



HAL
open science

Speech Maturity Dataset

William N Havard, Loann Peurey, Kasia Hitczenko, Alejandrina Cristia

► **To cite this version:**

William N Havard, Loann Peurey, Kasia Hitczenko, Alejandrina Cristia. Speech Maturity Dataset. Many Paths to Language (MPaL) 2023, Oct 2023, Nijmegen, Netherlands. . halshs-04294803

HAL Id: halshs-04294803

<https://shs.hal.science/halshs-04294803v1>

Submitted on 20 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



1. Introduction

- Children's spontaneous vocal productions become increasingly adultlike
 - shape (canonical, non-canonical)
 - phonetic and phonemic properties (frequency range, phonotactics, etc.)
- But... research has been limited to a narrow set of languages and communities
 - Indo-European languages
 - Western(ised) speaker communities
 - narrow age range: 0 - 24mo

2. Speech Maturity Dataset

- Superset of BabbleCor (Cychosz et al., 2019)
- 15k (Babblecor) → 258,914 clips (ours)
- 398 children (209 male, 186 female)
- Large age range: 2mo - 6yr
- 14 communities
 - rich industrialised societies
 - farmer-forager communities
- 25+ languages

3. Zooniverse: Citizen Science

- **Citizen Scientist:** Non-scientific volunteers who annotate and label scientific data
- Clip labels based on the majority vote of at least 3 citizen scientists
- Majority vote: at least 50% of the citizen scientists endorsed a particular label
- Labels (speaker type and sex for a subset of the clips only N=110,577)
 - Vocalisation Type: canonical, non-canonical, laughing, crying, junk
 - Speaker Type: baby (younger than 3 years), child (3-12 years), adolescent (12-18 years)
 - Sex: Female/Male (for adolescents and adults)

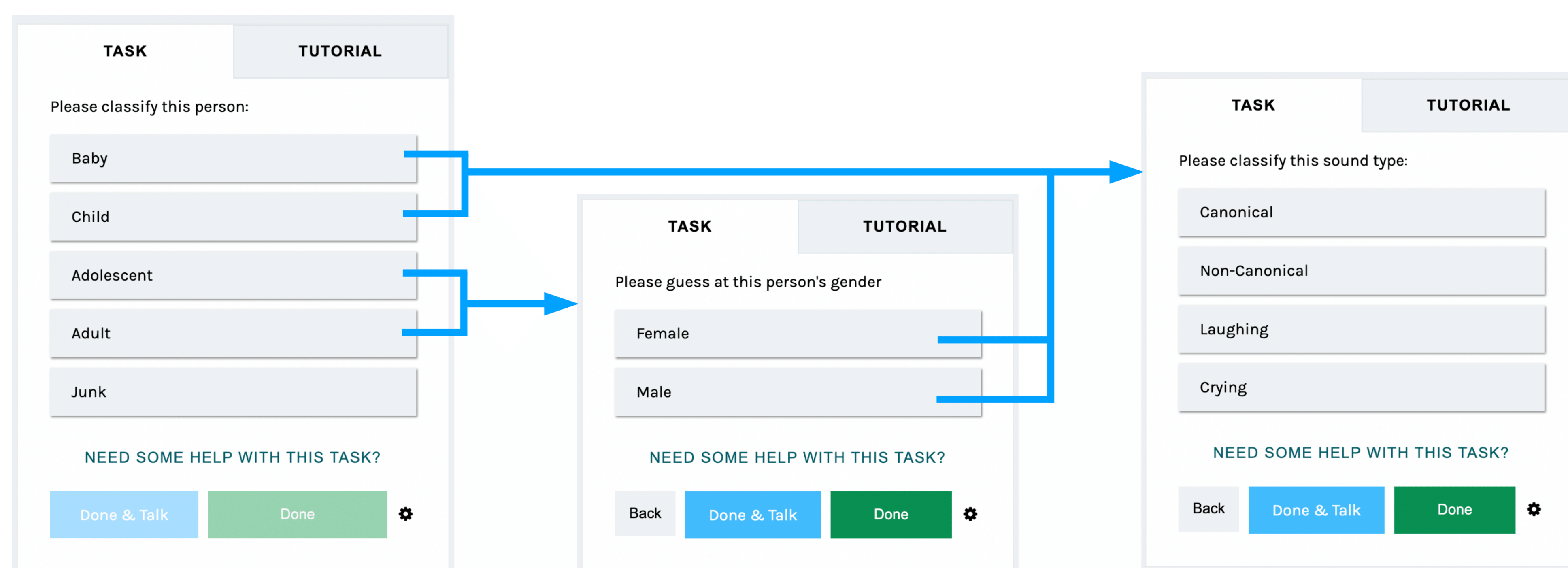


Figure 1: Zooniverse Pipeline

4. Metadata & Use Cases

- Wealth of metadata
 - Age
 - Sex
 - Linguistic Environment
 - Normativity
- Use Cases
 - Canonical/Linguistic Proportion
 - Train vocalisation-type classifiers

