Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 431 – 441

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

The OPT-ional Phenomenon in Singapore English: a Corpus-based Approach Using Time Annotated Corpora

Eric YongMing Lai, Liling Tan, Vincent Wong, Lenny Teng Tao Loke, Francis Bond*

Division of Linguistics and Multilingual Studies, Nanyang Technological University, 14 Nanyang Drive, Singapore 637332

Abstract

The Optional Omission of Past Tense (OPT) is prevalent in the colloquial register of Singapore English (SCE). This paper describes the investigation of the OPT phenomenon based on time annotated corpora. The Singapore and Hong Kong version of the International Corpus of English were extended with time annotation for this study. In Singapore English sentences that contain the perfective aspectual adverbs of already or yesterday, the OPT-ional phenomenon is found to be present 68.2% of the time. Although this phenomenon is also found in Hong Kong English, it is significantly more prominent in Singapore English ($p<0.05$, $z=6.27$). In SCE, there are also syntactic constraints that influence the OPT occurrences, where the omission of past tense occurs 39.8% more frequently in sentences with pre-verbal adverbs than post-verbal adverbs.

Selection and peer-review under responsibility of CILC2013.

Keywords: world Englishes; Singlish; Singapore English; Hong Kong English; corpus; annotation; tenses

1. Introduction

Singapore Standard English (SSE) and Singapore Colloquial English (SCE) are widely accepted as the two main registers of spoken English in Singapore (Platt, 1977; Gupta, 1994; Boa and Hong, 2006). Although SSE is the officially prescribed variety in the country, SCE is the predominant variety of day-to-day interactions (Cavallaro and Ng, 2009). With regard to intelligibility, SSE is not significantly different from the globalized varieties of English (i.e. British English and American English) (Gupta, 2005, Kirkpatrick and Saunders, 2005). SCE, however, has been richly influenced by substratum languages (Deterding and Poedjosoedarmo, 1998; Wee, 2003; Lim, 2007) and the

* Corresponding author. Tel.: +65-6592-1568 ; fax.: +65-6794-6303
E-mail address: bond@ieee.org

1877-0428
Selection and peer-review under responsibility of CILC2013.
doi:10.1016/j.sbspro.2013.10.666

resultant variety has undergone considerable phonological, morphological, and syntactic restructuring. This study investigates one of these restructured features in SCE viz. the Optional Omission of Past Tense (OPT) in SCE.

In literature on SCE, Crewe (1977) was among the first to observe non-standard tense marking. This was further discussed by Ho and Platt (1993) where the omission of the '-s' suffix in third-person singular present tense (e.g. '*he eat cheese*') was observed. Subsequently, Alsagoff and Ho (1998) noted an overgeneralization of the '-s' suffix in the first person present tense (e.g. 'I eats cheese'). Brown (2000) stated that in SCE "*present tense does not necessarily mean the present time... It can also be used, in certain circumstances, to refer to the past and the future*" (c.f. Sheng, 2007). Gut (2009) recognized this occasional lack of verbal past tense (e.g. 'I eat cheese just now') in SCE and suggested a shift in the function of present tense marking. In addition, Bao (1995) claimed that *already* in SCE marks the inchoative (1a) or perfective (1b) aspect and that the presence of already in a sentence allows SCE speakers to use a present tense verb to convey the past.

(1) My baby speak already. (c.f. Bao, 1995)
a. 'My baby has started to speak'
b. 'My baby has spoken'

The OPT phenomenon is a recurrent theme in studies on SCE (e.g. Alsagoff (2001), Deterding (2003), Ho (2003), Wee (2004), Fong (2004), Gut (2009)) that was investigated using word frequency and the rate of past tense marking. This study offers an alternative corpus extension approach by adding temporal annotations using the TimeML standards (Pustejovsky et al, 2003).

The OPT seems to be motivated by the presence of time-adverbials such as *yesterday* in an utterance (Bao 1998; Alsagoff 2001). We extend the notion of OPT to include the omission of past participle tense in the presence of completive-adverbial *already*. For example, from ICE-SIN S1A:051:127:1:B,

(2) *So they have ask for extension five times already.

This study hypothesizes a correlation between verbal tense and aspectual adverbs in SCE where the OPT is substantially instigated by the use of perfective adverbs in SCE sentences. Focusing on two particular perfective aspectual adverbs (*already* and *yesterday*) this study seeks to answer the following questions regarding OPT in SCE using a corpus based approach:

i. How often does the OPT phenomenon occur in SCE?
ii. Is the OPT phenomenon unique to SCE?
iii. Are there syntactic constraints governing the occurrence of OPT in SCE?

