

This Word Mean What: Constructing a Singlish Dictionary with ChatGPT

Chow Siew Yeng^{id}, Chang-Uk Shin^{id}, Francis Bond^{id}[1]

[1]Department of Asian Studies, Palacký University, Olomouc
siewyeng001@e.ntu.edu.sg, papower2@gmail.com, bond@ieee.org

Abstract

Despite the magnitude of recent progress in natural language processing and multilingual language modeling research, the vast majority of NLP research is focused on English and other major languages. This is because recent NLP research is mainly data-driven, and there is more data for resource-rich languages. In particular, Large Language Models (LLM) make use of large unlabeled datasets, a resource that many languages do not have. In this project, we built a new, open-sourced dictionary of Singlish, a contact variety that contains features from English and other local languages and is syntactically, phonologically and lexically distinct from Standard English (Tan, 2010). First, a list of Singlish words was extracted from various online sources. Then using an open Chat-GPT LLM API, the description, including the definition, part of speech, pronunciation and examples was produced. These were then refined through post processing carried out by a native speaker. The dictionary currently has 1,783 entries and is published under the CC-BY-SA license. The project was carried out with the intention of facilitating future Singlish research and other applications as the accumulation and management of language resources will be of great help in promoting research on the language in the future.

Keywords: Singlish dictionary, ChatGPT, Data Generation

1. Introduction

1.1. Purpose

In recent NLP research, studies that require a considerable amount of language and computational resources have become increasingly mainstream (Touvron et al., 2023). While BERT, a transformer encoder-based LLM released in 2018 (Devlin et al., 2019), learned from around 3.3 billion words, state-of-the-art models now use more than ten times that amount. For example, OpenAI's GPT3 and subsequent models have been trained on datasets of at least 500 billion tokens. In addition, the modeling performance of LLM seems to improve in proportion to the amount of data sets, the number of model layers, and the number of parameters (Kalyan et al., 2021).

However, most of these language resources are only available in the most popular languages. In particular, the largest amount of data is available for standard English, and the proportion of data available for languages such as Singlish is much smaller. This project seeks to fill this gap in the current state of language resources available. In the case of Singlish, due to its non-official status, there are no large dictionaries. Therefore, this project, inspired by the various language resource creating efforts through LLM (de Schryver, 2023; Elsner and Needle, 2023) in the last year, creates a dictionary through a pretrained LLM (ChatGPT) and manual editing in order to promote future research on Singlish. The generated responses by the LLM were used to create a rough draft of the dictionary,

speeding up the process as compared to writing a dictionary from scratch. While ChatGPT is not fine-tuned to Singlish, there are no LLMs trained specially for Singlish that are able to generate responses like ChatGPT.¹ Even though this effort is not sufficient for an LLM training, building this open-sourced dictionary that will allow for contributions from the public is a step in the right direction.

This dictionary (with the exception of its examples) is written in standard English and aims to describe the Singlish phrases and their usage to its readers. Each entry in the constructed dictionary contains the word or phrase's definition, example sentence, pronunciation, part of speech and alternate spellings and language of origin if applicable.

The results of this study are released under the open license 'CC-BY-SA' and are expected to enable further opportunities for future Singlish studies.

1.2. Singlish

Singlish is a contact language whose emergence can be attributed to the diversity of languages spoken in Singapore (Soh et al., 2022) such as Hokkien, Bazaar Malay, Cantonese (which only exerted more influence in recent years: Lim, 2011) and English. It has English as its superstrate language and has its lexicon, syntax and even prosody influenced by substrate languages like

¹There is [SingBert](#) which is a fine-tuned version of BERT on Singlish but it does not generate responses like ChatGPT, and is trained on a small and noisy corpus.

Baba Malay (Lim, 2011). In particular, a prominent feature that separates it from standard English is its extensive use of particles, on which, numerous studies have been conducted (Wong, 2005; Leimgruber, 2016). Singlish syntactic features include optional inflection of verbs (for third person singular subjects), optional articles and the lack of plural marking (Chow and Bond, 2022). While certain syntactic features are common in Singlish, the extent to which they are used by individual speakers varies with factors such as other languages they speak. Another feature that the variety has is tone. The tonality of Singlish has been extensively covered by Lim (2011). Particles, in particular, are distinguished through tone. Hence, it is an important feature and is taken into account in this dictionary. The local variety is important to the Singaporean identity and is used heavily in daily conversations (Li, 2021) despite the government's efforts to completely replace it with standard English (Cavallaro and Ng, 2009).

