

# Siyao (Logan) Peng

☎ +49-160-143-3593 | ✉ sp1184@georgetown.edu | 🏠 logan-siyao-peng.github.io | 🌐 logan-siyao-peng | 🎓 Siyao Peng

## Summary

- **Position:** Computational Linguist or NLP Specialist; **Start Time:** January 2023; **Location:** Munich, Germany
- Logan has a doctorate background in corpus and computational linguistics. He is enthusiastic about conducting language annotations and corpus analyses and employing linguistic insights to improve deep learning models.

## Education

### Georgetown University

Washington, D.C., USA

PH.D. CANDIDATE IN COMPUTATIONAL LINGUISTICS

Aug 2017 – May 2023

- Master of Science in Computational Linguistics conferred in May 2020
- Dissertation: Macro Structures in RST Corpora and Parsing for English and Chinese
- Advisors: Prof. Amir Zeldes & Prof. Nathan Schneider; Committee Member: Prof. Nianwen Xue
- Expected Defense Date: December 1, 2022; Degree Date: Early 2023; Diploma Date: May 2023

### SUNY - Stony Brook University

Stony Brook, NY, USA

PH.D. STUDENT IN LINGUISTICS

Aug 2016 – May 2017

### Leiden University

Leiden, the Netherlands

MASTER OF ARTS IN DIVERSITY LINGUISTICS

Aug 2015 – Aug 2016

### University of California – Berkeley

Berkeley, CA, USA

BACHELOR OF ARTS IN APPLIED MATHEMATICS, LINGUISTICS & FRENCH

Aug 2011 – May 2015

## Research

### Discourse Analysis

ADVISOR: DR. AMIR ZELDES

- Created the largest Chinese Rhetorical Structure Theory (RST) corpus with 50 annotated documents (total 60K+ tokens) from 5 genres which alleviated the lack of RST data and enabled training RST parsers in Mandarin Chinese and crosslingually.
- Improved RST parsing by joint training on multilingual datasets with additional pre-training and automatic EDU-wise translation.
- Conducted correlation studies between RST and paragraph structures for English and Chinese to strengthen the linguistic understanding of macro-level structures and further facilitate discourse parsing and downstream natural language understanding tasks.
- Computed disagreements between double annotations on two corpora to enhance the evaluation metrics for RST parsers.
- Evaluated the correlation of implicit relations in paralleled discourse corpora to parse and convert between discourse schemas.
- Experimented with regression models for discourse unit segmentation by engineering Universal Dependencies (UD) features to remedy the weakness of neural models on smaller datasets in an ensemble model.

### Multi-lingual & Multi-genre Semantic Annotation

ADVISOR: DR. NATHAN SCHNEIDER

- Evincing the cross-linguistic applicability of adpositional supersenses to Mandarin Chinese by annotating the Chinese translation of a parallel corpus – *The Little Prince* – and implemented quantitative and qualitative comparisons with annotated English corpora.
- Supervised semantic annotations for Reddit texts posted by L2 English speakers to analyze their L1 effects on preposition choice.

## Internship

### Baidu – PaddlePaddle Deep Learning Group

Beijing, China

NLP INTERN

Jan 2021 – June 2021

- Migrated a Pytorch text summarization model with pointer generator network and coverage mechanism into Baidu's deep learning platform – PaddlePaddle – and merged into the PaddleNLP library, filling the gap of text summarization examples.
- Acquired industry-level code standards through validating implemented models and reviewed bilingual read docs for PaddleNLP.
- Experimented with MarginRankingLoss, hard negative sampling, and few-shot learning on 9 Chinese FewCLUE benchmark tasks.

### Tencent – PCG AI Data Center

Beijing, China

APPLIED NLP INTERN

May 2020 – Sept 2020

- Ranked first (tied) in the Sentence-level Quality Estimation Shared Task on English-Chinese in the WMT 2020 conference by augmenting entities in the training data and by creating an ensemble of transformer- and XLM-based predictor-estimator models.
- Matched Kandian news to designated concepts and topics via textual similarities to enhance user portrait for recommendation.
- Outperformed Texsmart in coarse-grained entity recognition on Kandian news test set, followed by fine-grained entity classification.

