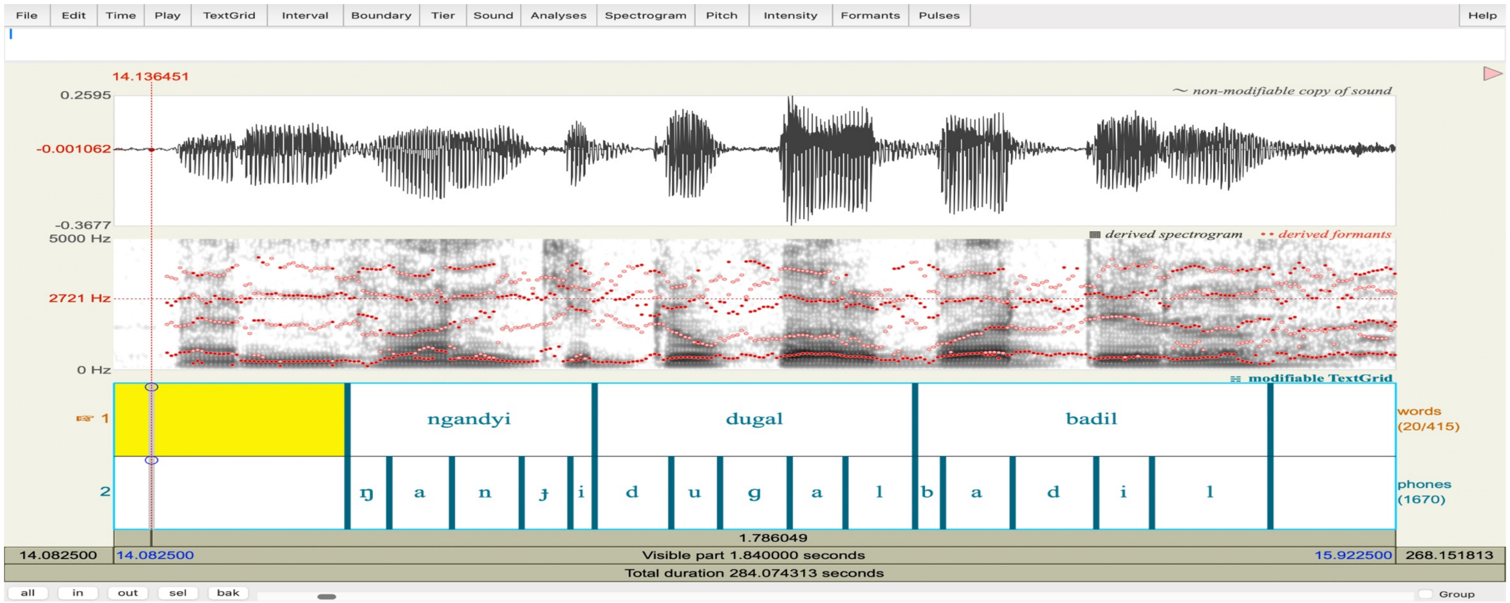


Multilingual MFA: Forced Alignment on Low-Resource Related Languages

Alessio Tosolini and Claire Bowern

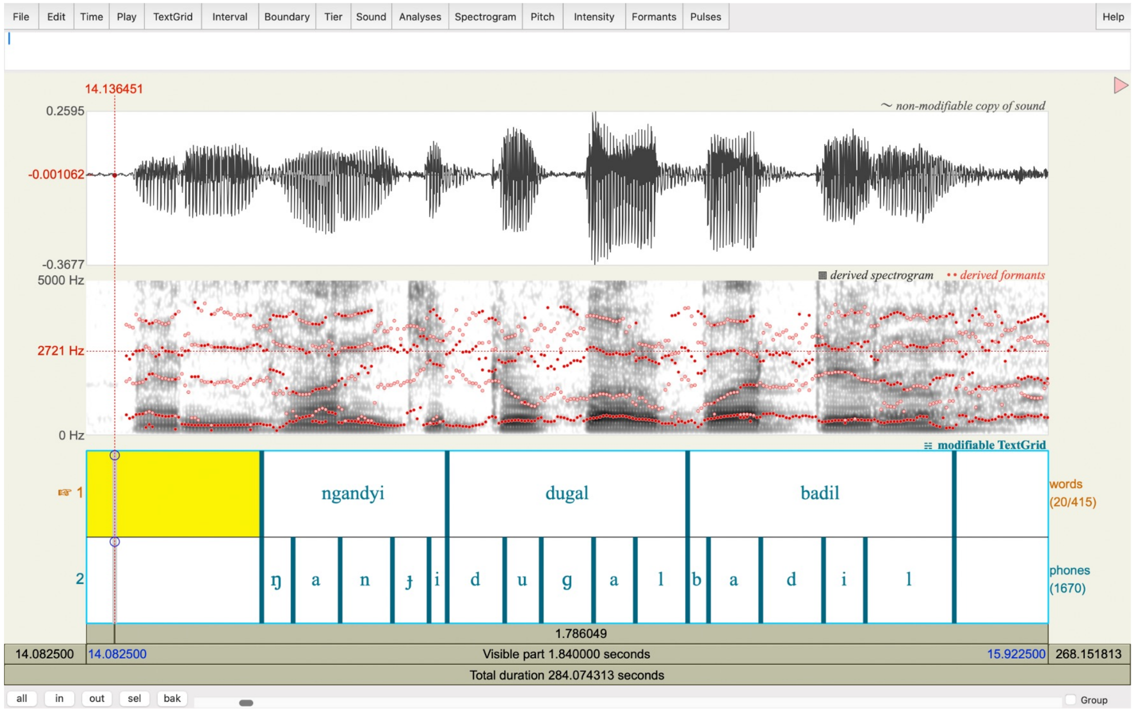
Yale University

ICLDC / ComputEl 2025

CHIRILA

Roadmap

- Introduction : underresourced languages and forced alignment
- Models and methodology
- Results
    - Accuracy of the models
    - Analyses of the models
- Main takeaways



Forced Alignment

Forced Alignment associates transcripts with audio and video (at utterance, word, or segment level).

It's incredibly useful for both linguistic research and community documentation projects.

Forced alignment requires an acoustic model and information about the grapheme to phoneme mappings (e.g. a dictionary of words and their phonemes). Acoustic model training is data hungry, and performance on languages across the world is very unequal.

McAuliffe et al 2017; Chodroff et al. 2024; DiCanio et al 2013; Babinski et al 2019

Various methods exist for increasing performance

- Use high resource model (e.g. English)
- Adapt high resource model
- Multilingual models

Here, we ask whether we can usefully combine language data from different languages from the same families (with very similar phoneme inventories) to get model improvement.

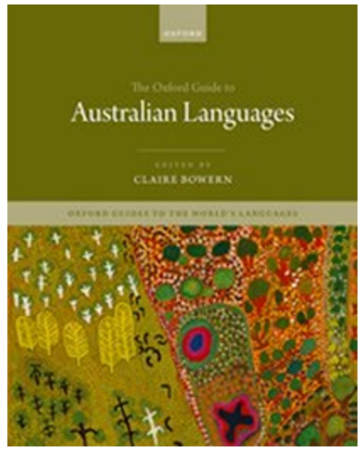
cf. San et al. 2021; Chodroff et al. 2024

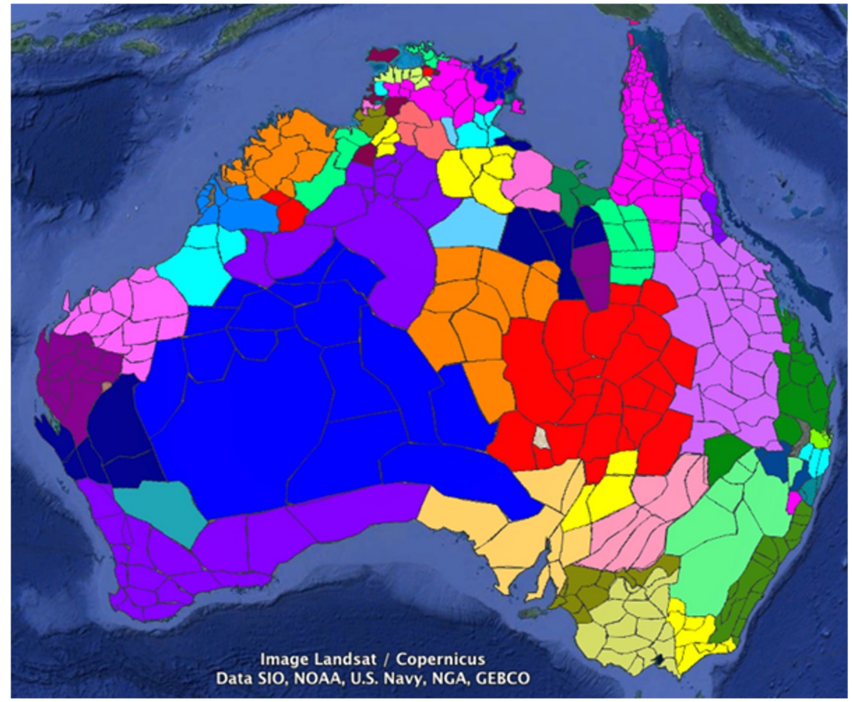
Australian Indigenous languages

c. 440 languages

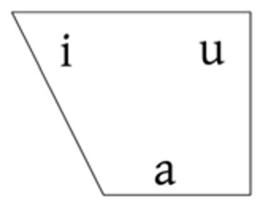
Similar phoneme inventories across the country

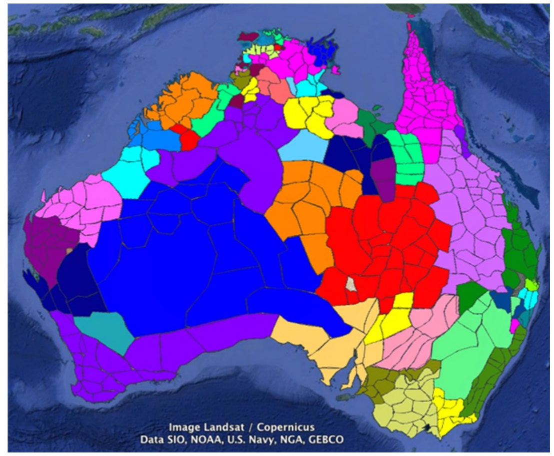
Chronically underrepresented in multilingual datasets (e.g. Common Voice)





	Labial	Lamino-dental	Apico-alveolar	Retroflex	Palatal	Velar
Nasal	m	n̪	n	ɳ	ɲ	ŋ
Stop	p	t̪	t	ʈ	c	k
Liquid		l̪	l ɹ r	ɭ	ʎ	
Glide	w		j			





Can we take advantage of multilingual datasets to train better forced aligners?

Models and methodology

Models:

- English adapted (base model trained on over 3000 hours of **global** data)
  - Yidiny
  - Big5 (Bardi, Gija, Ngaanyatjarra, Yan-nhangu, Yidiny)
  - Base model (McAuliffe and Sonderegger, 2024)
- From scratch
  - Yidiny
  - Big5 (Bardi, Gija, Ngaanyatjarra, Yan-nhangu, Yidiny)

Testing Datasets:

- Yidiny, seen data
- Yidiny, unseen data
- Kunbarlang

Models and methodology

Models:

- English adapted (base model trained on over 3000 hours of **global** data)
  - Yidiny
  - Big5 (Bardi, Gija, Ngaanyatjarra, Yan-nhangu, Yidiny)
  - Base model (McAuliffe and Sonderegger, 2024)
- From scratch
  - Yidiny
  - Big5 (Bardi, Gija, Ngaanyatjarra, Yan-nhangu, Yidiny)

Testing Datasets:

- Yidiny, seen data
- Yidiny, unseen data
- Kunbarlang

Language	Language Family	Reference	Collector	Minutes
Bardi	Nyulnyulan	A: Bowern_C05	Claire Bowern	108
Gija	Jarrakan	E: 0098MDP0190	Frances Kofod	157
Kunbarlang	Gunwinyguan	E: 0384SG0324	Isabel O'Keefe; Ruth Singer	16
Ngaanyatjarra	Pama-Nyungan	P: WDVA1	Inge Kral	53
Yan-nhangu	Pama-Nyungan	E: dk0046	Claire Bowern	290
Yidiny	Pama-Nyungan	A: A2616	R.M.W. Dixon	50

Table 1: Corpus information. A: AIATSIS; E: Elar; P: Paradisec

Evaluating an MFA model

- Method 1: Looking at how closely manual alignments match with MFA alignments
    - Can look at overall “accuracy”
    - Break it down by manner of articulation and place of articulation
        - Will English adapted models perform worse on e.g. nasals?

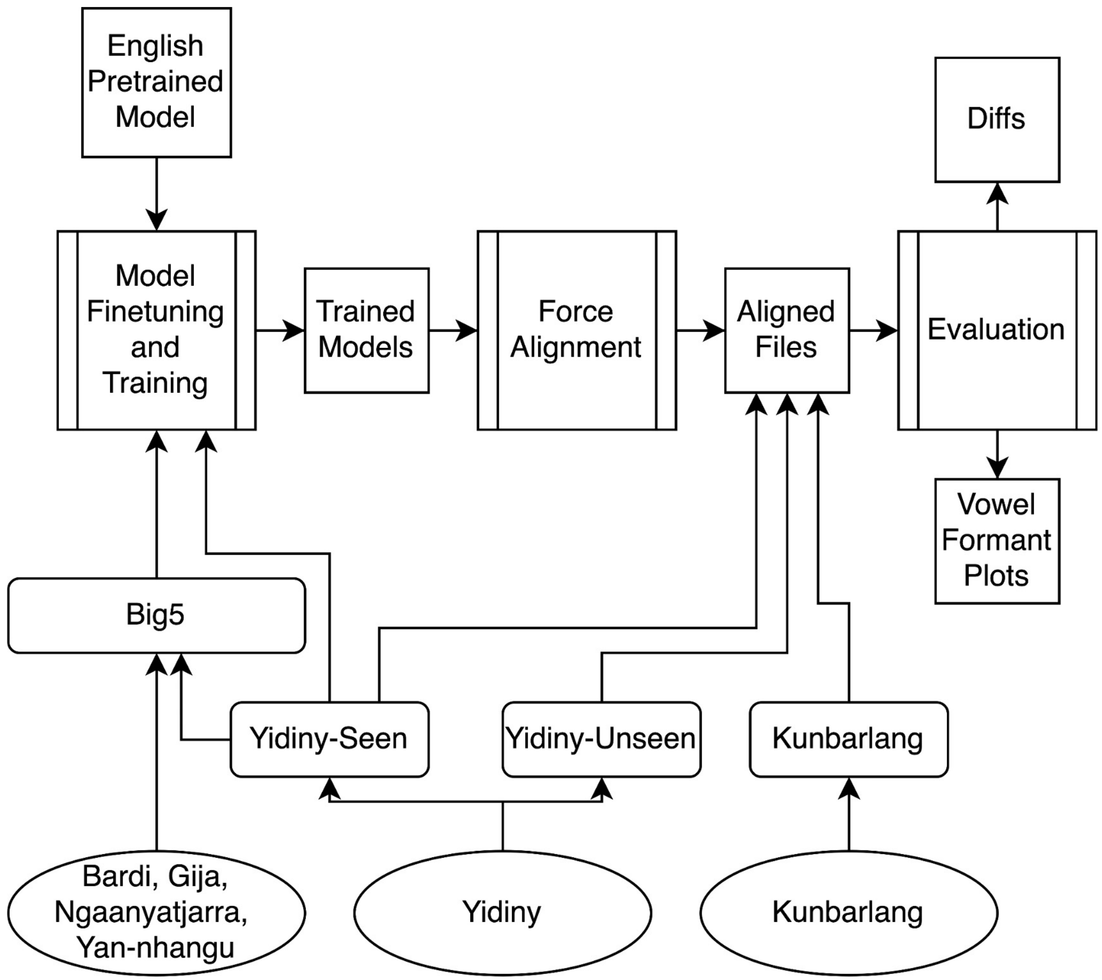
Evaluating an MFA model

- Method 1: Looking at how closely manual alignments match with MFA alignments
  - Can look at overall “accuracy”
  - Break it down by manner of articulation and place of articulation
    - Will English adapted models perform worse on e.g. nasals?
- Method 2: Looking at how closely manual *analyses* match with MFA alignments
  - We look at vowel formant plots

