Navigating the Iterative Process Between Machine Learning and Linguistics Research

Madeleine Guettler & Cameron Jester

Simmons University, Boston, MA, 02115

Abstract

Combine data science methodologies with interdisciplinary research leads to improved processes and further insights. A random forest model, created to uncover key attributes in linguistic mapping achieved 86% accuracy, yet uncovered anomalous f0 data, hinting at data annotation inaccuracies. Expert validation confirmed the inaccuracies, prompting corrective measures. Despite a 1.2% accuracy drop post-correction, model credibility increased. This iterative approach is vital for leveraging machine learning to further complex hypothesis.

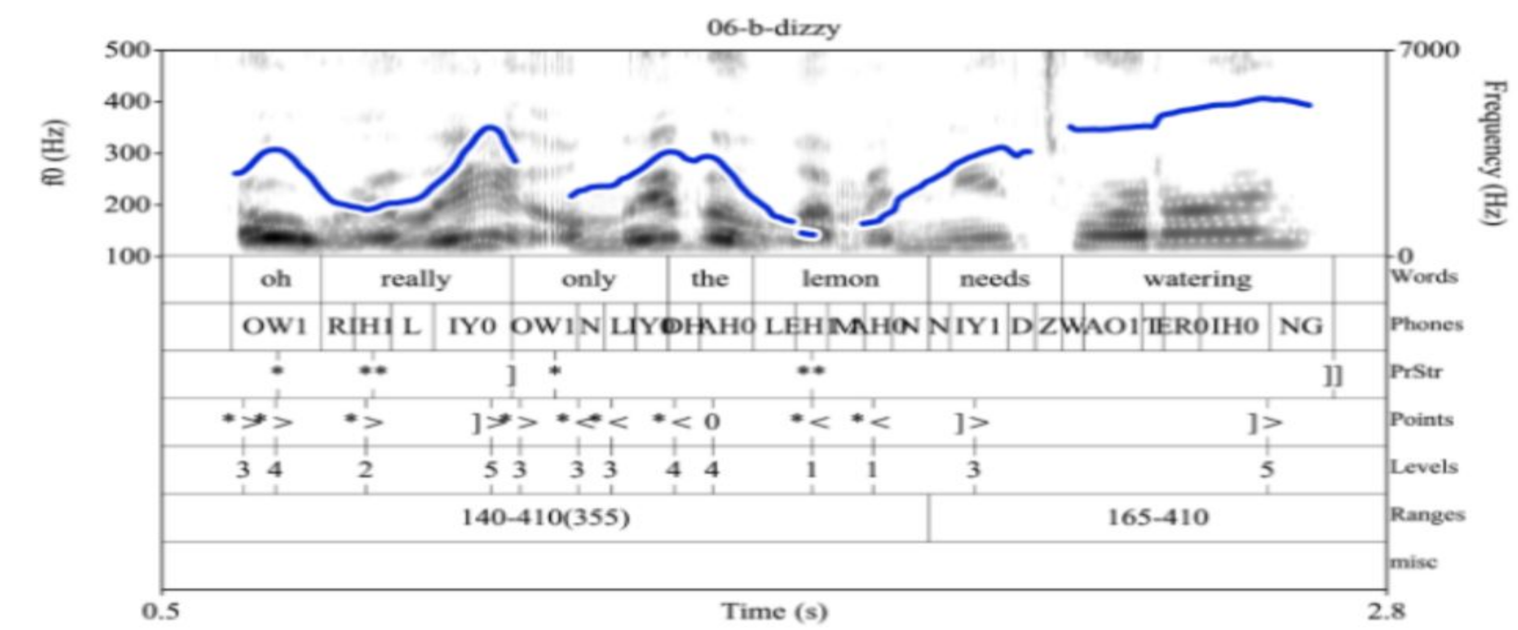
Methods

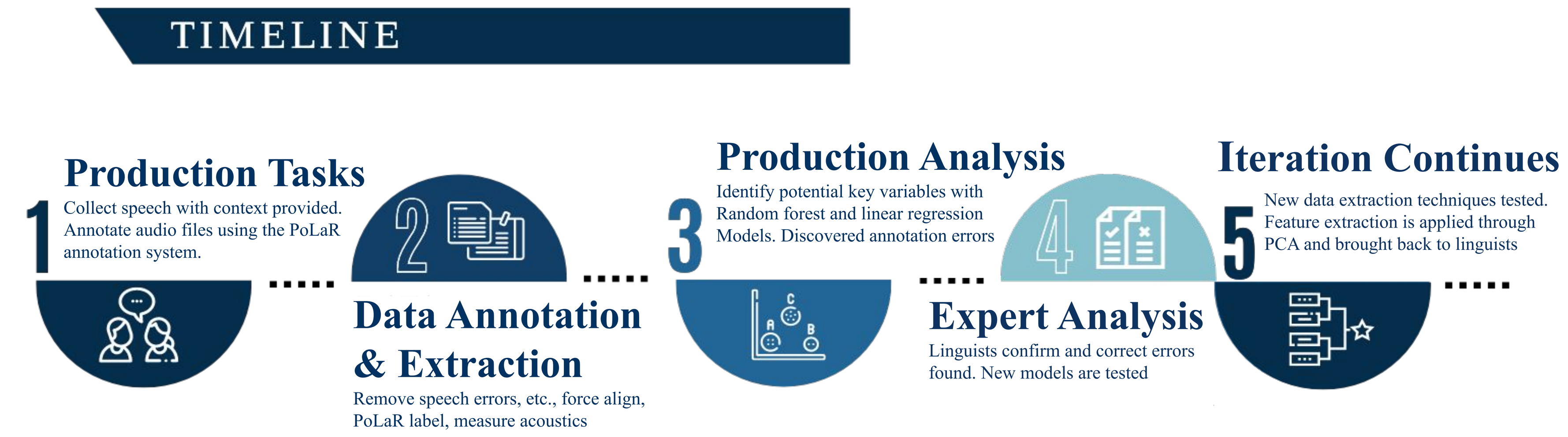
