A method for learning features from observed phonological classes

Connor Mayer[1] and Robert Daland[2]

[1]UCLA [2]Independent

Annual Conference of the Canadian Linguistics Association
University of British Columbia
June 3, 2019

Overview

This talk describes a fairly formal/computational paper.[1]

My goal is to **ignore as much of this formalism as possible!**

Instead, I hope to:

- Convince you that the theoretical approach is worthwhile.
- Give you a sense of how the computational bits work.
- Convince you that this is a useful tool, even if you're not a computationalist!
    - There is a freely available Python implementation of the algorithms I describe.

[1] Mayer and Daland resubmitted

Where do features come from?

Classic texts propose features are **universal**.[2]

- All languages can be described by the same finite set of features.
- Reflect phonetic properties of the vocal tract and perceptual system that facilitate categorical distinctions.
    - And not just in humans.[3]
- Phonological processes operate on the classes they define.

[2] e.g., Chomsky and Halle 1968
[3] e.g., Kuhl and Miller 1975

Phonetically disparate classes

Many classes **cannot be defined by a single set of phonetic properties**.

- **Classic example**: Sanskrit *ruki*.[4]
  - /s/ becomes retroflexed following {r, u, k, i}
- **Recent example**: Cochabamba Quechua.
  - Gallagher (2019) shows that /ʁ/ patterns as a voiceless stop.
- Mielke (2008) collected many such cases from grammars.
  - Only 71% of classes from 600 languages could be picked out by any feature system!

[4] e.g., Kiparsky 1973; Vennemann 1974

Alternative explanations for phonetic disparity

Maybe we need to expand our feature systems.

- e.g., /l/ seems to be [+continuant] in some languages and [–continuant] in others.[5]
- Should we add [midsagittal-continuant] and [parasagittal-continuant]?

It's unclear that this approach has explanatory value.

- How many additional features would we need?
- Could these features all be given a phonetic interpretation?
- How do learners determine which features are relevant for their language?

[5] e.g., Kaisse 2002; Mielke 2008

Learned features

Researchers have proposed that features may be **learned** and **language-specific**.[6]

- Feature systems are derived from perceived similarities between sounds.
- Typological patterns are by-products of general human cognitive capabilities, properties of human vocal tract and auditory system, channel bias, etc.

[6] e.g., Blevins 2004; Mielke 2008; Archangeli and Pulleyblank 2015; MacWhinney and O'Grady 2015; Archangeli and Pulleyblank 2018

Learned vs. universal features

Learned features are compatible with the basic motivation for universal feature theory!

- Certain phonetic properties are naturally salient.

A few additional assumptions:

- The *distribution* of sounds can inform their features.
    - Phonotactic properties, conditioning of processes, etc.
- No one dimension has primacy.
    - The learner uses the full range of available information.
- Features may vary cross-linguistically.

What do we gain from emergent features?

With universal features, we start from **typology** and then worry about exceptions.

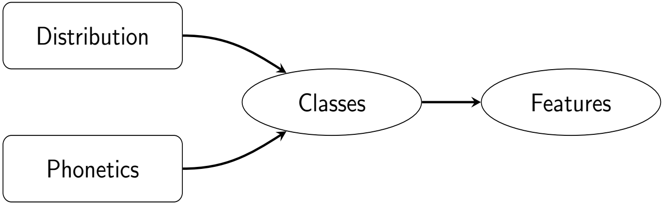
With emergent features, we focus on **learning mechanisms**.

- What are these mechanisms?
- How do they contribute to typological patterns?
- To what extent can distribution contribute to feature learning?[7]

[7] e.g., Moreton 2008; Hayes et al. 2009

Assumptions of our model

We assume a model of feature learning like below.



- The learner has converged on a segmental representation.[8]
- Some mechanism has identified a set of input classes.
  - Based on acoustic, articulatory, distributional similarity, etc.
- A feature system is derived from the set of input classes.

[8] e.g., Lin 2005; Feldman et al. 2013

Why start from classes?

It's unclear how features can be learned without being motivated by the classes they characterize.