Evaluating an MFA model

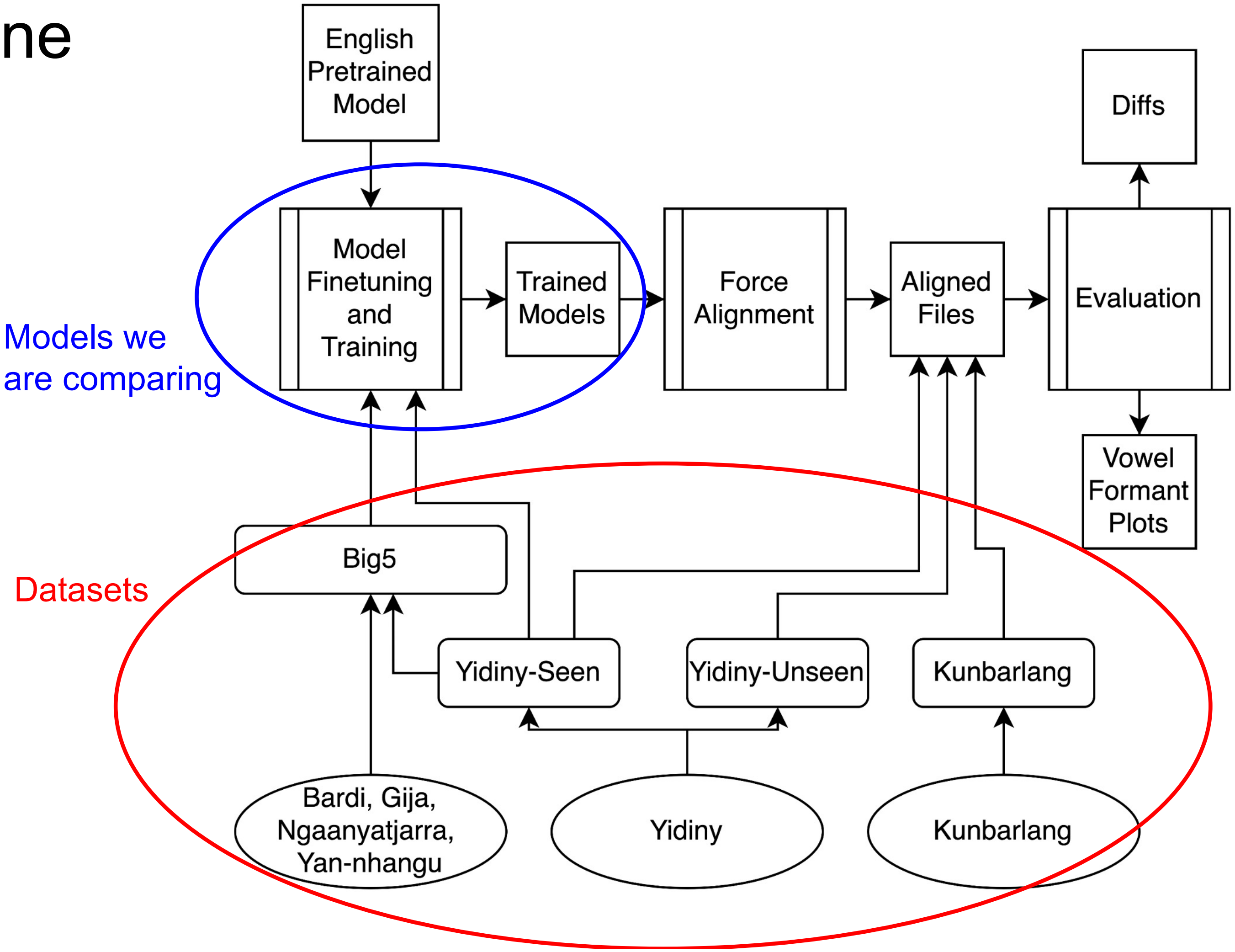
- Method 1: Looking at how closely manual alignments match with MFA alignments
  - Can look at overall “accuracy”
  - Break it down by manner of articulation and place of articulation
    - Will English adapted models perform worse on e.g. nasals?
- Method 2: Looking at how closely manual *analyses* match with MFA alignments
  - We look at vowel formant plots
- Human analyses vary! There is no one “gold standard”

Pipeline



English Pretrained Model
Model Finetuning and Training
Trained Models
Force Alignment
Aligned Files
Evaluation
Diffs
Vowel Formant Plots
Big5
Yidiny-Seen
Yidiny-Unseen
Kunbarlang
Bardi, Gija, Ngaanyatjarra, Yan-nhangu
Yidiny
Kunbarlang

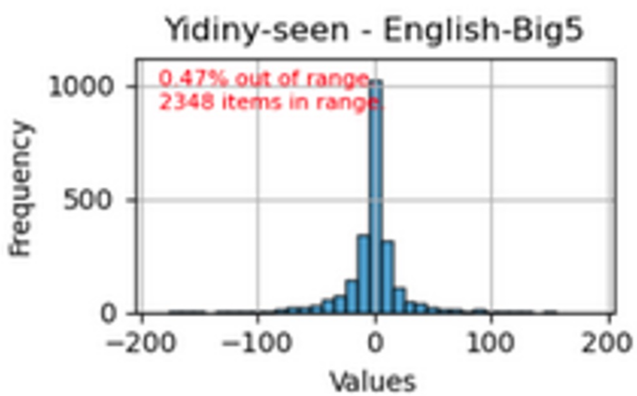
Pipeline


English Pretrained Model
Model Finetuning and Training
Trained Models
Force Alignment
Aligned Files
Evaluation
Diffs
Vowel Formant Plots
Models we are comparing
Datasets
Big5
Yidiny-Seen
Yidiny-Unseen
Kunbarlang
Bardi, Gija, Ngaanyatjarra, Yan-nhangu
Yidiny
Kunbarlang

Results: Accuracy

Things to note:

- They ideally should all have approx. **mean** 0 - which we see!
- The more spread (**sd**) a plot has, the more variation the forced aligner produces



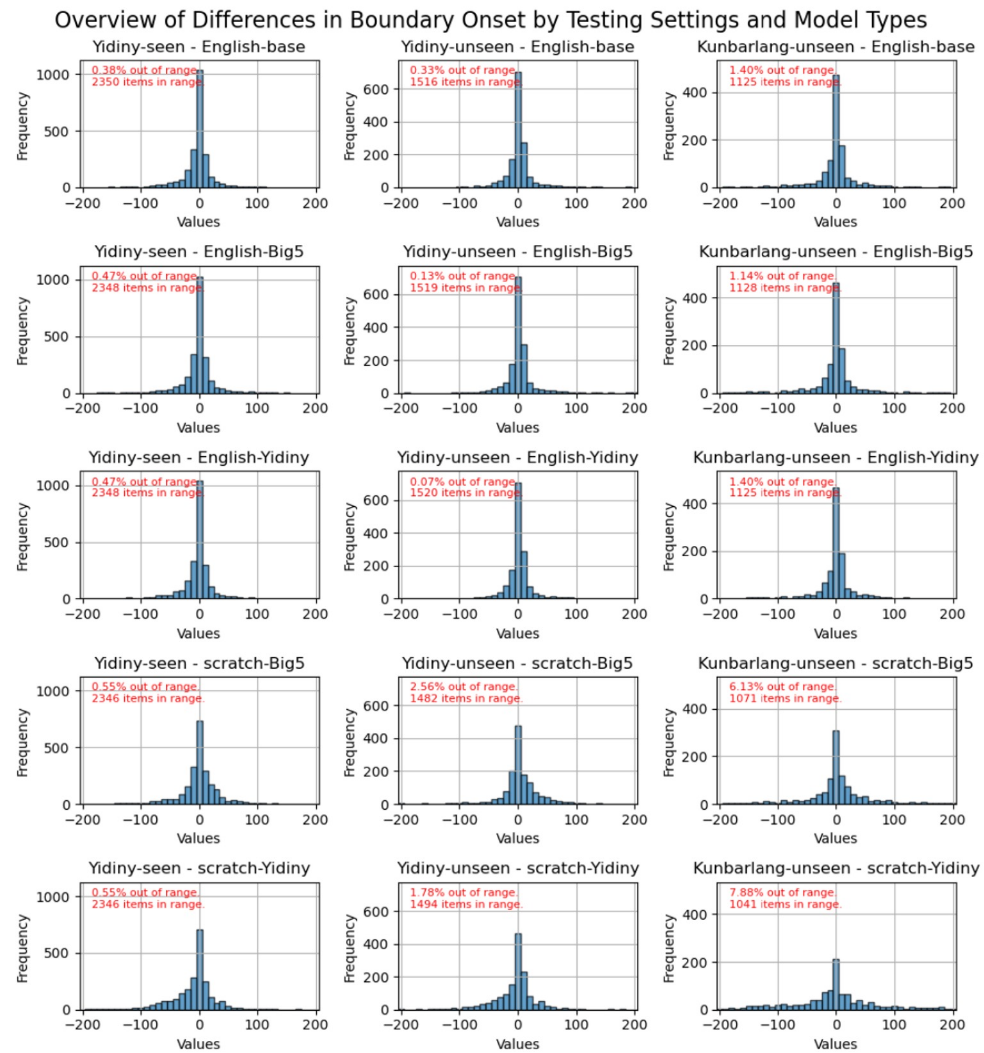


Figure 1: Onset boundary differences for all models across all testing settings.

Results: Accuracy

Things to note:

- They ideally should all have approx. **mean** 0 - which we see!
- The more spread (**sd**) a plot has, the more variation the forced aligner produces

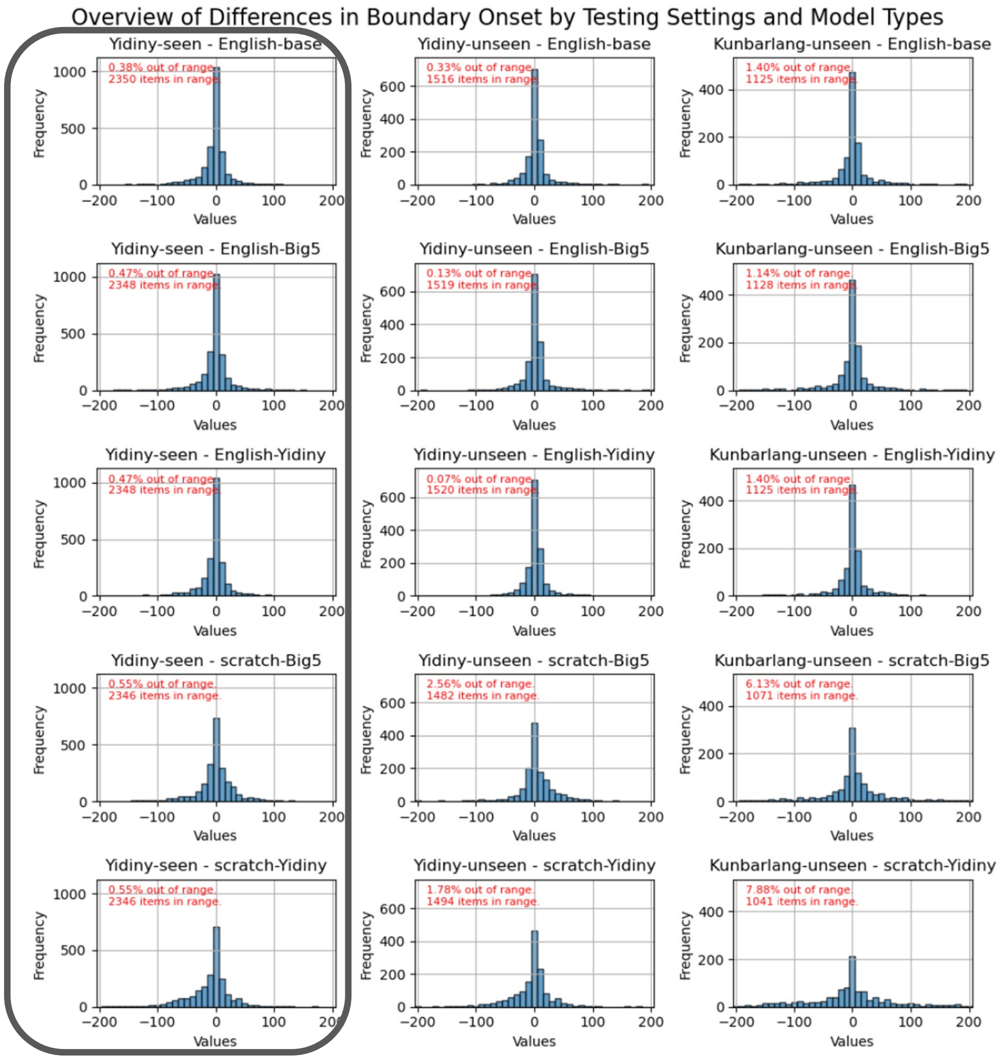


Figure 1: Onset boundary differences for all models across all testing settings.

Results: Accuracy

Things to note:

- They ideally should all have approx. **mean** 0 - which we see!
- The more spread (**sd**) a plot has, the more variation the forced aligner produces

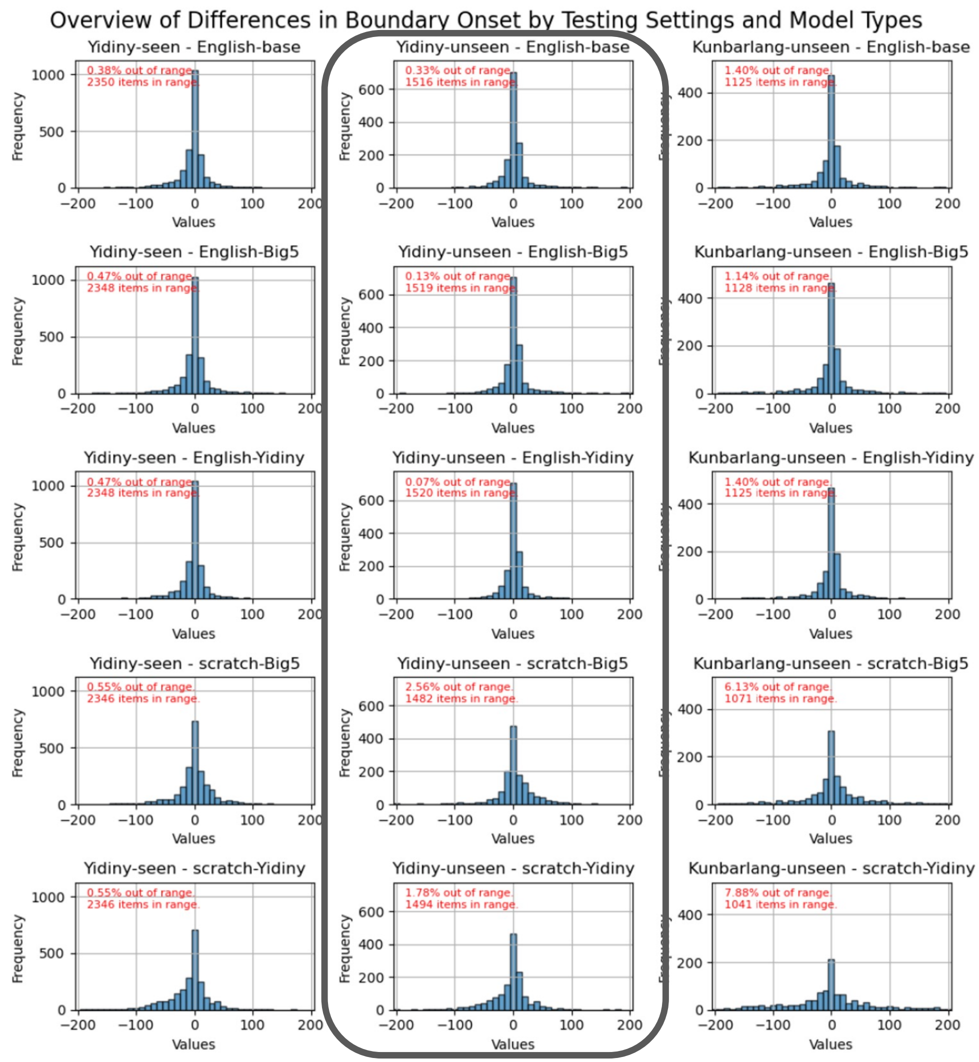


Figure 1: Onset boundary differences for all models across all testing settings.

