Word Association Network of English Words Unique to Singapore English

Jin Jye Wong [1] and Cynthia S. Q. Siew [1]

[1] *Department of Psychology, National University of Singapore, 9 Arts Link, Singapore, Singapore*

Abstract

This research describes a database of word association norms for words unique to the Singapore English dialect (e.g., *shiok* which means "great!" and *swaku* which refers to a country bumpkin). Because of the predominantly spoken nature of the basilectal form of Singapore English (colloquially referred to as "Singlish"), large-scale, local written corpora commonly used to compute semantic vector spaces of words in languages are uncommon. In order to study the semantic representations of uniquely Singapore English words, word associations were collected from native speakers of Singapore English via an online web application. When presented with a word (e.g. *shiok*), participants list the first words that come to their mind. The characteristics of the resulting association network are described.

Keywords

Semantic networks, word associations, cognitive networks, Singapore English, linguistics

Introduction

Singlish, a portmanteau of "Singapore English", is an informal, colloquial form of English used in Singapore [1]. As an English-based creole language, it incorporates elements of other languages commonly spoken in Singapore, such as Malay, Chinese dialects of Hokkien, Cantonese, and Teochew, and Indian languages such as Tamil [2], and deviates from English at both the lexical and grammatical levels [3]. This paper reports word associations collected from native speakers of Singlish in a paradigm similar to the Small World of Words project (SWOW) [4].

While natural language processing more commonly uses purely data-based approaches such as distributional semantics which are derived from textual analysis, use cases remain for word association approaches. For instance, word association-based models outperform text-based word co-occurrence models in predicting affective properties of words such as valence (the degree to which a word is positive or negative) [5], while also showing high correlation with human ratings [6]. Internal language models, which treat language as a body of knowledge residing in the brains of its speakers, which are also derived from word association data, also outperform external language models (such as those trained on corpora) that treat language as an extraneous object consisting of content produced by users of the language, in judging if words are related or similar [7]. It is worth noting that the set of word associations used to create the internal language models was smaller than the combined training corpora used by the external language models by several magnitudes.

Human-sourced word association data can also reinforce language models, especially in cases of languages where the available written corpora is smaller in size. SWOW data has already been used to improve downstream task performance on commonsense reasoning benchmarks [8].

For Singlish in particular, modern large language models are able to generate readable samples of text that adhere to Singlish grammar, but such snippets often only use the most common Singlish constructs and vocabulary (such as appending 'lah' or using 'makan' to substitute 'eat'), while simultaneously hallucinating Singlish words or constructs that do not exist (e.g. 'trafik' as an incorrect spelling of 'traffic', ungrammatically tacking on 'la(k)' to words within a sentence) [9].

Cognitive AI 2023,13th-15th November, 2023, Bari, Italy.
EMAIL: jin_jye@nus.edu.sg (A. 1); cynthia@nus.edu.sg (A. 2)

The aim of this study is to attempt to construct a large-scale free association network of Singapore English, the first of its kind. Such a network can provide more insight into the structure of the mental lexicon of the speakers of Singapore English and serve as a basis on which to conduct future psycholinguistic experiments.

Methods

2.1. Participant Details

Participants were recruited online using a crowd-sourced approach relying on social media, email, and word-of-mouth. To further incentivize participation, a lucky draw was held during the months of July to December 2022; 100 random participants were chosen every month to receive five Singapore dollars' (approximately 3.70 USD) worth of vouchers. Participation in the lucky draw was voluntary. This study was approved by the Institutional Review Board at the National University of Singapore.

This paper analyses data collected from 1st July 2022 to 16 January 2023. The dataset obtained consists of 1095 responses, of whom 614 (56.1%) identified as female, 458 (41.8%) identified as male, and 23 (2.1%) identified as neither. The average self-reported age was 31.9 years (SD = 12.72). Due to ethical considerations, no identifying data was collected, and it was possible for individuals to participate more than once.

