Investigating "Accuracy" of Small Phonetic Corpora: A Sampling Experiment

Coralie Cram & Claire Bowern   UCLA, Yale University :: Linguistics

The Problem

- Whalen et al. (2022): detailed phonetic descriptions are skewed towards a few regions and families.
- Many languages are only studiable using archival corpora (Whalen & McDonough 2015)
- They are attested through "multipurpose documentation" (not designed for phonetics and not collected with specific phonetic questions in mind)
- Usually unbalanced or missing crucial distinctions
- **How small a corpus can still capture features of the "language" (cf. Maddison 1999)?**
- We explore these questions by investigating differences in mean phonetic measures of increasingly smaller samples of the same dataset.

Results—Wordlist

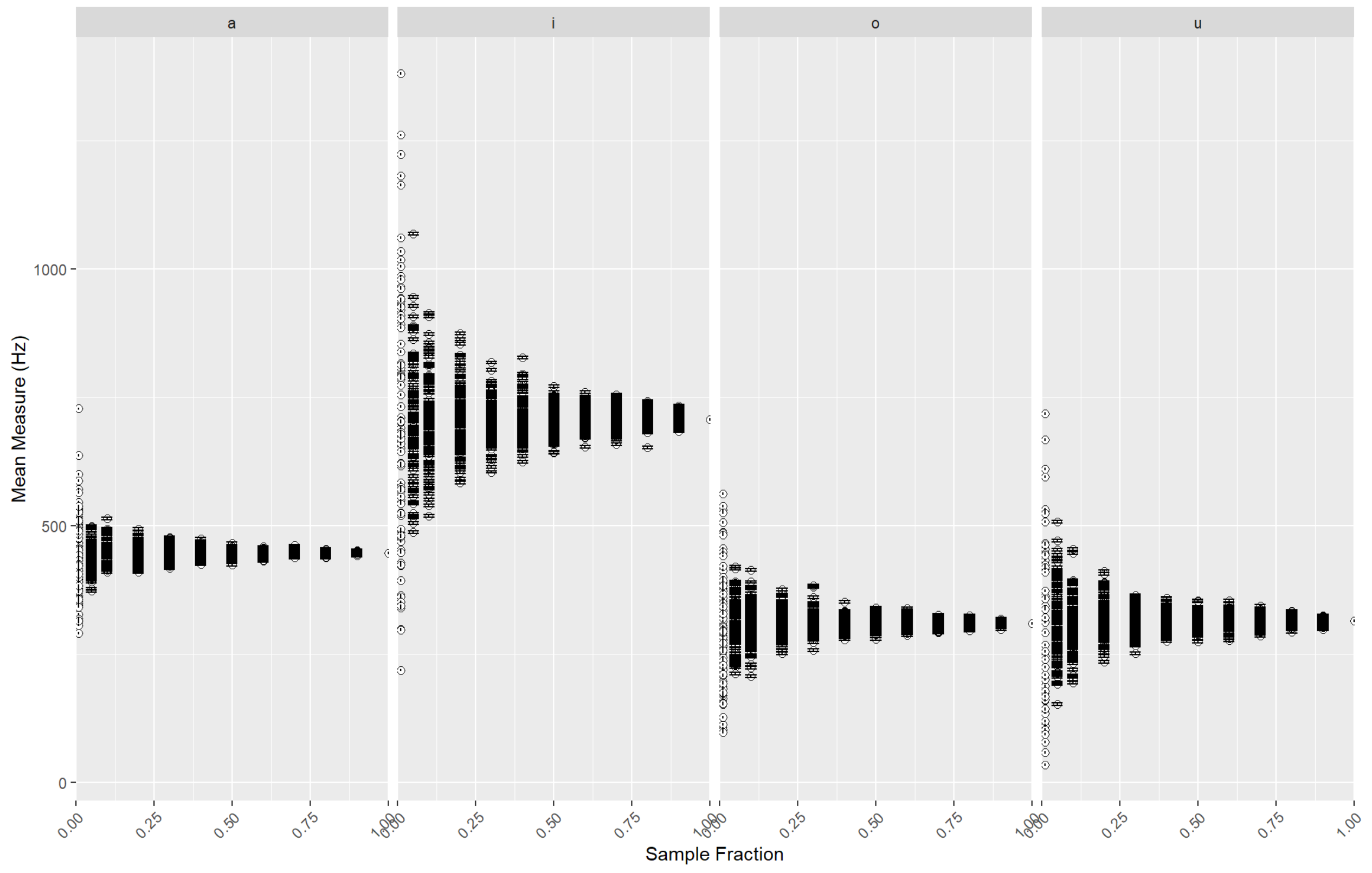

Figure 1: Mean Euclidean distance, resampled 100 times at 1-90% of full wordlist dataset