Results: Accuracy

Things to note:

- They ideally should all have approx. **mean** 0 - which we see!
- The more spread (**sd**) a plot has, the more variation the forced aligner produces

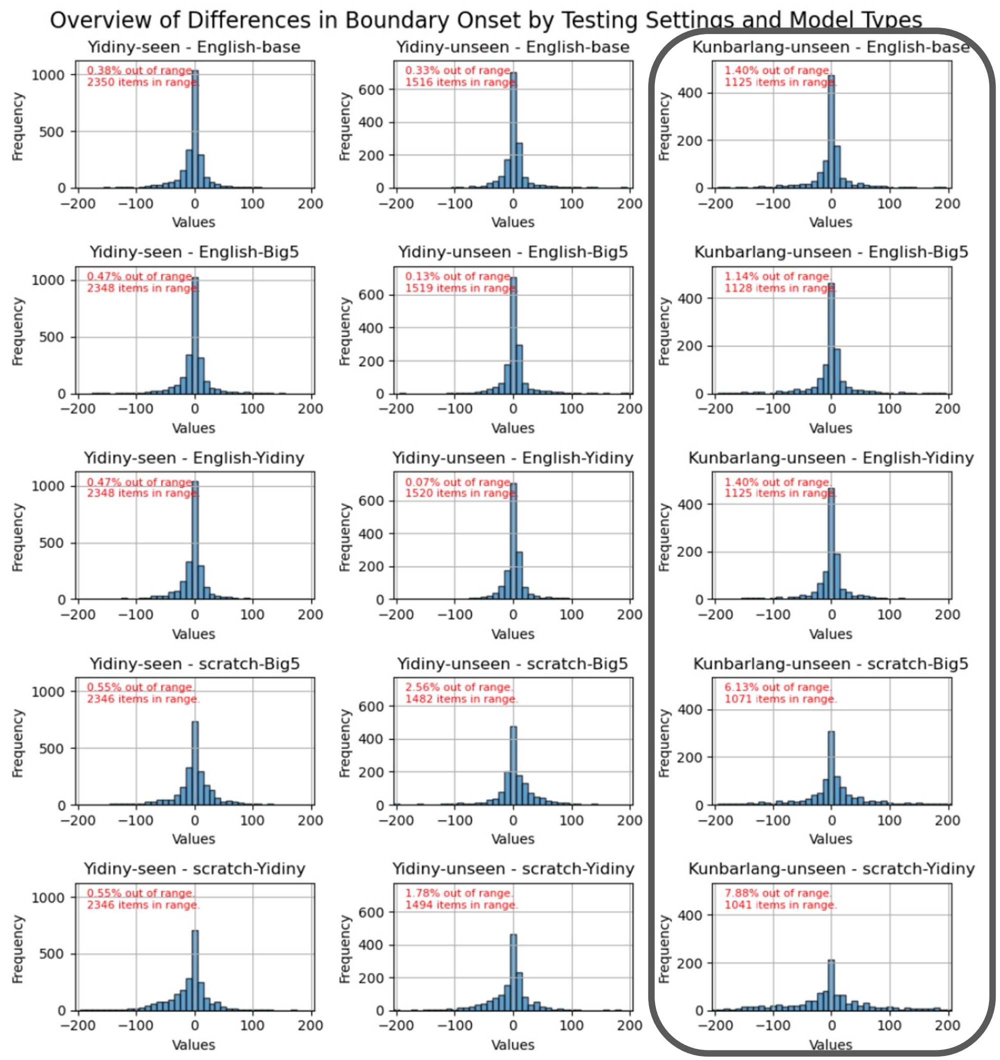
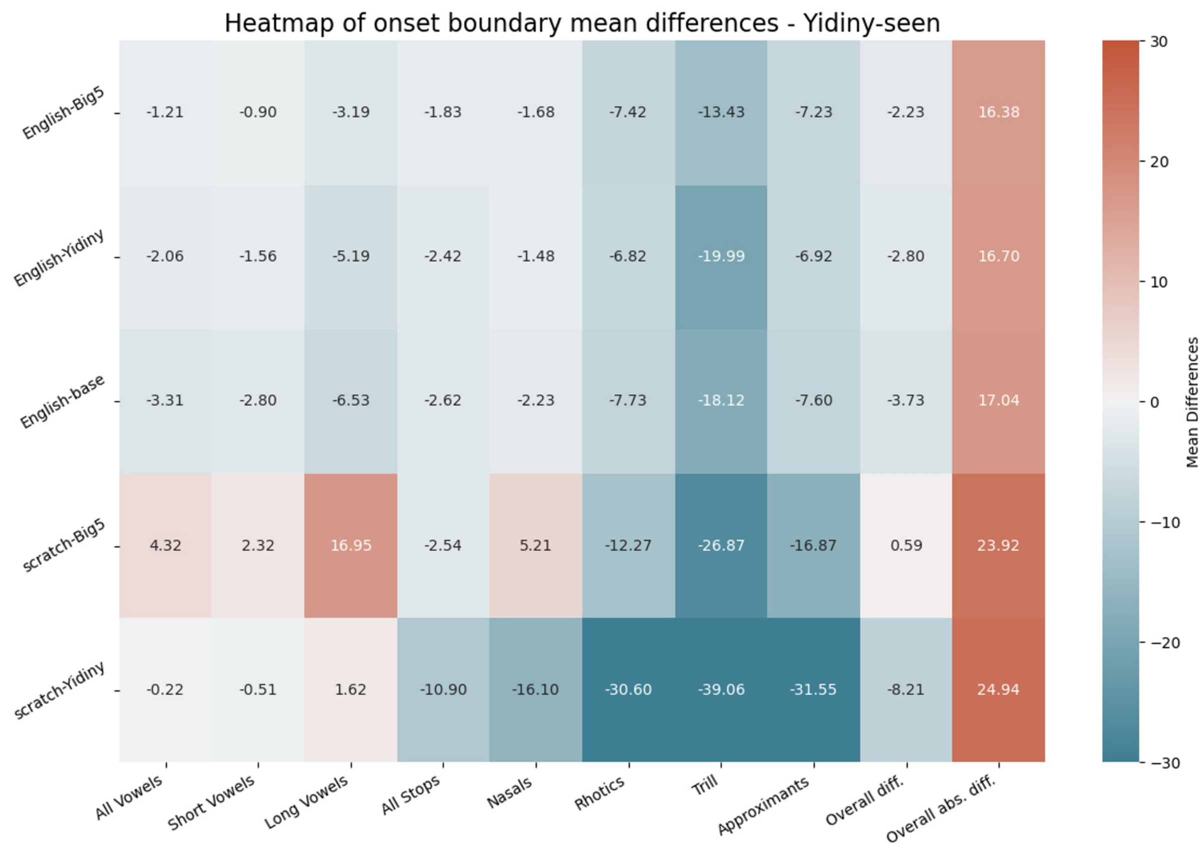
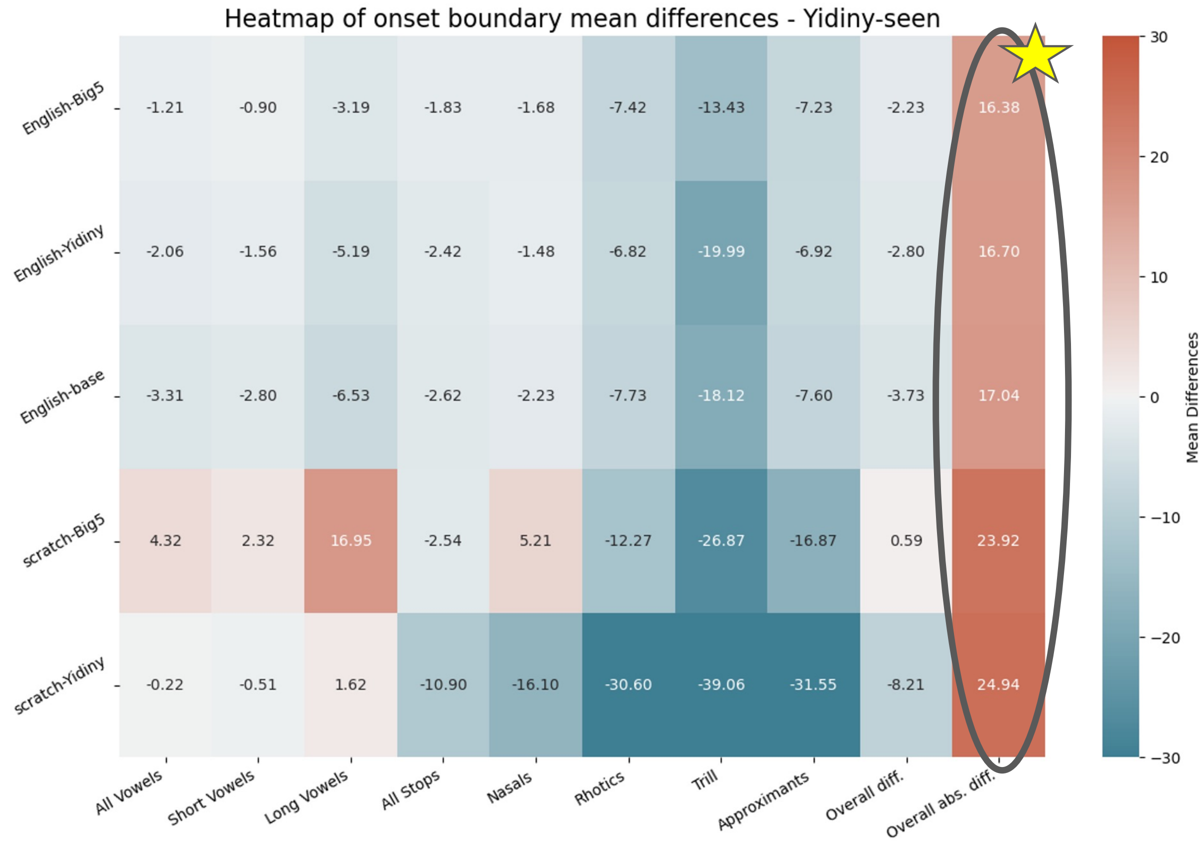


Figure 1: Onset boundary differences for all models across all testing settings.

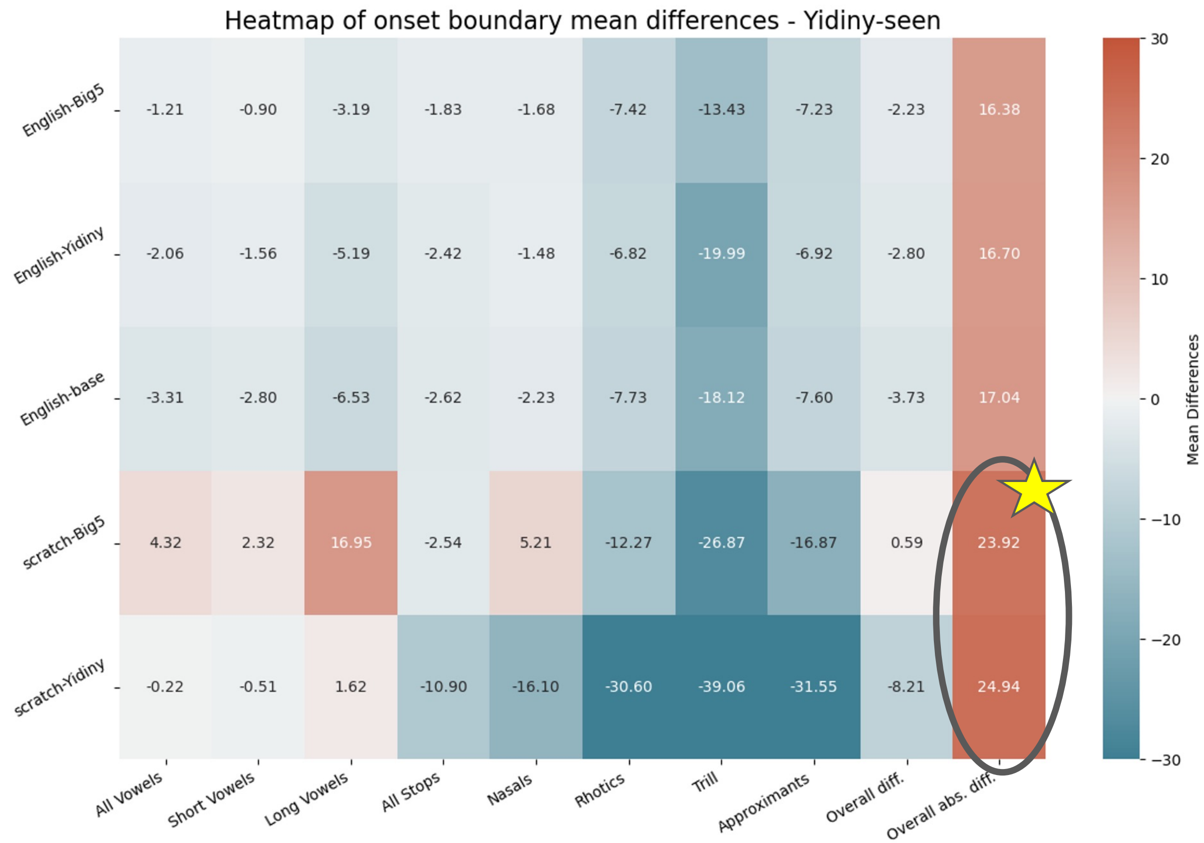
Results: Accuracy - Yidiny seen testing setting