Data was from PoLaR labeled textgrids were extracted using a robust pipeline using statistical software, R. Data manipulations were implemented to properly prepare for the machine learning process. Further manipulations were coded throughout the iterative process to meet the needs of linguistic inquiry.

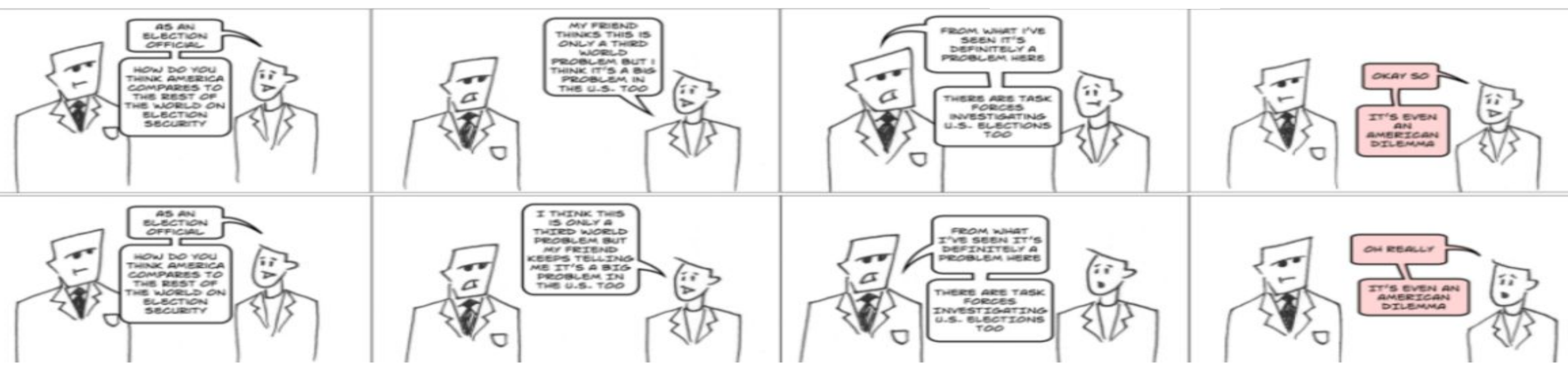
Process:

- Create hypothesis about important variables
- Munge data and apply appropriate preparatory steps to put into the following machine learning models:
  - Linear Regression
  - Random Forest
  - Principal Component Analysis
- Redeploy models with adjusted data values to account for annotation errors
- Recover more data through additional data manipulation
- Finalize machine learning models for important attributes in pragmatic meaning

Background







What is prosody?

