Using Jaccard Distance to Measure the Linguistic I-Proximity of Phonological Inventories in a Contrastive Hierarchy

Dr. John Archibald
University of Victoria, Dept. of Linguistics

L3 Workshop 2022

KING'S College LONDON

Measuring Proximity

- Typological distance (Rothman, 2015)
- Structural similarity (Westergaard, 2021)
- Wholesale (Schwartz & Sprouse, 2021)
- Property-by-Property (Archibald, 2021)

- What the field lacks is a way of reliably measuring linguistic similarity or proximity.

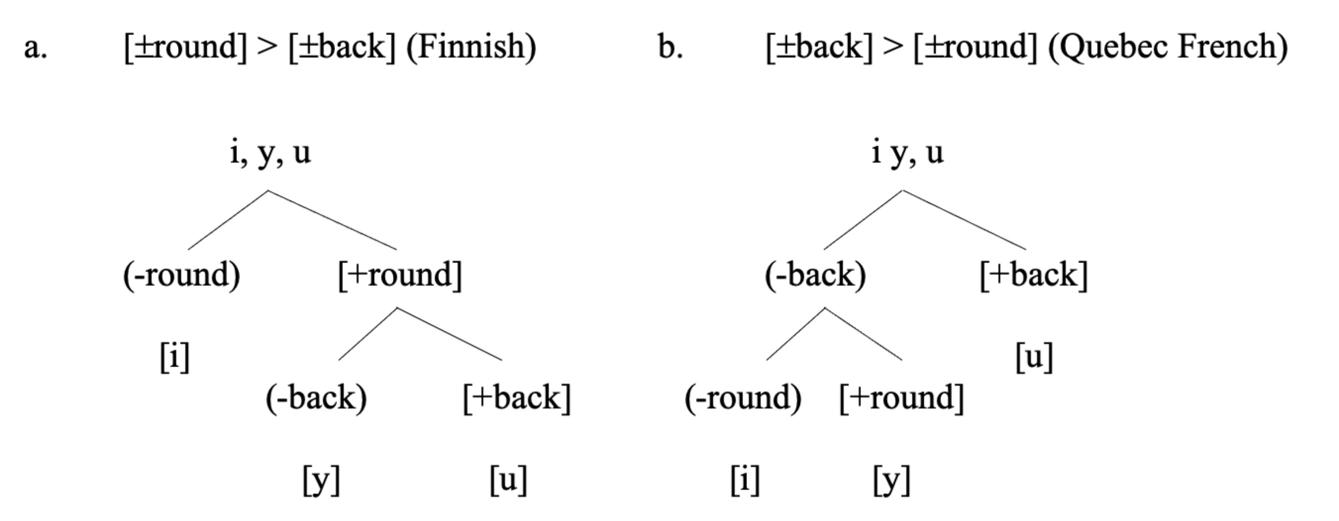
- In the phonetic domain, cross-linguistic comparisons proceed segment-by-segment (Flege & Bohn, 2021)
- much of L2 phonological research has demonstrated that L2/L3 phonology reveals *inventory* effects.

- In order to understand L2/L3 phonology, we need to look at the whole system (or inventory) not just individual vowels or consonants.

- Munro and Derwing (2008) showed that Mandarin learners of English vowels had trouble with the vowels [ɪ,ɛ,æ,ʌ,ʊ] vowels which form a natural class under feature theory.

