Yan Cong

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

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: Identifying and explaining constraints and universals in natural language semantics

02/2015 **M.A. in Language Studies, with merit**, *Hong Kong Baptist University*
Thesis: *The second language acquisition of the Mandarin potential complement construction*
Aim: Building statistical models to measure language competence and performance

Experience

Computational linguistics
07/2021-present

Postdoctoral research trainee, Feinstein Institutes for Medical Research, Northwell Health: **Use Natural Language Processing to identify speech and language biomarkers**

- ○ Deploying and fine-tuning transformer language models (GPT-3, BERT, T5-11b) to identify language disorganization. Deep learning toolkit: PyTorch
- ○ Developing NLP pipelines and scalable classifiers. Toolkit: Stanford CoreNLP (Semgrex, Dependency parse), Penn Discourse Treebank, WordNet-3.0, spaCy
- ○ Processing Large-scale text and speech dataset. Delivering on 5+ cross-team projects. Toolkit: NumPy, pandas. Resulting in 3 journal articles
- ○ Leveraging static models (word2vec, GloVe, LSA). Clustering, regression, and dimensionality reduction. Machine learning toolkit: scikit-learn

Transformer language models
09/2020-08/2021

Graduate student researcher, Department of Linguistics and Languages, MSU: **Perform error analysis, increase transformer language models' transparency, understand their shortcomings, incorporate language data in multiple languages**

- ○ Deployed cloud virtual machine instances to conduct task-based inference, analyzed neural language models' functionalities, and designed assessment algorithms
- ○ Developed Python and R programs to evaluate models' performance, resulting in 1 paper

Semantics and syntax
08/2016-08/2021

Data analyst / Graduate student researcher *Semantics & Syntax Lab*, Department of Linguistics and Languages, MSU: **Experiments in Linguistic Meaning and Structure**

- ○ Developed 5+ web-based acceptability surveys performed in Amazon Mechanical Turk
- ○ Led coordination of stimuli design and paradigm design
- ○ Provided coordination of data collection efforts (Electroencephalography (EEG) measurement for 50+ participants), resulting in 3 conference presentations
- ○ Trained and monitored 3 junior lab members in testing procedures

Speech
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
- ○ Modeled correlation of rhythm variation and speech perception in noise using R (packages: *lattice*, *ggplot*, *dplyr*, *tidyverse*, *igraph*, etc.), resulting in 1 manuscript

Language annotation
04/2015-07/2016

Project assistant *Joint Research Center on Chinese Linguistics*, Hong Kong Polytechnic University: **Corpus linguistics, ontology, multiple languages**

- Annotated 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

Technical 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 **bash** scripting, JavaScript, HTML, CSS, SQL

2017-present

Sample scripts

- Using Transformer Language Models (LMs) to detect language and speech disturbances in mental disorders `https://github.com/yancong222/SSD-LM-STanglab`
- Natural Language Processing (NLP): `https://github.com/yancong222/scripts`
- Others & Miscellaneous: `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). Basic familiarity with Microsoft Azure, Amazon AWS
- Acoustics software: OpenSmile (INTERSPEECH), Montreal Forced Aligner (MFA in *kaldi*), Audacity, Praat
- Psychology software: PsychoPy (Visual Paradigm), E-Prime

Publications

Commonsense reasoning

Cong, Yan. (2022) *Association for Computational Linguistics Workshop on Commonsense Representation and Reasoning*. Psycholinguistic Diagnosis of Language Models' Commonsense Reasoning. `https://csrr-workshop.github.io/`

Natural language semantics

Pandia, Lalchand; **Cong, Yan** and Ettinger, Allyson. (2021) *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*. Pragmatic competence of pre-trained language models through the lens of discourse connectives. `https://aclanthology.org/2021.conll-1.29/`

Transformer language models

Cong, Yan and Wolff, Phillip. (2022) *Annual Meeting of The Linguistic Society of America (LSA)*. Inferring markedness from semantic weight: An approach using the T5 language model. Washington, D.C.

Syntax of language that I do not speak myself

Cong, Yan and Ngonyani, Deogratias. *Descriptive and theoretical approaches to African linguistics: Selected papers from the 49th Annual Conference on African Linguistics*. Stative and Passive. Berlin: Language Science Press. `https://langsci-press.org/catalog/book/306`