# Research Statement

Benjamin King – University of Michigan

Natural Language Processing (NLP) is rapidly growing field that draws from Computer Science, Linguistics, Statistics, and Cognitive Science. It has found applications across many disciplines, including business, medicine, the humanities, and science. While the field continues to experience successes like IBM's Watson or Google Translate, NLP technologies remain out-of-reach for speakers of most of the world's languages.

My research aims to close the gap between commonly spoken languages like English or Chinese, for which there exist an abundance of NLP resources and technologies, and minority languages that often lack even the most basic NLP resources and tools. Basic NLP technologies such as part-of-speech tagging and parsing can be used in the development of tools such as grammar checkers and automatic translators that can help provide access to healthcare [4], promote literacy [10], and even enable enable Web usage.

While the focus of my thesis work is on enabling these basic NLP technologies for a broad number of languages, my research has also included work in a number of other areas of NLP and across disciplines. I've been fortunate to have many opportunities for collaboration with other researchers at other top universities in areas such as multilingual scientometrics, social networks, and socio-linguistics. In Michigan's NLP group, I've worked closely with fellow students on text similarity and collective intelligence. Finally, in internships, I've completed large projects in the areas of information retrieval and health informatics. I don't consider these areas of research to be separate from my core thesis work of bringing NLP to minority languages, but rather as areas of broad interest that can be studied in many languages as my work enables basic NLP for those languages.

## Language Identification

One of the most obvious barriers to minority language NLP is that minority language text is less common and less easily obtained than text in a majority language, e.g. English. Fortunately, large quantities of minority language text can often be found on the Web by using targeted crawling techniques such as [3] and [5], but this text comes at a price: it is noisy. In addition to common sources of noise like misspellings or slang, minority language text on the Web is seldom alone. More often than not, a Web page containing minority language text also contains text in at least one other language. In order to be used in downstream tasks, minority language text would need to be identified and separated from other text on the page. Because previous work in language identification assumed that documents were monolingual, I developed the first **word-level language identification methods**. [6]

Word-level language identification allows individual words in a document to each have their own language. This is appropriate for code-switched documents (where the author is fluent in more than one language), documents with inline foreign language translations, and many other types of documents commonly encountered on the Web. The methods I've developed use weakly-supervised learning, which leads to them being highly accurate even when the set of possible languages is very large and the known vocabulary for each language is very small.

## NLP Tools for Minority Languages

A second barrier to NLP for minority languages is that supervised techniques are often not applicable due to a lack of annotated resources and the difficulty in finding qualified annotators. To address this problem, my research uses a technique called **cross-lingual projected learning**, which requires only parallel bilingual text in both a majority language and a minority language. After inducing a word alignment, which indicates which words correspond to translations of one another, this method of learning runs an existing NLP tool on the majority side and projects its output across the word alignments onto the minority side.

Whereas previous work in this area of learning has projected from a single majority language and has suffered from poor coverage due to differences between the languages, my work extends this by projecting from multiple majority languages, increasing the diversity, resulting in better coverage and accuracy. To do this, I leverage a *massively parallel corpus*, which is a text collection with the same text translated into a large number of languages. Examples of such corpora include the Universal Declaration of Human Rights and the Bible. Multi-source cross-lingual learning using a massively parallel corpus not only allows for improved accuracy over existing methods, it is also robust across many different types of languages.

## Minority Language Server

The techniques of my thesis are implemented and demonstrated in an online system called the Minority Language Server[1]. This is a Big Data system that crawls the Web and automatically learns languages from the text that it retrieves. For each language, the system starts knowing just a few words in that language. It uses its vocabulary to craft Web search engine queries that are likely to return pages containing the language of interest. Additionally, it can use any parallel text that it has to transfer parsers and taggers to minority languages. The system's knowledge for each langauge is also available for download. This will take the form of language vocabularies, language models, and language-specific treebanks.

The system is designed to be a useful resource to researchers who want to do NLP work on minority languages. I am in the process of inviting language experts into a collaboration to help curate a specific language about which they are knowledgeable, by providing a small amount of guidance to the system. I am hoping to officially launch the site within the next month.