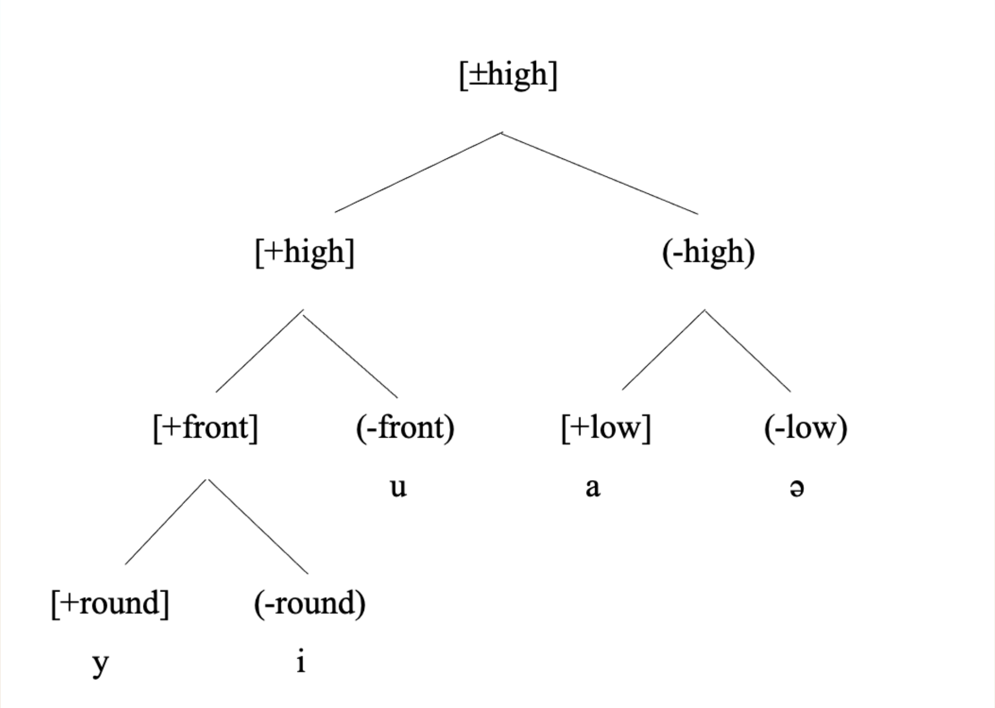
- Dresher’s (2009) Contrastive Hierarchy (CH) model of phonology is particularly well-suited to formalizing the notion of cross-linguistic similarity, and can be used productively to predict and explain the property-by-property transfer witnessed in L3 grammars.
- The CH has been used to successfully account for L1A (Bohn & Santos, 2018), and historical change (Oxford, 2015). It has also been used in the domain of morphosyntax (Cowper & Hall, 2019) and sociolinguistics (Natvig & Salmons, 2021).

- a 3-vowel system might have different underlying phonological structure in different languages.
- Finnish ranks the feature [round] above [back] while Quebec French ranks the feature [back] above [round].

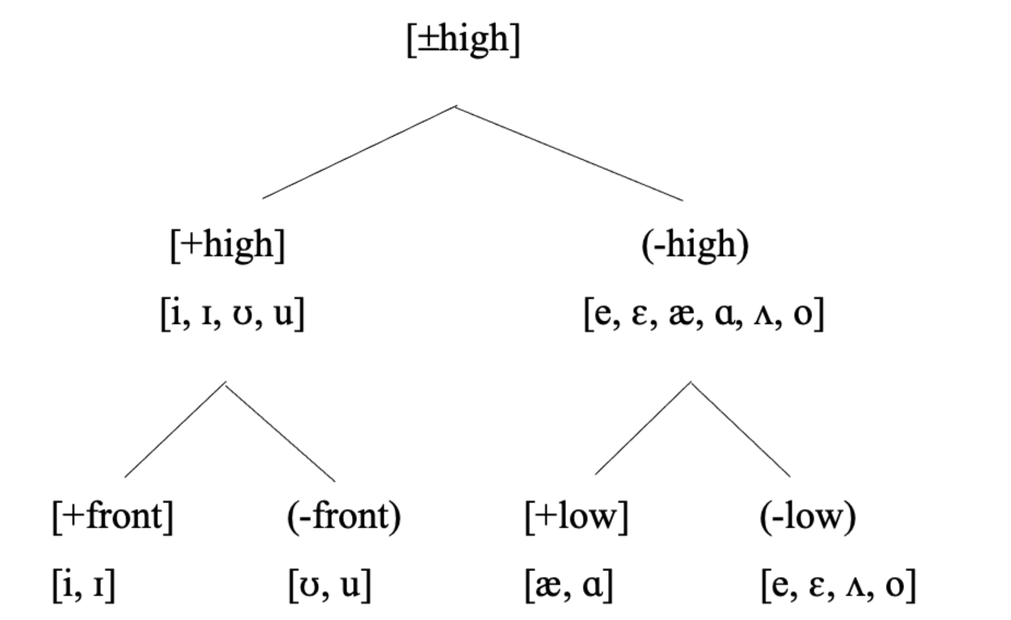


- In these models then a language is defined by both the features *and* their ranking. Using this type of model, we can explain the inventory effects such as Munro & Derwing (2008).

- Following Wu (2021) the CH for Mandarin vowels is given in Figure 2.