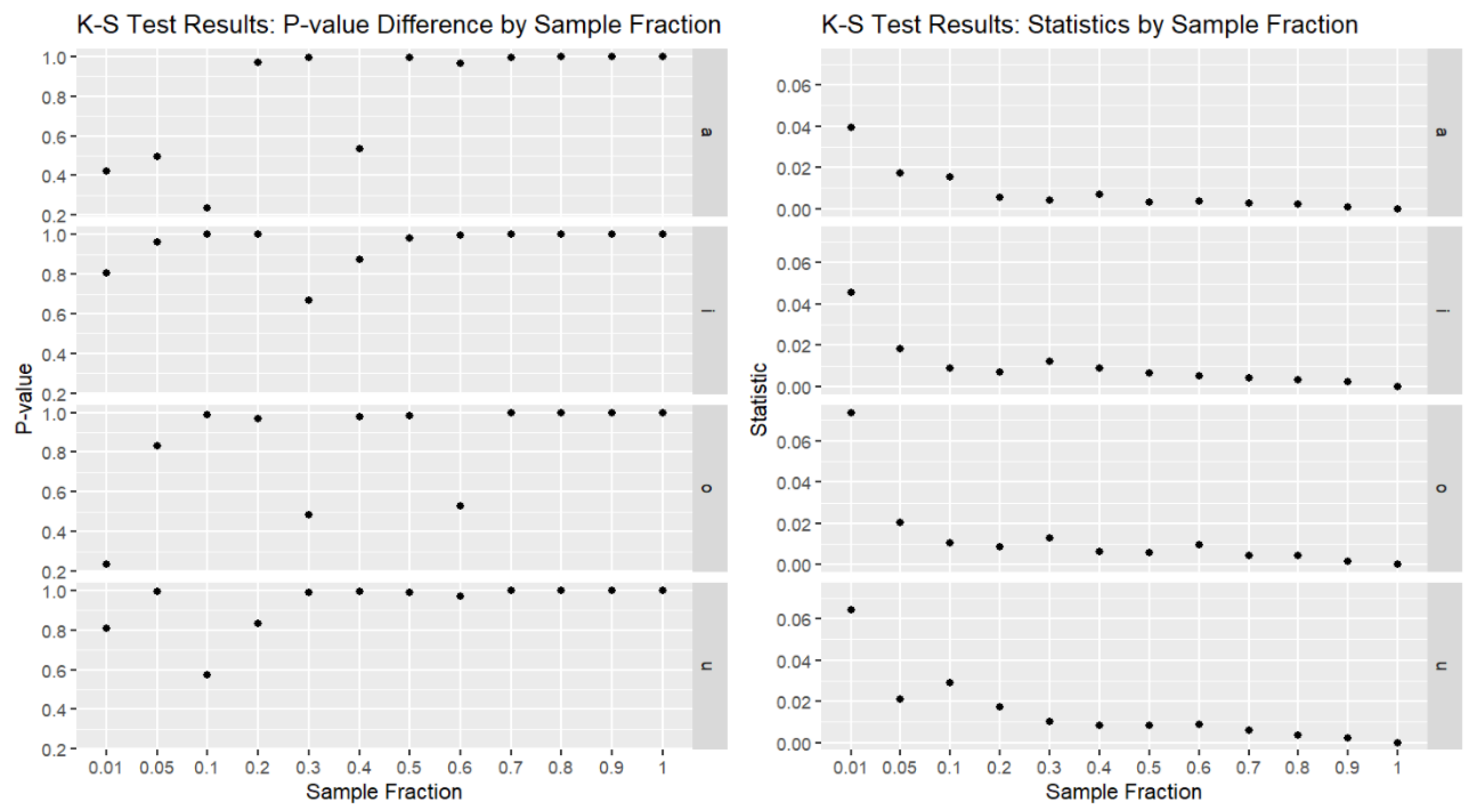

Figure 2: Kolmogorov-Smirnov results for wordlist data

Materials & Methods

- 2 female Bardi speakers (Nyulnyulan, Australia; 7 vowel system (/i(ː), a(ː), o, u(ː)/); wordlist data (928 tokens (short vowels, non-final))
- F1 & F2 of midpoint extracted using `forrest` in `wrassp` (Bombien & Winkelmann, 2023)
- Mean Euclidean distance measures ($d = 2\sqrt{(a_2 + b_2)}$) then randomly resampled from larger subsets (1%–90%) 100 times for each fraction & vowel
- Two-sided Kolmogorov-Smirnov test in `dgof` measures goodness-of-fit between means of subset and full dataset
- Results also compared to larger corpus of narrative data with 7836 tokens, 5 speakers