Due to the diversity in the sources for its lexicon and the informal nature of the language, its vocabulary is not well captured in standard dictionaries. Moreover, in many cases, the origin language and standard spelling (if it exists) of a word in Singlish may not be known to the average speaker.

- (1) *Dey, wǒ mén paktor always makan at*
 T Md C E M E
 Hey we date eat
kopitiam one.
 M+H S
 coffee shop PART

'Hey when we date we always eat at the coffee shop (one).'

For instance, a typical Singlish utterance is shown in 1 (Cheng, 2021)² where the second line represents the origin language of the term (T-Tamil, Md-Mandarin, C-Cantonese, E-English, M-Malay, H-Hokkien/Hakka, S-Singlish). A single utterance could easily involve words from multiple languages but is spoken, simply, as Singlish. It is not code-switching, because a speaker of Singlish typically will not speak all or even any of the non-English languages.

Hence, a Singlish dictionary functions to collect into the lexicon words originating across various languages and unique expressions that are used in this variety and explain them to the readers.

²While Cheng (2021) labelled 'one's origin to be unknown, in this paper, it is instead labelled as 'S' for Singlish as its usage has Chinese origins, but has taken its form in an English word. More on the origins of 'one' can be found in Wong (2005).

1.3. Existing Resources

Although Singlish is predominantly a spoken language, there are some resources in the forms of corpora and dictionaries. The different corpora of Singlish and Singapore English illustrate the usage of language in different mediums and time periods.

The NUS SMS corpus (Chen and Kan, 2015) is a corpus of 67,093 text messages focusing on English and Mandarin Chinese. The data was crowdsourced from Amazon's Mechanical Turk, Short-Task, ZhuBaJie and NUS students. Although this was not a Singlish centred project, 46.9% of the English SMS were contributed by people from Singapore (Chen and Kan, 2012, p. 18). With the sheer amount of data, this corpus contains a significant collection of Singapore English. It is a public corpus and can be used freely with citation.

The International Corpus of English, the Singapore Corpus (ICE-Singapore) is a record of spoken and written English text in Singapore with many subcategories including telephone calls, broadcast interviews, academic writing and creative writing. The corpus contains a lot of text data but as the data is collected across many different domains, Singlish is mainly found in private dialogues (categorised with the tag 'S1') and makes up only a small proportion of the total data.

The Red Dot Baby Talk wordlist is a list of words used in the Red Dot Baby Talk Quiz (Woon and Styles, 2021), a quiz made to help document 'baby talk' in Singaporean children and the age of acquisition of these words. The list consists over a hundred words and because of the nature of the quiz, are mostly basic words that are likely to be in the lexicon for young children or babies, including onomatopoeia. This list is published under the CC by NC 4.0 license and can be used for non-commercial purposes.

There have also been other non-official dictionaries for Singlish, each with distinct properties.

In 2001, the *Coxford Singlish Dictionary* (Goh, 2002) was first published. Its name is a play on the Oxford dictionary and it is, according to Huddart (2014, pp. 75), 'an amalgamation of satirical comment on Singaporean society and a source of linguistic data'. The dictionary is written in a playful tone and contains, in addition to Singlish words, Singlish pronunciations of words in standard English. It has 809 Singlish words and phrases.

The Dictionary of Singlish and Singapore English (Lee, 2004) (henceforth DSSE) is a substantial collection of over 1,000 entries providing their origin (or speculated origins), their meanings and real examples of usage. Visitors to the website are free to suggest contributions to the dictionary through an online form. Unfortunately, the dictionary was last updated in 2016 and the data on the

dictionary does not have an open licence.

Singlish Dictionary (hereafter singlish.net) is an online dictionary compiled from May 2017 to August 2018. Despite being active for a relatively short period of time, there are around 140 entries including phrases and common acronyms used in Singlish. The content of the website is shared “AS IS” with no warranties, and confers no rights.

Wiktionary is a freely available international dictionary that contains words in various languages. A portion of the words are tagged with the categories of Singlish and Singapore English and the entries for these words contain their pronunciations, alternative forms and etymologies. This dictionary is licensed under CC-BY-SA.