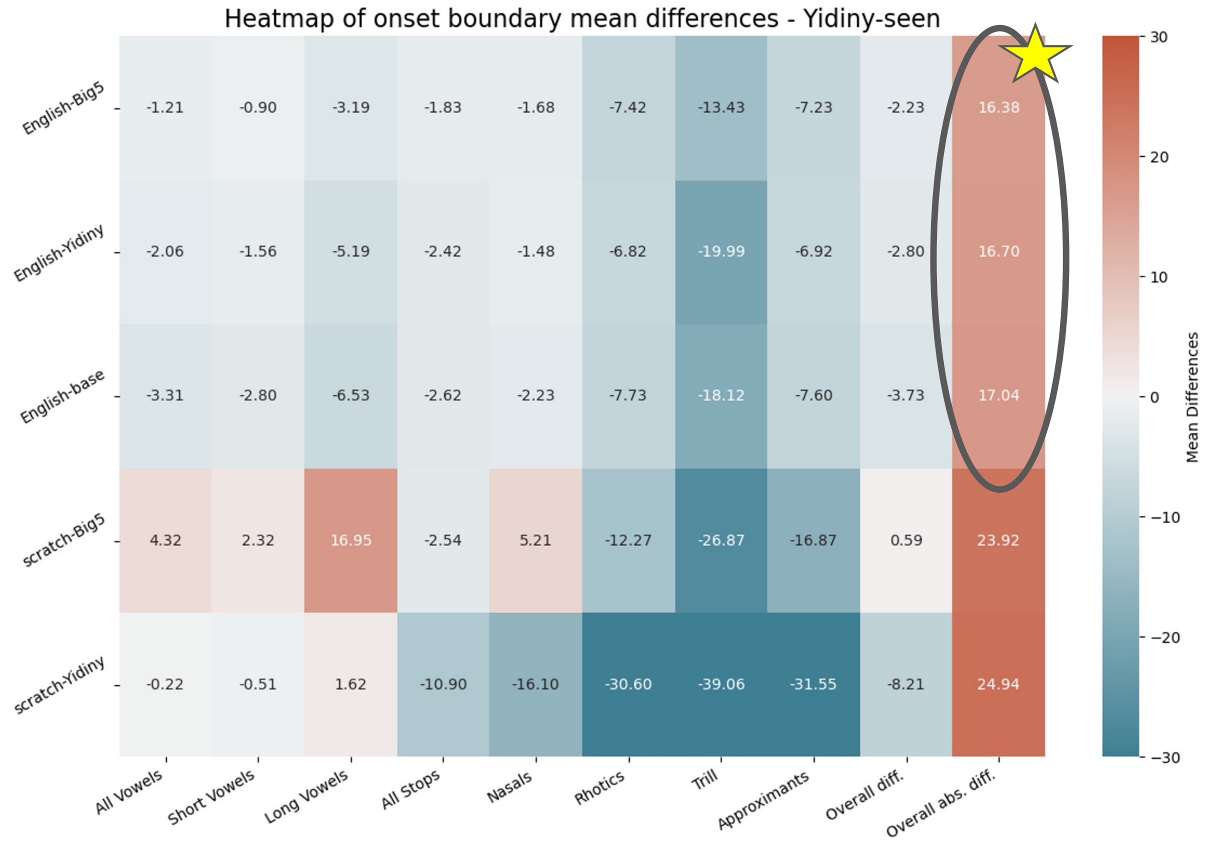
Results: Accuracy - Yidiny seen testing setting



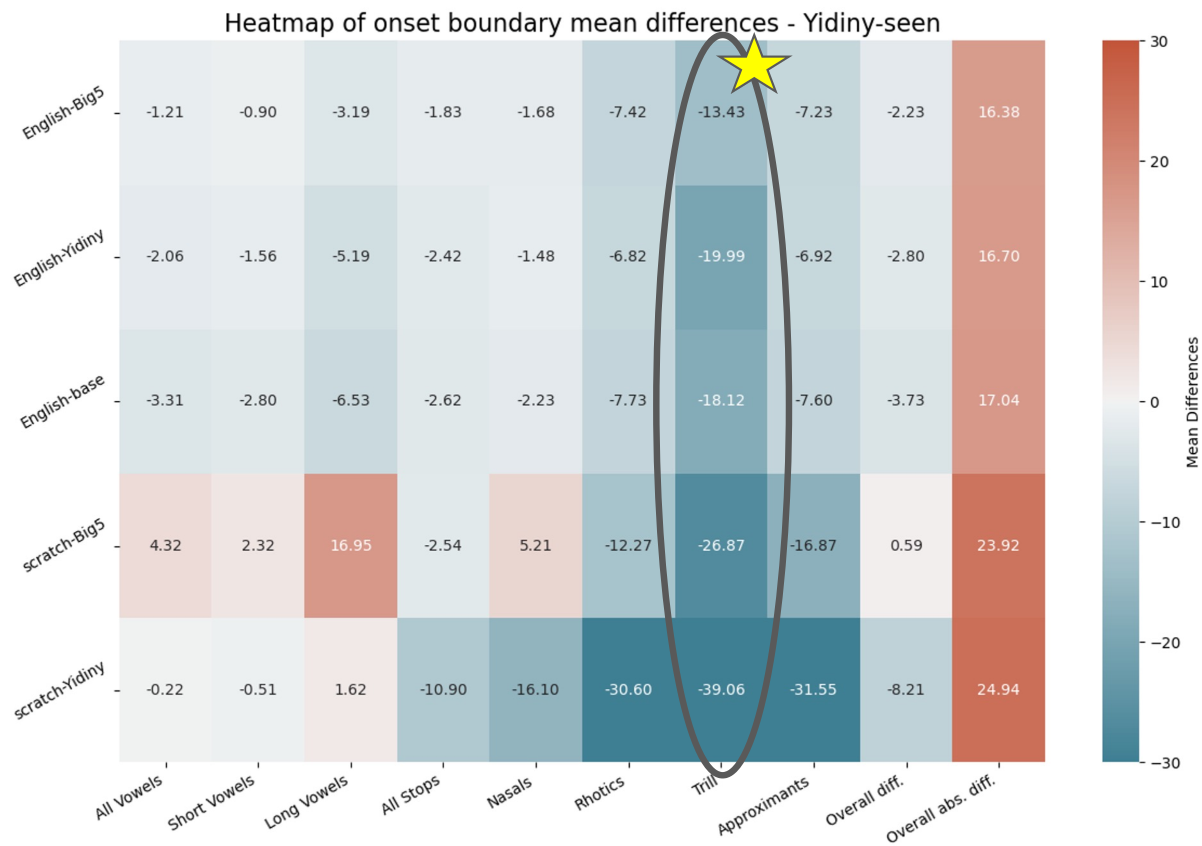
Results: Accuracy - Yidiny seen testing setting



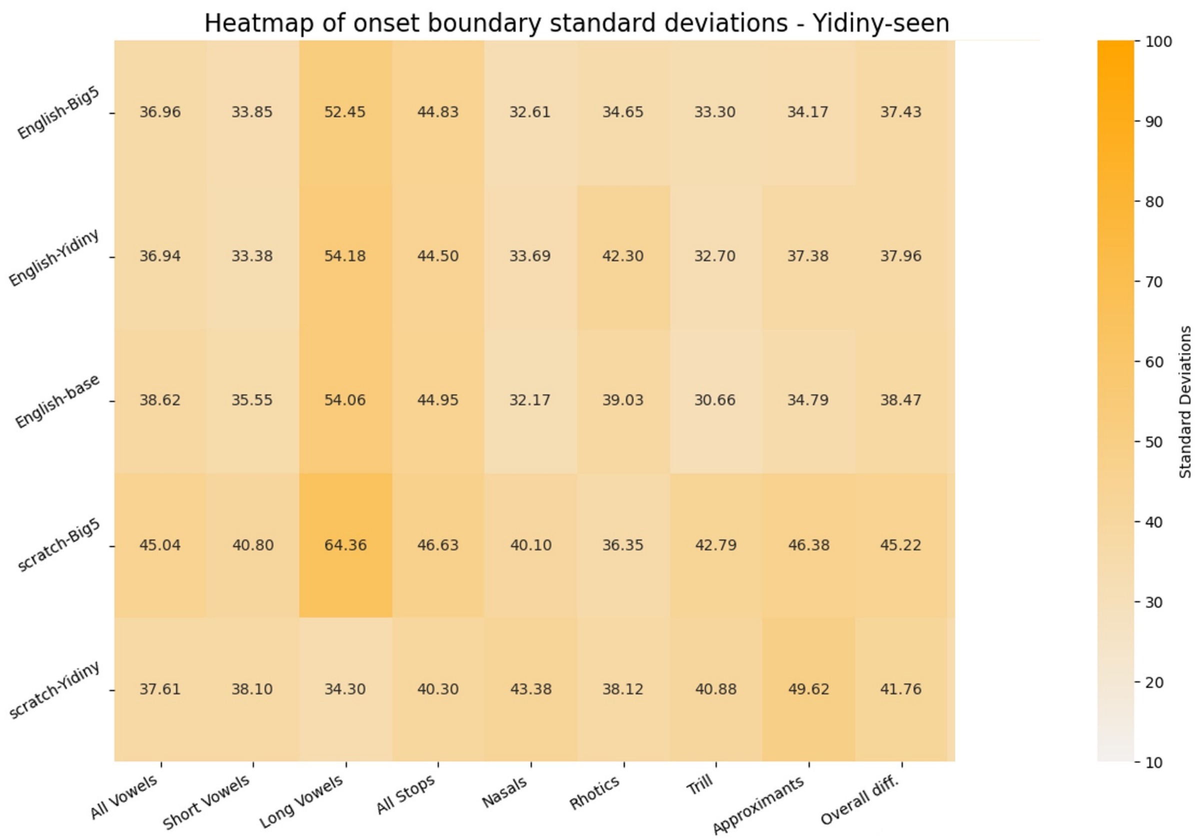
Results: Accuracy - Yidiny seen testing setting



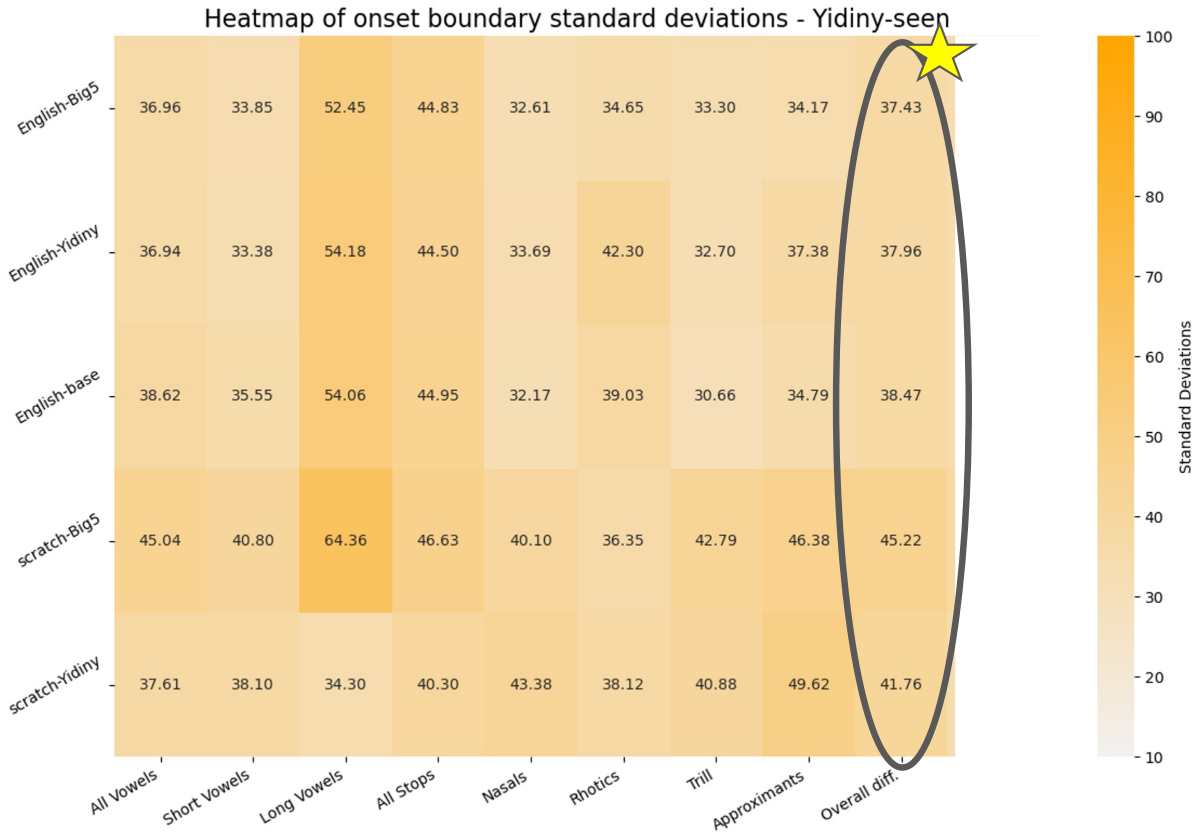
Results: Accuracy - Yidiny seen testing setting



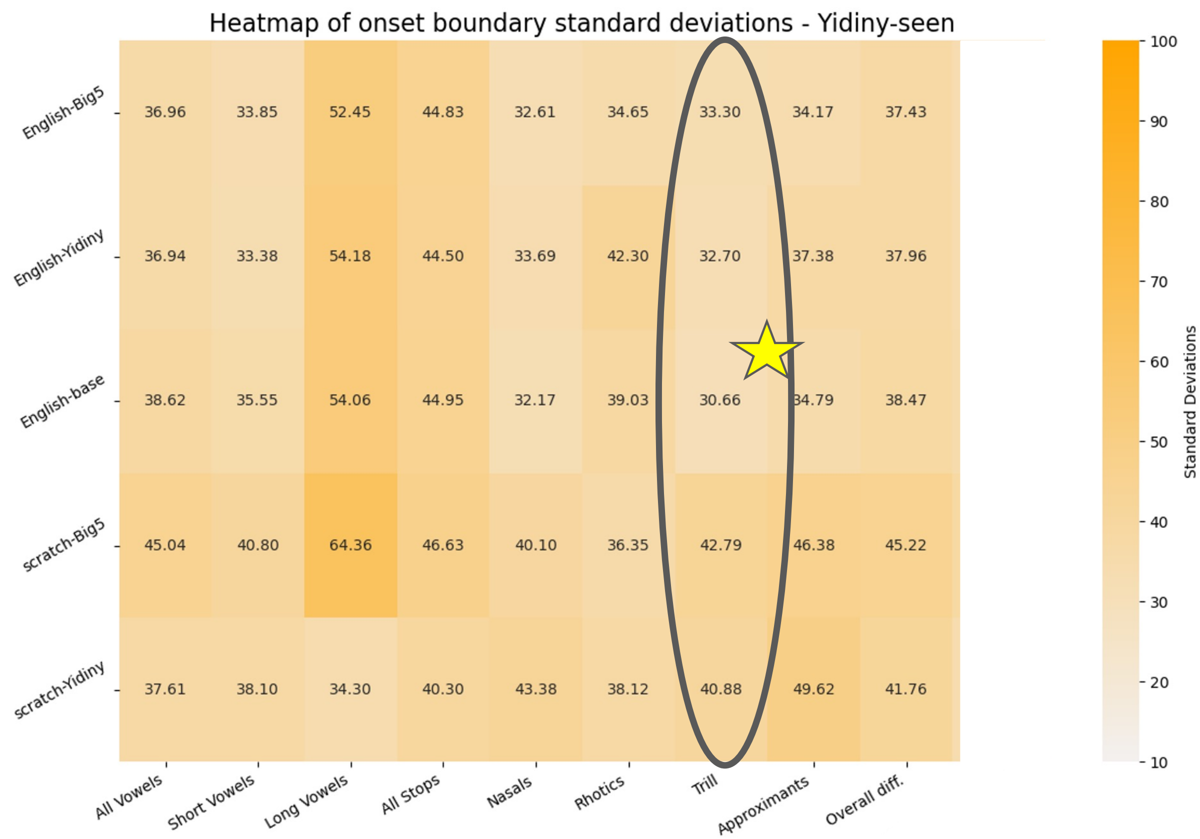
Results: Precision - Yidiny seen testing setting



Results: Precision - Yidiny seen testing setting



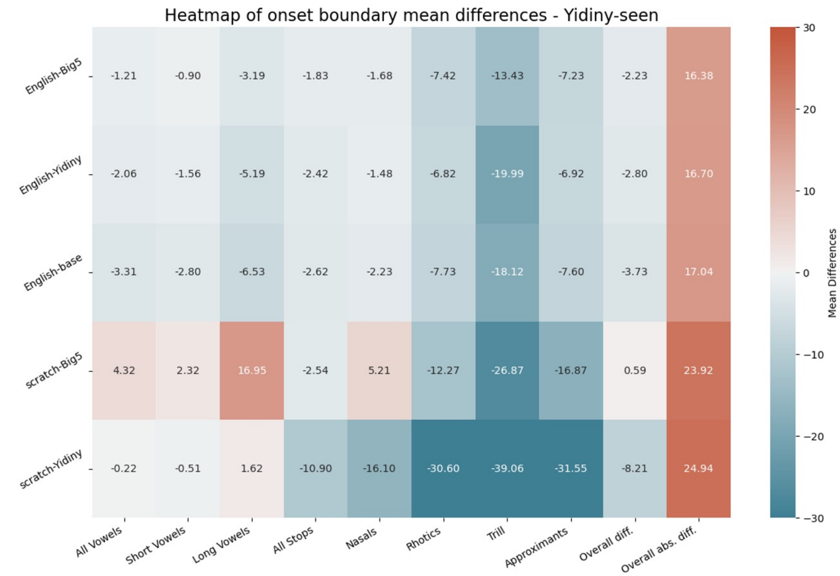
Results: Precision - Yidiny seen testing setting