Other than the syntactic/lexemic constraints of the OPT phenomenon, previous researches had also looked at the phonemic motivations in final plosive deletions due to consonant clusters coda (e.g. Deterding and Poedjosoedarmo (1998), Bao (1998), Lim (2004), Gut (2005)). However phonemic constraints for OPT will not be discussed in this study.

2. Corpus Based Approach

In the process of investigating the OPT phenomenon in SCE, existing corpora were extended with temporal annotations. To explore the nature of OPT occurrences, this study utilizes time tagged corpora, where the corpora contain a layer of time annotation that describes verbs and adverbs as events that embed information about the tense of a particular intention in each sentence. An OPT occurrence is identified as a sentence that is headed by a present tense marked verb and a perfective adverb conveying the past.

This paper describes the process of the corpora extension and concludes with the investigation of the OPT phenomenon in SCE. This paper is organized as follows; Section 3 is concerned with the choice of two corpora, one being the representation of SCE and the other as a cross English variety comparison. Section 4 describes the corpora extension task, which includes the removal of the original annotation and manual tagging of the corpora with

temporal tags. Section 5 discusses the OPT phenomenon with the results extracted from the time-tagged corpora. This paper concludes with Section 6, discussing future explorations of the OPT phenomenon.

3. Choice of Corpora

The chosen corpus to investigate the OPT phenomenon in SCE is the spoken section of the International Corpus of English – Singapore (ICE-SIN). The spoken section (~600,000 words) of the ICE-SIN emulates the colloquial register (SCE) of Singaporeans' day-to-day conversations; e.g. the usage of adverbs immediately after verbs, such as "finish already" (see Fig. 1). More specifically, the private dialogues stratum (~200,000 words, which includes face-to-face conversations and phone calls) within the spoken section was chosen as the data for temporal annotation extensions. Public dialogues (e.g. classroom lessons and business transactions) and monologues (e.g. legal presentations and spontaneous commentaries) were excluded as they constitute a relatively more formal speech setting, and speakers would have likely used the standard register (SSE) where OPT is theoretically absent as it is viewed as ungrammatical in SSE.

The written section (~400,000 words) of the corpus were also excluded in the annotation task as it represented edited English texts (e.g. academic writing, student writing and news reports) where the OPT realizations might have been sanitized (for detailed description of the ICE corpus design refer to Nelson, 1996a).

3.1. ICE-SIN corpus Representation

The ICE-SIN corpus is reasonably large in size, (~1,000,000 words) and it adopted the stratified sampling method (McEnery and Wilson, 2001) as its corpus design. Invariably, the stratified corpus lacks the proportional representation of the Singapore population as a whole. Nonetheless, the stratified sampling is sufficiently effective in representing the language use of SCE. The corpus inputs were from both male and female informants across a wide range of age groups, aged 18 and above. In addition, "all contributors to the corpus were educated through the medium of English and were either born in Singapore or have moved here at an early age and received education through English in Singapore" (Nelson, 1996: 35-53).

Short of building a similar and updated version of the spoken section of the ICE-SIN, extending the current ICE-SIN with temporal annotations is the most resource effective way to explore the OPT phenomenon.

3.2. ICE-HK as a comparable corpus

To verify the uniqueness of the OPT phenomenon, the International Corpus of English – Hong Kong (ICE-HK) was selected as the corpus for cross-corpora evaluation. The ICE corpora were built with the primary aim of establishing contrastive linguistics across different varieties of English worldwide. Every variety of English in the ICE corpus was compiled using the same corpus design (Leitner, 1992).

Moreover, Singapore and Hong Kong share similar historical backgrounds – both were former British colonies, with a Chinese-majority population where substrate influences from their Chinese 'dialects' were observed (Tan, 1997). The substratum characteristics of both Singapore and Hong Kong English (HKE) would have presented similar occurrences of the OPT phenomenon, where sociolinguistics justifications motivate these grammatical occurrences.

3.3. Availability of ICE-SIN

Other than the representation of the corpus, the availability of the corpora was also taken into consideration in the selection of the corpora to explore the OPT phenomenon. All ICE corpora are free to be downloaded and their usages are governed by the individual ICE license agreement crafted by the individual project heads of the respective Englishes. The ICE corpora chosen for this study can only be used for non-profit linguistic research purposes and the redistribution of the ICE texts is prohibited. Although the prohibition to redistribute the annotated text may limit the potential usage of this corpus extension effort, this data could be used to train an automatic

temporal-event tagger to tag future English speech data with the OPT phenomenon. An algorithm can be devised to time-tag SCE data showing the OPT phenomenon, which would provide an ideal model for a tagger to train on.