- If we apply these L1 features to English vowels we get the following parse:



- Note that the feature hierarchy cannot uniquely define the vowels [ɪ, ɛ, æ, ʌ, ʊ]; an inventory effect explained by phonological features.
- But what the field needs is a way to compare *inventories* (or hierarchies) such as English versus Mandarin.

- I explore using Jaccard Distance (Purnell, Raimy & Salmons, 2019) to do so. Jaccard Distance is a common way to compare the members of sets (Matthe et al. 2006). The formula is shown below:

$$d_J = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} = 1 - J(A, B)$$

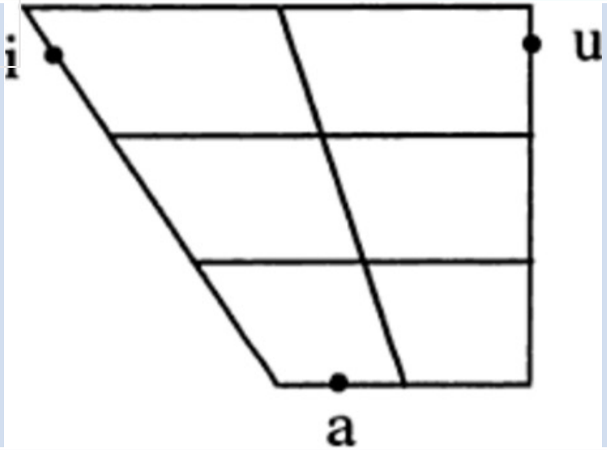
The numerator is the union minus the intersection while the denominator is the set union

- If both sets are identical then the Jaccard distance equals 0
- If there are no common elements then Jaccard distance equals 1

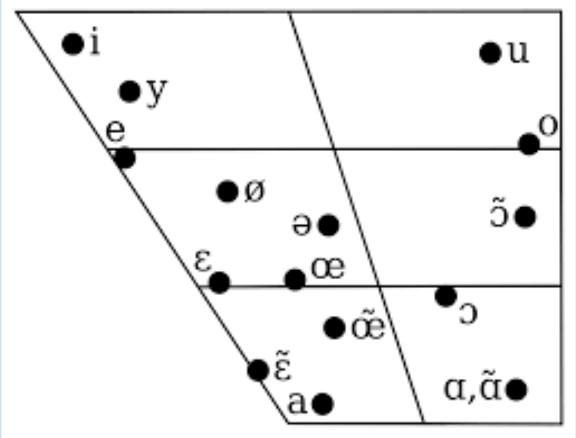
Four Vowel Inventories

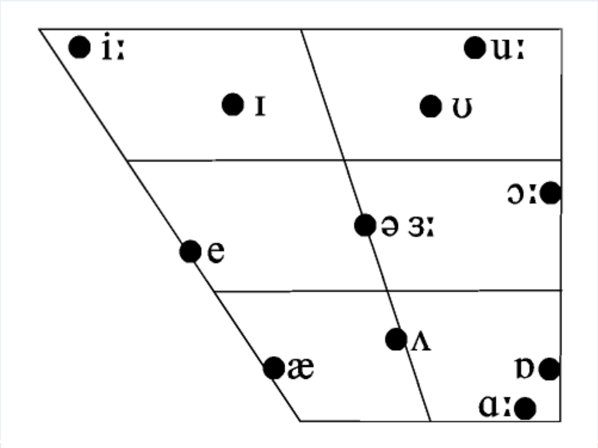
- Arabic
- French
- English
- Mandarin

Arabic

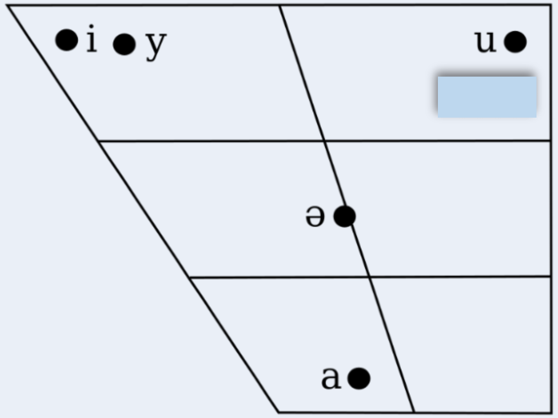


French





English



Mandarin

- So which inventories are most similar?

