Yan Cong

+1 (517) 940 1452
yancong222@gmail.com
https://yancong222.github.io/

Summary

Dedicated speech and language data researcher with 5+ years experience in both theoretical and experimental linguistics, specializing in the psychology and the neurology of language. Expertise in analysis and evaluation of NLP models and transformer language models.

Education

07/2021 **Ph.D. in Linguistics with Cognitive Science Specialization**, *Michigan State University*
Dissertation: *Competition in natural language meaning - The case of adjectival constructions*
Aim: Explaining cognitive constraints/universals in human **dialogue pragmatics**

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

Skills

Research and Analysis

Natural/controlled experiment design and the appropriate statistical tools, Quantitative methods (e.g., web-based survey, crowd-sourcing), Qualitative methods (e.g., **conversation** analysis), A/B testing

Technical

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 **bash** scripting, JavaScript, HTML, CSS, SQL
- **Links to sample scripts NLP:** `https://github.com/yancong222/scripts`
- **Links to sample scripts ML:** `https://github.com/yancong222/scriptscz`

2018-present Software development and implementation
- Development Environments: Visual Studio; RStudio; Anaconda
- Productivity Applications: Git/GitHub
- Cloud Service: Google Cloud Platform (GCP), Microsoft Azure
- Acoustics software: OpenSmile; Montreal Forced Aligner (MFA in **kaldi**); Audacity
- Psychology software: PsychoPy (Visual Paradigm); E-Prime

Experience

Computational linguistics 07/2021-present

Postdoctoral research trainee, Feinstein Institutes for Medical Research, Northwell Health: **Pre-processing real world datasets; use NLP methods to identify speech biomarkers**
- Develop NLP pipelines and scalable classifiers, leveraging CoreNLP (Semgrex, Dependency parse), Penn Discourse Treebank, WordNet-3.0
- Deploy pre-trained transformer language models (GPT-3, RoBERTa, BERT, T5-11b) to identify speech biomarkers. Deep learning toolkit: **PyTorch**
- Large-scale dataset processing. Metrics include similarity, relative probability. Methods include tf-idf, word2vec, next sentence probability, centroid. **Feature extraction**.

Transformer Language Models 09/2020-07/2021

Graduate student researcher, Department of Linguistics and Languages, MSU: **Perform data and error analysis in order to improve transformer language models and understand their shortcomings**

- 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: **Speech perception in noise; confusion matrix; sonority scales**

- 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

Psychology 08/2016-07/2021

Data analyst *Psycholinguistics Lab*, Department of Linguistics and Languages, MSU: **Using behavioral measures and neurology equipment to understand human language**

- Led coordination of stimuli design and auditory/visual paradigm design (4 team members)
- Analyzed real world speech/text dataset, resulting in 2 conference papers

Language data corpus 04/2015-07/2016

Project assistant *Joint Research Center on Chinese Linguistics*, Hong Kong Polytechnic University - Peking University: **Annotation; messy data; linguistic data consortium**

- 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 Chineses Variations and Chinese Linguistic KnowledgeNet

Selected publications

To appear **Cong, Yan** and Wolff, Phillip. *Proceedings of the 2022 Linguistics Society of America (LSA) Annual Meeting Conference*. Inferring markedness from semantic weight: An approach using the **T5 language model**

To appear Hansel, Katrin, **Cong, Yan**, Nikzad, Amir, Cho, Sunghye, Berretta, Sarah, Behbehani, Leily, and Tang, Sunny. *Special Issue of Schizophrenia Bulletin* (July 2022), **Latent Factors** of Speech and Language Disturbance and Relationships to **Acoustic and Lexical Computational Features**.

2021 Pandia, Lalchand, **Cong, Yan** and Ettinger, Allyson. *Proceedings of the 2021 SIGNLL Conference on Computational Natural Language Learning*. **Pragmatic competence of pre-trained language models through the lens of discourse connectives**. `https://arxiv.org/pdf/2109.12951.pdf`

Natural Language

Chinese (native), English (near-native), Cantonese (conversational), Korean (Limited working proficiency), Kiswahili (elementary level), Fijian (4 months of classroom learning), Turkish (2 months of fieldwork)