The time annotated data will be sent to the respective contact person for the ICE-SIN and ICE-HK corpora and the hosting of the extended corpora will be at the project teams' discretion. Permission would be sought to redistribute a sample of the time tagged data if the project teams reject the full redistribution of the temporal annotations.

4. Corpora Extensions

The objective of the annotation extension task was to provide more conclusive corpus-based results of when and how the OPT phenomenon sentences with perfective adverbs (*already* and *yesterday*) were filtered out for the task of temporal tagging, to reduce unnecessary annotations by human. The resulting outputs of the corpora extensions are stored in Time Markup Language (TimeML) format; TimeML is used in the temporal time annotation evaluation task (TempEval) of the computational semantic evaluation exercise SemEval (Pustejovsky et al, 2003; SemEval, n.d.). TimeML abides by the Extensible Markup Language (XML) format where the syntax of the ML can be checked for its well-formedness.

4.1. Cleaning up the original ICE annotations

The original ICE annotations were marked with extra-segmental annotations at the discoursal level and other English-related annotations. An example of the original tags is as follow (refer to the ICE corpus manual for the full ICE annotations guide; Nelson, 1996):

Table 1. Example of ICE original annotations.

Original ICE tags	Annotation description
<$[A-Z]>	Speaker identification
<#>	Text unit markers
<[>…</[>	Overlapping turns
<,>and<,,>	Short and long pauses
<foreign>…</foreign>	Foreign word
<unclear>…</unclear>	Unclear speech

As the OPT phenomenon occurs at the segmental level (i.e. the lexical, clausal and sentential level), all the original discoursal markups and English related tags were removed as they were of no value to the time tagged corpora and the OPT investigation. Future research may look into whether discoursal data will affect the temporal utterances. It must be noted that similarities between the original ICE tags and XML will result in loud technical 'noises' when processing the time-tagged data.

As for socially determined factors such as gender, age and education, the ICE-SIN has no metadata encoded in the text. However, the domain based strata that the corpus design is based on can be coded in the textfile name in accordance to how the individual sentences were tagged (e.g. in the text unit tag <ICE-SINS1A-049#83:1:C>, S1A refers to the Private Dialogue data and 049 points back to the Direct Conversations section of the corpus).

A text unit corresponds loosely to a sentence, though it may be syntactically incomplete. In the spoken sector of the ICE corpora, a change of speaker's turn always corresponds to a new text unit (Nelson, 1996). A python script was written to clean the original annotation and extract the text units with the target perfective adverbs *already* and *yesterday*. Sentences with <unclear>…</unclear> tags and mono-word sentences were excluded from the tagging task because even if the OPT phenomenon is present; they remain inconclusive.

4.2. Time tagging the ICE corpora

The time tagging of the ICE corpora adhered to the TimeML standards and to investigate the OPT phenomenon, only three basic tag types from the TimeML were used, viz. <EVENT>, <MAKEINSTANCE> and <TLINK> tags.

Although TimeML offered a variety of other tags that will produce finer-grained definition of temporal ordering, past linguistic researches on time relations had shown that it is more appropriate to capture time relations in natural language using a reduced set of temporal tags (Schilder, 1997; Freska, 1992). The reduced number of tags would also be preferred as it increases the reliability of the annotation by both computer and human (Verhagen, 2005).

Table 2. Reduced TimeML Annotation Schema.

Tags	Attribute	Values
EVENT	class	'ASPECTUAL' | 'OCCURRENCE' | 'STATE'
MAKEINSTANCE	pos	'ADJECTIVE' | 'VERB'
	tense	'PAST' | 'PRESENT' | 'NONE'
	aspect	'PROGRESSIVE' | 'PERFECTIVE' | 'PERFECTIVE_PROGRESSIVE | 'NONE'
TLINK	relType	'BEGINS' | 'DURING' | 'END'

The <EVENT> tags were used to annotate words in the sentence that mark semantic events. The tag contained the class attribute that determined the nature of the semantic events that were being tagged. The <EVENT> tags were applied on (i) the head verb of each sentence and (ii) the *already* or *yesterday* adverb.

For every <EVENT> tag there was a corresponding <MAKEINSTANCCILCE> tag that encoded the tense, aspect and its pos (part of speech). The pos attribute in <MAKEINSTANCE> was a coarse grained tag which only differentiate the main word classes (i.e. "VERB" or "ADVERB").

Table 3. Annotation of the tense and aspect attributes adheres to the following paradigm.