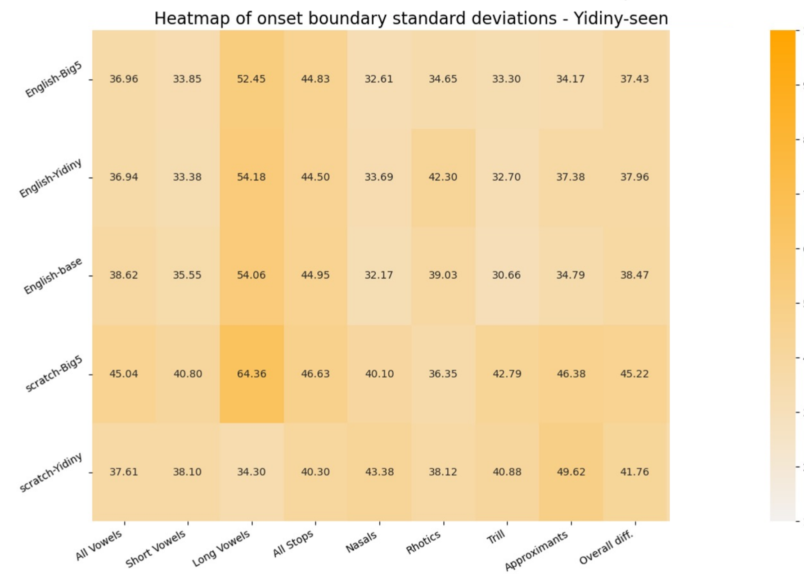


Precision Yidiny-seen Summary

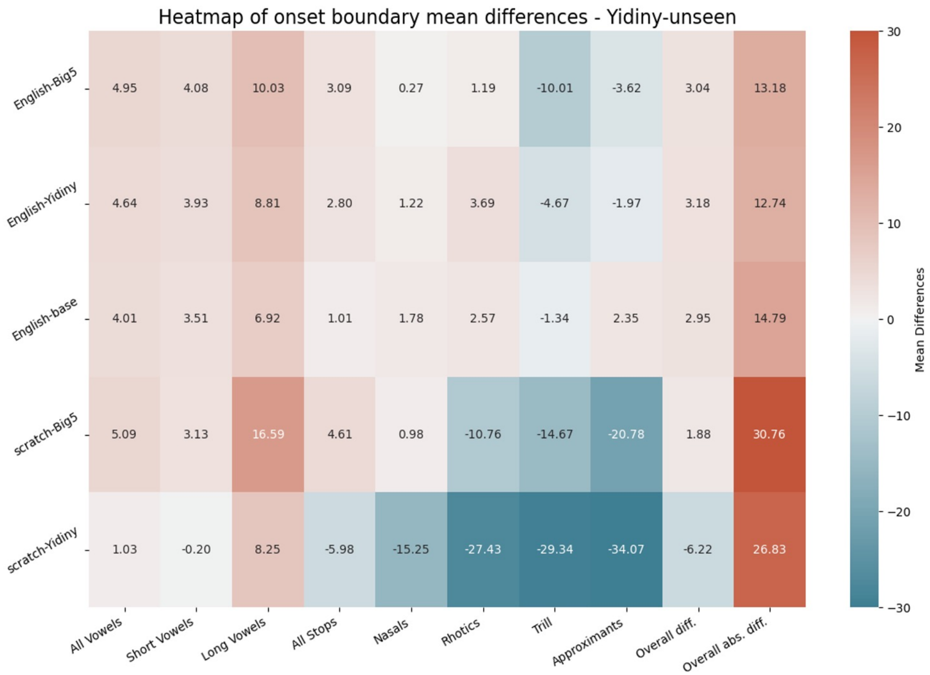
In the seen language seen data setting:

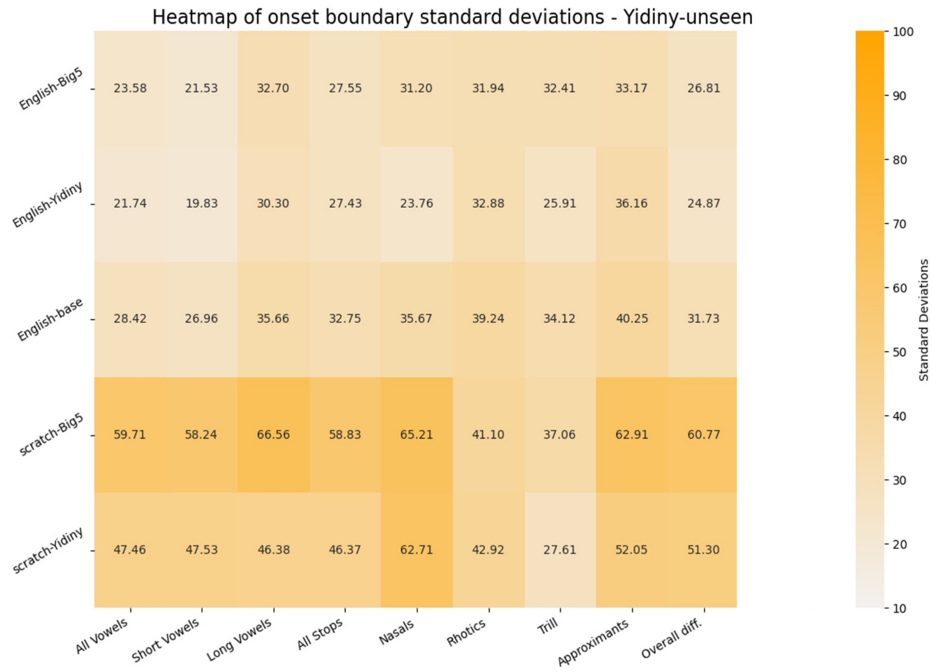
- Multilingual models show (slightly) higher precision than monolingual models
- Models trained from scratch are slightly less precise and less accurate than English-based models



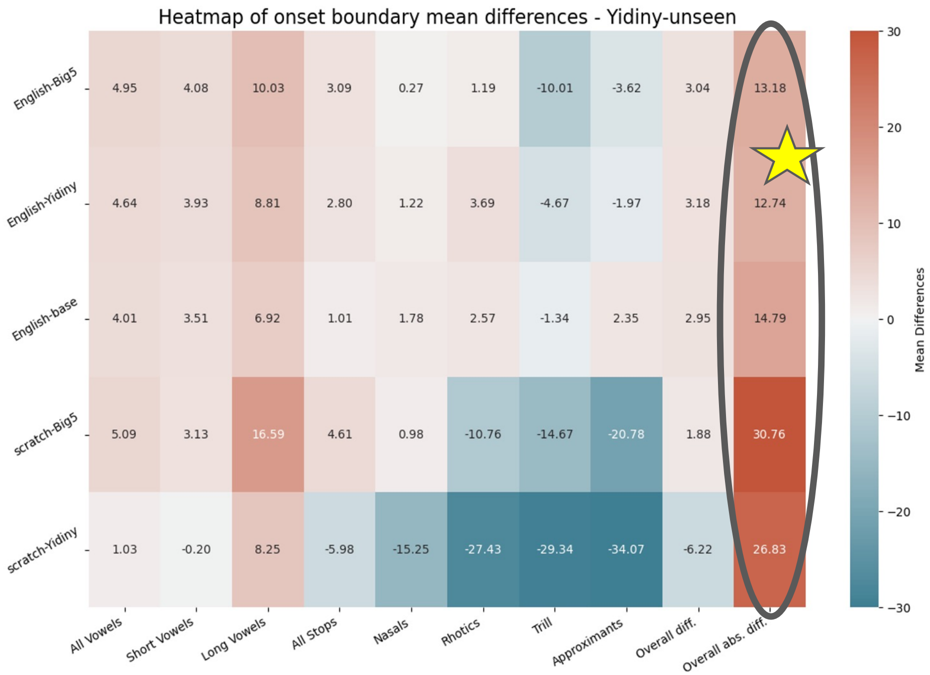


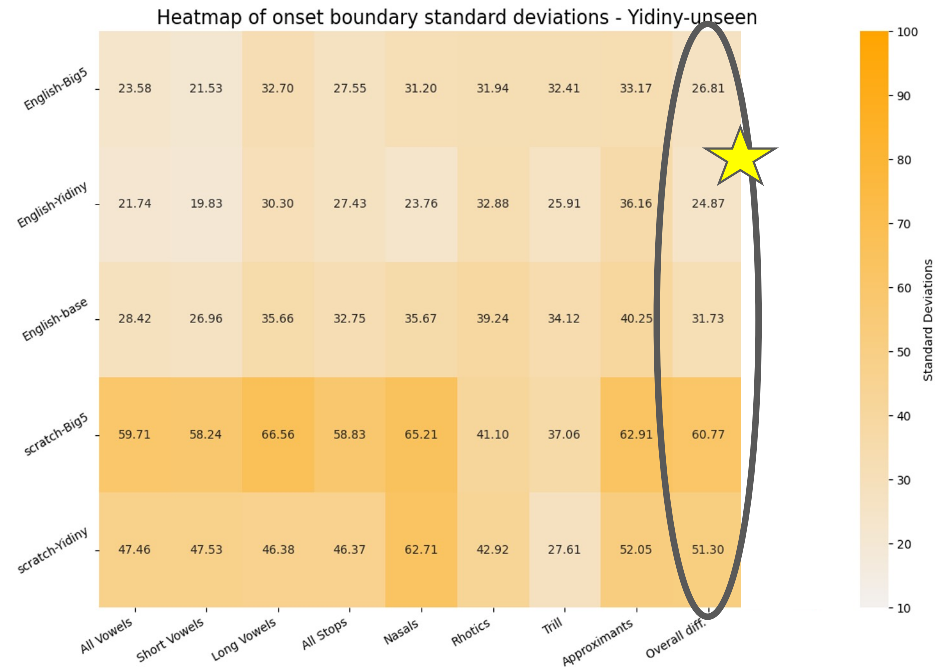
Results: Accuracy - Yidiny unseen testing setting



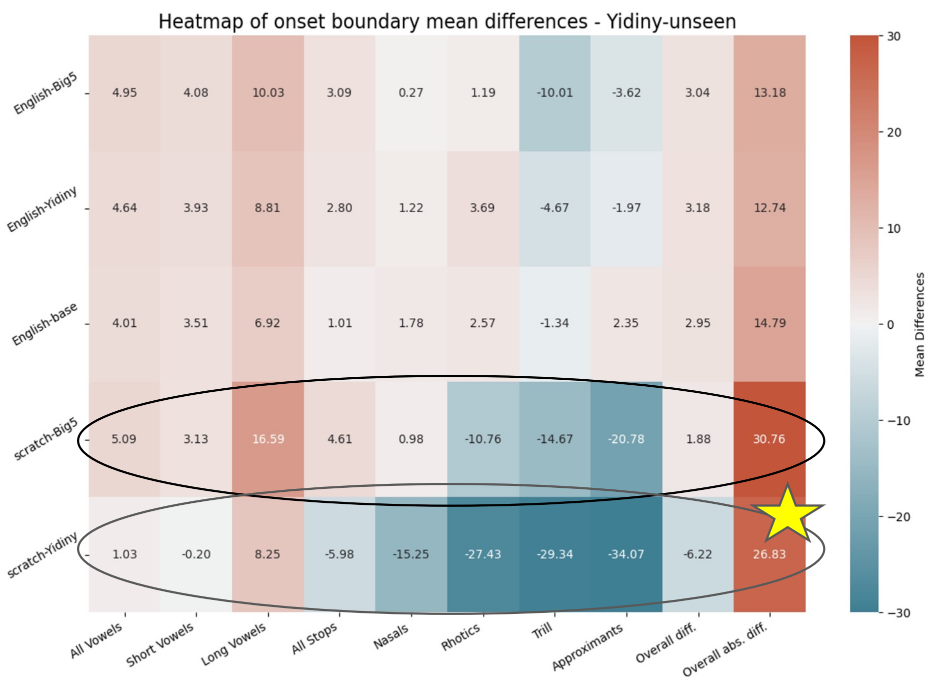


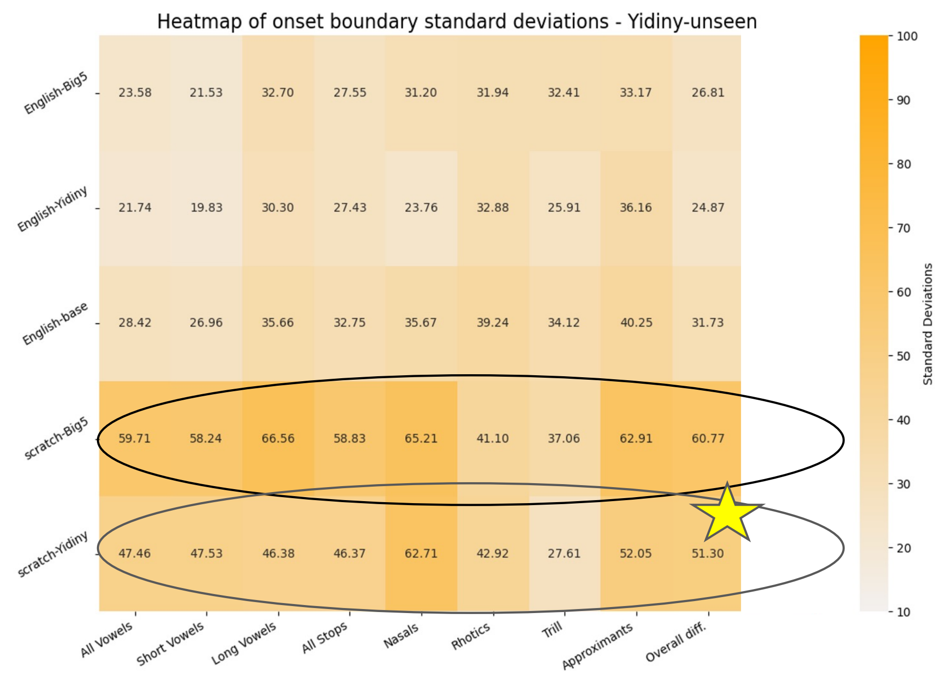
Results: Accuracy - Yidiny unseen testing setting



