Generating Feature Vectors from Phonetic Transcriptions in Cross-Liguistic Data Formats

Arne Rubehn[1], *Jessica Nieder*[1], *Robert Forkel*[2], *Johann-Mattis List*[1]

[1]Chair for Multilingual Computational Linguistics, University of Passau, Germany [2]DLCE, MPI-EVA, Leipzig, Germany

UNIVERSITÄT PASSAU

Introduction

tl;dr

We propose a new approach to dynamically generate phonological feature vectors for all sounds that are represented in a valid IPA notation.

Rationale

- Representing sounds as phonological feature vectors can enhance a wide range of tasks in CL and NLP
- Current approaches define feature representations over *fixed sets of sounds*
- With constantly increasing amounts of cross-linguistic data, unseen sounds are encountered more frequently (Moran, 2012)

👉 **Need for a more flexible and robust system that can analyze unseen sounds!**

Materials and Methods

Materials

- We use the *Cross-Linguistic Transcription System* (CLTS; Anderson et al., 2018) to robustly analyze and parse sounds
- Analyses on distinctiveness are performed on data aggregated in *Lexibank 1.0* (List et al., 2023)

Feature System

- Inventory of 39 fairly standard binary features
- 25 vocalic and consonantal features from Zsiga (2013)
- Extended by 14 complementary features for representing diphthongs, clicks, and tones (from Mortensen et al., 2016 and Rubehn, 2022)

Workflow

- Parse sounds using CLTS
- Retrieve canonical IPA description of sounds
- Map descriptive features onto binary features
  - *e.g.* 'fricative': [-son, +cont]
- Hierarchical mapping ensures correct handling of complex sounds