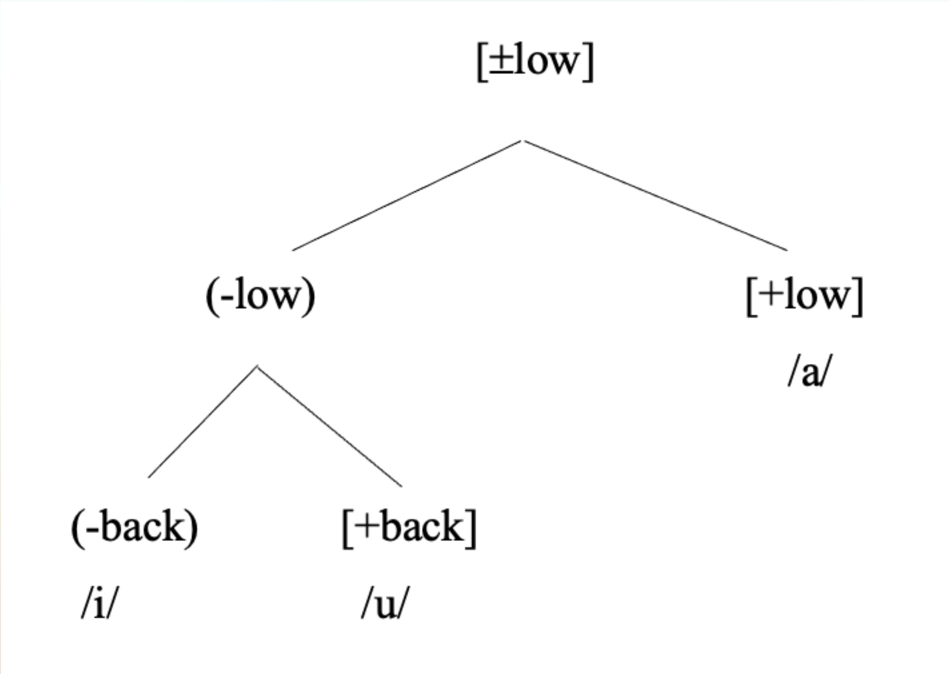
- Archibald (2022ab) reanalyzed Benrabah’s (1991) data to explain why learners transferred French vowels (and not Arabic vowels) into their L3 English.
- Jaccard Distance allows us a way to formalize these comparisons (with Mandarin added just for fun).
- Identical = 0.

Languages	Distance
Arabic:English	(11-1)/11= .9
French: English	(24-9)/24= .6
Mandarin: English	(17-3)17= .8

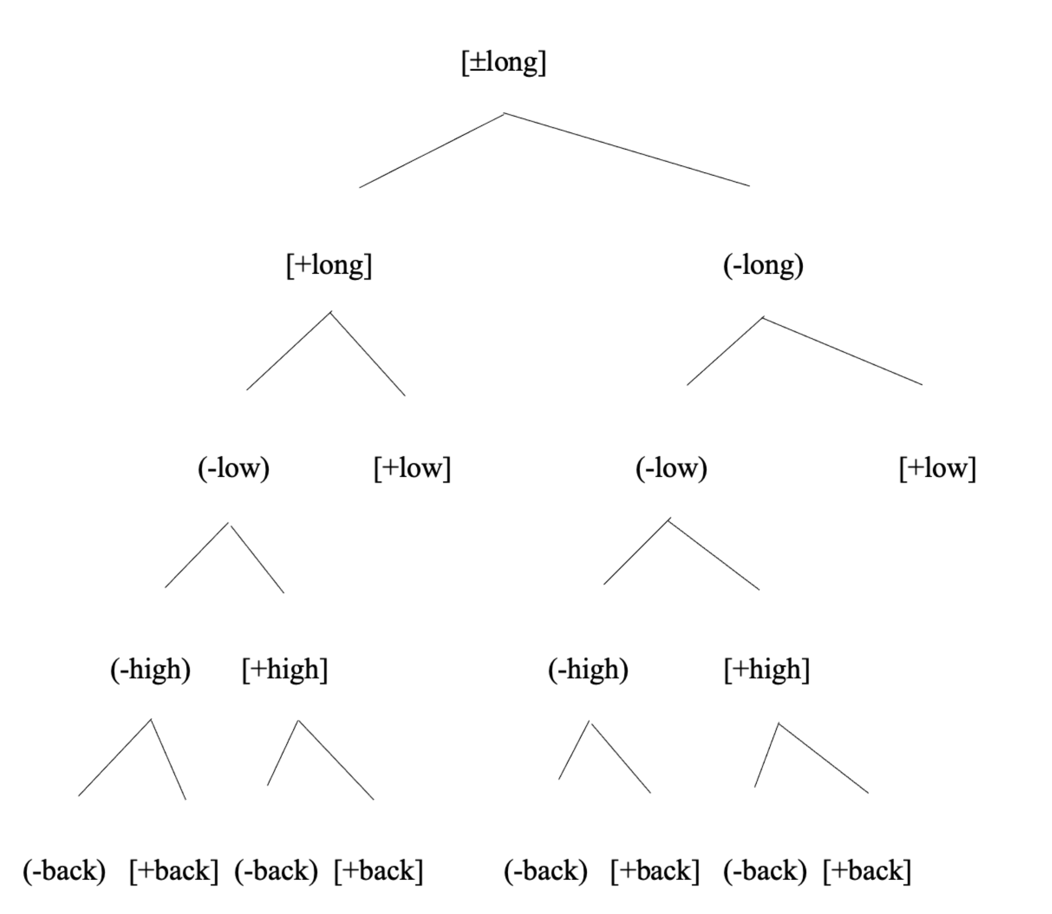
- With respect to the vocalic domain, French is the closest to English, then Mandarin, then Arabic.
- Jaccard Distance involves comparing *sets* not *members* of sets and thus allows us to compare phonological inventories (and explain bilingual inventory effects) as well as explain the property-by-property transfer shown in Archibald (2022).