As of 2022, there were 27 words marked as Singlish in the Oxford English Dictionary (OED): very incomplete coverage.³

The various resources differ in terms of the licensing rights and comprehensiveness of their lexicon. Both the Coxford dictionary and DSSE have more entries, but are not open source or actively maintained. singlish.net and the OED are neither large nor open. On the other hand, while Wiktionary is open, it has relatively fewer Singlish entries. Currently, there is no Singlish dictionary that is open-source, has a relatively large lexicon and takes note of particle tones. Therefore, this project will build a new, large, open lexicon to support future research through fully open-source data with a new methodology.

2. Methodology

2.1. Outline

The process of the project can be divided into the following steps:

1. **Collect a list of Singlish words** from DSSE, Red Dot Baby Talk (Woon and Styles, 2021) and Wiktionary (Wiktionary.org).
2. **Generate draft entries** for these words using ChatGPT (OpenAI, 2023).
3. **Manually check the entries** by the principal annotator through cross validation with existing sources and native speaker intuition to improve the entries (e.g., changing to standard spelling if it exists: *chochok* to *cocok*).

At the same time, 4 other annotators were recruited for the purpose of getting an agreement score for the tags chosen for each of the responses as part of evaluation.

³<https://www.smartsinga.com/facts-about-singapore/>, accessed 2023-10-23

4. **Normalize the pronunciations** of the words in X-SAMPA notation.⁴

2.2. Details

To elaborate on the process, firstly, a script was written to scrape the lexical entries from DSSE and wiktionary. The words from Red Dot Baby Talk were taken from the source data (a spreadsheet). We did not take the categorised examples and descriptions in order to keep our license open. These lists were then merged.

Through the ChatGPT API, a draft of description, pronunciation, POS and alternate spellings was then produced for the collected words. The following prompt was given:

You are writing a Singlish dictionary. Now, please write a dictionary entry of ‘WORD’. With a full description of the word, its POS, pronunciation and alternate spellings if there are.

In the prompt, ‘WORD’ was replaced with each word in the word list. An example of a query and response pair is shown below for the entry of ‘bo hiu’.

Query: “You are writing a Singlish dictionary. Now, please write a dictionary entry of ‘bo hiu’. With a full description of the word, its POS, pronunciation and alternate spellings if there are.”

Response: “Sure, here’s a dictionary entry for ‘bo hiu’ in Singlish:

Word: Bo hiu

Pronunciation: boh hiu

Definition: Bo hiu is a Singlish phrase that translates to “don’t care” or “not bothered”. It is often used to indicate a lack of interest, indifference or apathy towards a person or situation. It can also be used to express nonchalance or to shrug off something that might be seen as important or serious.

Example:

- A: “Eh, why didn’t you answer my call last night?”

- B: “Bo hiu lah, I was sleeping.”

Overall, ‘bo hiu’ is a common Singlish phrase that is used to express a casual attitude towards something or someone.

⁴X-SAMPA is a computer-readable representation of the IPA that uses only ASCII. It was proposed as an extension of SAMPA (Wells, 1995).

The prompts were submitted from May to September 2023. All the responses were then collated into a single document for ease of evaluation. After that, the entries were manually checked by the principal annotator (the first author). In cases where their intuition was insufficient to make a judgement, ChatGPT's response was cross-validated against other sources such as the previously mentioned dictionaries. After comparing against other sources and native knowledge of the variety, each entry was marked with a tag of 'Yes', 'No', 'Not sure', 'Partial-pron/etymology' or 'Partial-example/spelling, etc.', indicating the acceptability of the description given by ChatGPT and if inadequate, the nature of its inadequacy. The full list of tags and their examples can be found in Table 1 under Section 3. In some cases where the tag was a form of 'Partial' or 'No', an updated description was then written if the principal annotator's knowledge was enough; other dictionaries were not referred to in such cases. This precaution was taken to avoid any form of copying from other resources that are not specified to be free to use. In cases where the description was tagged 'Yes' or 'Not sure', the original descriptions were kept⁵ but the entries differ in the dictionary through an additional tag which indicates whether they have been verified.