## Related Areas

In addition to my thesis's central topics, I've also had the ability to collaborate with other researchers in a number of diverse areas. First, as a researcher for IARPA's "Foresight and Understanding in Scientific Exposition" (FUSE) and "Socio-cultural Content in Language" grants, I worked with prominent NLP faculty and students at seven other major universities. FUSE offered the oppotunity to study the problems of **scientometrics** and **bibliometrics** using Big Data. As a team, we built a system to predict future citations in a corpus of more than 50 million scientific documents. In my work on this project, I also explored the use of heterogeneous networks (networks with multiple types of nodes in the same network) as a representation for scientific social communities, finding that such a representation allows one to easily improve on the state of the art in scientometric impact prediction. [7]

The SCIL grant studied language use in **social networks**. In my time on the grant, I created a system that could analyze discussions in Web forums and determine what subgroups

[1] `http://clair.si.umich.edu/minlang`

of users existed. This worked by analyzing sentiment toward both other users and toward topics of discussion. [2]

Working in a large NLP group like Michigan's CLAIR group also allows for collaboration with fellow students. One of this lab's areas of study is a branch of collective intelligence and human computation known as **collective discourse**. This describes the phenomenon of multiple people writing independently about the same event. An interesting behavior that emerges from collective discourse are that people try to balance multiple objectives in their speech act, trying to get their message across while being as diverse as possible from others in the way that it is said. I was able to lead a research project in this area studying two unique datasets in this area of collective discourse, crossword clues and caption contest entries, and developing lexical network-based random walk models that are appropriate for multi-label clustering of this type of discourse. [8]

In two summer internships at Cengage Learning, a maker of online reference databases, I led a research project in **information retrieval** and **health informatics**. I created a search engine for medical records that participated in the Text REtrieval Conference's medical track. The system used information extraction, medical ontologies, and keywords to build models of patients based on their conditions, treatments, and vital statistics, matching these with partial templates built from queries. [9] The system was the highest performing in the conference and since this paper's publication, many similar systems have adopted this model.

## Directions for Future Work

In the long term, there are several directions I plan to take my research. One obvious direction is to apply my techniques toward higher-level NLP applications, such as machine translation, which can be directly useful to a more people. Almost all modern machine translation relies on large amounts of parallel text for training, texts that are not as widely available for minority languages. Making use, however, of linguistic knowledge and NLP tools such as parsing allows for high quality machine translation with less example text. Areas such as information retrieval and health informatics could stand to benefit enormously from minority language NLP as machine translation can help to enable access to the Web and healthcare, respectively.

My work also has natural synergies with the work of linguists, especially those studying endangered languages. The Human Language Project [1] proposes that documentary and computational linguists collaborate more closely in creating documentation in an electronically readable format. Making use of language documentation and language experts will require a shift of paradigms from the purely data-driven approach that is normally taken in NLP. Learning methods for minority language NLP should be able to as easily exploit partial annotations from experts as they are fully labeled data. I hope to develop learning methods that are appropriate for the practical realities of minority languages and are able to combine information from many diverse sources.

Finally, I plan to continue to research ways to better solve my current research problems of transfering NLP tools to minority languages by making use of Big Data from the Web and by advancing the state of the art in cross-lingual learning. NLP has seen its greatest successes in applications using supervised learning with labeled data, but it has seen less success with unsupervised learning with unlabeled data, and arguably even less with semi-supervised learning, making use of both labeled and unlabeled text. For minority language NLP to approach the quality of NLP that majority languages currently experience, we must learn to make better use of unlabeled text. I believe that this and my other research goals will best be accomplished through collaboration with other researchers, both within NLP and in other disciplines.

## References


[1] Steven Abney and Steven Bird. The human language project: building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 88–97. Association for Computational Linguistics, 2010.

[2] Amjad Abu-Jbara, Diab Mona, **King, Ben**, and Dragomir R. Radev. Identifying opinion subgroups in arabic online discussions. In *Proceedings of the 51st Annual meeting of the Association for Computational Linguistics*, volume 2, pages 249–254.

[3] Rayid Ghani, Rosie Jones, and Dunja Mladenic. Building minority language corpora by learning to generate web search queries. *Knowledge and information systems*, 7(1):56–83, 2005.

[4] Gurdeeshpal Randhawa, Mariella Ferreyra, Rukhsana Ahmed, Omar Ezzat, and Kevin Pottie. Using machine translation in clinical practice. *Canadian Family Physician*, 59(4):382–383, 2013.

[5] Kevin P Scannell. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15, 2007.

[6] **King, Ben** and Steven Abney. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119.

[7] **King, Ben**, Rahul Jha, and Dragomir R Radev. Heterogeneous networks and their applications: Scientometrics, name disambiguation, and topic modeling. *Transactions of the Association ofr Computational Linguistics*, 2:1–14, 2014.

[8] **King, Ben**, Rahul Jha, Dragomir R. Radev, and Robert Mankoff. Random walk factoid annotation for collective discourse. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 249–254.

[9] **King, Ben**, Lijun Wang, Ivan Provalov, and Jerry Zhou. Cengage learning at trec 2011 medical track. In *Proceedings of the 20th Text REtrieval Conference*.

[10] Lawrence Williams. Web-based machine translation as a tool for promoting electronic literacy and language awareness. *Foreign Language Annals*, 39(4):565–578, 2006.