Active Voice	Passive Voice	Tense	Aspect
teach	is taught	PRESENT	NONE
is teaching	is being taught	PRESENT	PROGRESSIVE
has taught	has been taught	PRESENT	PERFECTIVE
has been teaching	-	PRESENT	PREFECTIVE_PROGRESSIVE
taught	was taught	PAST	NONE
was teaching	was being taught	PAST	PROGRESSIVE
had taught	has been taught	PAST	PERFECTIVE
had been teaching	-	PAST	PREFECTIVE_PROGRESSIVE

For every pair of head verb and adverb <EVENT> tags, the annotators added a <TLINK> tag that encoded the temporal relation type between the <EVENT> tags. The <TLINK> tag had the relation type (relType) attribute that specify whether an event had begun ("BEGINS"), had ended ("ENDS") or the event was ongoing ("DURING") at the time when the sentence was recorded.

Figure 1 shows an example of the original source textfile S1A-049.txt from the ICE-SIN and the resultant time annotated output in TimeML format. From sentence *ICE-SIN:S1A-049#83:1:C*, the head verb *Finish* was tagged with the <EVENT> tag and for its <MAKEINSTANCE> the annotator respectively assigned PRESENT and NONE to its tense and aspect attribute. For the adverb *already*, it was tagged with a second <EVENT> tag and the annotator assigned PAST and PERFECTIVE to its tense and aspect attribute. Finally the annotator linked the events with the <TLINK> tag and assigned the ENDs value to the relType attribute to specify that the event when the sentence was recorded. With such annotations we identified the OPT phenomenon by looking at the tense and aspect disparity between the <EVENT> tags.

```
314  <$B>
315  <ICE-SIN:S1A-049#81:1:B>
316  Ya because now it is the mid-
     year exam so I have to mark
     quite
317  a lot of papers
318
319  <$A>
320  <ICE-SIN:S1A-049#82:1:A>
321  How many classes you teaching
322
323  <$C>
324  <ICE-SIN:S1A-049#83:1:C>
325  Finish already
326
327  <$B>
328  <ICE-SIN:S1A-049#84:1:B>
329  I have four classes
```

```
1  <?xml version="1.0"?>
2  <!DOCTYPE TimeML SYSTEM "TimeML.dtd">
3  <TimeML>
4  <EVENT eid="e1" class="OCCURRENCE">
5  Finish
6  </EVENT>
7  <MAKEINSTANCE eiid="ei1" eventID="e1"
   pos="VERB" tense="PRESENT" aspect="NONE"/>
8  <EVENT eid="e2" class="ASPECTUAL">
9  already
10 </EVENT>
11 <MAKEINSTANCE eiid="ei2" eventID="e2"
   pos="ADVERB" tense="PAST"
   aspect="PERFECTIVE"/>
12 <TLINK eventInstanceID="ei1"
   relatedToEventInstance="ei2"
   relTypes="ENDS"/>
13 </TimeML>
```

Fig. 1. Sample text from S1A-049.txt (left); Time tagged sample of sentence *ICE-SIN:S1A-049#83:1:C* (right)

The original selected stratum of the ICE-SIN and ICE-HK contain 30,117 and 63,075 text units respectively. The corpus design ensures that the number of words is ~200,000 from the *private dialogue* stratum. The difference in the number of text units meant that the ICE-HK text units are on average two times shorter than those in the ICE-SIN.

After cleaning the original ICE annotation and removing mono-word and <unclear> text units, the extracted counts of ICE-SIN and ICE-HK text units are 21,401 and 38,910 respectively. The drop from 63,075 to 38,910 means that approximately half of the ICE-HK data are made up of mono-word sentences or unclear content, reflecting a relatively poorer quality of the corpus content and speech recordings.

The remaining cleaned and extracted text units that contained the selected perfective adverbs *already* and *yesterday* were manually time annotated. The manual annotations resulted in 286 meaningful occurrences of the OPT phenomenon in the extracted ICE-SIN texts and 187 such occurrences in the ICE-HK texts.

4.3. Inter-Annotation Agreement (IAA) of the time annotations

A subset of four time-tagged textfiles was annotated by all four annotators (two were randomly selected from the ICE-SIN and ICE-HK respectively). Due to the involvement of more than two annotators, the common double annotation IAA calculation was inappropriate and the traditional Cohen kappa contingency table would have over-generalized the fine differences between the annotators' annotations. Also, the IAA standards set by the TimeBank 1.2 were based on the IAA of two annotators (Pustejovsky et al, 2003). Since the annotation task in this study involved 4 raters, the kappa coefficient was used as a gauge for the IAA score across all four annotators.

