<a rel="license" href="<http://creativecommons.org/licenses/by-nc/4.0/>"><img alt="Creative Commons License" style="border-width:0" src="![img](https://i.creativecommons.org/l/by-nc/4.0/88x31.png)" \*></a><br />This work is licensed under a <a rel="license" href="<http://creativecommons.org/licenses/by-nc/4.0>\*">Creative Commons Attribution-NonCommercial 4.0 International License</a>.

# Table of Contents

1. [About Small World of Words project (SWOW) & SWOW-ZH](#orga7a1ee6)

2. [Instructions to the repository](#org6423954)

1. [Obtaining the data](#org008856a)

2. [Raw data](#org1a6cd57)

3. [Preprocessing scripts](#orgc0ba651)

4. [processing scripts](#orga56d790)

1. [Associative frequencies and graphs](#org8eea69f)

2. [Derived statistics](#orgdff87ab)

3. [Centralities and similarities](#org11b9387)

5. [Applicability in other SWOW lexicons](#org511e306)

3. [Publications based on SWOW](#org124b364)

<a id="orga7a1ee6"></a>

# About [Small World of Words project (SWOW)](https://smallworldofwords.org/project/) & SWOW-ZH

The small world of words project is a large-scale scientific study that aims to

build a mental dictionary or lexicon in the major languages of the world and

make this information widely available <sup><a id="fnr.1" class="footref" href="#fn.1" role="doc-backlink">1</a></sup>.

In contrast to a thesaurus or dictionary, we use word associations to

learn about what words mean and which ones are central in the human mind. This

enables psychologists, linguists, neuro-scientists and others to test new

theories about how we represent and process language. This knowledge could also

be applied in a variety of ways, from learning about the difference between

cultures, to learning (or forgetting) new words in a first or a second language.

SWOW-ZH is a daughter project of SWOW to map mental lexicon in Chinese, as the

suffix `ZH` stands for \*Zhongwen\* (中文, \*Chinese\*). It was initiated to provide a

comprehensive framework to measure the mental lexicon with regard to the Chinese

culture and people, and the bases for comparable studies between Chinese and

other languages.

The participant task we used is called \*multiple response association\* <sup><a id="fnr.2" class="footref" href="#fn.2" role="doc-backlink">2</a></sup>.

The methodology is based on a continued word association task, in which

participants see a cue word and are asked to give three associated responses to

this cue word. As the number of participants increases, the lexicon becomes

comprehensive and efficient in representing mental lexicon. Therefore, it

focuses on the aspects of word meaning that are shared between people without

imposing restrictions on what aspects of meaning should be considered.

Chinese is a demographically and culturally complex language, whose dialects and

writing systems are difficult to exhaust. In the SWOW-ZH project, we primarily

\*focused on Mandarin Chinese\* (普通话, Putonghua), which is the \*de facto\* standards

for the language used in most regions of Chinese mainland. Additionally, the

native dialect of the participants was collected as a complementary information.

We didn't include Cantonese, but it was collected in another daughter project of

SWOW [SWOW-HK](https://smallworldofwords.org/hk).

The study was conducted in Professor CAI Qing's lab at the School of Psychology

and Cognitive Science, East China Normal University (华东师范大学心理与认知科???

学院，蔡清教授团???), in collaboration with Dr. Simon de Deyne at Melbourne

University. Please address questions and suggestions to:

- DING Ziyi 丁子益 [ziyi.ecnu@gmail.com](mailto:ziyi.ecnu@gmail.com)

- LI Bing 李兵 [lbing314@gmail.com](mailto:lbing314@gmail.com) [github](https://github.com/lib314a)

<a id="org6423954"></a>

# Instructions to the repository

In this repository you will find a basic analysis pipeline for the

Chinese SWOW project which allows you to import a preprocessing the data

as well as compute some basic statistics.

<a id="org008856a"></a>

## Obtaining the data

In addition to the scripts, you will need to retrieve the word

association data. Currently word association and participant data is

available for 10,192 cues. The data consists of over 2 million responses

collected between 2016 and 2023. They are currently submitted for

publication. Note that the final version is subject to change. If you

want to use these data for your own research, you can obtain them from

the Small World of Words research page

(<https://smallworldofwords.org/project/research/>).

To start the pileline, SWOW-ZH\_raw.mat should be put into the data folder.

Please note that data themselves are licensed under Creative Commons

Attribution-NonCommercial-NoDerivs 3.0 Unported License

(<http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en\_US>). They

cannot be redistributed or used for commercial purposes.

While the majority of the data was collected on [the SWOW platform (ZH)](https://smallworldofwords.org/zh),

a subset was collected on another China-based surveying platform [NAODAO (脑岛)](https://www.naodao.com) using the same tasks with the same inclusion standards.

This presumably won't detriment the reliability of the data.

