occurs in. Grammatical borrowings can be more challenging but are not always so. More precisely, the most difficult features for the parser are function words that are absent in AmE (e.g. sentence-final particles) or constructions that lack analogues in AmE (e.g. copula omission). The parser has no prior exposure to such features, and so struggles with them.

These conclusions, which we expect to generalize beyond SgE, are useful because they clarify the circumstances under which off-the-shelf parsers can be feasibly used for the study of New Englishes. They indicate which aspects of creating parses and treebanks will require more attention from trained annotators and which aspects could use less. This knowledge is important, considering how labor-intensive and costly annotation can be. To that end, we also highlighted strategies for improving parser accuracy. However, a more important point that we make is that despite obvious differences, there are still likely to be enough overlaps in the linguistic features of AmE and New Englishes, such that parsers developed for AmE can be productively used for the latter.

Finally, to illustrate the value of parses, we presented an exploratory analysis of NPs in SgE, which is best performed using a treebank instead of, for example, a concordance. We showed how our analysis, even though preliminary, can help identify additional questions for future research. We hope that our discussion illustrates the potential contribution of treebanks and parses for research into World Englishes and will encourage the development of high-quality treebanks of New Englishes.

## Funding

The authors wish to acknowledge research funding support from the following sources: the Humanities and Social Science Research Fund and a faculty start-up grant, National University of Singapore; DSO National Laboratories, Singapore; and the Social Science Research Council, Singapore.

## Acknowledgments

We are grateful to two anonymous reviewers for detailed comments that have helped improve this paper and to Wong Zi Shu for editorial assistance.

## Sources

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Greenbaum, Sidney, and Gerald Nelson. 1996. "The International Corpus of English (ICE) project." *World Englishes* 15: 3–15.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. "Building a large annotated corpus of English: The Penn Treebank." *Computational Linguistics* 19: 313–330.

Taylor, Ann, Mitchell P. Marcus and Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeillé, ed., *Treebanks*. Dordrecht: Springer, 5–22.

Huang, Nick, Jia Wen Hing, Li Lin, and Zhiming Bao. 2022. "A tagged and annotated corpus of Singapore English." National University of Singapore manuscript.

## References

Bao, Zhiming. 1995. "*Already* in Singapore English". *World Englishes* 14: 181–188.

Bao, Zhiming. 2001. "The Origins of Empty Categories in Singapore English". *Journal of Pidgin and Creole Languages* 16: 275–319.

Bao, Zhiming. 2009. "*One* in Singapore English". *Studies in Language* 33: 338–365.

Bao, Zhiming. 2014. "*Got* in Singapore English". In Eugene Green, and Charles F. Meyer, eds. *The Variability of Current World Englishes*. Berlin: De Gruyter Mouton, 147–165.

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

Bao, Zhiming, and Hui Min Lye. 2005. "Systemic Transfer, Topic Prominence, and the Bare Conditional in Singapore English". *Journal of Pidgin and Creole Languages* 20: 269–291.

Bao, Zhiming, and Lionel Wee. 1999. "The Passive in Singapore English". *World Englishes* 18: 1–11.

Crewe, William J. 1977. *Singapore English and Standard English: Exercises in Awareness*. Singapore: Eastern Universities Press.

Davies, Mark, and Robert Fuchs. 2015. "Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-based English Corpus (GloWbE)". *English World-Wide* 36: 1–28.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. "AllenNLP: A Deep Semantic Natural Language Processing Platform". *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 1–6.

Gonzales, Wilkinson Daniel Wong. 2023. "Broadening Horizons in the Diachronic and Sociolinguistic Study of Philippine English with the Twitter Corpus of Philippine Englishes (TCOPE)". *English World-Wide* 44: 403–434.

Gonzales, Wilkinson Daniel Wong, Mie Hiramoto, Jakob R. E. Leimgruber, and Jun Jie Lim. 2023. "The Corpus of Singapore English Messages (CoSEM)". *World Englishes* 42: 371–388.

Greenbaum, Sidney. 1988. "A Proposal for an International Computerized Corpus of English". *World Englishes* 7: 315.

Gupta, Anthea Fraser. 1992. "The Pragmatic Particles of Singapore Colloquial English". *Journal of Pragmatics* 18: 31–57.

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

Hiramoto, Mie, and Yosuke Sato. 2012. "*Got*-interrogatives and Answers in Colloquial Singapore English: Aktionsart and Stativity". *World Englishes* 31: 186–195.

Honnibal, Matthew, and Mark Johnson. 2015. "An Improved Non-monotonic Transition System for Dependency Parsing". *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1373–1378.

Huang, Nick. 2018. "Control Complements in Mandarin Chinese: Implications for Restructuring and the Chinese Finiteness Debate". *Journal of East Asian Linguistics* 27: 347–376.