- In spoken language it's thought to convey meaning
- It's not what you say, but *how* you say it through alterations in pitch, duration, and intensity
- Prosody maps to meaning, and here our meaning is what we call mirativity: the idea of being surprised

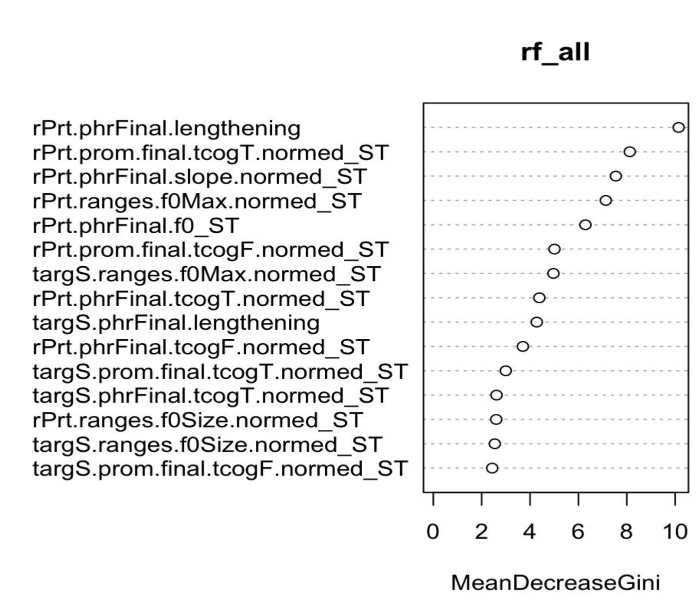
Question

- What are the best attributes to classify pragmatic meaning?

Results

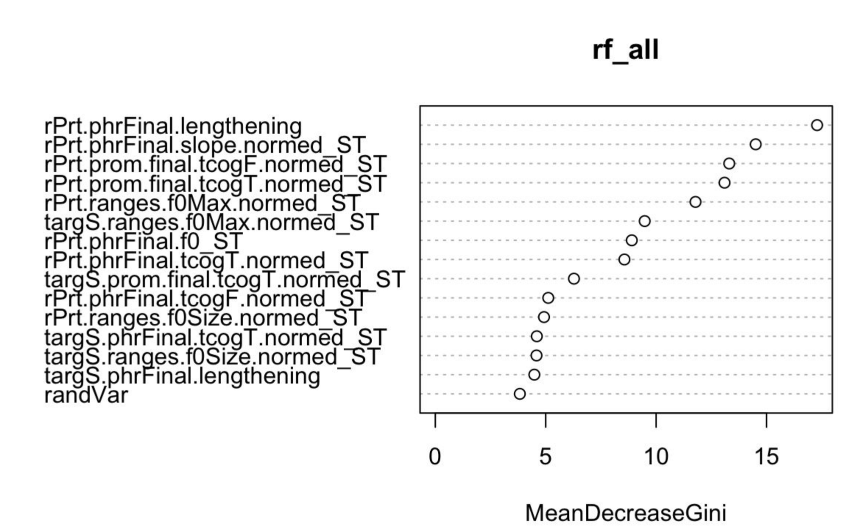
Confusion matrix for rf-all: rPrt + targS features: 84.8% accurate N = 79 (lost to na.omit)