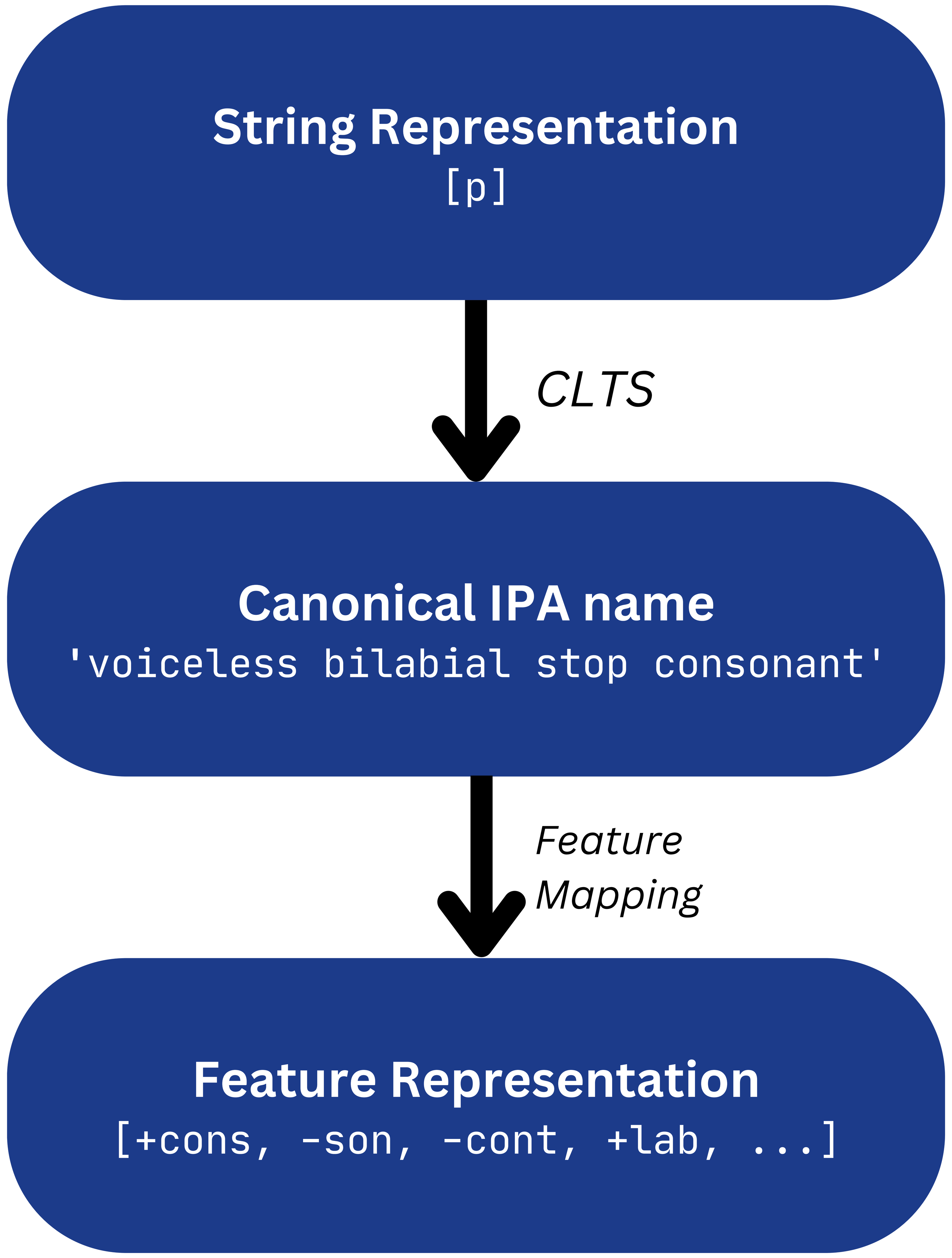

Figure 1: Workflow of vector creation.

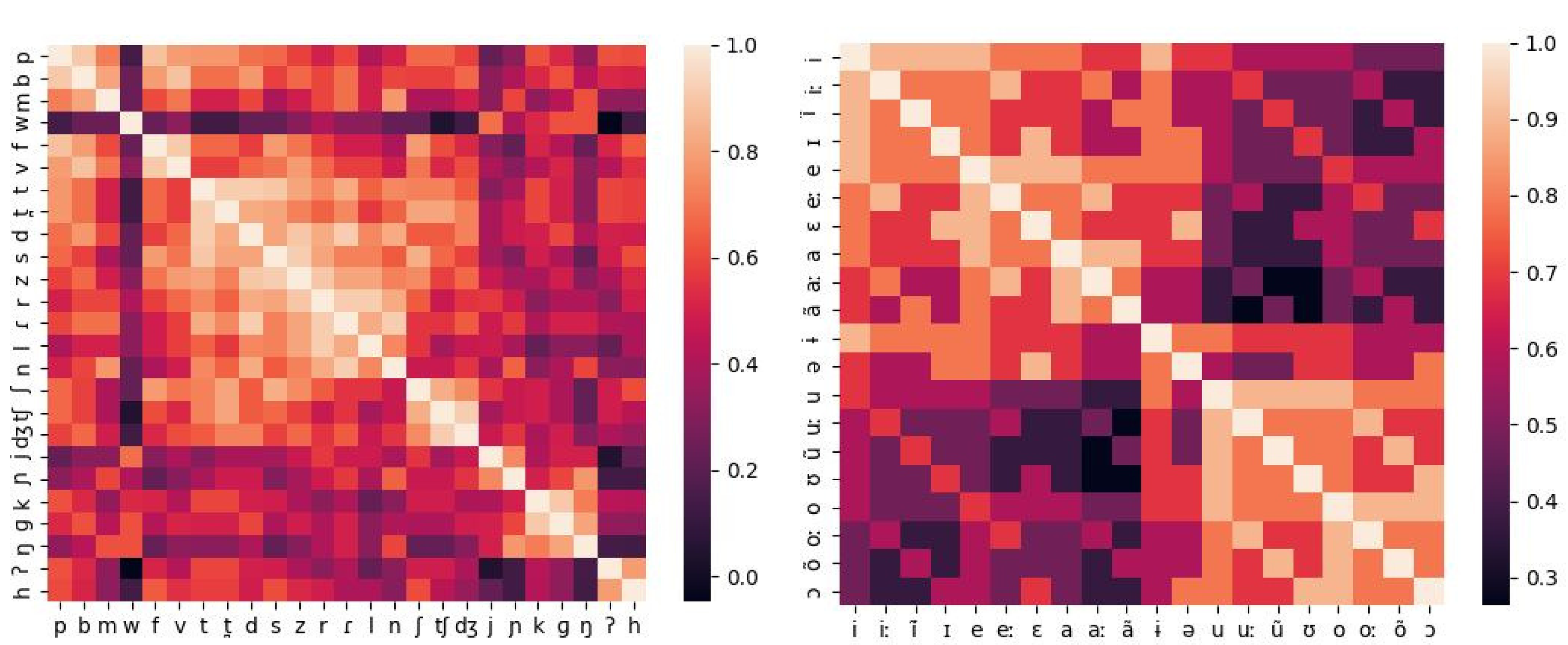

Figure 2: Cosine similarities between consonant (left) and vowel (right) vectors generated with our model.