- Designed an argumentation schema for a wide range of low-stakes high school and college essays based on claim versus evidence distinctions, by conducting three rounds of pilot annotations and think-aloud experiments.
- Extracted tree-ngrams and computed their KNN and tree edit distance features from automatically parsed RST trees and incorporated them into a Random Forest to predict the organization score of higher-education essays.

## Publication

- S. Peng**, Y. Liu, and A. Zeldes. GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing. In *Proc. of ACL-IJCNLP 2022*
- L. Gessler, S. Behzad, Y. Liu, **S. Peng**, Y. Zhu, and A. Zeldes. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proc. of DISRPT@EMNLP2021*
- L. Gessler, **S. Peng**, Y. Liu, Y. Zhu, S. Behzad, and A. Zeldes. Overview of AMALGUM – large silver quality annotations across English genres. In *Proc. of SCIL2021*
- M. Kranzlein, E. Manning, **S. Peng**, S. Wein, A. Arora, and N. Schneider. PASTRIE: A Corpus of Prepositions Annotated with Super-sense Tags in Reddit International English. In *Proc. of LAW@COLING2020*
- H. Wu, Z. Wang, Q. Ma, X. Wen, R. Wang, X. Wang, Y. Zhang, Z. Yao, and **S. Peng**. Tencent submission for WMT20 Quality Estimation Shared Task. In *Proc. of WMT@EMNLP2020*
- S. Peng**, Y. Liu, Y. Zhu, A. Blodgett, Y. Zhao, and N. Schneider. A Corpus of Adpositional Supersenses for Mandarin Chinese. In *Proc. of LREC2020*
- L. Gessler, **S. Peng**, Y. Liu, Y. Zhu, S. Behzad, and A. Zeldes. AMALGUM – A Free, Balanced, Multilayer English Web Corpus. In *Proc. of LREC2020*
- Y. Yu, **S. Peng**, and G. Yang. Modeling Long-Range Context for Concurrent Dialogue Acts Recognition. In *Proc. of CIKM2019*
- Y. Yu, Y. Zhu, Y. Liu, Y. Liu, **S. Peng**, M. Gong, and A. Zeldes. GumDrop at the DISRPT2019 Shared Task: A Model Stacking Approach to Discourse Unit Segmentation and Connective Detection. In *Proc. of DISRPT@NAACL2019*, pages 133–143
- Y. Zhu, Y. Liu, **S. Peng**, A. Blodgett, Y. Zhao, and N. Schneider. Adpositional Supersenses for Mandarin Chinese. In *Proc. of SCIL@LSA2019*, pages 334–337
- S. Peng** and A. Zeldes. All Roads Lead to UD: Converting Stanford and Penn Parses to English Universal Dependencies with Multi-layer Annotations. In *Proc. of LAW-MWE-CxG@COLING2018*, pages 167–177

## Teaching

FA 2021 Computational Corpus Linguistics  
 FA 2020 Introduction to Languages  
 SP 2020 Analyzing Language data with R  
 SP 2020 Statistical Machine Translation  
 SP 2019 Analyzing Language data with R  
 FA 2018 Intro: Natural Language Processing  
 SP 2017 Languages in the United States  
 FA 2016 Languages of the world

## Awards

2017-2022 Georgetown Ph.D. Assistantship Stipend  
 2018-2021 Georgetown Ling. Dept. Student Travel Grants  
 2019 Georgetown Grad. School Student Travel Grants  
 2016-2017 Stony Brook Ph.D. Assistantship Stipend

## Coursework

NLP Natural Language Processing, Corpus Linguistics, Discourse Analysis, Semantic Representation, Machine Learning, Dialogue System, Machine Translation

CS Data Structure & Algorithm, Structure and Interpretation of Computer Programs

Math Calculus, Probability, Linear Algebra, Discrete Mathematics, Differential Equations

## Skills

NLP Scikit-learn, Numpy, Pandas, StanfordNLP, Pytorch, Tensorflow, Flair, Keras, PaddlePaddle

Programs Python, R, Bash, MATLAB, SQL, Haskell

Tools Linux, Bash, Google Cloud, AWS, Git, L<sup>A</sup>T<sub>E</sub>X

Languages Mandarin Chinese (native), English (fluent), French (intermediate), German (beginner)