To cite these data: 【待定???

If you find any of this useful, please consider sharing the word

association study (<https://smallworldofwords.org/zh/project>).

<a id="org1a6cd57"></a>

## Raw data

Since this is an ongoing project, data is regularly updated. Hence, all

datafiles refer to a release date in its filename.

Current release is 【待定???.

1. sequenceNumber: ascending sequence from 1 to the end

2. sheetID: unique identifier for sheets, each sheet includes one cue

and three responses given by one participant

3. participantID: unique identifier for the participant

4. created<sub>at</sub>: time and date of participation

5. age: age of the participant

6. nativeLanguage: native language from a short list of common

languages

7. gender: gender of the participant (Female / Male / X)

8. education: Highest level of education: 1 = None, 2 = Elementary

school, 3 = High School, 4 = College or University Bachelor, 5 =

College or University Master

9. city: city (city location when tested, might be an approximation)

10. country: country (country location when tested)

11. section: identifier for the snowball iteration (set1-10 = Ten sets

collected in the SWOW platform, NAODAO = One set collected in the

NAODAO platform (<https://www.naodao.com/research/project>))

12. cue: cue word

13. R1Raw: raw primary associative response

14. R2Raw: raw secondary associative response

15. R3Raw: raw tertiary associative response

16. R1: primary associative response

17. R2: secondary associative response

18. R3: tertiary associative response

<a id="orgc0ba651"></a>

## Preprocessing scripts

To avoid possible mistakes when read Chinese strings in MATLAB, we

recommend that all the data should be loaded and saved as mat format. We

also provide data in csv format for the users of other programming

languages.

The preprocessing scripts consist of wordCleaning.m,

participantCleaning.m and dataFiltering.m scripts.

wordCleaning.m: Problematic cue words and responses are marked or

modified according to the dictionaries. The dictionaries could be found

in the data/dictionaries folder. The input of the script,

SWOW-ZH<sub>raw.mat</sub>, should be put in the data folder.

participantCleaning.m: Problematic participants are deleted. The script

could take a day to compare every response with a Chinese wordlist.

dataFiltering.m: Remain 55 participants for each cue words. The output

of the script is written to data/SWOW-ZH<sub>R55.mat</sub>. The participants were

selected to favor participants with less missing responses and Mandarin

speakers. The preprocessed data could be found in the Small World of

Words research page

(<https://smallworldofwords.org/project/research/>).

<a id="orga56d790"></a>

## processing scripts

The preprocessing scripts consist of networkGeneration.m,

frequencyCalculating.m, centralityCalculating.m and

similarityCalculating.m scripts.

【还有R的代码需要加在这里，可以发我我来写进去???

<a id="org8eea69f"></a>

# Associative frequencies and graphs

networkGeneration.m: The preprocessed data is used to derive the

associative frequencies (i.e., the conditional probability of a response

given a cue) and saved in the output folder named as assocFrequency<sub>R1</sub>

or \\_R123, where the first column contains cue words, the second column

contain responses, the third column contains associative frequencies

between them. Use associative frequencies to extract the largest

strongly connected component for graphs based on the first response (R1)

or all responses (R123). The graphs are written to data/

SWOW-ZH<sub>network.mat</sub>. And the adjacency matrices are written to output

folder named as adjacencyMatrix<sub>R1</sub> or \\_R123 and consist of directed

weighted matrices, where each row labeled by N cue words and each column

labeled by N responses. Then, the N×N matrices are filled by normalized

associative strengths. In most cases, associative frequencies will need

to be converted to associative strengths by dividing with the sum of all

strengths for a particular cue. Vertices that are not part of the

largest connected component are listed in a report in the output folder

named as lostNodes<sub>R1</sub> or \\_R123.

<a id="orgdff87ab"></a>

# Derived statistics

frequencyCalculating.m: The script is used to describe the

characteristics of responses, cue words and participants.

1. Response statistics

> Currently the script calculates the number of cue words where a response is reported

> as their types, tokens and hapax legomena responses (responses that only occur once).

> The results can be found in the output folder named as resStats.

2. Cue statistics

> Only words that are part of the strongly connected component are

> considered. Results are provided for the R1 graph and the graph with

> all responses (R123). The results can be found in the output folder

> named as cueStats<sub>R1</sub> or \\_R123. The file includes the following:

- Coverage: How many of the responses are retained in the graph after

removing those words that aren't a cue or aren't part of the strongest

largest component.

- Unknown: The number of unknown responses

- R1missing: The number of missing R1 responses

- R2missing: The number of missing R2 responses

- R3missing: The number of missing R3 responses

> A histogram of the response coverage for R1 and R123 graphs can be

> obtained from the frequencyCalculating.m script. Vocabulary growth

> curves can be obtained with plotVocabularyGrowth.R.

> 【如果coverage是用R算的这里需要改一下???