Besides gender and age, information about race, birth country, current country of residence, and spoken languages were also collected. Regarding race, 828 (75.6%) identified as Chinese, 87 (7.94%) identified as Malay, and 67 (6.12%) identified as Indian, which approximately reflects Singapore's racial profile. Of the remainder, 45 (4.11%) identified as none of the above three races, while 68 (6.21%) did not respond. Most participants indicated their birth country as either Singapore (n = 1027, 93.8%), with the next largest subset being Malaysia (n = 16, 1.46%). The vast majority also indicated that their current country of residence was Singapore (n = 1068, 97.5%). Additionally, participants were also asked if they were native (L1) English speakers, and 950 (86.8%) indicated that they were, while the remainder were not. Other languages spoken by participants include Mandarin Chinese (n = 800, 73.1%), Malay (n = 96, 8.76%), and Tamil (n = 16, 1.46%). Participants could indicate more than one non-English language.

2.2. Data Collection Procedure

Upon clicking the button to begin the study, participants were instructed to fill in their demographic information. Next, they were presented with 20 cues (Singlish words/phrases) each from a list of approximately 4,500 Singlish words or phrases manually compiled from various dictionaries (e.g. the Coxford Singlish Dictionary [10]) and were instructed to provide up to three responses to each cue. If a participant was unable to provide any responses, they could proceed to the next cue regardless. After 20 cues, participants were directed to a landing page which contained a debriefing and the unique identifier code for their response which they could email to the researchers to participate in the lucky draw. The stimuli presented were selected pseudo-randomly; cues that had the fewest responses in the current iteration were more likely to be presented.

2.3. Quality Control

Following criteria similar to that of the original SWOW project [4], responses from some participants were excluded from analysis. First, to ensure that participants had sufficient experience with Singlish, participants who indicated their country of birth or current residence to not be Singapore were excluded. This removed 76 (6.94%) responses out of the total number of original 1,095 responses, leaving us with 1,019 responses. Next, participants for whom less than 75% of the responses were unique were also excluded (i.e. participants who gave the same response to many different cue words). This subset consisted of 3 (<1%) participants. Finally, participants with 75% or fewer of their responses appearing on a compiled Singlish dictionary were removed; this comprises of participants that responded with non-English words or phrases as well as participants that responded with

unrecognizable strings of English alphabet. This removed 15 (1.37%) of the participants. As a result, a total of 1,900 responses are not further considered, and the final dataset consists of 1,001 participants and 19,999 responses.

Data cleaning and analysis was conducted using R [11]. Verbose responses (i.e. responses that consisted of more than one word) were spilt into their constituent words, with each word treated as a separate responses from the same participant.

For spelling correction, the *hunspell* R package [12] was used. A custom Singlish dictionary was created by adding the list of Singlish cues to the default British English dictionary. Since many cues were phrases consisting of more than one word, the individual component words of each phrase were added manually as separate entries. For example, the cues *buay pai* ("not bad") and *buay tahan* ("can't endure") would result in the monograms *buay*, *pai*, and *tahan* being added. In the automated pipeline, words deemed to be misspelled were automatically replaced with the first word suggested by *hunspell*, with the function *hunspell_suggest*. This also entails that words spelled differently in American English would be corrected to their British English equivalents; consequently, different spellings of a word (e.g. *color* and *colour*) are treated as the same word (*colour*).

To convert words to their standard forms, lemmatization was performed using the *lemmatize_words* function from the *textstem* R package [13]. This maps inflections of any word back to their root word (e.g. *incurred* and *incurs* would be corrected to *incur*). Finally, any remaining word-final punctuation (e.g. full stops, tags, and quotes) as well as non-English characters, were removed.

Qualities of the Singlish Network

A word association network was created using the *igraph* R package [14]. Vertices of the graph consist of the Singlish cues and the split responses provided by participants. An edge connects two nodes if one of the cues produced the other word as a response. Edges are weighted, with edge lengths determined by taking the reciprocal of the number of times the response was produced when the cue word was presented; a short edge thus denotes a stronger relation between the two vertices it connects.

The final network was constructed by considering the first responses (R1) given by all participants, the rationale being that second and third responses, where given, might be susceptible to influence by any responses that precede them. The decision was made to construct an undirected instead of a directed graph given that the set of cues contained only uniquely Singlish words or phrases, while the set of responses contained both Singlish words and words found in the conventional English lexicon.

