Speech technology and Argentinean Welsh

Elise Bell
The University of Texas at El Paso
eabell2@utep.edu

19 August, 2019

Speech language technology

Requirements

- Spoken and/or written data sets
- Sufficient data to make generalizations
- Products include ASR, TTS, parsers

Minority languages

- Lesser-resourced in general
- Strategies include crowd-sourcing, rule-based approaches
- May exclude speakers who deviate from the norm

Current speech tech resources for Welsh
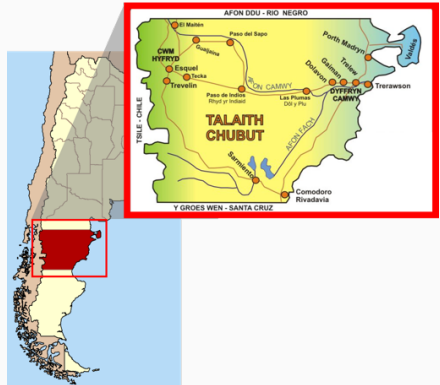
Products

- Website vocabulary plugin
- Spelling & grammar checker
- Part of speech tagger
- Digital assistant

Resources

- Trydar corpus of social media
- CEG corpus of written Welsh
- Proceedings of the National Assembly for Wales (bilingual)
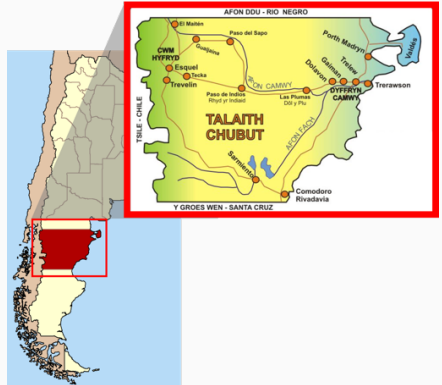- Spoken corpora (Siarad, Patagonia)

Modern Welsh in Argentina

- Revitalization of inter-generational transmission
- L2 Welsh education for all ages
- Ongoing teacher exchange program with Wales
- Spoken Welsh domains include chapel, private homes, Eisteddfodau

Modern Welsh in Argentina

- Under-resourced, under-researched
- Differs from other Welsh dialects
- All speakers bilingual with Spanish (at least)
- Many different Welsh acquisition pathways

Linguistic divergence

Many factors predict differences between Argentinean Welsh and other dialects.

- Welsh-speaking populations have had sparse contact for most of the last century
- Cross-linguistic influence from competing languages Spanish and English
- L2 acquisition effects
- Historical language divergence
- Sociolinguistic pressures and incentives

Hypothesis Welsh VOT (voiceless unaspirated vs. voiceless aspirated) affected by contact with Spanish (voiced vs. voiceless unaspirated)

Method Measurements in conversational speech from two Welsh bilingual corpora (Spanish, English)

Results Argentinean Welsh uses more Spanish-like VOT, compared to English-like VOT of control group

Previous experimental work: Bell (2018)

Hypothesis Duration and quality of Welsh vowels affected by contact with Spanish (little allophonic vowel variation)

Method Experimentally elicited vowel productions from Welsh-Spanish and Welsh-English bilinguals

Results Argentinean Welsh maintains phonemic contrasts between long and short vowels, but has greatly reduced allophonic variation in vowel duration and quality

Dialectal variation in speech technology

Previous research on speech recognition of dialect and accent differences has shown:

- Systems trained on multiple dialects perform better than those trained on a single dialect
- Including accent classification in training improved later classification of both accent-classified and accent-unclassified datasets

Challenge Lack of data

Challenge Lack of data

- Knowledge-based approaches
- More data (natural & lab produced)

Challenges and suggestions

Challenge Lack of data

- Knowledge-based approaches
- More data (natural & lab produced)

Challenge Accessing a sparse speaker community

Challenge Lack of data

- Knowledge-based approaches
- More data (natural & lab produced)

Challenge Accessing a sparse speaker community

- Outreach at community events
- Existing social networks (older speakers)
- Online recruitment (younger speakers)

Conclusion

1. Argentinean Welsh is different from other dialects of Welsh in the way speakers acoustically realize underlying phonemic contrasts

Conclusion

1. Argentinean Welsh is different from other dialects of Welsh in the way speakers acoustically realize underlying phonemic contrasts

2. Speech technologies benefit from the inclusion of individual and dialectal variation in the source data

Conclusion

1. Argentinean Welsh is different from other dialects of Welsh in the way speakers acoustically realize underlying phonemic contrasts
2. Speech technologies benefit from the inclusion of individual and dialectal variation in the source data
3. Welsh in Argentina presents a valuable resource for the continuing development of Welsh speech technology

Conclusion

Inclusion of the Argentinean Welsh community in the development of Welsh speech technology will:

- Provide greater acoustic variation for model input
- Yield models that are more accessible to non-mainstream Welsh speakers
- Strengthen ties between speakers in Wales and Argentina
- Support goals of language revitalization

- Welsh colony established in 1865
- Several thousand colonists emigrated 1865-1914
- Welsh used in public & private life
- Decline began due to Spanish education in 1890s