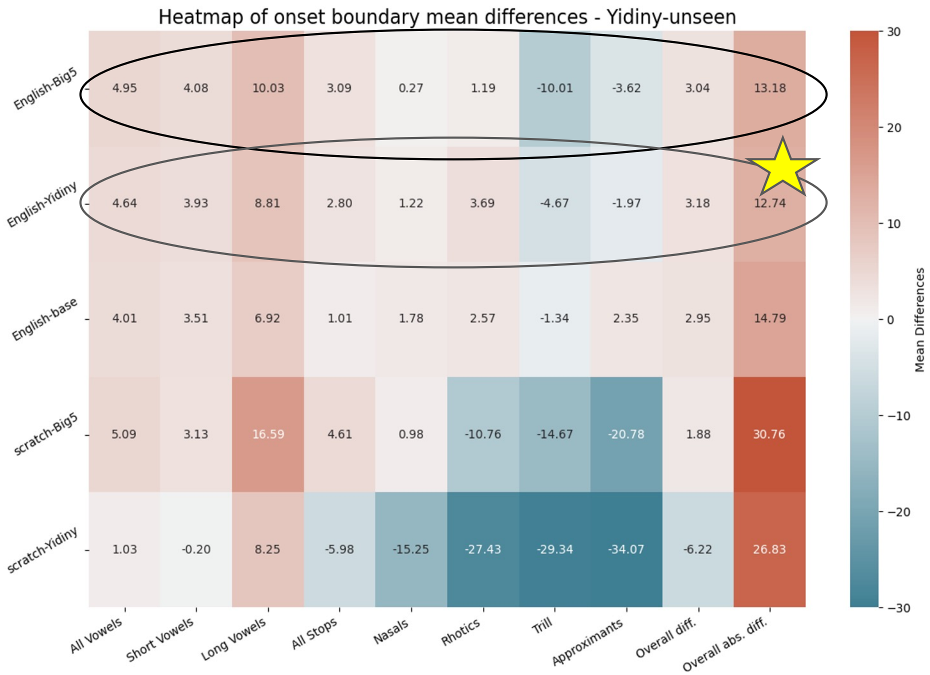


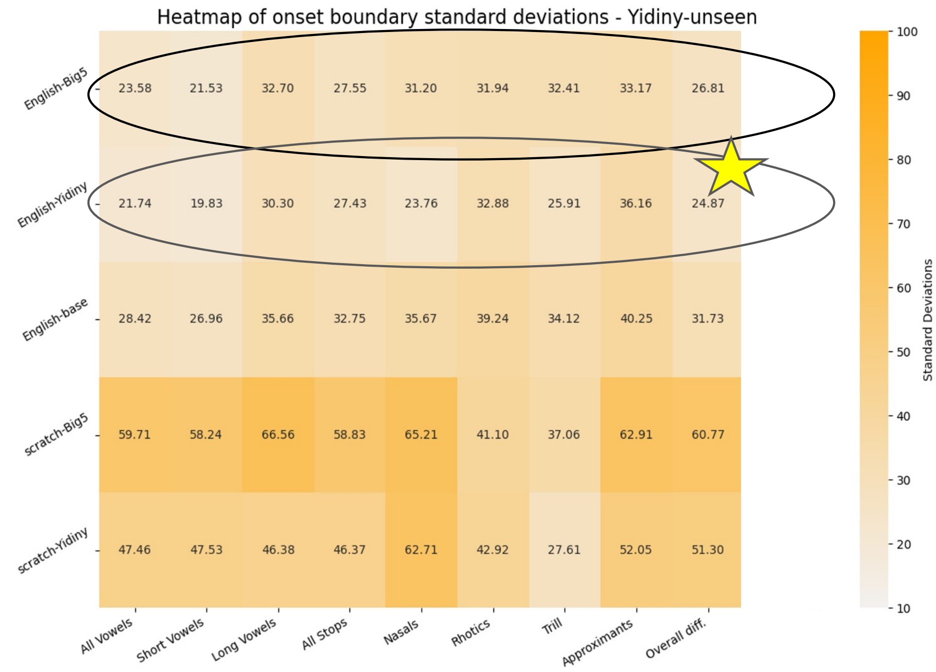
Results: Accuracy - Yidiny unseen testing setting





Results: Accuracy - Yidiny unseen testing setting

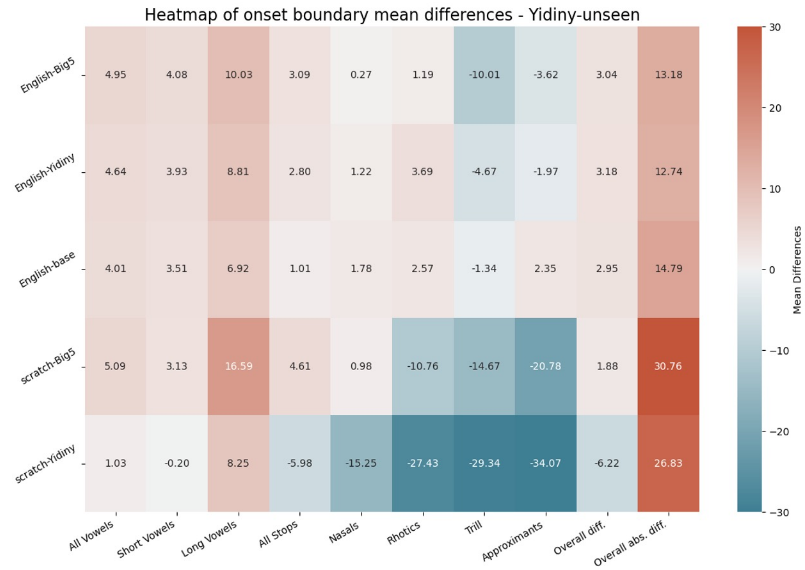


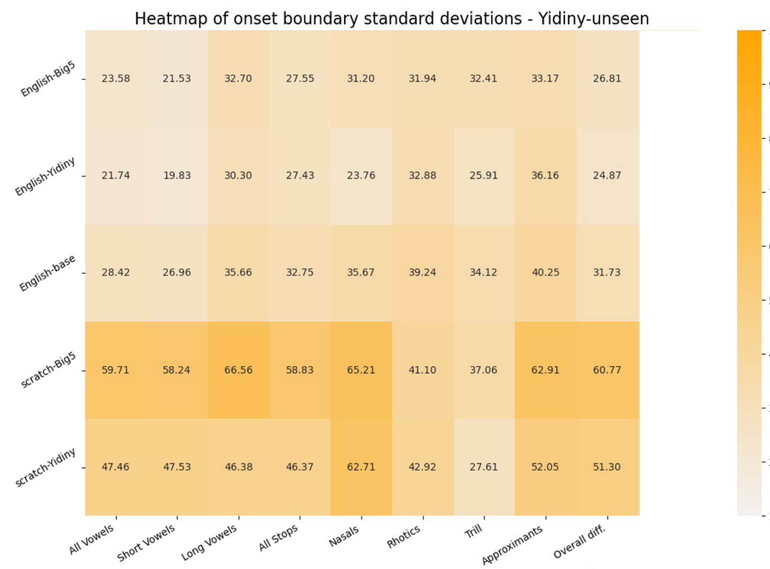


Precision Yidiny-unseen Summary

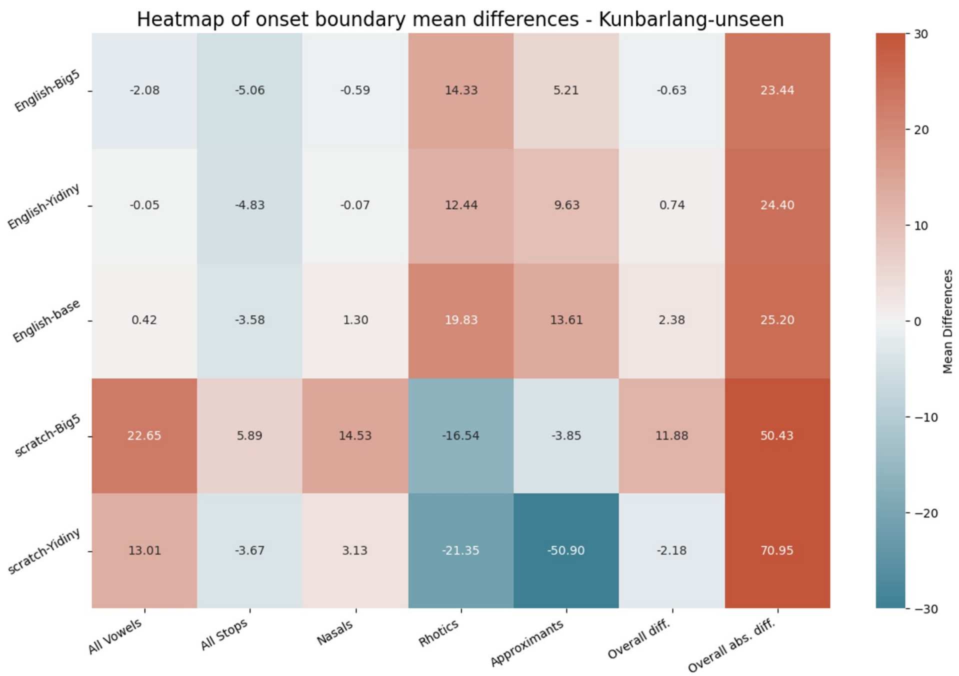
In the seen language unseen data setting:

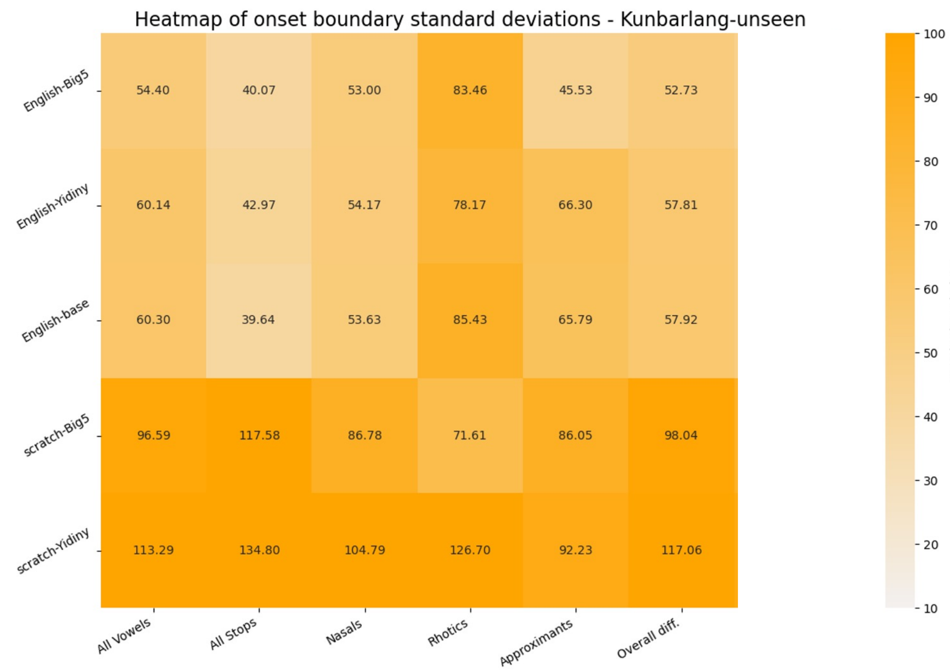
- Multilingual models show (slightly) lower precision than monolingual models
- Models trained from scratch are <u>quite</u> less precise and less accurate than English-based models



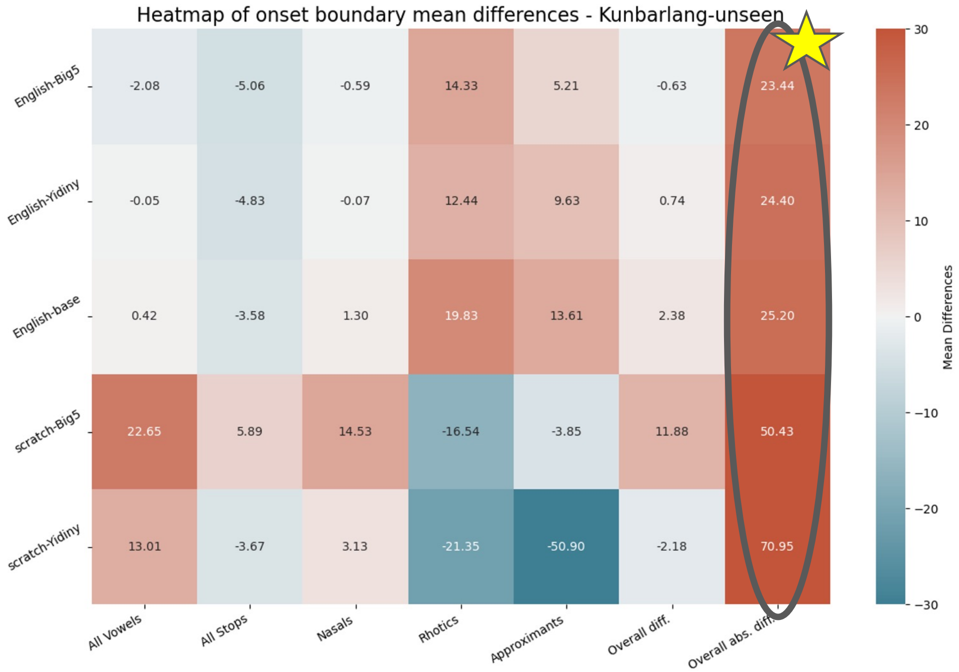


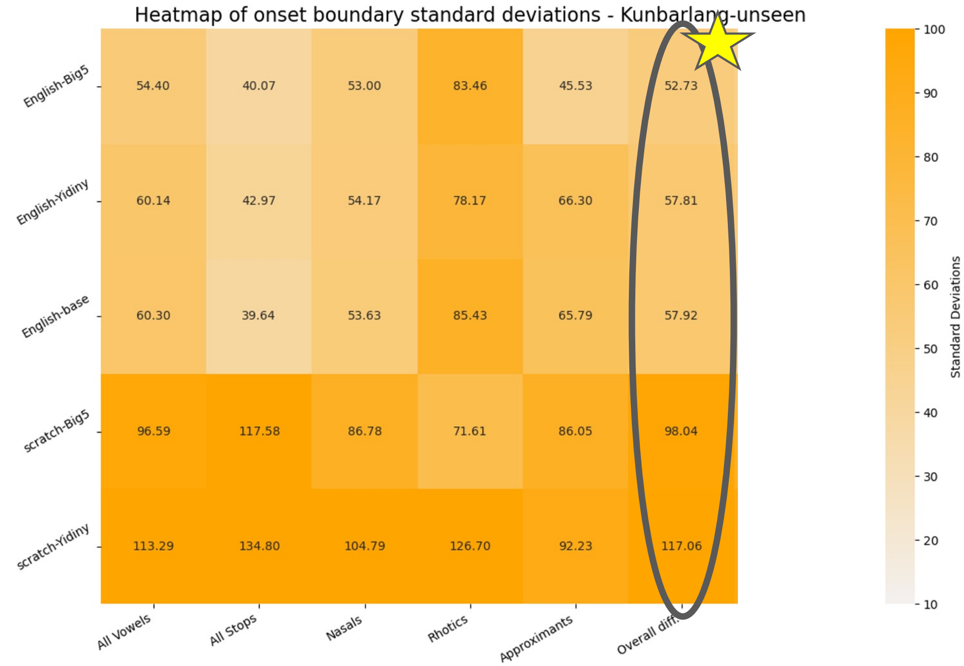
Results: Accuracy - Kunbarlang (unseen) testing setting