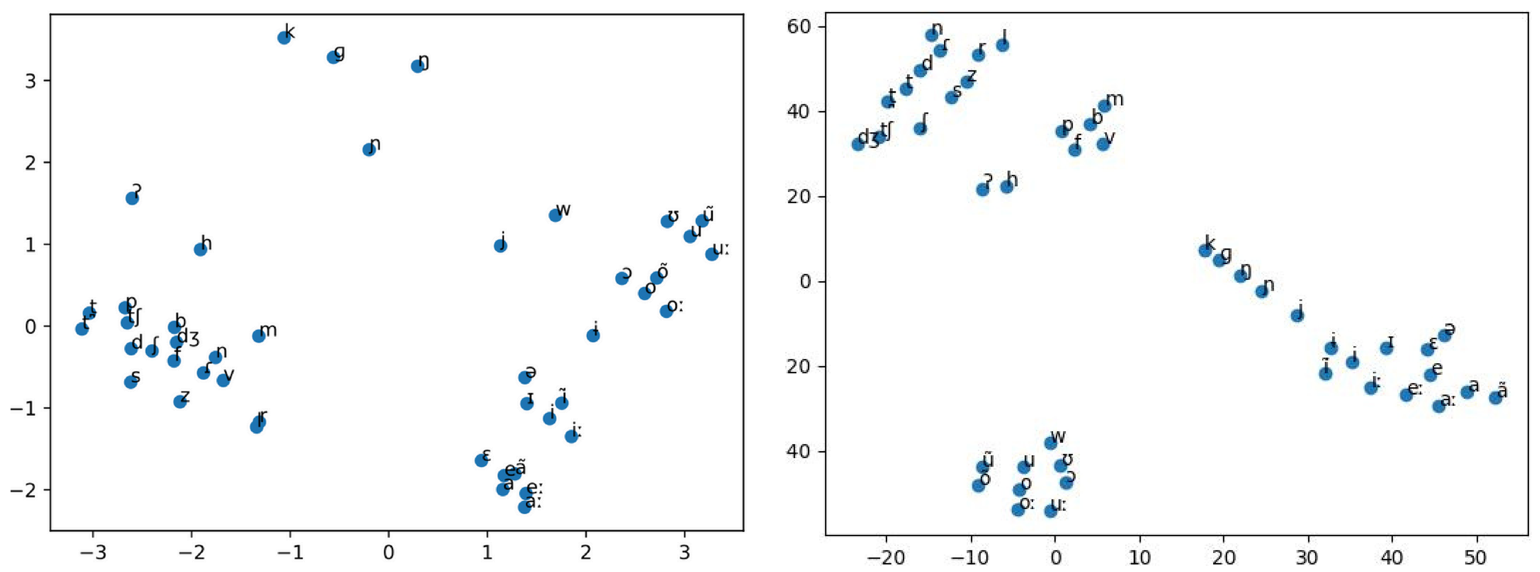

Figure 3: Two-dimensional reduction of feature vectors using PCA (left) and t-SNE (right).

n confused sounds	*n* varieties	Portion
0	2,376	0.818
≤1	2,567	0.884
≤2	2,648	0.912
≤3	2,689	0.926
≤4	2,841	0.978

Table 1: Number of language varieties in Lexibank 1.0 with at most *n* confused sounds.

annoy	e	ɱ	f a ð a ɾ
be annoyed	e	ɱ	f a ð a ɾ s e
be ill	e s t a ɾ + e	ɱ	f e ɾ m o
fall ill	k a e ɾ + e	ɱ	f e ɾ m o
illness	e	ɱ	f e ɾ m e ð a ð
sick, ill	e	ɱ	f e ɾ m o
December	d i θ j e	m	b ɾ e
March		m	a ɾ θ o
May		m	a ʝ o
November	n o β j e	m	b ɾ e
September	s e β t j e	m	b ɾ e
Sunday	d o	m	i ŋ g o

Figure 4: Concordance line for Spanish transcriptions featuring [ɱ] or [m].

Results

Evaluation

- Figures show that similarity patterns between vectors align with established phonological classifications
- Feature representations are highly disctinctive on Lexibank
- Distinctions that are lost can be mostly explained by allophonic variation

Code and Data

```
pip install soundvectors
```


References: Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., and List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting,* 4(1):21-53. · List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., Gray, R. D. (2023). *Lexibank [Database, Version 1.0].* Max Planck Institute for Evolutionary Anthropology, Leipzig. · Moran, S. (2012). *Phonetics Information Base and Lexicon.* PhD, University of Washington. · Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (2016). Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING2016, the 26th International Conference on Computational Linguistics: Technical Papers,* pages 3475-3484. · Rubehn, A. (2022). A feature-based neural model of sound change informed by global lexicostatistical data. Master's thesis, University of Tübingen. · Zsiga, E. (2013). *The Sounds of Language: An Introduction to Phonetics and Phonology,* volume 7. John Wiley & Sons.

Acknowledgements: This project was supported by the ERC Consolidator Grant ProduSemy (PI Johann-Mattis List, Grant No. Grant No. 101044282, see https://doi.org/10.3030/101044282). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

European Research Council
Established by the European Commission