Results—Narratives

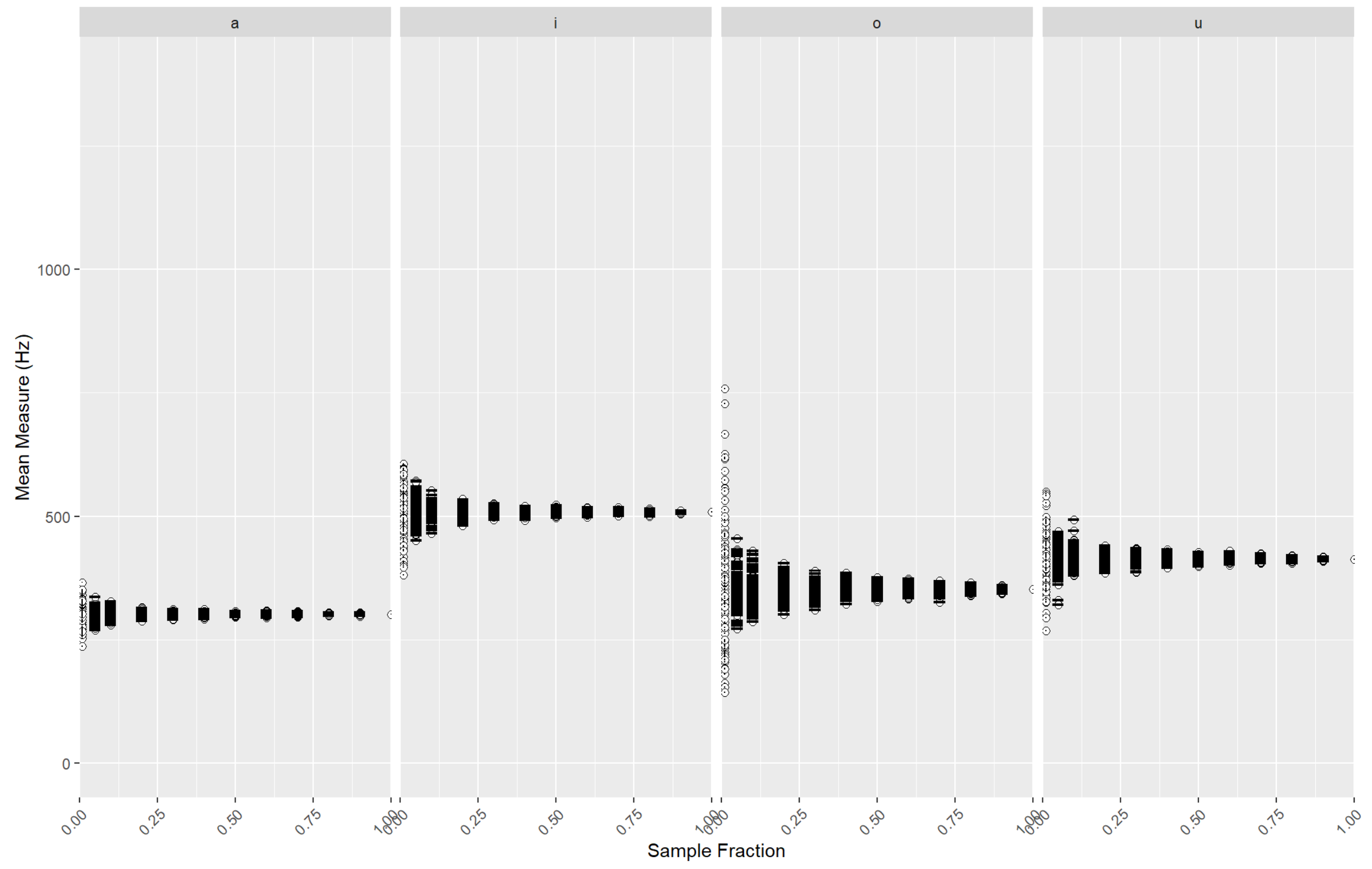

Figure 3: Mean Euclidean distance, resampled 100 times at 1-90% of full narrative dataset