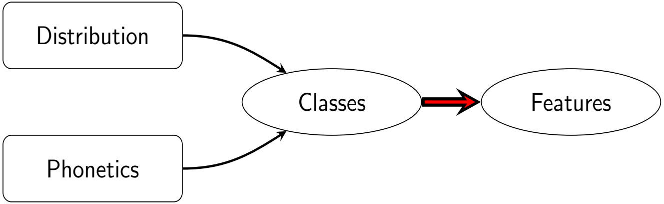
Past approaches to learning phonological categories learn **classes**.

- From phonetic data.[9]
- From distribution.[10]

Class learning likely involves integration of multiple such sources of information.

[9] e.g., Lin 2005; Mielke 2012
[10] e.g., Calderone 2009; Goldsmith and Xanthos 2009; Mayer submitted

Scope of this work



We focus on how to get from classes to features.

Terminology

A **class system** is a set of classes.

A **feature system** is a set of features over a segmental inventory, and the values they can take.

- Features are **functions** that map segments to values.

- **V**owels
- **G**lides
- **L**iquids
- **N**asals
- Obstruen**T**s

σ	syl	cons	apprx	son
V	+	−	+	+
G	−	−	+	+
L	−	+	+	+
N	−	+	−	+
T	−	+	−	−

Featural descriptors

A **featural descriptor** is a set of feature/value pairs.

- *Intensional* description of a class.

$$\begin{bmatrix} +\text{son} \\ +\text{apprx} \end{bmatrix}$$

- Its *extension* is a set of segments.

$$\left\langle \begin{bmatrix} +\text{son} \\ +\text{apprx} \end{bmatrix} \right\rangle = \{\text{V, G, L}\}$$

σ	syl	cons	apprx	son
V	+	−	+	+
G	−	−	+	+
L	−	+	+	+
N	−	+	−	+
T	−	+	−	−

A feature system **covers** a class system if there is (at least) one unique featural descriptor for (at least) every class.

Combining featural descriptors

Intersection of classes corresponds to **union** of featural descriptors.

$$\langle\ [+\text{son}]\ \rangle = \{\text{V, G, L, N}\}$$

$$\langle\ [+\text{cons}]\ \rangle = \{\text{L, N, T}\}$$

$$\left\langle\ \begin{bmatrix} +\text{son} \\ +\text{cons} \end{bmatrix}\ \right\rangle = \{\text{L, N}\} = \{\text{V, G, L, N}\} \cap \{\text{L, N, T}\}$$

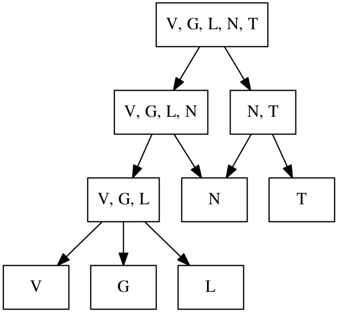
σ	syl	cons	apprx	son
V	+	−	+	+
G	−	−	+	+
L	−	+	+	+
N	−	+	−	+
T	−	+	−	−

Parent/child relationships

Class systems are **hierarchical**.

Parent/child relationships are one way of expressing this hierarchy.

- **Crucial** for deriving feature systems.

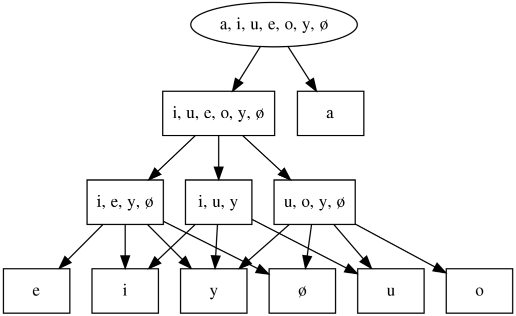


Parent/child relationships

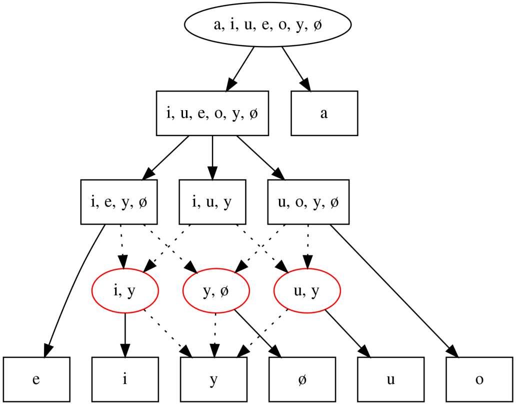
X is a parent of Y iff Y is a subset of X, and no class intervenes between the two.

