UD_Gwichin-TueCL: Annotating an endangered polysynthetic Athabascan language in UD

Matthew Andrews

University of Tübingen, Germany

matthew.andrews@student.uni-tuebingen.de

Athabascan, endangered languages, polysynthetic, Universal Dependencies

Introduction

Gwich'in is an endangered language spoken in the U.S. state of Alaska and the Yukon and Northwest Territories of Canada. This project aims to annotate enough sentences in the Universal Dependencies framework to be useful to both researchers and developers interested in the study of Athabascan languages and language revitalization projects. The focus is on the Alaskan variant of the language or Western Gwich'in, also known as Dinjii Zhuh K'yaa.

Gwich'in belongs to the Athabascan language family of North America. It is one of several Athabascan languages spoken in Alaska today by a few hundred people. Language revitalization efforts are underway with Gwich'in being taught as a subject at school and university. The applicant has created an online dictionary with roughly 3,000 words to assist with the annotation project. This dictionary is now also available for students and other researchers of the language to use.

The materials used come from the Alaska Native Language Archive to which I hope to contribute in addition to the Universal Dependencies project.

Motivation

I am a student of computational linguistics who is also interested in language documentation and linguistic fieldwork. I began working on the annotation of Gwich'in (Dinjii Zhuh K'yaa) during a semester course to develop tools for low resource languages. I also began working on an online dictionary for Gwich'in in a course on lexicography simultaneously. This tool has proven useful in assisting with the annotation process.

I chose to work with Gwich'in having lived in Alaska for many years and having some familiarity with Athabascan languages. As an outsider, I am aware of the care needed to annotate Gwich'in sentences following UD guidelines while also remaining faithful to previous analyses made in conjunction with native speaker consultants.

The next stage of the project will be to complete the validation step for which I will need to register language-specific features. The training provided during UniDive will help me to confidently complete this step. In addition, the training will help me to connect with other researchers who may be interested in Athabascan or other Native American languages and who may have suggestions or ideas on how to improve the treebank for the community. The skills I learn at the training will also enable me to better assist and empower the community in future revitalization and documentation projects. I am also curious to learn about SUD and the potential suitability of this annotation scheme for Gwich'in and other Athabascan languages.

Challenges

Polysynthetic languages have complex morphology and can often express an entire English sentence with a single word. These languages are challenging to annotate in the Universal Dependencies framework which is lexicalist in nature. Different solutions to this problem have been proposed by the annotated treebanks of Abaza, Chukchi, and Yupik, among others. However, there are currently no agreed upon standards for how to best represent polysynthetic languages in the UD framework. Furthermore, there are currently no Athabascan languages with publicly available treebanks for comparison to my knowledge.

Specific challenges I have faced include annotating aspects of the Gwich'in language for which I have yet to find a comparable feature elsewhere in UD. These include so-called classifiers and classificatory verbs as well as directionals.

Of particular interest to the study of Athabascan languages are the classificatory verbs stems, which encode both the verbal action as well as the characteristics of the object. Many of these verb stems occur together with an incorporated noun to form a verb word. In addition, directionals encode three pieces of information: the direction itself, the orientation, and the distance from the speaker. I have yet to reach a satisfactory conclusion on how to best represent these, although my hunch is with the enhanced representation.

These topics would be interesting to discuss and brainstorm with colleagues about how to best represent them while remaining faithful to both the language and UD guidelines.

Current and future work

I have currently annotated roughly 300 sentences. Approximately a third of these are one word sentences taken from introductory grammar sheets to illustrate verb paradigms in Gwich'in. These are supplemented with morphological features, many of which are language specific. The remaining sentences mostly come from academic dissertations on classificatory verbs and directionals. A small subset of sentences are my own effort at translating or constructing simple Gwich'in sentences.

UD_Gwichin-TueCL is ready for validation. In addition, the applicant intends to add enhanced representations to the project in the near future. The applicant is also interested in beginning a treebank of Upper Tanana, a closely related Athabascan language, perhaps as a thesis project.

It is my hope that the Gwich'in treebank will spark interest among other Athabascanists to contribute additional Athabascan languages to the Universal Dependencies Project.

References

https://shoowadoo.github.io/Dictionary-projects/gwichin-dictionary/

https://github.com/UniversalDependencies/UD_Gwichin-TueCL