**SWOW-ZH**

The Small World of Words project (SWOW) (<https://smallworldofwords.org/project/>) project is a scientific project to map word meaning in various languages. In contrast to dictionaries, it focuses on the aspects of word meaning that are shared between people without imposing restrictions on what aspects of meaning should be considered. 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.

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.

Suggestions are always appreciated, and do not hesitate to get in touch if you any questions. Any questions and suggestions could be send to [ziyi.ecnu@gmail.com](mailto:ziyi.ecnu@gmail.com). 【Help: 不是很确定邮箱应该写在哪里】

**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/>).

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.

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>).

**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\_at: 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

**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\_raw.mat, 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\_R55.mat. 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/>).

**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】

**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\_R1 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\_network.mat. And the adjacency matrices are written to output folder named as adjacencyMatrix\_R1 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\_R1 or \_R123.

**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 types, tokens and hapax legomena responses (responses that only occur once). The results can be found in the output folder named as resStats.

1. 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\_R1 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】

1. Participant statistics

Only participants remain after preprocessing are included in the demographic statistics. The results can be found in the output folder named as ppStats\_R1 or \_R123. The file includes the following: age, native language, gender, level of education, city and country.

**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\_R1 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\_R1 or \_R123.

**Applicability in other SWOWs**

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\_network.mat. 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.