

## Summary

Dedicated language data researcher with 5+ years experience in both theoretical and experimental linguistics, specializing in the psychology and the neurology of language. Expertise in neuro-behavioral-linguistics experiment design, analysis and evaluation of NLP models and representations, and computational modeling. Outstanding analytical and teamwork skills.

## Education

07/2021 **Ph.D. in Linguistics with Cognitive Science Specialization**, Michigan State University, (MSU) U.S.

**Dissertation:** *Competition in natural language meaning - The case of adjectival constructions*

**Aim:** Identifying and explaining cognitive constraints/universals in human language

02/2015 **M.A. with merit**, Hong Kong Baptist University, H.K.

**Major:** Language Studies.

**Thesis:** *The second language acquisition of the Mandarin potential complement construction*

**Aim:** Building statistical models to measure language competence and performance

## Computer Skills

2017-present Programming Languages

- Design, implement and debug Python programs
- Object-centered design and implementation in C++
- Statistical testing, modeling, advanced graphics in R and MATLAB
- Basic familiarity with Shell script, Java, HTML, CSS, SQL
- Development Environments: Visual Studio; RStudio; Anaconda; Visual Paradigm
- Productivity Applications: Git/GitHub; Cloud Service (Virtual Machines, Colab)
- Links to sample scripts NLP:** <https://github.com/yancong222/scripts>
- Links to sample scripts ML:** <https://github.com/yancong222/scriptscz>

## Experience

Computational linguistics Postdoctoral trainee, Feinstein Institutes for Medical Research, Northwell Health

07/2021-present

- Develop scalable classifiers and tools leveraging regression, and rules-based models

07/2021-present

- Work with 5 team members from different cultural and educational background, including medical, linguistics, computational, and acoustics

07/2021-present

- Deploy pre-trained transformer language models (including GPT-3, RoBERTa, BERT, T5-11b) to identify speech biomarkers, with accuracy improved by  $\approx 7\%$

07/2021-present

- Large-scale dataset pre-processing and analyzing, including 3 metrics (similarity, cross-entropy loss, and relative probability) and over 5 methods (tf-idf, word2vec, glove, random walk, next sentence probability, centroid, etc.).

- Natural Language Processing 09/2020-07/2021 Graduate student researcher, Department of Linguistics and Languages, MSU
- Deployed Google Cloud virtual machine instances to conduct task-based inference, analyze neural language models' functionalities, and design assessment algorithms
  - Developed Python programs and R programs to evaluate transformers' performance, designing metrics such as *accuracy*, *cross-entropy loss*, *HITS@K*, *relative rank*
  - Identified areas for transformer models improvement, through case studies of RoBERTa-large and text-to-text-transfer-transformer (T5)
- Acoustic analysis 08/2019-07/2021 Lab member *Timing, Attention, and Perception Lab*, Department of Psychology, MSU
- Developed R scripts for confusion matrices analysis of speech perception in noise: Multidimensional Scaling (MDS) using the `cmdscale()` function in R
  - Implemented and plotted MDS (package *igraph*) as `layout.mds` in R
  - Used MATLAB to manipulate data and generate confusion matrix
  - Modeled correlation of rhythm variation and speech perception in noise using SPSS and R (packages: *lattice*, *ggplot*, *dplyr*, *tidyverse*), resulting in 1 manuscript
- Neurology 08/2016-07/2021 Lab member *Neurolinguistics Lab*, Department of Linguistics and Languages, MSU
- Led coordination of stimuli design and auditory/visual paradigm design (4 team members)
  - Provided coordination of data collection/acquisition efforts: Electroencephalography (EEG) measurement for 60+ participants, resulting in 3 conference presentations
  - Trained junior members in acquisition and graph analysis of event-related potentials
  - Professional proficiency with EEG, limited working proficiency with eye-tracking, and basic familiarity with fMRI and PET
- Psychology 08/2016-07/2021 Lab member / Data analyst *Semantics & Syntax Lab*, Department of Linguistics and Languages, Michigan State University (MSU)
- Designed 2 artificial language learning experiments and 1 lexical decision task implemented in PsychToolbox (MATLAB), PsychoPy, IBEX farm
  - Developed 4 web-based acceptability judgement surveys performed in Amazon Mechanical Turk and Prolific, and 1 lab-based self-paced reading task
  - Analyzed speech/text data, using repeated measures ANOVAs (GG corrected) and pairwise comparisons. These studies led to publications
- Linguistic data corpus 04/2015-07/2016 Project assistant *Joint Research Center on Chinese Linguistics*, Hong Kong Polytechnic University - Peking University
- Annotated and extracted dataset on Balanced Corpus, Web-based Corpus, and Inter-language Corpus
  - Assisted annotation, classification, and statistical modeling for 2 ontology projects on World Chinese Variations and Chinese Linguistic KnowledgeNet
- Linguistic theory 02/2015-04/2015 Senior Research Assistant, *Department of English language and literature*, Hong Kong Baptist University
- Research studies, with a focus on the argument structure of Chinese
  - Seminar series entitled as "the C-Command Club" on the essence of generative linguistics