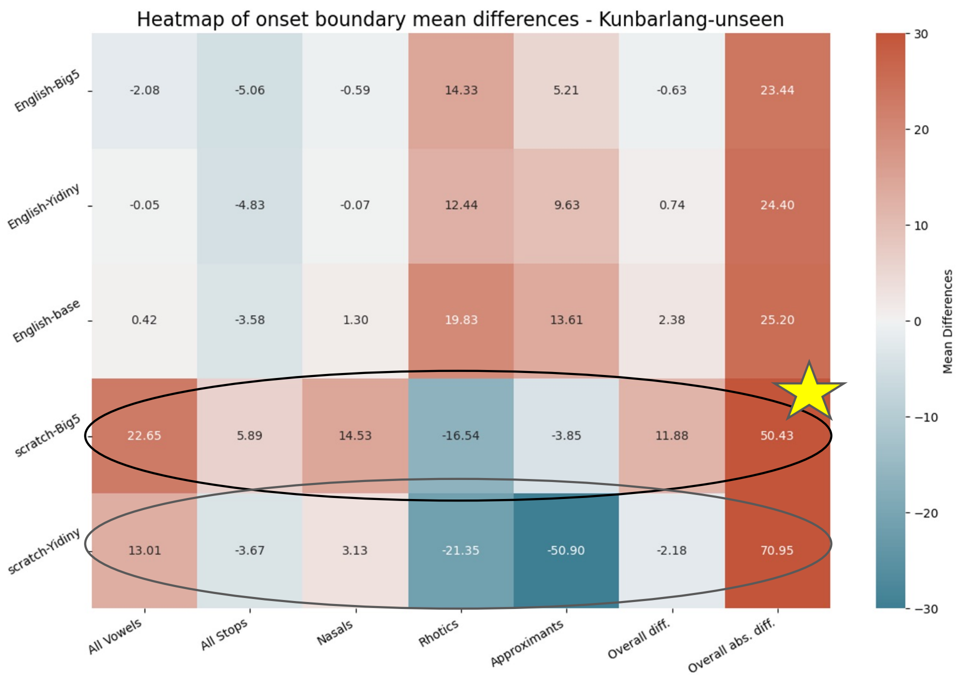


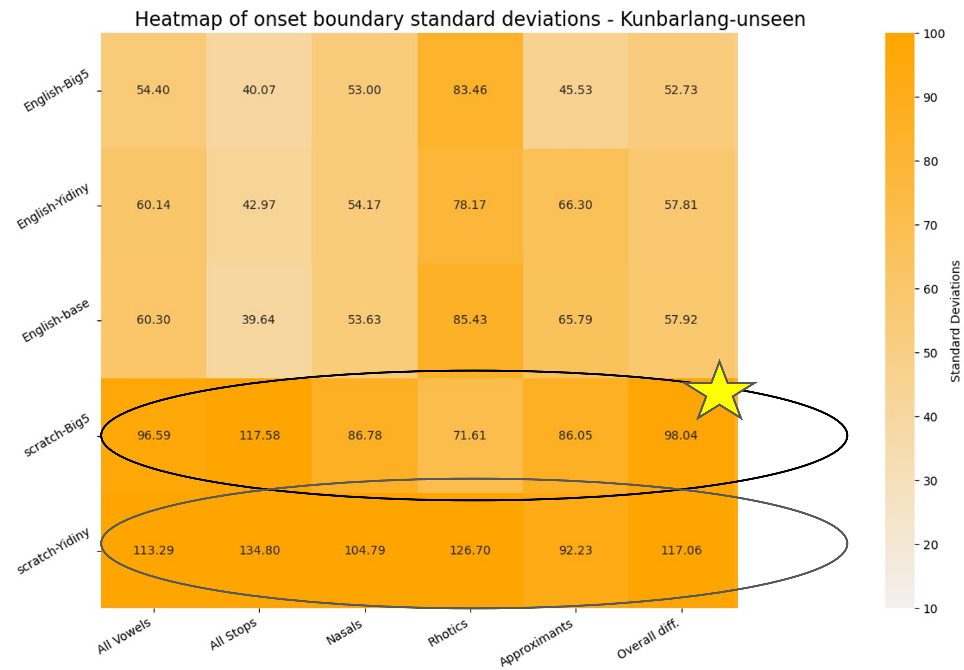
Results: Accuracy - Kunbarlang (unseen) testing setting



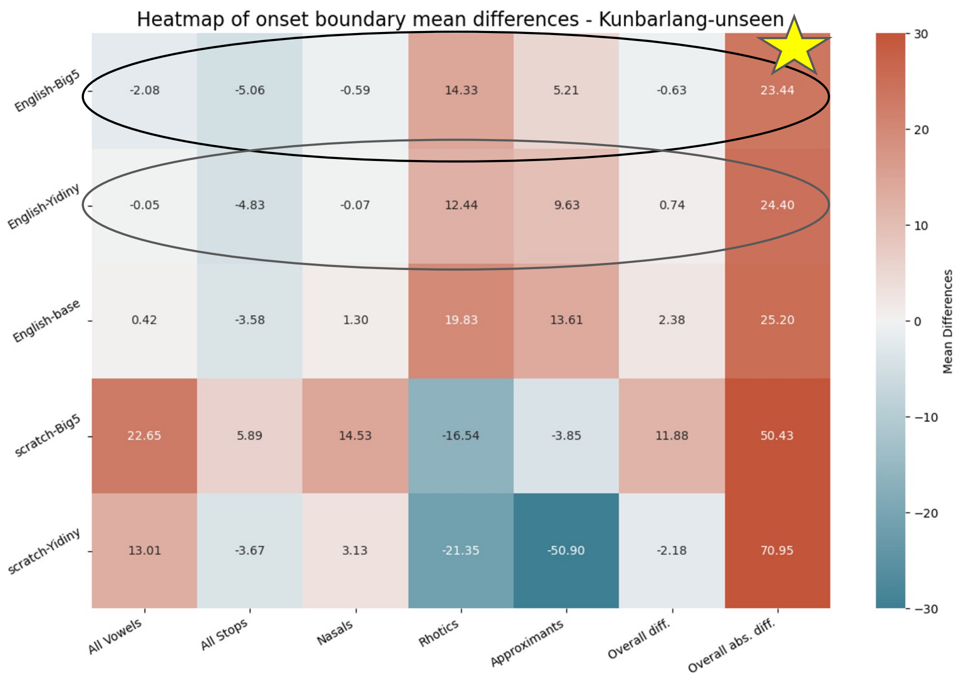


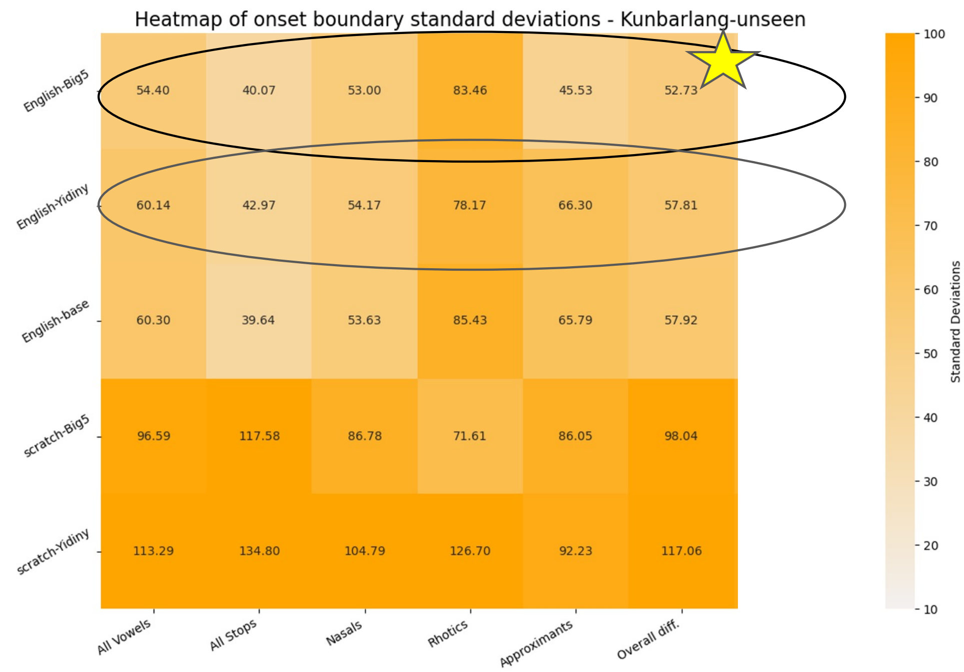
Results: Accuracy - Kunbarlang (unseen) testing setting





Results: Accuracy - Kunbarlang (unseen) testing setting

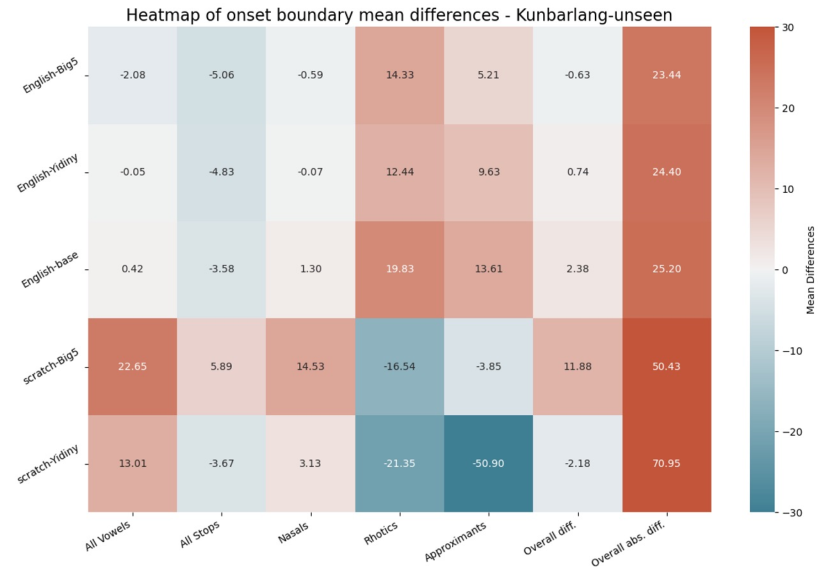


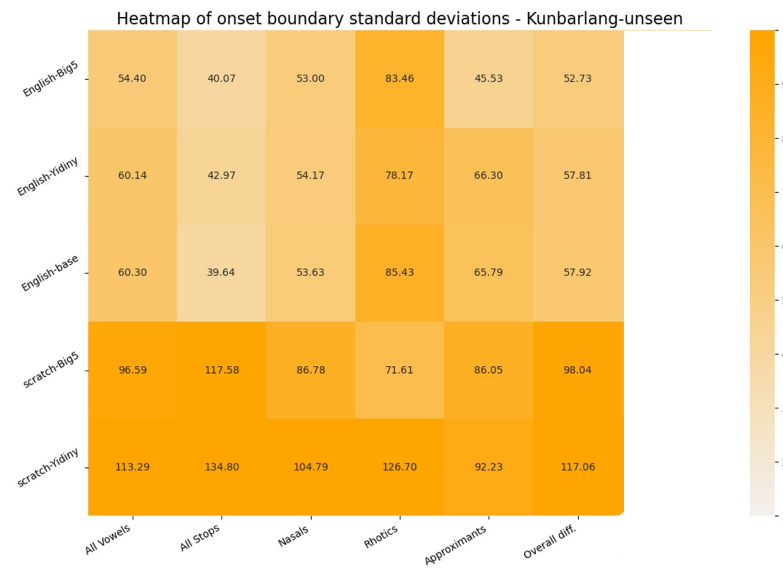


Precision Kunbarlang Summary

In the unseen language seen data setting:

- Multilingual models show more precision than monolingual models
- Models trained from scratch are <u>much</u> less precise and less accurate than English-based models



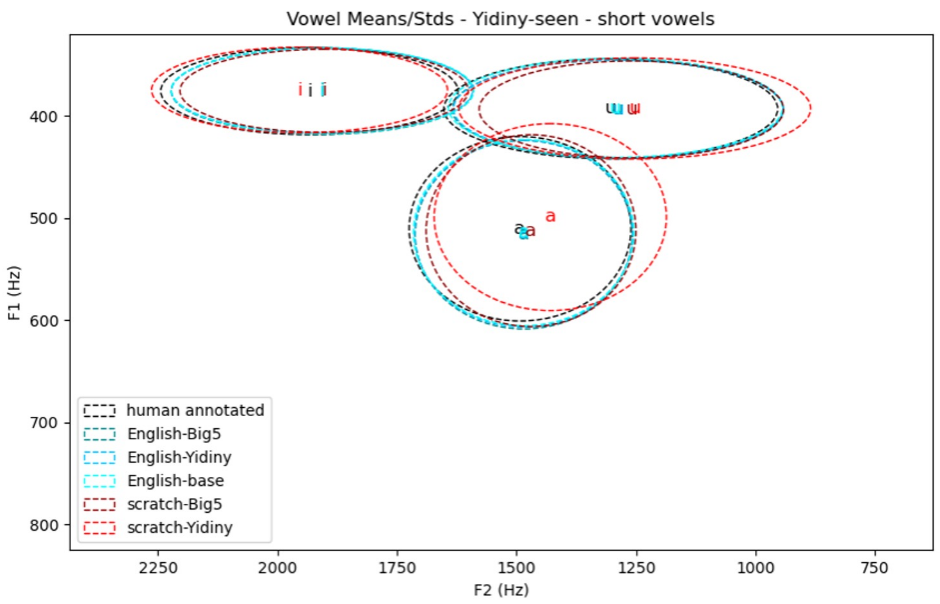


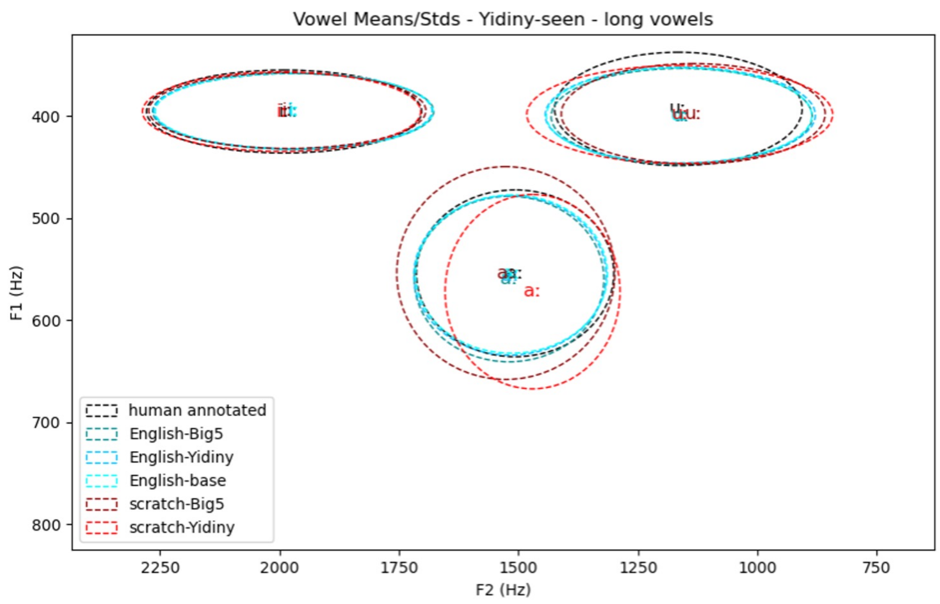
Precision and Accuracy Summary

Across all settings:

- Precision and accuracy seem highly correlated
- Multilingual training data:
    - improved performance in the Yidiny-seen and Kunbarlang-unseen settings
    - slightly decreased performance in the Yidiny-unseen setting
- Models trained from scratch consistently perform worse than English-based models
    - The performance gap was Kunbarlang >> Yidiny-unseen >> Yidiny-seen
    - In the Kunbarlang setting, multilingual training data had the biggest (positive) impact

Comparison of Analyses: Yidiny seen



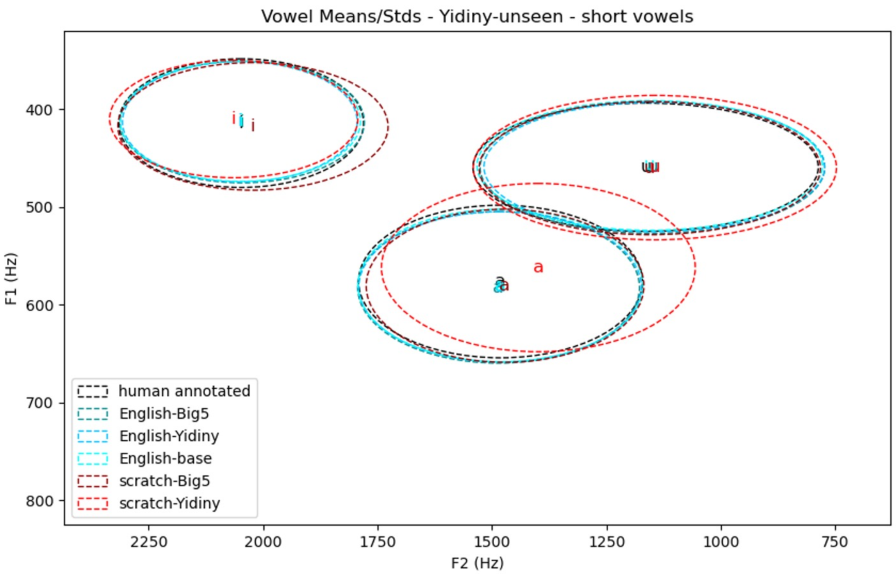


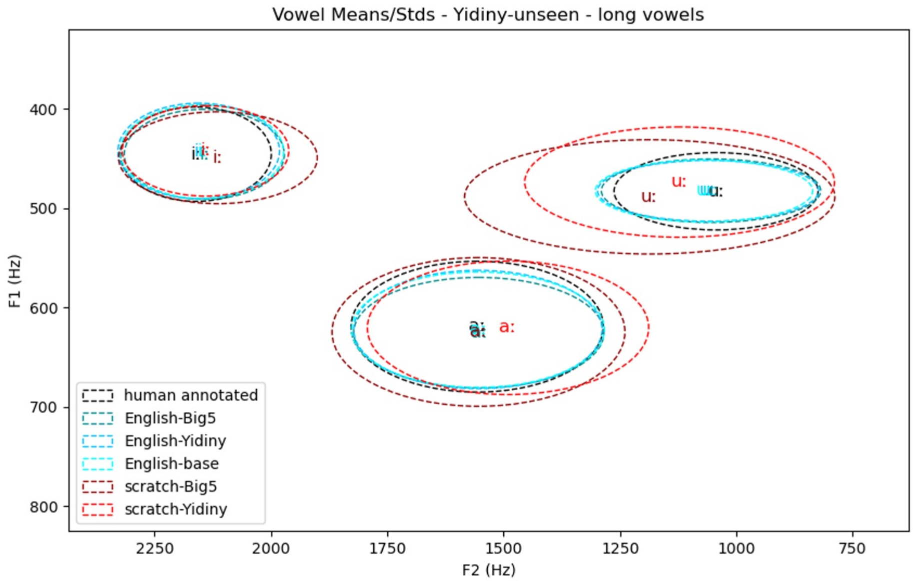
All analyses are similar, except the scratch-Yidiny model.