Jurafsky, Dan, and James H. Martin. 2023. *Speech and Language Processing* (3rd ed.).

Kachru, Braj B., ed. 1982. *The Other Tongue*. Urbana: University of Illinois Press.

Kirk, John M. 2017. "Developments in the Spoken Component of ICE Corpora". *World Englishes* 36: 371–386.

Kirk, John M., and Gerald Nelson. 2018. "The International Corpus of English project: A Progress Report". *World Englishes* 37: 697–716.

Kitaev, Nikita, Steven Cao, and Dan Klein. 2019. "Multilingual Constituency Parsing with Self-attention and Pre-training". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3499–3505.

Klein, Dan, and Christopher D. Manning. 2003. "Accurate Unlexicalized Parsing". *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.

Lee, Nala H., Ai Ping Ling, and Hiroki Nomoto. 2009. "Colloquial Singapore English *got*: Functions and Substratal Influences". *World Englishes* 28: 293–318.

Lee, Si Kai. 2022. "On Agreement-drop in Singlish: Topics Never Agree". *Glossa: A Journal of General Linguistics* 45: 1–27.

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

Leuckert, Sven. 2019. *Topicalization in Asian Englishes*. New York: Routledge.

Li, Charles N., and Sandra A. Thompson. 1976. "Subject and Topic: A New Typology of Language". In Charles N. Li, ed. *Subject and Topic*. New York: Academic Press, 457–489.

Lim, Lisa. 2007. "Mergers and Acquisitions: On the Ages and Origins of Singapore English Particles". *World Englishes* 27: 446–473.

Lin, Li. 2022. "A Corpus-based Grammar of Singapore English: Description and Change". Ph.D. Dissertation, National University of Singapore.

Lin, Li, Kunmei Han, Jia Wen Hing, Luwen Cao, Vincent Ooi, Nick Huang, and Zhiming Bao. 2023. "Tagging Singapore English". *World Englishes* 42: 624–641.

Nelson, Gerald, Sean A. Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Platt, John T. 1975. "The Singapore English Speech Continuum and its Basilect 'Singlish' as a 'Creoloid'". *Anthropological Linguistics* 17: 363–374.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". *Association for Computational Linguistics (ACL) System Demonstrations*, 101–108.

Sato, Yosuke. 2016. "Remarks on the Parameters of Argument Ellipsis: A New Perspective From Singapore English". *Syntax* 19: 392–411.

Sekine, Satoshi, and Michael John Collins. 2013. "Evalb". Available at https://nlp.cs.nyu.edu/evalb/

Tay, Mary W. J. 1979. "The Uses, Users and Features of English in Singapore". In Jack C. Richards ed. *New Varieties of Englishes*. Singapore: SEAMEO Regional Language Centre, 91–111.

Teo, Ming Chew. 2020. *Crosslinguistic Influence in Singapore English: Linguistic and Social Aspects*. London: Routledge.

Tongue, Ray K. 1979. *The English of Singapore and Malaysia*. Singapore: Eastern University Press.

Wallis, Sean. A., and Gerald Nelson. 2000. "Exploiting Fuzzy Tree Fragments in the Investigation of Parsed Corpora". *Literary and Linguistic Computing* 15: 339–361.

Wee, Lionel. 2018. *The Singlish Controversy: Language, Culture, and Identity in a Globalizing World*. Cambridge: Cambridge University Press.

Ziegeler, Debra. 2000. *Hypothetical Modality: Grammaticalization in an L2 Dialect.* Amsterdam: John Benjamins.

Ziegeler, Debra. 2015. *Converging Grammars: Constructions in Singapore English*. Berlin: De Gruyter Mouton.

## Address for correspondence

Nick Huang
Department of English, Linguistics, and Theatre Studies
National University of Singapore
Block AS5, 7 Arts Link
Singapore, 117570
Singapore

znhuang@nus.edu.sg
https://orcid.org/0000-0001-8022-1097

## Co-author information

Li Lin
East China University of Political Science and Law
3136@ecupl.edu.cn

Kunmei Han
National University of Singapore
kunmei.han@u.nus.edu
https://orcid.org/0000-0001-5986-9913

Jia Wen Hing
National University of Singapore
jw_hing@nus.edu.sg
https://orcid.org/0000-0001-8181-2905

Luwen Cao
National University of Singapore
luwencao@nus.edu.sg
https://orcid.org/0009-0006-1794-1271

Vincent Ooi
National University of Singapore
vinceooi@nus.edu.sg
https://orcid.org/0000-0002-8005-7407

Zhiming Bao
National University of Singapore
bao@nus.edu.sg
https://orcid.org/0000-0002-7859-2654

## Publication history

Date received: 4 December 2023
Date accepted: 5 February 2025
Published online: 21 March 2025