The kappa coefficient calculations were based on Warren's (2010) percentage of overall agreement of the fixed-marginal multirater kappa (Siegel and Castellan, 1988) and the free-marginal multirater kappa (Randolph, 2005). The value of the percentage of overall agreement (Po) kappa ranges from -1.0 to 1.0, with kappa scores of:

- -1.0 indicating the perfectly true negatives (i.e. perfect disagreement between raters is below chance),
- 0.0 indicating the annotators' agreement is equal with chance and,
- 1.0 indicating perfectly true positives (i.e. the perfect agreement of all raters is above chance).

Following the TimeBank 1.2 corpus standards for temporal annotation IAA, the Po kappa was calculated for each type of TimeML tags (i.e. the <EVENT>, <MAKEINSTANCE> and <TLINK> tags) involved in the annotation extension task. The multirater Po kappa calculations require the number of cases and the number of possible categories. For example, when considering the IAA of the <EVENT> tag, the number of cases refer to the number of <EVENT> tags that the annotators are required to tag; for each textfile there are two <EVENT> tags within a sentence – referring to the verb and adverb respectively. The category number referred to the number of possible

differences, e.g. an annotator can tag an <EVENT> tag with any of the 3 categories; class="STATE" or class="ASPECTUAL" or class="OCCURRENCE". Table 4 summarized the IAA calculation criterion. Since the <MAKEINSTANCE> tags had two attributes that were crucial in the OPT analysis, the IAA calculation of the <MAKEINSTANCE> tag is based on the tense and aspect attributes rather than <MAKEINSTANCE> tag as a whole.

Table 4. IAA Calculation Criterion.

Tag Types	No. of cases	Categories of tag types
<EVENT>	3	class ::= "STATE"|"OCCURRENCE"|"ASPECTUAL"
tense	3	tense ::= "PAST"|"PRESENT"|"NONE"
aspect	3	aspect ::= "PERFECTIVE"|"PROGRESSIVE"|"NONE"
<TLINK>	3	relType::= "BEGINS"|"ENDS"|"DURING"

The IAA scores for the time annotation task were above the widely accepted inter-rater agreement threshold, i.e. >0.70 (Warrens, 2010). The fixed marginal score on the <TLINK> tags were extremely low because there was one sentence instance with omitted past tense that can be of perfective or inchoative intention and the annotators differ in their tagging (2 annotators tagged the <TLINK> with the ENDS attribute, 1 annotator tagged it as BEGINS and the last annotator tagged it as DURING). The 0.08 score on the fixed marginal kappa measured how the extent of the annotators' variation but the effect was counter-balanced by the free marginal kappa that weighs their similarity.

Table 5. IAA of the Time Annotation Task.

Tag Types	Overall Po	Fixed Marginal	Free Marginal
<EVENT>	**0.75**	0.57	0.63
tense	**0.81**	0.43	0.77
aspect	**0.85**	0.70	0.78
<TLINK>	**0.79**	**0.08**	0.69

5. The OPT-ional Phenomenon

To differentiate the sentences displaying the OPT phenomenon from other sentences in the time-annotated corpora, the tense disparity between a verb <EVENT> and an adverb <EVENT> was considered. A verb would be annotated with the <EVENT> tag and assigned either the STATE or OCCURRENCE value for its class attribute. The class attribute for an adverb <EVENT> tag would be assigned the ASPECTUAL value. Since all the <MAKEINSTANCE> tags for the perfective adverbs contained the PAST tense attribute, we identified an OPT occurrence when the verbal <MAKEINSTANCE> tag contained the PRESENT tense attribute. The contingency table to define sentences with and without the occurrence of the OPT phenomenon is summarized below (see Table 6). Although the <TLINK> tags were not utilized by this study, the relType="BEGINS" and relType="ENDS" attributes can be used to explore the incohative and perfective aspects of *already* as Bao (1995) had observed.

Table 6. Contingency Table for OPT Phenomenon

EVENT (verb)	EVENT (adverb)	MAKEINSTANCE (verb)	MAKEINSTANCE (adverb)	OPT phenomenon
class::= "state"| "occurrence"	class::= "aspectual"	tense::= "PRESENT"	tense::= "PAST"	**OPT present**
class::= "state"| "occurrence"	class::= "aspectual"	tense::= "PAST"	tense::= "PAST"	OPT absent

5.1. How often does the OPT phenomenon occur in SCE?

Table 7 presents the likelihood of OPT occurrences from the time annotated sentences; V-RB refers to the sentences which contain post-verbal adverbs and RB-V refers to the sentences which contain pre-verbal adverbs. The OPT phenomenon manifested 68.18% of the time when an *already* or *yesterday* adverb was present in the sentence in the ICE-SIN.

