Research Grants – 2019- 2021

Title: Documenting word order variation in Mayan languages: A collection of Ch'ol narratives

Investigators: Jessica Coon

Duration: 2018–2019

Funding Agency: National Geographic Society, Explorers Grant

Summary:

Today there are thirty distinct Mayan languages spoken in Mexico, Guatemala, and Belize. As descendants of a common ancestor language, known as Proto-Mayan, these contemporary languages have for millennia maintained many of the distinctive grammatical characteristics of their common ancestor language, and have managed to do so in the face of centuries of contact with the dominant colonial language of the region, Spanish. However, with an increase in globalization and exposure to media of all forms, language diversity worldwide is diminishing at an alarming rate. Unless radical changes are made, more than half of the world's currently spoken languages are predicted to vanish by the end of this century. Maintaining linguistic diversity is important not only to the health and wellbeing of indigenous communities worldwide, but also to the scientific study of language. In order to fully understand the unique human capacity for language, it is critical that linguists build and test theories not just on majority languages like English, but also on under documented languages like those of the Mayan family. The present project will contribute to linguistic research and language conservation through the creation of an annotated corpus of Ch'ol oral tradition. This corpus will be made available online to researchers interested in Mayan languages and culture, as well as in a site designed to engage the public in issues surrounding language conservation. The creation of this corpus will further foster capacity building and collaborative research with native speaker linguists and trainees in Mexico, the US, and Canada.

Title: **What makes us flexible? The role of cognitive control and sensory representations in spoken word recognition**

Investigators: PI Meghan Clayards, collaborators, Shari Baum, Morgan Sonderegger & Rachel Theodore

Duration: 2020-2023

Funding Agency: **SSHRC**

Summary:
One of the most important things we do every day is understand spoken language. We effortlessly handle variability in different talkers and contexts with more flexibility than any automatic speech recognition system. However, listeners are themselves variable. In the past 10 years there has been an explosion in interest in individual differences in speech perception. However, as this field is still in its infancy, research is fragmented. Some dimensions of individual variability have been identified but we don't know what underlies them. The central goal of this project is to test proposals about how these dimensions or 'speech perception profiles' are supported by sensory abilities and cognitive control.

Title: **Cognitive predictors of second language learning success**

Investigator: **Meghan Clayards**

Duration: **2021-2026**

Funding Agency: **NSERC**

Summary:
Learning to distinguish between nonnative speech sounds can be challenging, especially when those sounds make distinctions that are not part of the native language or rely on particular acoustic-phonetic cues in a way that is not part of the native language. While there is generally learning over time, these problems can persist even after months or years of learning in an immersion environment. Exposure to natural variation in speech sounds from multiple talkers has been hypothesized to facilitate learning, however it's not clear how critical this variation is or whether the role depends on the abilities of the learner. At the individual level, large individual differences have been shown in amount of learning and degree of ultimate attainment even under controlled laboratory training paradigms. And while there are indications that initial aptitude predicts learning, it is not clear what predicts initial aptitude. Some studies have found associations between long term outcomes and cognitive factors such as nonspeech auditory processing or attention. However, it's not clear if these factors are predictive of learning or if exposure to a second language may affect cognitive skills. The longterm objective of this research program is to understand the cognitive factors affecting learning of second language speech sounds. This includes both individual trait-level factors and training-level factors.

Title: **"Question embedding and ignorance inferences",**

Investigators: **Bernhard Schwarz (PI) Rajesh Bhatt, David Oshima, Maribel Romero, Yael Sharvit, Michael Wagner (collaborators).**

Duration: 2019-23

Agency: Social Sciences and Humanities Research Council (SSHRC)

Summary:

This project aims at furthering our understanding of language's inherent logic by zeroing in on the phenomenon of question embedding, as in, "Ann knows whether Ben is invited", where "know" embeds the polar question "whether Ben is invited". Such ignorance implications form a very pervasive data pattern, in English and cross-linguistically, involving a whole class of embedding predicates and a range of types of embedded questions. This data pattern raises theoretical questions that go to the very heart of current debates about the semantics of questions and the semantics-pragmatics distinction, and present a rich source of data for the investigation of language's inherent logic.

Title: **Compositional reasoning and OOD generalization in multimodal transformer models**

Investigators: **Yash Goyal, Aishwarya Agarwal, Siva Reddy, Aaron Carouville (PI)**

Durations: **2021–2023**

Agency: **Samsung-Mila Collaboration Grant**

Summary: Recent large scale transformer based language models have demonstrated that they can learn image representations from scratch using natural language supervision and can achieve remarkable zero-shot image-classification performance. However, in spite of being pretrained on large scale image-caption datasets and having millions of parameters, it is being revealed that such multimodal models lack essential capabilities that we would expect in an AI agent, such as understanding and reasoning about the compositional nature of language and vision data, and the ability to generalize to out-of-distribution (OOD) data.