The R1 network has a total of 8215 vertices (mean degree = 4.27, SD = 7.74), with the vertex *no* having the highest degree of 191. For edges, the R1 network has a total of 17545 edges (mean edge length = 0.924, SD = 0.194), with the shortest edge (strongest association) being the edge from the Singlish cue phrase *monyet see monyet do* to the response *monkey*, which has a weight of 0.111. The network is not fully connected, and consists of 62 connected components, with the largest connected component consisting of 8071 vertices (98.2% of all vertices); this component has a global clustering coefficient of 0.00326 and an average local clustering coefficient of 0.0118.

The average shortest path length (ASPL) of the network is 1.73 and the network diameter is 8; these measure provide an indication of how efficient the network is overall [15]. The Small-World Index of the network is 0.44, showing that it is less interconnected than an Erdös-Rényi random graph with the same number of vertices and edges [16].

Discussion and Future Directions

The data collected represents the first-ever word association dataset dedicated to Singlish. Network construction notwithstanding, it can be used in tandem with existing Singapore English resources such as the Auditory English Lexicon Project (AELP) [17] to further examine the Singlish creole. Completeness of the network can be improved by further studies that examine Singlish responses to English cues, as well as consideration of non-English responses.

Other possible avenues of investigation include comparison of the network with similar word association networks of other languages. The constructed network can also be used as a starting point to predict properties of less common Singlish words.

References

[1] T. R. Yeo, Singlish, 2010. URL: https://eresources.nlb.gov.sg/infopedia/articles/SIP_1745_2010-12-29.html.

[2] J. T. Platt, The Singapore English speech continuum and its basilect 'Singlish' as a 'Creoloid', Anthropological Linguistics (1975): 363–374

[3] J. R. E. Leimgruber, Singapore english. Language and Linguistics Compass, 5.1 (2011):47–62.

[4] S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, and G. Storms, The "Small World of Words" English word association norms for over 12,000 cue words. Behavior research methods 51 (2019), 987-1006.

[5] H. Vankrunkelsven, S. Verheyen, G. Storms, and S. De Deyne, Predicting Lexical Norms: A Comparison between a Word Association Model and Text-Based Word Co-occurrence Models. Journal of Cognition 1, (2018), 45.

[6] B. van Rensbergen, S. De Deyne, and G. Storms, Estimating affective word covariates using word association data. Behavior Research Methods 48 (2008), 1644–1652. https://doi.org/10.3758/s13428-015-0680-2

[7] S. De Deyne, A. Perfors, and D. J. Navarro, in: Predicting human similarity judgments with distributional models: The value of word associations. COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers, 2016, pp. 1861–1870. https://lirias.kuleuven.be/2324421

[8] C. Liu, T. Cohn, and L. Frermann, Commonsense Knowledge in Word Associations and ConceptNet. 2021. arXiv preprint arXiv:2109.09309.

[9] Z. X. Yong, R. Zhang, J. Z. Forde, S. Wang, S. Cahyawijaya, H. Lovenia, L. Sutawika, J. C. B. Cruz, L. Phan, Y. L. Tan, and A. F. Aji, Prompting Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages, 2015.

[10] C. Goh, Y. Y. Woo, The Coxford Singlish Dictionary, 2nd. ed, Angsana Books, 2009

[11] R Core Team, R: A language and environment for statistical computing, 2019. URL: https://www.R-project.org/. version 4.1.2 (2021-11-01)

[12] J. Ooms, hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker, 2020. URL: https://CRAN.R-project.org/package=hunspell

[13] T. W. Rinker, {textstem}: Tools for stemming and lemmatizing text, 2018. URL: http://github.com/trinker/textstem

[14] G. Csardi and T. Nepusz, The igraph software package for complex network research. Interjournal, Complex Systems 2006, p. 1695. URL: https://igraph.org

[15] C. S. Q. Siew, D. U. Wulff, N. Beckage, and Y. Kenett, Cognitive Network Science: A review of research on cognition through the lens of network representations, processes, and dynamics. Complexity (2019) 1-24.

[16] M. D. Humphries and K. Gurney, Network 'small-world-ness': a quantitative method for determining canonical network equivalence. PloS one 3.4 (2008)

[17] W.D. Goh, M. J. Yap, M. and Q. W. Chee, The Auditory English Lexicon Project: A multi-talker, multi-region psycholinguistic database of 10,170 spoken words and nonwords. Behavior Research Methods 52.5 (2020): 2202-2231.