Table 7: OPT Occurrences in sentences containing *already* and *yesterday* in the ICE-SIN

Syntactic Structure	OPT present	OPT absent	Total no. of sentences
V-RB	30	56	86
RB-V	165	35	200
Total occurrences	195 (68.18%)	91 (31.89%)	286

To verify whether the OPT phenomenon occurred in sentences without perfective adverbs, we extracted a random subset of the 50 untagged sentences from the spoken private dialogue stratum of the ICE-SIN to check whether OPT occurred in the absence of adverbs.

In contrast to the sentences with aspectual adverbs, these sentences lacked the temporal information needed to determine the intended tense. Thus it was not possible to annotate temporal information to these sentences intra-sententially. The Time ML required a minimum of two events for the temporal relation between the two events to be meaningfully annotated. Hence, a different methodology was used to determine the OPT occurrence – an annotator was required to determine the occurrence of OPT in non-adverbial sentences. The tenses of randomly selected sentences were determined from the context of three sentences before and after the selected text units. Each sentence was checked for its tense and was subsequently categorized as (a) unknown (sentences that are ambiguous or cannot be determined by the context), (b) positively present, (c) positively past or (d) OPT occurrence. To avoid this meticulous categorization, future annotation tasks that seek to further discuss the OPT phenomenon should consider a different methodology of inter-sentential time annotation.

Disregarding the unknown sentences from the categorization task, the OPT phenomenon was present 17.14% of time in the sentences without perfective adverbs however the OPT occurrences in these sentences were significantly lower ($p<0.05$, $z=5.71$) than OPT occurrences in sentences with *already* and *yesterday*.

5.2. Is the OPT-ional phenomenon uniquely Singaporean?

Using the same definition of the OPT phenomenon in Table 6, the omission of past tense was also present in the time annotated ICE-HK; the OPT phenomenon was present in 36.55% of the sentences with the *already* and *yesterday* adverb (see Table 8).

Kortmann and Lunkenheimer (2011) observed that the morphological simplification is a recurrent feature across World Englishes, where they are often motivated by the lack of morphology in contact varieties. In the case of Singapore and Hong Kong Englishes, it is possible that typologically isolating Chinese substrates (e.g. Mandarin, Cantonese, Min, etc.) had influenced the occurrences of the OPT phenomenon in SCE and Hong Kong English. The rate of occurrence in the ICE-SIN (68.18%) is above chance (50%) while the rate (31.89%) in the ICE-HK is below the opportunistic threshold. Though the OPT is not absolutely unique to SCE, it is more salient as compared to Hong Kong English.

Table 8: OPT Occurrences in *already* and *yesterday* sentence in ICE-HK

Syntactic Structure	OPT	non-OPT	Total no. of sentences
V-RB	27	37	64
RB-V	45	78	123
Total occurrences	72 (36.55%)	115 (63.45%)	187

5.3. Are there syntactic constraints to the occurrences of OPT in SCE?

From Table 7 and 8, both Englishes reflected a higher rate of occurrences of the OPT phenomenon in sentences with pre-verbal adverbs. However, it remained uncertain as to why the pre-verbal adverbs stimulate the OPT phenomenon more extensively than post verbal adverbs. From the *already* examples provided in Bao (1995) and the anecdotal experiences of the authors of this paper, it went against the common perception that the OPT phenomenon is more common in sentences with post-verbal adverbs. The current layer of temporal annotation is insufficient to explain this incongruity.

5.4. Relevance of the investigated OPT in SCE today

While the ICE-SIN is representative of SCE in size and stratification, it must be noted that the data was collected in the 1990s (more than 20 years ago from the time of this writing). The linguistic variation of SCE can be said to be comparatively more volatile than the relatively well-established British English, given the observation of a sharp language shift where the usage of English at home in Singapore has increased in the late 2000s (Vaish, 2008). Therefore, the corpus data might not be representative of the current variety of SCE spoken in Singapore. Nevertheless, time disparity is often disregarded in corpus linguistics studies when recently built corpora were crossed referenced with the de facto British National Corpus (2007); e.g. Ferraresi et al. (2008)'s cross-corpora evaluation of ukWaC corpus with the BNC.

Still, the question remains; *is the OPT-ional phenomenon still relevant today?* Anecdotally, as speakers regularly exposed to Singapore English, we are positive that the OPT phenomenon is present in SCE today. However, from an empirical perspective, we can only be as certain as what the data had presented.