This project proposes to systematically evaluate and improve the compositional reasoning and OOD generalization abilities of one of the flagship systems – CLIP (Computational Linguistics and Psycholinguistics)

Title: **Advanced computing infrastructure for integrating machine learning and linguistics,"**

Investigators: **Tim O'Donnell (PI), Siva Reddy**

Durations: **2021**

Agency: **CFI JELF Innovation Fund**

Summary: In this project, a high-performance computing cluster will be built to support basic research on the computational systems which underlie language. These resources will support research broadly organized into two themes. In the first theme, it will support the development of universal, linguistically informed representations of language that can enhance the state of the art in natural language processing. The second research theme will apply the universal representations developed in the first theme to a number of problems in the science and engineering of language including: engineering of natural language knowledge bases queries, predicting human sentence processing times, and connecting the meaning of sentences to situations in the world.

This research has a number of practical benefits to Canada. This infrastructure will greatly increase local research capacity and thus foster the training of high-quality personnel and the transfer of technology from academia to industry through the applicant's contacts with local technology companies.

Title: **Investigating Combinations of Neural Networks and Constraint Programming for Structured Prediction**

Investigators: **Siva Reddy, Sarath Chandar, Gilles Pesant (PI)**

Durations: **September 2020 – August 2022**

Agency: **IVADO – Fundamental Research Grant**

Summary: The main research objective in this project is to improve the state of the art for structured prediction using neural networks. The use of constraint programming during the training and inference phases of neural networks (NNs) will be investigated. For this project, Natural Language Processing (NLP) is considered as domain rich in structured prediction problems, and the focus is on parsing natural language to semantic structures such as semantic role labeling (the problem of parsing text to who did what to whom) and knowledge graph query prediction (the problem of parsing language to machine-executable representations) in order for the research project can have practical impact. In particular, the focus is on texts in Finance, Health and News domains for semantic role labeling, and Encyclopedic texts for knowledge graph queries.

Title: **Access to resources on COVID-19 through a chatbot**

Investigators: **Dialogue, Mila, Nu Echo, Samasource, Google, Johns Hopkins University and Data performers**

Durations: **May 2020 – Sep 2020**

Agency: **Scale AI**

Summary: This project is to build, deploy and scale the chatbot infrastructure and interaction designers to create effective user experiences. The current chatbot's capacity is to be augmented by the expertise available at Nu Echo. Nu Echo has available capacity to significantly contribute to this project, and the track record of building production-ready conversational systems using technology in use by Dialogue (Rasa platform).

A large-scale language model (BERT) is pre-trained on COVID content and fine-tuned on question-answering datasets. The answers are encoded using the language model in dense vectors. The questions are also encoded and the answer candidates are ranked in order of similarity and the closest answer is selected. The model is trained on Q&A datasets to predict the passages that contain the answer. The resulting passage will be used as the utterance of the chatbot.

Title: **Robust conversational models for accessing the world's knowledge**

Investigators: **Siva Reddy**

Durations: **April 2020 – March 2025**

Agency: **NSERC Discovery Grant and Supplement**

Summary: This project focuses on bias in large deep learning models. Training deep learning models on large social datasets crawled from the web or derived from social media sources often leads to models with unintended social biases. These include representational biases that can lead to demographically

skewed predictions, recommendations, or stereotypical and socially unacceptable outputs. This project's goal is to identify, characterize and control bias in these models.

Title: **Open-Domain Conversational Question Answering**

Investigators: **Kaheer Suleman, Siva Reddy (PI)**

Durations: **April 2020-Mar 2021**

Agency: **MSR-Mila Grant**

Summary: In this project, next-generation search engines which have the ability to answer conversational questions are to be developed. The main challenge is building a joint inference model that understands conversations, performs information retrieval and reads documents to answer questions.

Title: **Representation Learning for Natural Language Processing**

Investigators: **Siva Reddy**

Durations: **2020-2024**

Agency: **Facebook CIFAR AI Chair, CIFAR Pan-Canadian AI Strategy**

Summary: Enabling ubiquitous machines, such as smartphones, smart home appliances, self-driving cars and robots with natural language understanding abilities opens up potential opportunities for the broader society to benefit from, e.g., in accessing the world's knowledge, or in controlling complex machines with little effort.

In this project, the focus is on the task of accessing knowledge stored in knowledge-bases and text documents in a colloquial manner.

The scientific questions addressed: 1) Are linguistically-informed models better than uninformed models? 2) How can inductive biases help machine learning? and 3) What are the challenges in enabling conversational interactions?