Another possibility was for an entry to be merely an alternative spelling of another. In that case, only one description was preserved and the other similar entries were then annotated with the 'Redirect to' tag, with an indication of the chosen entry with the description. For instance, 'balukoo' and 'baluku' (both meaning 'bruise'(noun)) were in the original word list but 'baluku' was chosen as the dictionary entry and 'balukoo' was then redirected to the other spelling. Each final description was also split into its components such as definition, part of speech and example.

At the same time, 4 other annotators were recruited and initially grouped into 2 groups. Annotators in group 1 were each tasked with tagging the first 50 entries in the list of ChatGPT's responses provided on separate google sheets while annotators in group 2 were tasked with the next 50 entries. The final tags chosen by all the annotators (including the principal annotator) are compared in Section 3.

Lastly, pronunciations were added to the entries in X-SAMPA. Although ChatGPT provided pronunciations of the words in a majority of its answers, they were written in a variety of formats, from IPA to phonetic spelling and as compared to the meanings of the words, are more often inaccurate. Hence, they were all re-annotated and standard-

⁵Small edits such as removing the mention that the phrase 'is a Singlish phrase' were made.

ised in X-SAMPA.

On average, the descriptions from ChatGPT were evaluated and annotated at a rough rate of 30 words/hour including cases where the tag 'No' was given and no corrected description is given. This rate varied significantly based on the accuracy of ChatGPT's descriptions which affected the need for updated definitions and examples.

2.3. Annotators

The 5 annotators involved in this project range from 25 to 30 years old. They are all Singaporean Chinese native speakers of Singlish and 3 out of the 5 have gone through some form of military service.⁶ During the initial split, it was ensured that both groups had members with military service experience. All annotators are unpaid volunteers.

3. Results

3.1. Data

At this point in time, the dictionary has 1,783 words/phrases in its lexicon, including some which have been redirected to an entry of an alternate spelling. 138 of those entries originated from the Red Dot Baby Talk word list, 1,201 entries came from DSSE and the other 462 entries from Wiktionary. The duplicates in the entries have been deleted in the portion of the data that has been annotated hence the total is less than the sum of the entries in the resources. This collection of Singlish vocabulary with definitions is currently the largest Singlish dictionary in existence.

3.2. Detailed Annotation

Out of the 579 entries that have been manually annotated by the principal annotator, 46.2% of ChatGPT's descriptions were completely satisfactory (tagged with 'Yes'), 26.8% were unsatisfactory (tagged with 'No'), 4.9% were partially correct but unsatisfactory information on pronunciation or etymology, 8.0% were partially correct but unsatisfactory in areas such as example, spelling or etc. and 6.8% were tagged 'Redirect-to', which means that the word is an alternative spelling of another one (which has been chosen as the entry) listed on the vocabulary. The last 7.3% were tagged 'Not sure' which means that the annotator did not know the word well enough to evaluate the correctness of the description and, just in the detailed annotation, that the description given by ChatGPT was similar enough to that in the online resource but also,

⁶This may be a factor in their knowledge of the Singlish words collated considering that many Singlish words are more often used in the military context.

in most cases, contained something additional that has not been verified. An example of each of these tags is shown in Table 1.

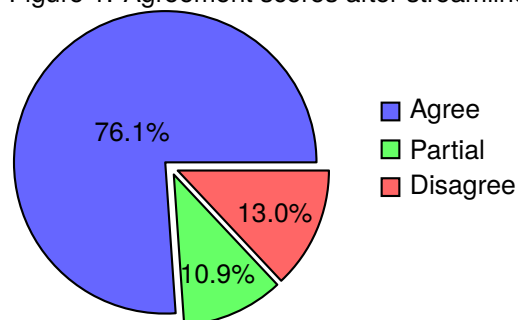
The description for 'leh' was wrong as it is not used to seek confirmation. The rest of the responses gave accurate definitions of the word but included additional less accurate information, with most of the inaccuracies lying in the description of the pronunciation. For example, both 'a's in 'atas' should be pronounced as /a/. On the other hand, in the entry 'Ah Long', its etymology as well as alternate spellings are wrong. In the case of 'blank', the description largely matches the native speaker's knowledge and the other sources, but a particular part - 'a lack of emotion' - has yet to be verified and hence, it was tagged with 'Not sure'. The last row of Table 1 shows 'balukoo' which is redirected to another entry 'baluku'. In cases like the last, the accuracy of the description is not taken into account in the tag chosen.⁷

A sample of updated entries after this process of detailed annotation (excluding their tags and pronunciations) is shown below.

