

README File

Supporting data for the thesis “Words in Minds and Machines: A Computational Characterization of Chinese Mental Lexicon” are made available on HKU DataHub in the following folders.

- Chapter 2
- Chapter 3
- Chapter 4
- Chapter 5

Chapter 2: Data associated with the chapter “Composition as Nonlinear Combination in Semantic Space: A Computational Characterization of Compound Processing”, which includes the following subfolders.

- Notch_Model
- Dataset
- Statistical_Analysis

Notch_Model

- train_val.py: Codes to train and validate the Notch model

Dataset

- LDT.xlsx: Response latencies and error rates, as well as word- and character-level features are extracted from the MELD-SCH database (Megastudy of Lexical Decision in Simplified Chinese; Tsang et al., 2018), with computed metrics SimCL and DistC plugged in for each two-character word and nonword. Type-based and position-specific family size are computed for each constituent character. See description for each column in the table below.
- Eye-Tracking.xlsx: First fixation duration and total fixation duration are extracted from the database of eye-movement measures on words in Chinese reading (Zhang et al., 2022), with word- and character-level features retrieved from MELD-SCH (Tsang et al., 2018). The computed metrics SimCL and DistC are plugged in for each two-character word. Type-based and position-specific family size are computed for each constituent character. See description for each column in the table below.
- Gao.xlsx: List of real words, pseudowords, and nonwords are identified by Gao et al. (2022). The computed metric DistC is plugged in for each item.

Variable Name	Description
<i>Items</i>	
Word	Real word or nonword
C1	The first constituent character
C2	The second constituent character
<i>Columns specific to the LDT dataset</i>	
Real	Lexicality of the item (1 for real words; 0 for nonwords)
zRT	Mean standardized reaction time across participants
ERR	Mean error rate across participants
<i>Columns specific to the Eye-Tracking dataset</i>	
FFD	First fixation duration
TFD	Total fixation duration
<i>Columns shared by the LDT and Eye-Tracking dataset</i>	
Stroke	Total number of strokes of the whole word
LogWF	Word frequency based on SUBTLEX-CH
LogCF	Character frequency based on SUBTLEX-CH
LogFS.Type	Character's type-based and position-specific family size
LogNoM	Character's number of meanings
LogNoP	Character's number of pronunciations
SimCL	Similarity between the compositional and the lexicalized meaning representations
DistC	Distinctness of the compositional meaning representation

¹ Log indicates that log-transformation is implemented on the variable

² C1 or C2 indicates that the variable is for the first or the second constituent character

Statistical_Analysis

- analysis_real_zRT.R: Analysis of the LDT data (zRT of real words)
- analysis_real_ERR.R: Analysis of the LDT data (ERR of real words)
- analysis_non_zRT.R: Analysis of the LDT data (zRT of nonwords)
- analysis_non_ERR.R: Analysis of the LDT data (ERR of nonwords)
- analysis_FFD.R: Analysis of the Eye-Tracking data (FFD)
- analysis_TFD.R: Analysis of the Eye-Tracking data (TFD)
- analysis_Gao.R: Analysis of Gao et al's (2022) data

Note: To replicate the analysis, please set directory to the path containing the dataset and make sure that packages are properly installed before running the R code.

Chapter 3: Data associated with the chapter “Probing Lexical Ambiguity in Chinese Characters via Their Word Formations: Convergence of Perceived and Computed Metrics”, which includes the following subfolder.

- Metrics

Metrics

- metrics.xlsx: Computed metrics (cDoM and cRoM) and perceived metrics (pNoM and pRoM) for the 4,344 characters.

Chapter 4: Data associated with the chapter “The good, the bad, and the ambivalent: Extrapolating affective values for 38,000+ Chinese words via a computational model”, which includes the following subfolders.

- Code
- Norms

Code

- train_test.py: codes used to implement training and validation across folds
- extrapolate.py: codes used to implement extrapolation

Note:

1. Please make sure that dependencies are properly installed before running the codes. For your reference, we used the following packages to implement the experiment.
 - keras 2.10.0
 - keras-mdn-layer 0.3.0
 - tensorflow 2.10.0
 - tensorflow-probability 0.18.0
2. Please try different hyperparameters (i.e., number of hidden units, number of Gaussian components) to optimize your own model.

Norms

- norms.xlsx: standardized rated norms and extrapolated norms summarized in one file

Sheets and Columns	Description
Rated	
Word	Word sample
z.valence.mean	Mean of standardized* valence ratings
z.valence.var	SD of standardized* valence ratings
valence.N	Number of valid valence ratings
z.arousal.mean	Mean of standardized* arousal ratings
z.arousal.var	SD of standardized* arousal ratings
arousal.N	Number of valid arousal ratings
Extrapolated	
Word	Word sample
z.valence.mean	Extrapolated z.valence.mean
z.valence.var	Extrapolated z.valence.var
z.arousal.mean	Extrapolated z.arousal.mean
z.arousal.var	Extrapolated z.arousal.var

* Standardized transformation was based on individual rater's mean and SD.

Chapter 5: Data associated with the chapter “A Generative Approach to Extrapolate Word Concreteness Ratings”, which includes the following subfolders.

- Code
- Results
- Norms

Code

- train_test.py: codes used to implement training and validation across folds
- extrapolate.py: codes used to implement extrapolation

Note:

1. Please make sure that dependencies are properly installed before running the codes.
For your reference, we used the following packages to implement the experiment.
 - keras 2.10.0
 - keras-mdn-layer 0.3.0
 - tensorflow 2.10.0
 - tensorflow-probability 0.18.0
2. Please try different hyperparameters (i.e., number of hidden units, number of

Gaussian components) to optimize your own model.

Results

- result.xlsx: complete results of the five-fold cross-validation for predictions based on MDN and KNN

Norms

- norms.xlsx: rated norms and extrapolated norms summarized in one file

Sheets and Columns	Description
Rated	
Word	Word sample
con.mean	Mean of concreteness ratings
con.var	SD of concreteness ratings
con.N	Number of valid concreteness ratings
Extrapolated	
Word	Word sample
ext.con.mean	Extrapolated con.mean
ext.con.var	Extrapolated con.var

We appreciate that the use of the codes and the database is acknowledged and cited properly. We also invite you to share any amendments that you may have made on the computational model. Do not hesitate to contact us should you have any questions or specific requests.

Tianqi Wang

tianqi93@connect.hku.hk