Intersectional closure

The **intersectional closure** of a class system is the set of all intersections of its classes.



Input



Intersectional closure of input

Bringing it all together

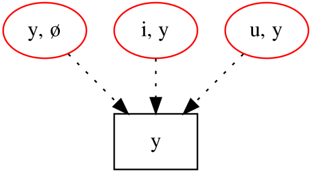
Intersectional closure covering theorem

If a feature system covers a class system, it also covers its intersectional closure.

- e.g., if you have [+high] and[+front], you can't help getting $\begin{bmatrix} +\text{high} \\ +\text{front} \end{bmatrix}$.

Multiple parenthood theorem

If a class in the intersectional closure has more than two parents, it is *exactly equal* to the intersection of any two of its parents.



- This entails that **any class with more than one parent can be uniquely identified by the union of the features of any two of its parents!**

Learning features from classes

Now we can turn to the main question:

How do we learn features from classes?

Central insight

We need a feature/value pair for every class that has a single parent in the intersectional closure.

- If a class has no parents, it's the segmental inventory.
- If a class has more than one parent, it can be picked out by the union of its parents' features.

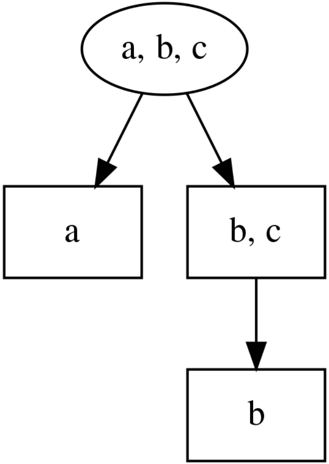
Algorithms for learning features from classes

We will use a toy class system for expository purposes.

- A more complex example is provided later.

This class system:

- is intersectionally closed.
- does not contain all singleton classes.



Privative specification

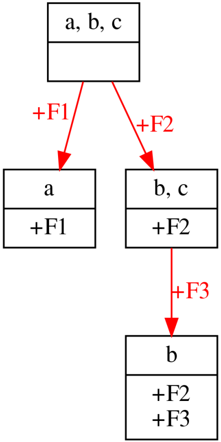
Privative specification algorithm

Assign a new $+f$ feature/value pair to all segments in each class that has exactly one parent.

Assigns **privative** values: $\{+, 0\}$

Compatible with theories that consider all features privative.[11]

σ	F1	F2	F3
a	$+$	0	0
b	0	$+$	$+$
c	0	$+$	0



[11] e.g., Anderson and Ewen 1987; Avery and Rice 1989; Lahiri and Marslen-Wilson 1991; Frisch 1996

Complementary specification

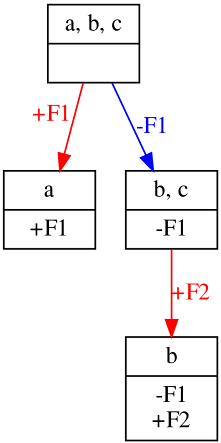
Complementary specification algorithm

Assign a new $+f$ feature/value pair to each class that has exactly one parent, and a $-f$ feature/value pair to *the complement of that class with respect to its parent* if it is present in the input.

Assigns **contrastive** values: $\{+, -, 0\}$

There are theoretical reasons to allow '$-$' feature values.[12]

σ	F1	F2
a	$+$	0
b	$-$	$+$
c	$-$	0



[12] e.g., Archangeli and Pulleyblank 1994; Archangeli 2011

Inferential complementary specification

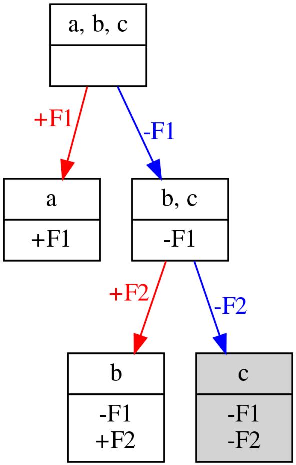
Inferential complementary specification algorithm

Assign a new $+f$ feature/value pair to each class that has exactly one parent, and a $-f$ feature/value pair to the complement of that class with respect to its parent *even if it is not present in the input.*

Assigns **contrastive** values: $\{+, -, 0\}$

Assumes limited generalization based on input classes.