- **leh** (Particle: -) Leh is a Singlish particle.
 1. The high level tone leh (tone 1) is used to ask questions. It has a similar meaning to "what about" except it occurs after the noun.
 2. The mid level tone leh (tone 3) is used usually with the intention of persuasion and to sound more convincing. — 1. *We're about to go already. Your friend leh1? 2. Eh, don't like that leh3. We need you here to play mahjong. (translation: Don't be like that. We need you here to play mahjong)*
- **atas** (Adj: Malay) Atas means 'haughty' or 'snobbish', often used to describe someone who is acting or behaving in an elitist or pretentious manner. Additionally, it can also refer to something that is high-end or posh, such as a luxurious restaurant or an expensive brand. Note: Atas is derived from the Malay language, where it means 'upper class' or 'high society'. In Singlish, it has taken on a slightly negative connotation due to its association with snobbery and elitism. — *Wow, your friend is so atas. She only wants to eat at Michelin-starred restaurants!*
- **Ah Long** (N: Hokkien) Ah long is a term used to refer to loan sharks. They are typically unlicensed moneylenders who charge very high interest rates and use harassment, intimidation, and violence to collect payments from borrowers who are unable to pay off their debts — *I heard he borrowed money from an Ah Long to pay off his gambling debts.*

⁷The description for 'balukoo' is actually inaccurate.

Figure 1: Agreement scores after streamline



3.3. Inter-annotator Agreement

During the annotation process, some annotators were particularly enthusiastic and tagged more than the required amount. Thus, at the end, a total of 334 entries were tagged by at least 2 annotators and their agreement scores were calculated. With the predefined categories, it was found that 40.0% of the annotations had complete agreement (all annotators assigned the same category for the entry), while 13.9% had partial agreement (at least 2 but not all annotators agreed) and in 46.1% of entries, there was no agreement.⁸ The agreement score is low but this can largely be attributed to the granularity of the tags. For instance, a majority of disagreements was associated with the 'Partial' tags including example, spelling, pronunciation. A breakdown of the distribution in scores and entries is shown in Table 2

In addition, of the annotations with complete agreement, 51.1% were annotated as 'Yes', 48.9% as 'No' and there were no agreement for 'Partial' tags. Annotations without complete agreement were relatively evenly spread: 39.7% as 'Yes', 19.8% as 'No', and 40.5% as 'Partial'. The annotators annotated differently for the descriptions generated by LLM despite all being native speakers of the variety. We posit that the high level of disagreement is due to the fact that the responses contain a lot of information, and that the speakers, having their own idiolects, may interpret the expressions differently.

As part of further processing the data, we streamlined the categories into three: 'Useful definition' (comprising 'Yes' and both 'Partial's), 'Not sure' and 'Wrong definition'. This categorization yielded 76.1% complete agreement, with 74.3% of those agreeing that ChatGPT's outputs gave the correct definition. More detailed dissections of the scores are shown in Figure 1 and Figure 2.

⁸In this calculation, we treated 'Not sure' like 'NA' and did not count those annotations. As such, the agreement scores were only tabulated for entries which had at least two tags excluding 'Not sure' and this was 230 entries.