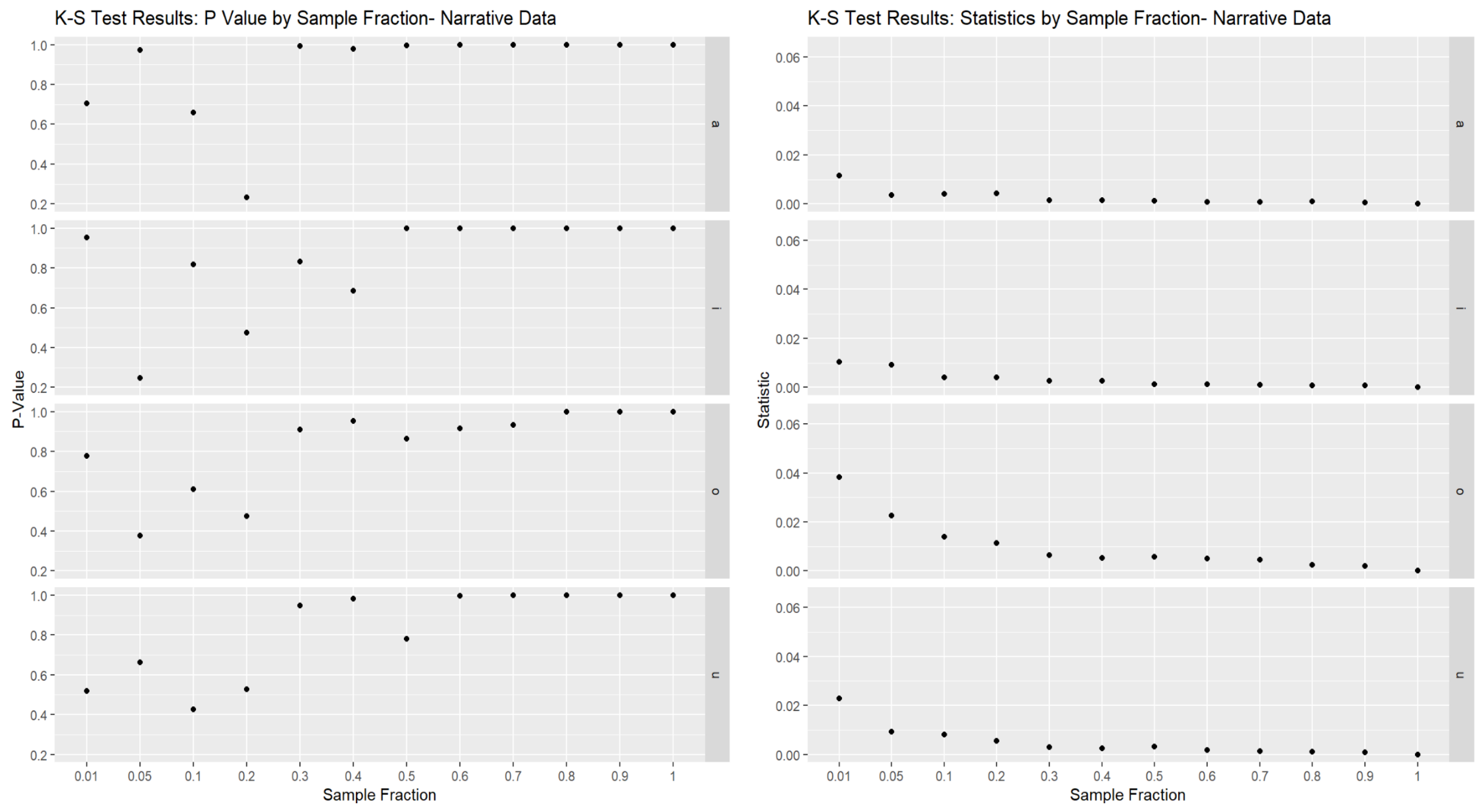

Figure 4: Kolmogorov-Smirnov results for narrative data

Discussion

- Mean measures increasingly converge towards the full sample mean for each vowel for larger subsamples.
- The K-S test shows a <0.05 statistic measure for all samples above the 1% subset, <0.025 for larger narrative dataset, and a 1.0 overlap (i.e, rejection of the null hypothesis for different samples) in p-values for all but one vowel and sample in sizes above 40% (narrative)/ 50% (wlist).
- The majority of results remain above significant overlap (p>0.05) for all sample sizes. Suggests most samples appear identifiably representative of the larger sample, even for the smaller wordlist dataset.
- Wordlist results demonstrate differences between vowels that correspond to their dispersion; e.g., /i/ has the widest distribution, which is reflected in the much wider shift in measures at smaller subsets of the data.
- This dispersion effect goes away for narrative data; e.g., /i/ measures are more stable than /o/ and /u/.
- (NB: testing only for sample replication, not controlling for mis-tracked formants, etc.)

Conclusion

- **Results tentatively indicate a high level of validity for small datasets.**
- Wider differences in dispersion might impact the validity of distributional and means-based analysis, though results from the narrative data suggest sample size alone might be most important.
- Mirroring Dockum & Bowern (2017) for phonotactics, **c. 300–400 tokens** is a safe minimum.
- Results also demonstrate value in using archival narrative corpora for phonetic research.

Contact Information

Email ccram@g.ucla.edu
Email claire.bowern@yale.edu