σ	F1	F2
a	+	0
b	-	+
c	-	-



Full specification

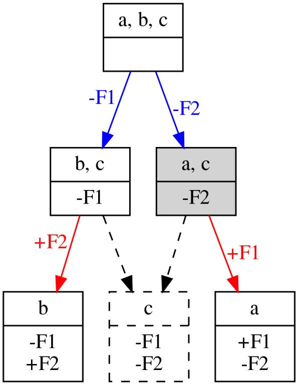
Inferential complementary specification algorithm

Assign a new $+f$ feature/value pair to each class that has exactly one parent, and a $-f$ feature/value pair to the complement of that class with respect to the *full segmental inventory* even if it is not present in the input.

Assigns **full** values: $\{+, -\}$

Prohibits underspecification.[13]

σ	F1	F2
a	+	-
b	-	+
c	-	-

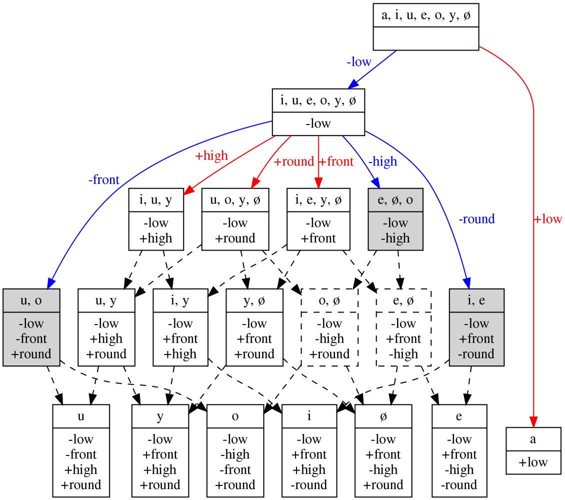


[13]e.g., Chomsky and Halle 1968

More realistic input

alphabet	{a, i, u, e, o, y, ø}
non-low	{i, u, e, o, y, ø}
high	{i, u, y}
front	{i, e, y, ø}
round	{u, o, y, ø}
singletons	{a}, {i}, {u}, {e}, {o}, {y}, {ø}

σ	low	front	high	round
a	+	0	0	0
i	−	+	+	−
u	−	−	+	+
e	−	+	−	−
o	−	−	−	+
y	−	+	+	+
ø	−	+	−	+



Summary

A symbolic feature system can be learned from a set of input classes.

- Done without reference to phonetic properties.

We presented four algorithms that differ in assumptions about

- what feature values are permitted.
- whether there is generalization from the input classes.

They operate based on insights into the hierarchical structure of class systems.

Applications to future research

We freely provide the code for use and extension in future research.

Input:

- Set of classes.
- Choice of featurization algorithm.

Output:

- A feature system that covers those classes.

Can be used as a component in phonological learning systems.

- E.g., systems that take us from a data set to a phonological grammar.

Applications to future research

We can revisit cases of **phonetically disparate classes**.

- Understand the feature systems that underpin these.

We can think more about **underspecification**.

- These algorithms *deterministically* predict underspecification.
- Testable against the literature and artificial grammar learning[14] studies.
- Allow us to avoid "opportunistic underspecification."[15]
- Similar to contrastive hierarchies,[16] but with different inputs and less stipulation.

We can study how learners **generalize from the input**.

- Provides testable predictions about how learners might generalize across classes.
- Also testable with artificial grammar learning studies.

[14] e.g., Moreton and Pater 2012
[15] Steriade 1995
[16] Dresher 2003; Hall 2007

Acknowledgements

Thanks to Bruce Hayes, Kie Zuraw, Tim Hunter, and the members of the UCLA phonology seminar for their valuable questions and insights. This research was supported by a Social Sciences and Humanities Research Council of Canada Doctoral Award to the first author.