5. Preliminary analysis

- Canonical (CP) and Linguistic proportion (LP)

$$CP = \frac{\text{Canonical}}{\text{Canonical} + \text{Non-Canonical}}$$

$$LP = \frac{\text{Canonical} + \text{Non-Canonical}}{\text{Canonical} + \text{Non-Canonical} + \text{Cry} + \text{Laugh}}$$

- Linear Mixed-Effect Model

- Predict LP and CP
- Fixed Effects: age, sex, monolingualism
- Random Effects: child ID nested in corpus
- Significant positive effect of age
- No significant effect of age or monolingualism

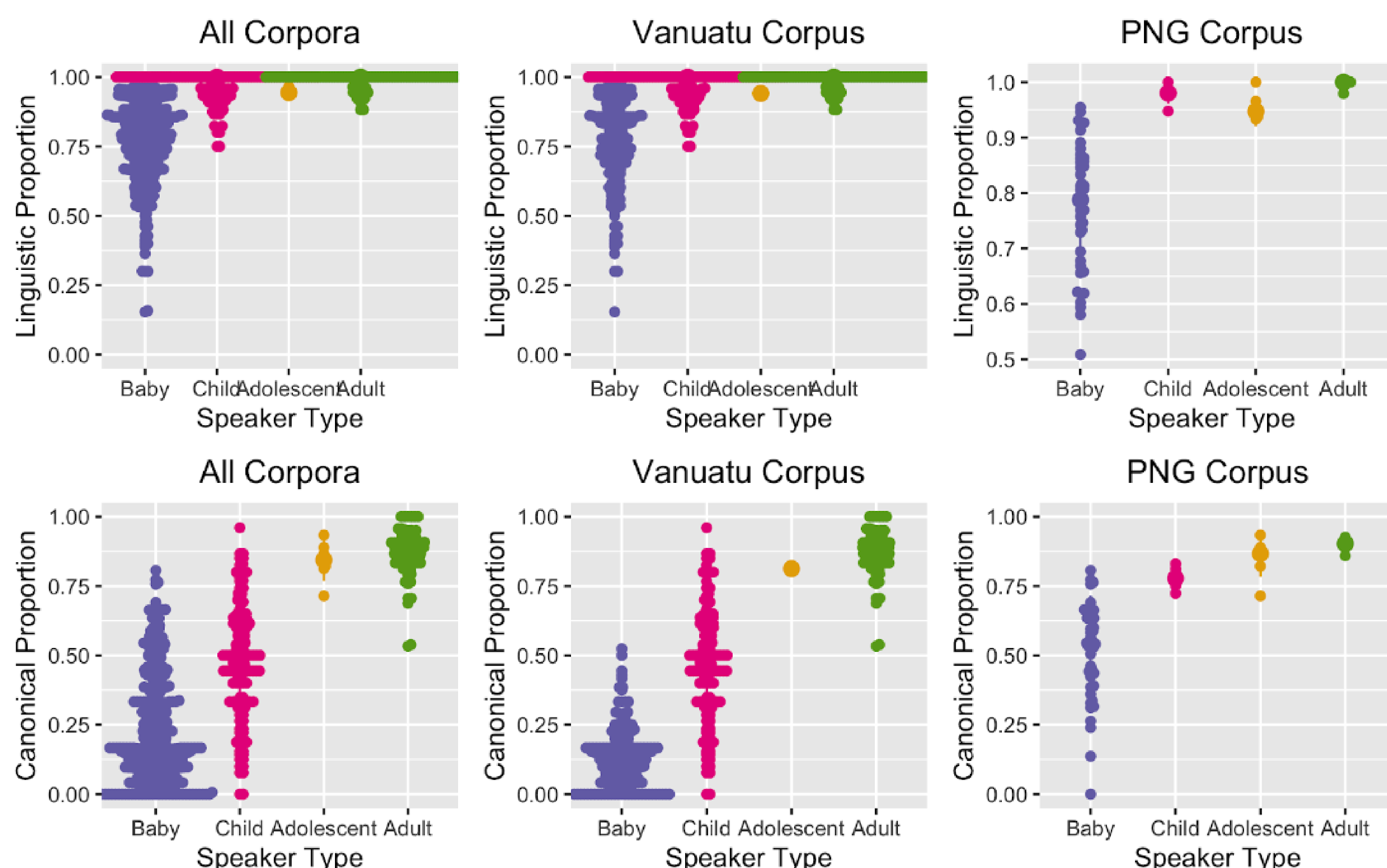


Figure 2: Linguistic proportions (top) and canonical proportions (bottom) by speaker age