6. Conclusion

This paper described the investigation of the OPT phenomenon in SCE and HKE. In the process, the ICE-SIN and ICE-HK were extended with TimeML temporal tags.

From the temporal annotations, we observed that the rate of the occurrence of the OPT phenomenon was 68.2% in SCE sentences and 36.6% in HKE sentences with the perfective adverbs *already* or *yesterday*. Although the OPT phenomenon was not unique to Singapore as it is also found in Hong Kong English, it was more salient in Singapore English. By looking at the pre/post-verbal position of the perfective adverb, we observed that the OPT phenomenon was more prominent in sentences where the adverb occurs before the verb.

Future researchers could attempt to further extend the corpora with full grammatical sentence parses (e.g. Context-Free Grammar or Dependency Grammar parses) or look into the phonological environment where the OPT phenomenon occurs.

Finally, this paper also discusses the time disparity limitations in cross-corpora comparison. The time disparity between built corpora and the time of linguistic investigation urges the need for corpora to be self-sustainable so as to be relevant for continual linguistic research. Otherwise, corpus-based researches would be considered historical as the data may no longer be relevant to the linguistic phenomenon under investigation. The task of bridging the gap between a natural language phenomenon and pre-dated corpus representation is indeed a conundrum, but this must not deter future corpora-based researches.

References

Alsagoff, L. & Ho, C.L. (1998). The Grammar of Singapore English. In J. Foley, T. Kandiah, Z. Bao, A. Gupta, L. Alsagoff, C. L. Ho, L. Wee, I. S. Talib & W. Bokhorst-Heng (Eds.), *English in New Cultural Contexts: Reflections from Singapore* (pp. 127–151). Singapore: Oxford University Press.

Alsagoff, L. (2001). Tense and aspect in Singapore English. In V. Ooi (Ed.), *Evolving Identities: The English Language in Singapore and Malaysia* (pp. 79–88). Singapore: Times Academic Press.

Bao, Z. (1995). Already in Singapore English. *World Englishes*, *14*(2), 181–188.

Bao, Z. (1998). The sounds of Singapore English. In J. Foley, T. Kandiah, Z. Bao, A. Gupta, L. Alsagoff, C.L. Ho, L. Wee, I. S. Talib & W. Bokhorst-Heng (Eds.), *English in New Cultural Contexts: Reflections from Singapore* (pp. 152–174). Singapore: Oxford University Press.