Long vowels show more variation for models trained from scratch.

Size of ellipses is large because of difficulty in formant extraction (noisy data)

Comparison of Analyses: Yidiny unseen





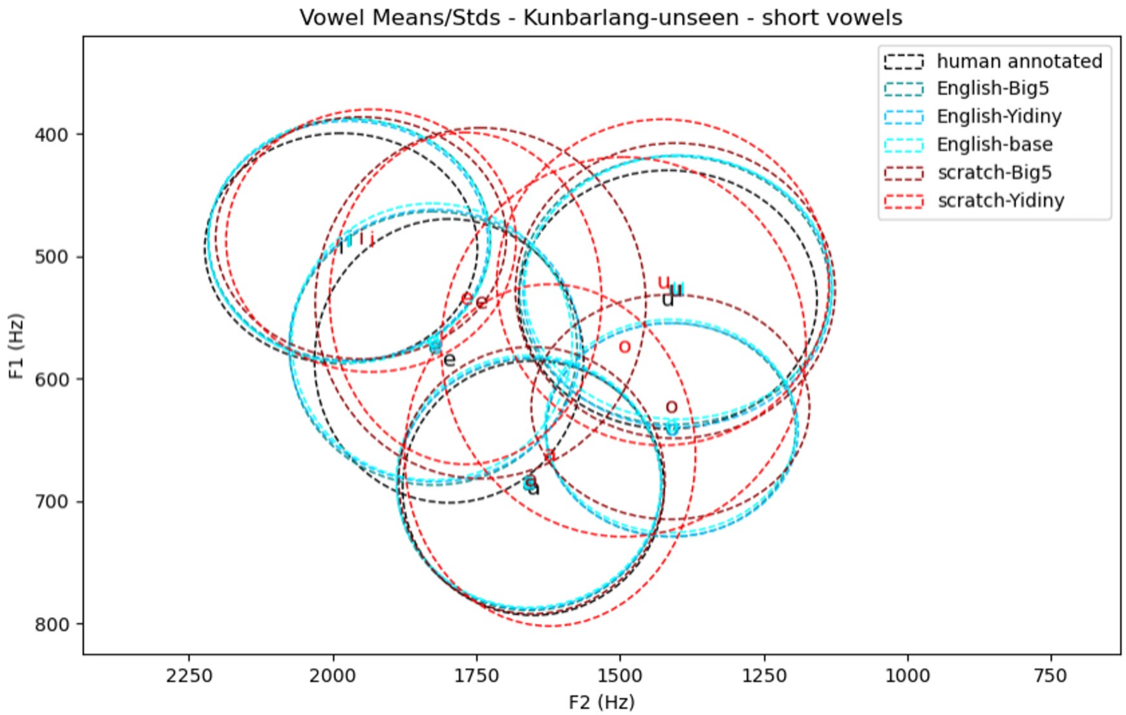
All analyses are similar, except the scratch-Yidiny model.

Long vowels show more variation for models trained from scratch.

Comparison of Analyses: Kunbarlang (unseen)

Models from scratch:

- Struggle with /e o/
    - /e/ absent in all training data, /o/ only present, but rare in one language of Big5



Comparison of Analyses Summary

In the Yidiny-seen and Yidiny-unseen settings:

- All models, except the model trained from scratch on only Yidiny data, gave approximately the same analyses
- Multilingual models provided better analyses

In the Kunbarlang setting:

- English-based models gave analyses nearly identical to the manually annotated analysis
- Models trained from scratch struggle with phones not in the training data (the mid-vowels)

In all settings:

- English-based models > scratch-Big5 >> scratch-Yidiny

Main takeaways

- Which models work best?
    - English based models work best. Even the off-the-shelf English model was better than all the models trained from scratch in almost all ways
    - Models trained from scratch were *almost* as good as the English models in the Yidiny-seen setting
- Does multilingual training data improve MFA?
    - Yes!
        - Slight improvements for English models
            - Biggest improvements by natural class came from natural classes not in the training data (e.g. trills for English)
        - Larger improvements for models trained from scratch
- How do analyses compare?
    - English based models ≈ manual annotation
    - The multilingual model trained from scratch also ≈ manual annotation for tokens it has trained extensively on

Low-Resource Forced Alignment and Future work

Low-Resource Forced Alignment:

- The Yidiny-seen setting is thus most similar to real settings
  - Smallest difference between the English-based and from-scratch models
- All models worked quite well in the Yidiny-seen setting
- **Recommendation: use an adapted English based model if you have less than 30 minutes of training data**

Future work:

- Optimizing forced alignment for settings like “Yidiny-seen”
- Hyperparameter tuning
- Data augmentation

References + Contact

Chodroff, Eleanor, E. Ahn, and Hossep Dolatian. 2024. Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. Language Documentation & Conservation

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2006.07264

McAuliffe, Michael, and Morgan Sonderegger. 2024. English mfa acoustic model v3.1.0.Technical report, https://mfa-models.readthedocs.io/acoustic/English/EnglishMFAacousticmodelv3_1_0.html.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In Interspeech, volume 2017, pages 498–502.

Wu, H., Yun, J., Li, X. *et al.* Using a forced aligner for prosody research. *Humanit Soc Sci Commun* **10**, 429 (2023). https://doi.org/10.1057/s41599-023-01931-4

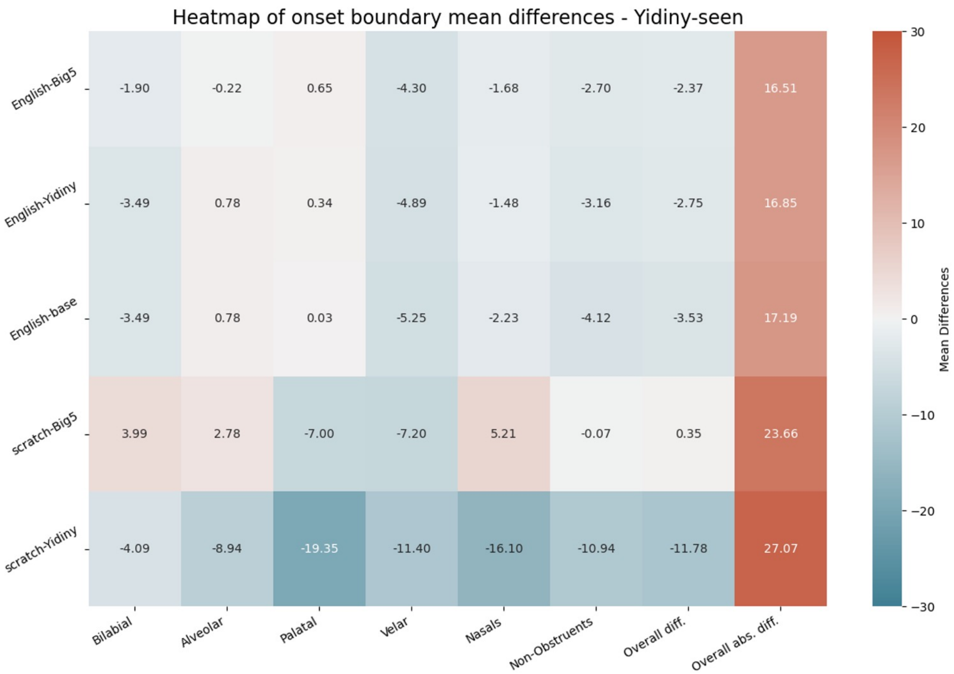
Young, N. J. and M. McGarrah, “Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype,” *Nordic Journal of Linguistics*, vol. 46, no. 1, pp. 105–131, May 2023.

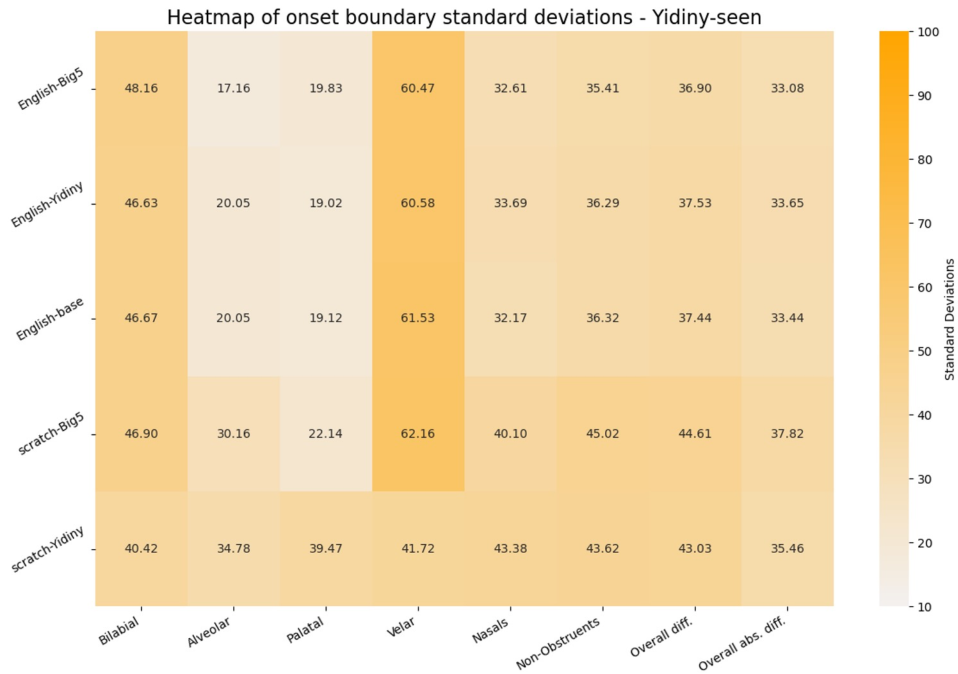
alessio.tosolini@yale.edu claire.bowern@yale.edu

Acknowledgements

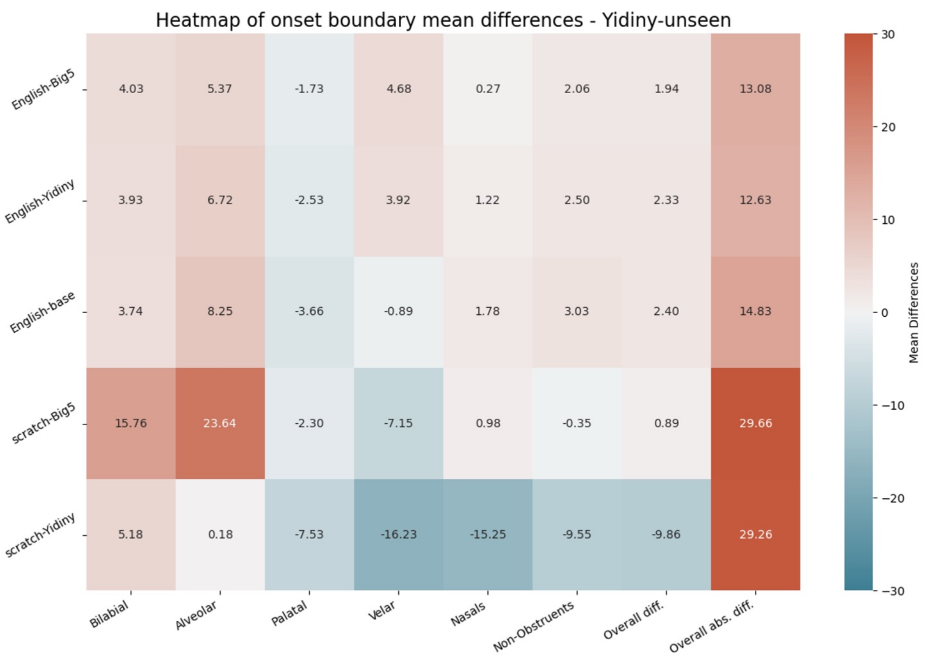
- ELAR, Paradisec, and AIATSIS archives
- Audiences at Yale and SYNC for feedback
- The elders and others who worked with linguists to make records of their languages
- National Science Foundation’s Linguistics and DEL-DLI programs for long-term funding of research on language and with language communities, particularly in their support for broader impacts (earlier work funded by BCS-0844550, BCS-1423711, and BCS-2116164)

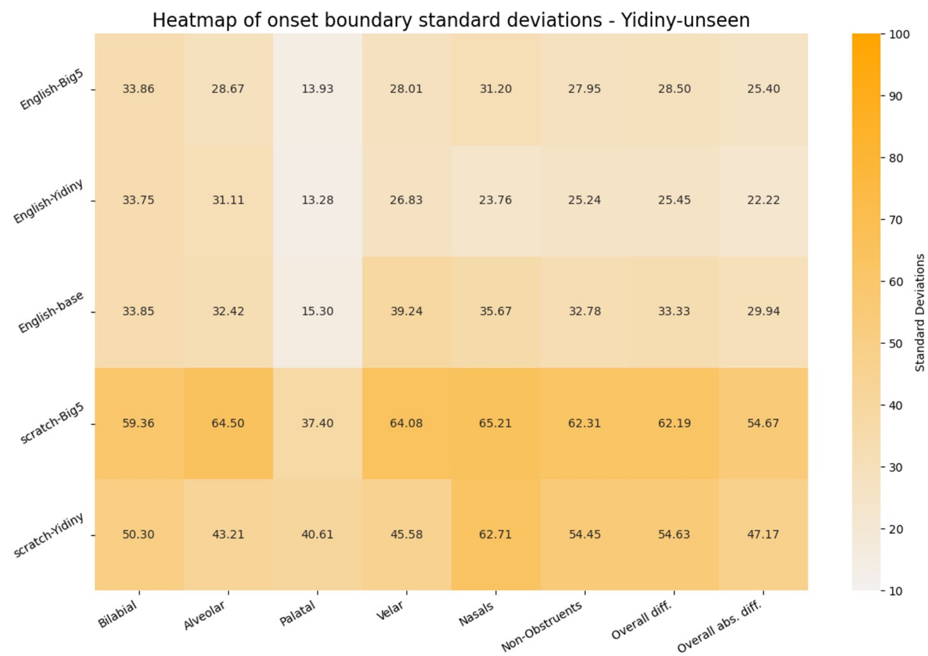
Appendix A: PoA Accuracy Yidiny seen





Appendix A: PoA Accuracy Yidiny unseen





Appendix A: PoA Accuracy Kunbarlang (unseen)