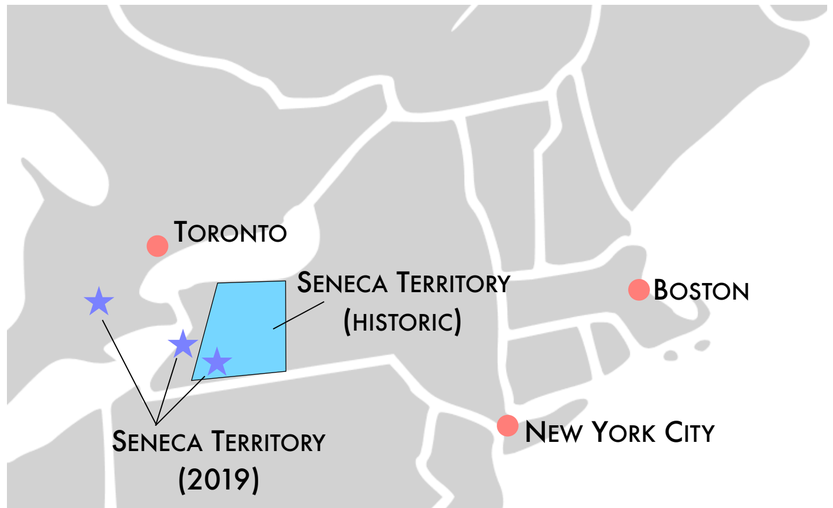
On The Promise And Pitfalls Of Repurposing Existing Language Technologies For Endangered Language Documentation

Emily Prud'hommeaux[2,4] Robbie Jimerson[1,2] Richard Hatcher[3] Raymond Ptucha[2] Karin Michelson[3]

[1]Seneca Nation of Indians, [2]Rochester Institute of Technology
[3]State University of New York at Buffalo, [4]Boston College, USA



ONÖDOWA'GA:' GAWËNÖ'

- **Seneca**: Haudenosaunee (Iroquoian) language family
- **~50** first-language speakers, **~100+** learners
- **11,000+** Seneca people living in USA and Canada
- Language spoken in **Eastern USA and Canada**
- **Polysynthetic** morphology, highly agglutinative
- **Active immersion programs** for children and adults
- Urgent need for **documentation and preservation**

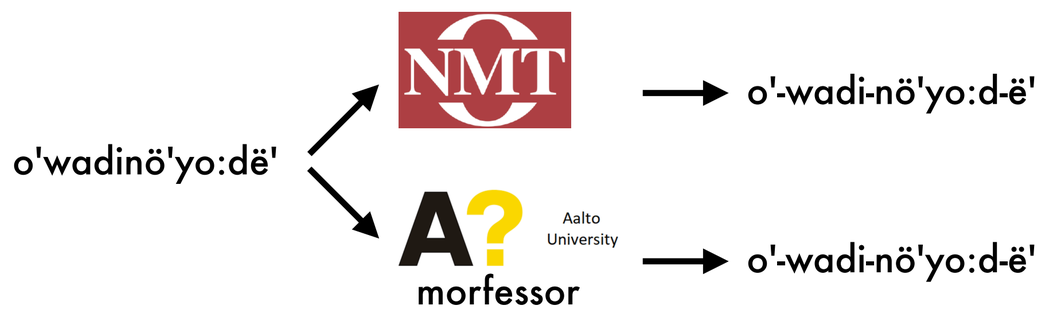
SPEECH RECOGNITION FOR TRANSCRIPTION

- Adult immersion students and instructors record elders regularly to produce instructional materials.
- Dozens of hours or recordings remain untranscribed.
- Automatic speech recognition (ASR) could aid in transcription of this data.
- Three ASR toolkits used to build models and generate transcripts of 2 hours of recordings.

- All three toolkits were challenging to install and run.
- All three resulted in poor accuracy (40-90% error) when run with default configurations.
- Even with extensive modification of training data and architectures, results were poor.
- Existing toolkits are designed for large corpora of high-quality recordings.
- In-house neural architecture for low-resource, variable quality data outperformed (20% error).

AUTOMATIC MORPHOLOGICAL PARSING

- Complex morphology and morphophonology in Seneca
- Immersion programs require accurate parses and examples of inflection in real world contexts.
- Two common packages for generating morphological parses from manually parsed examples.



- Both packages yielded weak accuracy (50-70% error rate).
- Rule-based FST required detailed prior knowledge of Seneca and much more human annotation and labeling but resulted in substantially lower error rates (10-20%).

CONCLUSIONS

- State-of-the-art methods require more data than is typically available for an indigenous language.
- Needs of indigenous language communities are very different from industry academic computing researchers.
- Tools, methods, and architectures designed specifically for this purpose and this kind of data are needed.
- Close collaboration between indigenous communities, linguists, and computer scientists is crucial.

ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 1761562.

We are grateful for the support and generosity of the Seneca elders and the Seneca community.