Bao, Z. & Hong, H. (2006). Diglossia and register variation in Singapore English. *World Englishes, 25*(1), 105–114.
Cavallaro, F. & Ng, B. C. (2009). Between status and solidarity in Singapore. *World Englishes, 28*(2), 143–159.
Crewe, W. (1977). The English Language in Singapore. Singapore: Eastern Universities Press.
Cruz-Ferreira, M. (2005). Past tense suffixes and other final plosives in Singapore English. In D. Deterding, A. Brown & E. L. Low (Eds.), *English in Singapore: Phonetic Research on a Corpus* (pp. 26–36). Singapore: McGraw-Hill Education (Asia).
Deterding, D. & Poedjosoedarmo, G. (2000). To what extent can the ethnic group of young Singaporeans be identified from their speech? In A. Brown, D. Deterding & E. L. Low (Eds.), *The English Language in Singapore* (pp. 1–9). Singapore: Singapore Association of Applied Linguistics.
Deterding, D. (2003). Tenses and will/would in a corpus of Singapore English. In D. Deterding, E. L. Low & A. Brown (Eds.), *English in Singapore: Research on Grammar* (pp. 31–38). Singapore: McGraw-Hill Education (Asia)
Deterding, D. & Poedjosoedarmo, G. (1998). The Sounds of English: Phonetics and Phonology for English Teachers in Southeast Asia. Singapore: Prentice Hall
Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff and S. Sharoff (Eds.), *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google* (pp. 47–54). Marrakech, Morocco.
Freksa, C. (1992). Temporal Reasoning Based on Semi-Intervals. *Artificial Intelligence, 54*(1), 199–227.
Gupta, A. (1994). *The Step-Tongue: Children' s English in Singapore*. Clevedon, UK: Multilingual Matters.
Gupta, A. (2005). Inter-accent and inter-cultural intelligibility: a study of listeners in Singapore and Britain. In D. Deterding, A. Brown & E. L. Low (Eds.), *English in Singapore: Phonetic Research on a Corpus* (pp. 138–152). Singapore: McGraw-Hill Education (Asia)
Gut, U. (2005). The realisation of final plosives in Singapore English: Phonological rules and ethnic differences. In D. Deterding, A. Brown & E. L. Low (Eds.), *English in Singapore: Phonetic research on a Corpus* (pp. 14–25). Singapore: McGraw-Hill.
Gut, U. (2009). Past tense marking in Singapore English verbs. *English World-Wide, 30*(3), 262– 277.
Halácsy, P., Kornai, A. & Oravecz, C. (2007). HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. (pp. 209–212). Prague, Czech Republic: Association for Computational Linguistics.
He, J. (2007). Grammatical Features of Singapore Colloquail English: A Corpus-based Variation Study (Unpublished doctoral dissertation). National University of Singapore, Singapore. Retrieved from https://scholarbank.nus.edu.sg/bitstream/handle/10635/16215/contents.pdf?sequence=1.
Ho, M. & Platt, J. (1993). *Dynamics of a Contact Continuum: Singaporean English*. Oxford: Clarenden Press.
Ho, M. (2003). Past tense marking in Singapore English. In D. Deterding, E. L. Low & A. Brown (Eds.), *English in Singapore: Research on Grammar* (pp. 39–47). Singapore: McGraw-Hill Education (Asia)
Kirkpatrick, A. and Saunders, N. (2005). The intelligibility of Singaporean English: a case study in an Australian university. In D. Deterding, A. Brown and E. L. Low (Eds.), *English in Singapore: Phonetic Research on a corpus* (pp. 153–162). Singapore: McGraw-Hill Education (Asia).
Kortmann, B. and Lunkenheimer, K. (Eds.) (2011). The electronic World Atlas of Varieties of English [eWAVE]. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://www.ewave-atlas.org/.
Leitner, G. (1992). International Corpus of English: Corpus design – problems and suggested solutions. In G. Leitner (Ed.), *New directions in English language corpora: methodology, results, software developments* (pp. 33–64). Berlin: Mouton de Gruyter.
Leech, G. (2005). Adding Linguistic Annotation. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, (pp. 17–29). Oxford: Oxbow Books.
Lim, L. (2004). Sounding Singaporean. In L. Lim (ed.), *Singapore English: A grammatical description* (pp. 19–56). Amsterdam/ Philadelphia: John Benjamins.
Lim, L. (2007). Mergers and acquisitions: On the ages and origins of Singapore English particles. *World Englishes*, 27(4), pp. 446–473
Nelson, G. (1996a). Markup Systems. In S. Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English* (pp. 36–53). Oxford: Clarendon Press.
Nelson, G. (1996b). The Design of the Corpus. In S. Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English* (pp. 27–35). Oxford: Clarendon Press.
Payne, T. (1997). *Describing morphosyntax: A guide for field linguists*. Cambridge: Cambridge University Press.
Pustejovsky, J., Castano, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., & Katz, G. (2003). TimeML: Robust specification of Event and Temporal expressions in Text. Proceedings of the *Fifth International Workshop on computational Semantics* (IWCS-5) (pp. 1–11). Tilburg, Netherlands.
Platt, J. (1977). A model for polyglossia and multilingualism (with special reference to Singapore and Malaysia). *Language in Society. 6*(3), 361–378.
Randolph, J. (2005). Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. Paper presented at the *Joensuu University Learning and Instruction Symposium 2005*. Joensuu, Finland, October, 14-15.
Schilder, F. (1997). *Temporal Relations in English and German Narrative Discourse*. (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh, UK.
SemEval. (n.d.). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/SemEval.
Siegel, S. & Castellan, J. (1988). *Nonparametric statistics for the social sciences* (2nd ed.). New York: McGraw-Hill.
Tan, J. (1997). Education and Colonial Transition in Singapore and Hong Kong: comparisons and contrasts. *Comparative Education, 33*(2), 303–312.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/.

Vaish, V. (2008). Mother Tongues, English, and Religion in Singapore. *World Englishes, 27*(3-4), 450–464.

Verhagen, M. (2005). Temporal Closure in an Annotation Environment. *Language Resources and Evaluation, 39*(2), 211–241.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., & Pustejovsky, J. (2009). The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation, 43*(2), 161–179.

Warrens, M. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification, 4*(4), 271–286.

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

Wee, L. (2004). Singapore English: Morphology and syntax. In B. Kortmann, K. Burridge, R. Mesthrie, E. Schneider & C. Upton (Eds.), *A Handbook of Varieties of English*, Vol. 2*: Morphology and Syntax* (pp. 1058–1072). Berlin: Mouton de Gruyter.