Word	Response	Tag
chai poh	Certainly, here's a dictionary entry for "chai poh": Word: Chai Poh Pronunciation: chai-poh / tʃɑɪ pɒh Alternate spellings: Cai Poh, Chye Poh Description: Chai poh is a term used in Singlish to refer to preserved radish that's commonly used in local dishes such as fried carrot cake, chee cheong fun, and oyster omelette. The radish is finely sliced, mixed with salt and spices, and then left to ferment for a couple of weeks, which gives it a pungent aroma and a slightly salty taste. Some people also enjoy eating chai poh on its own as a snack. Example: "Can you please add extra chai poh to my fried carrot cake?"...	Yes
leh	Sure, I'd be happy to help! Leh (pronounced "luh" or "leh" with a rising tone) is a common Singlish particle that is used to soften a request, statement, or command. It typically follows a sentence and is added to the end to express a mild suggestion or to seek agreement or confirmation. It can also be used to express surprise or disbelief. Here are some examples of how "leh" is used in sentences: "Can you pass me the salt, leh?" ...	No
atas	Sure, here's the dictionary entry for "atas": - Word: Atas - Pronunciation: a-tos (with emphasis on the first syllable) Definition: Atas is a Singlish term that means "haughty" or "snobbish," often used to describe someone who is acting or behaving in an elitist or pretentious manner. Additionally, it can also refer to something that is high-end or posh, such as a luxurious restaurant or an expensive brand. Example of usage: "Wow, your friend is so atas. She only wants to eat at Michelin-starred restaurants!"...	Partial-pron/etymology
Ah Long	- Pronunciation: ah long Alternate spellings: Ah Beng, Ah Long, Ah Beng Long Definition: Ah Long is a term used in Singlish to refer to an illegal moneylender or loan shark. It is derived from the Hokkien dialect, with "Ah" being a common prefix used to address someone, and "Long" meaning a dragon. Together, the term suggests that these moneylenders are often ruthless and possess an intimidating presence, just like a dragon. Ah Longs typically operate through illegal channels, offering high-interest loans to individuals who are in need of immediate cash. They generally target vulnerable borrowers, often using intimidation tactics or violence to enforce repayment. Example: "I heard he borrowed money from an Ah Long to pay off his gambling debts."...	Partial-example/spelling, etc.
blank	Sure, here's the dictionary entry for 'blank': Word: Blank Pronunciation: /blæŋk/ Definition: An expression used to describe a state of mental confusion or inability to think of a response to a question or situation. It can also be used to describe a lack of emotion or enthusiasm towards something. Example: "Huh, sorry ah, I go blank already. What was the question again ah?" (Translation: "Sorry, I'm feeling confused and I can't remember the question. Can you repeat it?")...	Not sure
balukoo	Balukoo is a Singlish slang word that means 'nonsense' or 'rubbish'. It is commonly used to describe something that is untrue or foolish. It is pronounced as buh-loo-koo with emphasis on the first syllable. Example: "Don't listen to him, what he said is just balukoo."	Redirect to

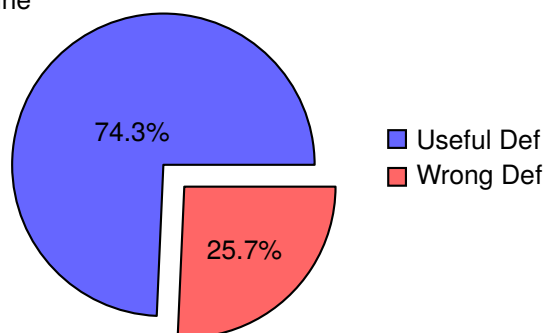
Table 1: Responses with tags

The responses here are shortened versions that only give up to the first example.

Table 2: Inter-Annotator Agreement

Entries	# Annotators	Agreement Scores		
		Complete	Partial	Disagree
1-53	4	4	24	25
54-100	3	13	26	7
101-334	2	65	-	170

Figure 2: Breakdown of agreed tags after streamline



4. Discussion

The scores in Figure 1 and Figure 2 indicate that for over half the samples, all annotators agreed that ChatGPT's responses were useful in terms of generating definitions for the entries. The significant difference between the proportion of 'Yes' annotated answers (which ranged around 25-60%) and the proportion of answers with just the correct definition shows that ChatGPT often outputs inaccuracies when providing additional information.

The fact that the model produces responses with at least a correct definition a majority of the time makes it helpful as a starting point for generating a dictionary though these outputs have to be checked before the dictionary can be reliable.

Although the initial granularity of the annotation tags resulted in a more rigorous annotation process and a lower agreement score, the complexity, together with how responses were not overwhelmingly of one level of correctness forces the participants to read each response carefully. It is intended to produce a more informative and accurate response. As a result, we were able to infer with confidence that approximately half of the samples have an agreed upon correct definition provided in the response. This translates into a significantly lowered cost for creating a new dictionary.