See **connormayer.com** for the full paper and code.

References I

Anderson, J. M. & Ewen, C. J. (1987). *Principles of Dependency Phonology*. Cambridge: Cambridge University Press.

Archangeli, D. (2011). Feature specification and underspecification. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 148–170). Oxford: Wiley-Blackwell.

Archangeli, D. & Pulleyblank, D. (1994). *Grounded phonology*. Cambridge, MA: MIT Press.

Archangeli, D. & Pulleyblank, D. (2015). Phonology without universal grammar. *Frontiers in Psychology*, *6*, 1229.

Archangeli, D. & Pulleyblank, D. (2018). Phonology as an emergent system. In S. Hannahs & A. Bosch (Eds.), *The Routledge Handbook of Phonological Theory* (pp. 476–503). London: Routledge.

Avery, P. & Rice, K. (1989). Segment structure and coronal underspecification. *Phonology*, *6*, 179–200.

References II

Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Calderone, B. (2009). Learning phonological categories by independent component analysis. *Journal of Quantitative Linguistics*, *16*(2), 132–156.

Chomsky, N. & Halle, M. (1968). *The sound pattern of english*. New York: Harper & Row.

Dresher, E. (2003). The contrastive hierarchy in phonology. In D. C. Hall (Ed.), *Toronto Working Papers in Linguistics 20: Special issue on contrast in phonology*. University of Toronto.

Feldman, N., Griffiths, T., Goldwater, S., & Morgan, J. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Frisch, S. A. (1996). *Similarity and frequency in phonology*. (Doctoral dissertation, Northwestern University).

References III

- Gallagher, G. (2019). Phonotactic knowledge and phonetically unnatural classes: The plain uvular in cochabamba quechua. *Phonology*, *36*, 37–60.
- Goldsmith, J. & Xanthos, A. (2009). Learning phonological categories. *Language*, *85*(1), 4–38.
- Hall, D. C. (2007). *The Role and Representation of Contrast in Phonological Theory*. (Doctoral dissertation, University of Toronto).
- Hayes, B., Zuraw, K., Siptar, P., & Londe, Z. (2009). Natural and unnatural constraints in hungarian vowel harmony. *Language*, *85*, 822–863.
- Kaisse, E. M. (2002). Laterals are [-continuant]. MS, University of Washington.
- Kiparsky, P. (1973). Phonological representations. In O. Fujimura (Ed.), *Three Dimensions of Linguistic Theory* (pp. 1–136). Tokyo: TEC Co.
- Kuhl, P. K. & Miller, J. D. (1975). Speech Perception by the Chinchilla: Voiced-Voiceless Distinction in Alveolar Plosive Consonants. *Science*, *190*, 69–72.

References IV

- Lahiri, A. & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, *38*(3), 245–294.
- Lin, Y. (2005). *Learning features and segments from waveforms: A statistical model of early phonological acquisiton*. (Doctoral dissertation, UCLA).
- MacWhinney, B. & O'Grady, W. (Eds.). (2015). *The Handbook of Language Emergence*. Chichester: John Wiley & Sons.
- Mayer, C. (submitted). An algorithm for learning phonological classes from distributional similarity.
- Mayer, C. & Daland, R. (resubmitted). A method for projecting features from observed sets of phonological classes.
- Mielke, J. (2008). *The emergence of distinctive features*. Oxford: Oxford University Press.
- Mielke, J. (2012). A phonetically-based metric of sound similarity. *Lingua*, *1222*, 145–163.

References V

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*, 83–128.

Moreton, E. & Pater, J. (2012). Structure and substance in artificial phonology learning. part i: Structure, part ii: Substance. *Language and Linguistics Compass*, *6*(11), 686-701 and 702-718.

Steriade, D. (1995). Markedness and underspecification. In J. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 114–175). Oxford/Cambridge, MA: Blackwell.

Vennemann, T. (1974). Sanskrit *ruki* and the concept of a natural class. *Linguistics*, *130*, 91–97.