		Predicted A	Predicted B
Actual	A	22	7
	B	5	45



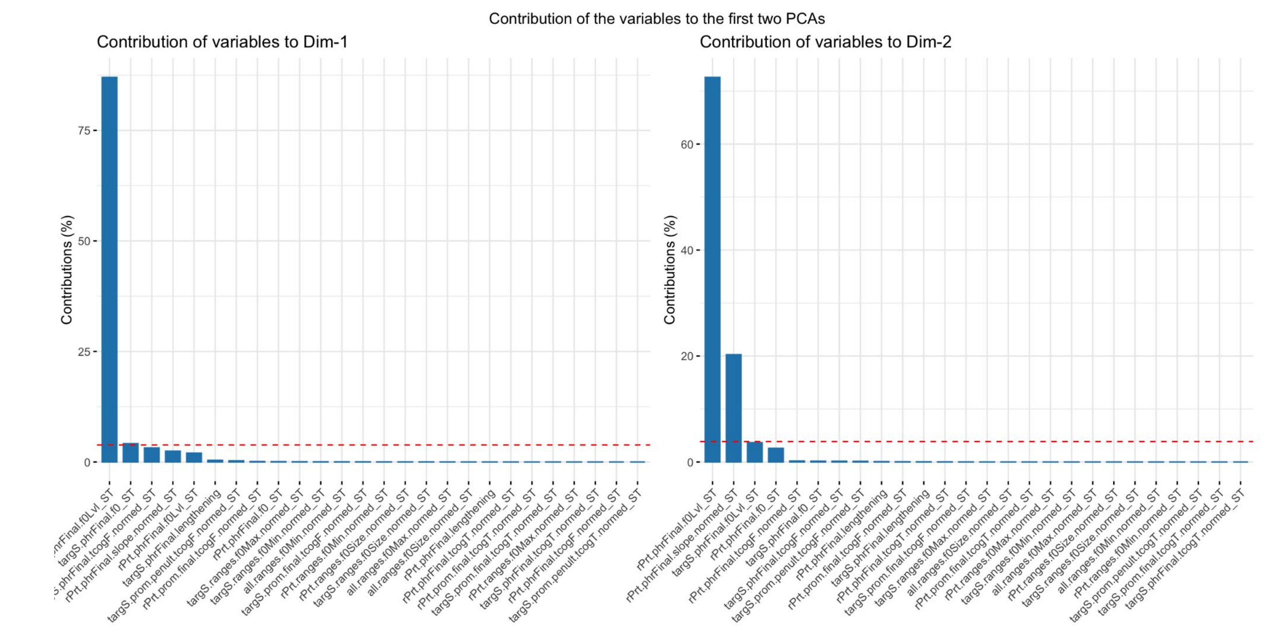
Confusion matrix for rf-all: rPrt + targS features: 89.8% accurate

		Predicted A	Predicted B
Actual	A	55	5
	B	8	60



PCA Feature Importance using first 7 dimensions: 79.7% accurate

		Predicted A	Predicted B
Actual	A	49	12
	B	13	49



- Because of the complexity of linguistic data, linguists have found it difficult to pinpoint which acoustic features help predict meaning.
- Machine learning models show that prosodic cues to meaning are distributed across the utterance and not purely restricted to any one prosodic element.
- Machine Learning can suggest features (and combinations of features) with high importance which in turn can guide linguistic inquiry
- In particular the models suggest that phrase boundary features provide cues to the two conditions: speaker's initial belief confirmed (condition A) or corrected (Condition B)

References

Selected References:

- Beckman & Hirschberg. 1994. The ToBI annotation conventions.
- Ahn et al. 2021. PoLaR Annotation Guidelines (version 1.0). Available at https://osf.io/usbx5.
- Rett & Sturman. 2021. Prosodically marked mirativity. In Proceedings of WCCFL 37.
- Barnes, Veilleux, Brugos, & Shattuck-Hufnagel. 2012. "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology." *Laboratory Phonology* **3(2)**, pp. 337-383.

Acknowledgments: This material is based upon work supported by the National Science Foundation under Grant No. 2042694, 2042702, 204274. Additionally, we would like to thank Nanette Veilleux, Byron Ahn, Alejna Brugos, and Stefanie Shattuck-Hufnagel and Simmons University for guidance and support.