As part of the process of checking, ChatGPT's results were compared with DSSE, the Red Dot baby Talk and multiple online resources. We found no examples of existing text being reproduced exactly. In general, the examples given seem very different and although the same words occur in

both definitions in a few entries, they are not close enough to be considered copying. For example-consider the following

- **ChatGPT** Agak agak is a Singlish phrase that derives its roots from the Malay language. It is typically used to express a rough estimation or a guess. The phrase can be translated to mean "roughly" or "approximately" in English. It is commonly employed in everyday conversations to describe a haphazard estimation of measurements, time, or quantities when precise information is not available.
- **DSSE** agak /ah-gah(k), aga(k̄)/ n. & v. [Mal., conjecture, guessing; agak-agak approximately, as far as one can guess, more or less] Also agak-agak. A n. A guess, an estimation: Agaration. B v. Guess, estimate.

5. Current State

Currently, 579 of the entries have been annotated with tags and 399 of the entries are complete with POS, description, example, origin (if known⁹) and alternate spellings. The remaining 68% (1,204/1,783 entries) contain the unchecked ChatGPT descriptions as of now. The state of the entry (verified or unverified) will be displayed in the dictionary through a *verified* tag. We are going through the entries at the rate of around 100 per month.¹⁰

As noted in Lim (2011)'s paper, tone plays an important role in Singlish, especially for sentence final particles. As such, the tonal descriptions for the pragmatic particles¹¹ were added by hand. The tonal numbers were given through an approximate matching with the first six tones used in jyutping with the addition of the quick falling tone used by the particle *lah*. In that case, the diacritic used for the Mandarin fourth tone (lāh) was used.

⁹This relies solely on the principal annotator's knowledge

¹⁰The difference between annotation and completion numbers is caused by the annotator's incomplete knowledge of the vocabulary. For instance, they may be unsure about the correctness of the response or they may know it is wrong but cannot describe the term well.

¹¹The tonal particles in the dictionary are *ah, hah, hor, leh, lah, lor, mah, meh, sia, what, wor*

The breakdown of words according to their POS and Origin is given in Table 5.

POS	Count	%
Noun	253	64.7%
Verb	55	14.0%
Adjective	48	12.3%
Interjection	12	3.1%
Other	23	5.9%

Origin	Count	%
Malay	97	32.5%
English	77	25.8%
Hokkien	72	24.2%
Cantonese	17	5.7%
Other	35	11.8%

Table 3: POS / Origin of Annotated Words

6. Future Steps

The current dictionary has certain limitations and potential for further expansion.

We will tap into other Singlish language resources to continue NLP research, and strive to increase the dictionary’s coverage and utility. A method that is being considered is the augmentation of the dictionary automatically. For every description given by ChatGPT, an automated process can be created that checks through every word inside it. Words in the description that are neither in standard English dictionaries nor already inside this Singlish dictionary can then be fed into the prompt mentioned earlier, continuously expanding the Singlish dictionary.

Subsequently, the Singlish words in the corpora such as the NUS SMS corpus can also be semi-automatically fed into ChatGPT to further expand the dictionary. These sentences can then be used as examples in existing or future entries. However, due to the format of the text messages, a certain amount of additional processing might be needed to ensure that the new entries and examples obtained are of a suitable format for the dictionary.

As an extension of this study, we could also compare the results of multiple prompts. This serves to increase our understanding of how useful ChatGPT is in this task, and also perhaps, generate more satisfactory responses. Another approach in this sense is to run multiple LLMs to generate multiple descriptive hypotheses that can be used together to build one complete entry.

The issue of variation could also be further explored. Variation in the annotator’s knowledge of Singlish may account for the high level of disagreement. Singlish is not taught formally and the lexicon used varies largely depending on factors like

home language, social circle, age, etc. It is possible to recruit annotators with more similar intuitions e.g., through selecting ones from a specific background (or perhaps even choosing only those who have a similar confidence/description of their own Singlish). However, we do not wish to pick one ‘standard’ variety. Instead, another approach could be to accept the description or a slight variation of it if at least one annotator puts ‘yes’, and note that it may not be universally accepted.

Finally, we would like to link the entries to wordnet senses, so that they can easily be translated into other languages through the Open Multilingual Wordnet (Fellbaum, 1998; Bond and Foster, 2013).

7. Conclusion

This study shows the relative effectiveness of using an LLM to create more resources for a low-resource language. For Singlish, while the accuracy of the descriptions are far from perfect, more than half were deemed by the principal annotator to be accurate and comprehensive, and the generation of the entries in general provided a baseline that facilitated the building of the dictionary. As compared to having to write all the entries manually, or even through crowd sourcing, this method is an efficient and low cost way of creating and expanding a dictionary. Overall, we have found ChatGPT a useful tool to make draft entries: around 50% of samples were usable as is. Correcting entries with some errors is still faster than writing descriptions from scratch, significantly reducing the amount of work.