<a id="org11b9387"></a>

# Centralities and similarities

centralityCalculating.m: Based on the largest strongly connected

component for graphs, the script calculates centrality-related

indicators including: types and tokens, in-degree, out-degree, PageRank,

centrality and betweenness. The scrip inserts some functions from the

Brain Connectivity Toolbox (BCT)

(<http://www.brain-connectivity-toolbox.net>). The output is written

in the output folder named as centrality<sub>R1</sub> or \\_R123.

similarityCalculating.m: Based on the largest strongly connected

component for graphs, the script calculates four kinds similarity

including: cosine similarity only (AssocStrength), positive pointwise

mutual information (PPMI), random walk (RW) and word embedding after

random walk (RW-embedding). The script is adapted from SWOW-EN and

SWOW-RP. The output is written in the output folder named as

similarity<sub>R1</sub> or \\_R123.

<a id="org511e306"></a>

## Applicability in other SWOW lexicons

Since other SWOWs are mainly processed by R scripts, a MATLAB scrip is

provided thus other SWOWs could be processed by MATLAB. The SWOWs.m is

used to count associative frequencies and generate graphs, and calculate

in-degrees of other SWOWs. The inputs of the script are preprocessed

data of other SWOWs put in the data/SWOWs folder. The outputs of the

script are the graphs written to data/SWOWs/SWOW-XX<sub>network.mat</sub>. While

the XX could be substituted by EN (American English), DU (Dutch) and RP

(Rioplatense Spanish). The outputs could be loaded as inputs into

centralityCalculating.m and similarityCalculating.m.

<a id="org124b364"></a>

# Publications based on SWOW

Following is an exhaustive list of the publications based on or used part of the lexicons:

- Journal articles

- Cox, C. R., & Haebig, E. (2022). Child-oriented word associations improve models of early word learning. Behavior research methods, (), 1???22.

- De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. Cognitive Science, 45(1), 12922.

- De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. Journal of Experimental Psychology: General, 145(9), 1228???1254. <http://dx.doi.org/10.1037/xge0000192>

- Jana, A., Haldar, S., & Goyal, P. (2022). Network embeddings from distributional thesauri for improving static word representations. Expert Systems with Applications, 187(), 115868.

- Johnson, D. R., & Hass, R. W. (2022). Semantic context search in creative idea generation. The Journal of Creative Behavior, 56(3), 362???381.

- Kumar, A. A., Balota, D. A., & Steyvers, M. (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. Journal of Experimental Psychology: Learning, Memory, and Cognition, 46(12), 2261.

- Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). Semantic memory search and retrieval in a novel cooperative word game: a comparison of associative and distributional semantic models. Cognitive Science, 45(10), 13053.

- Maxwell, N. P., & Buchanan, E. M. (2020). Investigating the interaction of direct and indirect relation on memory judgments and retrieval. Cognitive Processing, 21(1), 41???53.

- Meersmans, K., Bruffaerts, R., Jamoulle, T., Liuzzi, A. G., De Deyne, S., Storms, G., Dupont, P., ??? (2020). Representation of associative and affective semantic similarity of abstract words in the lateral temporal perisylvian language regions. Neuroimage, 217(), 116892.

- Meersmans, K., Storms, G., De Deyne, S., Bruffaerts, R., Dupont, P., & Vandenberghe, R. (2022). Orienting to different dimensions of word meaning alters the representation of word meaning in early processing regions. Cerebral Cortex, 32(15), 3302???3317.

- Melvie, T., Taikh, A., Gagn\\'e, Christina L, & Spalding, T. L. (2022). Constituent processing in compound and pseudocompound words. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie exp{\\'e}rimentale, (), .

- Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: a comprehensive model comparison. Cognitive Science, 45(8), 13030.

- Sarkar, S., Bhagwat, A., & Mukherjee, A. (2022). A core-periphery structure-based network embedding approach. Social Network Analysis and Mining, 12(1), 1???12.

- Valba, O., & Gorsky, A. (2022). K-clique percolation in free association networks and the possible mechanism behind the $ $7$\backslash$pm 2$$7$&plusmn;$2 law. Scientific reports, 12(1), 1???9.

- Valba, O., Gorsky, A., Nechaev, S., & Tamm, M. (2021). Analysis of english free association network reveals mechanisms of efficient solution of remote association tests. PloS one, 16(4), 0248986.

- Vankrunkelsven, H., Vankelecom, L., Storms, G., De Deyne, S., & Voorspoels, W. (2021). Guessing Words. In (Eds.), Cognitive Sociolinguistics Revisited (pp. 572???583). : De Gruyter Mouton.

- Verheyen, S., De Deyne, S., Linsen, S., & Storms, G. (2020). Lexicosemantic, affective, and distributional norms for 1,000 dutch adjectives. Behavior research methods, 52(3), 1108???1121.