The development of the largest Singlish open sourced dictionary and the first to provide a tonal description of particles in this project is a step towards collecting more Singlish data and improving the resource available for this variety. We expect to have checked and, as necessary, rewritten a majority in the dictionary by May 2024. Just like other open dictionaries such as Urban Dictionary have contributed towards training specialised embedding models for NLP (Wilson et al., 2020), we hope that this can contribute towards and encourage Singlish language research, especially since large-scale language resources are becoming increasingly prevalent in the field of NLP. Although Singlish is used in everyday speech in Singapore, it does not enjoy the status of being an official language and there is no representative dictionary. To fill this void, through the use of LLM (ChatGPT) and other online resources, we have created a new, completely open-source Singlish dictionary.

The resource described in this study is published on [Github](#) with a CC-BY-SA license.

8. Acknowledgements

We would like to thank the reviewers for providing constructive feedback on how to improve our paper. Additionally, we would also like to thank Professor Suzy Styles for sharing her vocabulary lists.

9. Bibliographical References

- Francis Bond and Ryan Foster. 2013. [Linking and extending an open multilingual wordnet](#). In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362, Sofia.
- Francesco Cavallaro and Bee Chin Ng. 2009. [Between status and solidarity in Singapore](#). *World Englishes*, 28(2):143–159.
- Tao Chen and Min-Yen Kan. 2012. [Creating a live, public short message service corpus: the NUS SMS corpus](#). *Language Resources and Evaluation*.
- Renaë Cheng. 2021. [10 bizarre things singaporeans do that the rest of the world won't understand](#). [Online; posted 14-June-2021].
- Siew Yeng Chow and Francis Bond. 2022. [Singlish where got rules one? Constructing a computational grammar for Singlish](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5243–5250.
- Gilles-Maurice de Schryver. 2023. [Generative AI and Lexicography: The Current State of the Art Using ChatGPT](#). *International Journal of Lexicography*, 36(4):355–387.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Micha Elsner and Jordan Needle. 2023. [Translating a low-resource language using GPT-3 and a human-readable dictionary](#). In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.
- Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Colin Goh. 2002. *The Oxford Singlish dictionary*, second edition. Angsana Books.
- David Huddart. 2014. [Declarations of Linguistic Independence: The Postcolonial Dictionary](#). In *Involuntary Associations*. Liverpool University Press.
- Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. [Ammus : A survey of transformer-based pretrained models in natural language processing](#).
- Jack Tsen Ta Lee. 2004. *A Dictionary of Singlish and Singapore English*.
- Jakob RE Leimgruber. 2016. [Bah in Singapore English](#). *World Englishes*, 35(1):78–97.
- Zhuoyang Li. 2021. [An analysis of the linguistic characteristics of Singlish](#). *Journal of Contemporary Educational Research*, 5(3).
- Lisa Lim. 2011. [Tone in Singlish: Substrate features from Sinitic and Malay](#). *Substrate Features in Creole Languages*, pages 271–288.
- Ying Qi Soh, Junwen Lee, and Ying-Ying Tan. 2022. [Ethnicity and Tone Production on Singlish Particles](#). *Languages*, 7(3):243.
- Teresa Rebecca Tan. 2010. [Singlish](#). *Singapore Infopedia*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Rangan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- John C Wells. 1995. [Computer-coding the IPA: a proposed extension of SAMPA](#). *University College of London*.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. [Urban dictionary embeddings for slang NLP applications](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4764–4773, Marseille, France. European Language Resources Association.

Jock Wong. 2005. [“Why you so Singlish one?” A semantic and cultural interpretation of the Singapore English particle one](#). *Language in Society*, 34(2):239–275.

10. Language Resource References

Tao Chen and Min-Yen Kan. 2015. [The National University of Singapore SMS Corpus](#).

ICE-Singapore. [International corpus of english — singapore](#).

OpenAI. 2023. [ChatGPT: Large-scale GPT-3.5 for Conversational AI](#).

Fei Ting Woon and Suzy Styles. 2021. [Red-Dot Baby-Talk Quiz](#).