- Wong, T. Y., Fang, Z., Yu, Y. T., Cheung, C., Hui, C. L., Elvev\aag, Brita, De Deyne, S., ??? (2022). Discovering the structure and organization of a free cantonese emotion-label word association graph to understand mental lexicons of emotions. Scientific Reports, 12(1), 1???12.

- Wulff, D. U., De Deyne, S., Aeschbach, S., & Mata, R. (2022). Using network science to understand the aging lexicon: linking individuals' experience, semantic networks, and cognitive performance. Topics in Cognitive Science, 14(1), 93???110.

- Wulff, D. U., & Mata, R. (2022). On the semantic representation of risk. Science Advances, 8(27), 1883.

- Proceedings, pre-prints etc

- Ashok Kumar, A., Garg, K., & Hawkins, R. (2021). Contextual flexibility guides communication in a cooperative language game. In , Proceedings of the Annual Meeting of the Cognitive Science Society (pp. ). : .

- Berger, U., Stanovsky, G., Abend, O., & Frermann, L. (2022). A computational acquisition model for multimodal word categorization. arXiv preprint arXiv:2205.05974, (), .

- Branco, Ant\\'onio, Rodrigues, Jo\\~ao, Salawa, Ma\lgorzata, Branco, R., & Saedi, C. (2020). Comparative probing of lexical semantics theories for cognitive plausibility and technological usefulness. arXiv preprint arXiv:2011.07997, (), .

- Du, Y., Wu, Y., & Lan, M. (2019). Exploring human gender stereotypes with word association test. In , Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6133???6143). : .

- Han, Z., & Truex, R. (2020). Word association tests for political science. Available at SSRN 3701860, (), .

- Kovacs, C. J., Wilson, J. M., & Kumar, A. A. (2022). Fast and frugal memory search for communication. In , Proceedings of the Annual Meeting of the Cognitive Science Society (pp. ). : .

- Liu, C., Cohn, T., De Deyne, S., & Frermann, L. (2022). Wax: a new dataset for word association explanations. In , Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (pp. 106???120). : .

- Liu, C., Cohn, T., & Frermann, L. (2021). Commonsense knowledge in word associations and conceptnet. arXiv preprint arXiv:2109.09309, (), .

- Nedergaard, J., Smith, K., & Smith, K. (2020). Are you thinking what i'm thinking? perspective-taking in a language game. In , CogSci (pp. ). : .

- Nighojkar, A., Khlyzova, A., & Licato, J. (2022). Cognitive modeling of semantic fluency using transformers. arXiv preprint arXiv:2208.09719, (), .

- Petridis, S., Shin, H. V., & Chilton, L. B. (2021). Symbolfinder: brainstorming diverse symbols using local semantic networks. In , The 34th Annual ACM Symposium on User Interface Software and Technology (pp. 385???399). : .

- Rodrigues, Jo\\~ao, Branco, R., & Branco, Ant\\'onio (2022). Transfer learning of lexical semantic families for argumentative discourse units identification. arXiv preprint arXiv:2209.02495, (), .

- Rotaru, A. S. (2020). Computational explorations of semantic cognition (Doctoral dissertation). UCL (University College London), .

- Salawa, Ma\lgorzata (). Word embeddings from lexical ontologies: a comparative study. , (), .

- Sarkar, S., Bhagwat, A., & Mukherjee, A. (2018). Core2vec: a core-preserving feature learning framework for networks. In , 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 487???490). : .

- Siow, S., & Plunkett, K. (2021). Exploring the variable effects of frequency and semantic diversity as predictors for a word's ease of acquisition in different word classes. In , Proceedings of the Annual Meeting of the Cognitive Science Society (pp. ). : .

- Thawani, A., Srivastava, B., & Singh, A. (2019). Swow-8500: word association task for intrinsic evaluation of word embeddings. In , Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP (pp. 43???51). : .

- van Paridon, J., Liu, Q., & Lupyan, G. (2021). How do blind people know that blue is cold? distributional semantics encode color-adjective associations. In , Proceedings of the Annual Meeting of the Cognitive Science Society (pp. ). : .

- Wulff, D. U., Aeschbach, S., De Deyne, S., & Mata, R. (2022). Data from the myswow proof-of-concept study: linking individual semantic networks and cognitive performance. Journal of Open Psychology Data, 10(1), .

- Yang, W., & Ma, X. (). Building knowledge graphs of experientially related concepts. , (), .

# Footnotes

<sup><a id="fn.1" href="#fnr.1">1</a></sup> De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. Behavior Research Methods, 45(2), 480???498. <http://dx.doi.org/10.3758/s13428-012-0260-7>

<sup><a id="fn.2" href="#fnr.2">2</a></sup> Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? Memory \\& cognition, 28(6), 887???899.