- I investigate whether Jaccard Distance is a plausible way to calculate linguistic I- proximity (as it is based on internal representations) and will discuss whether this is a feasible mechanism to model actual L3 learner behaviour.

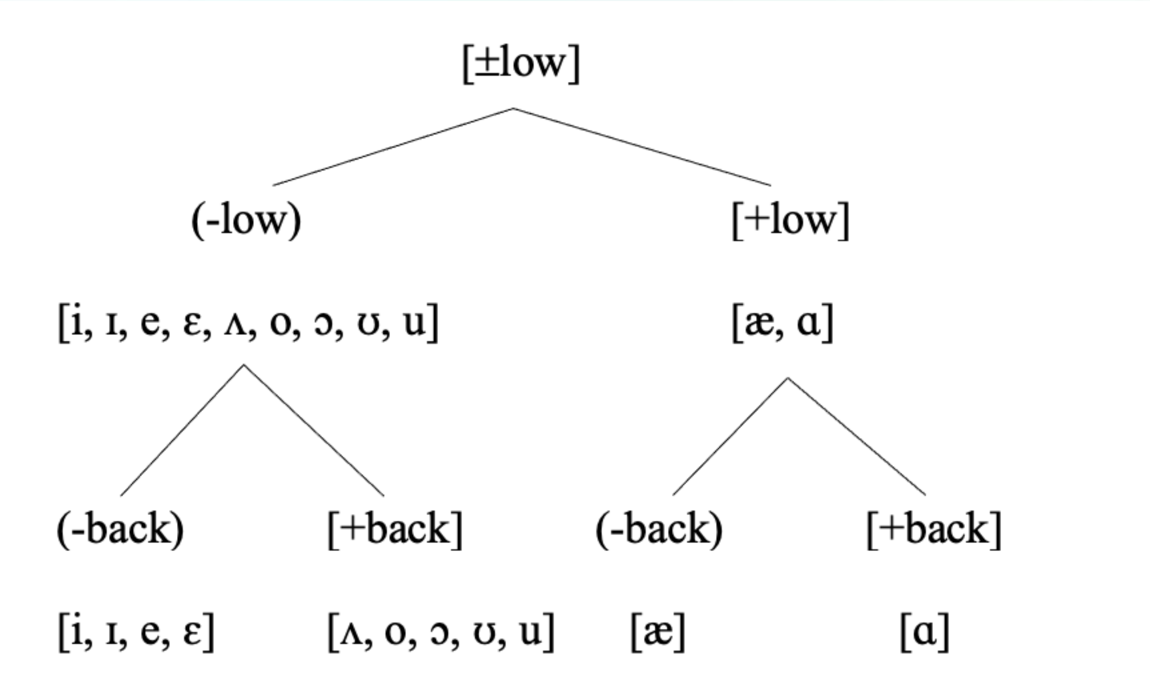
Arabic Hierarchy



French Hierarchy

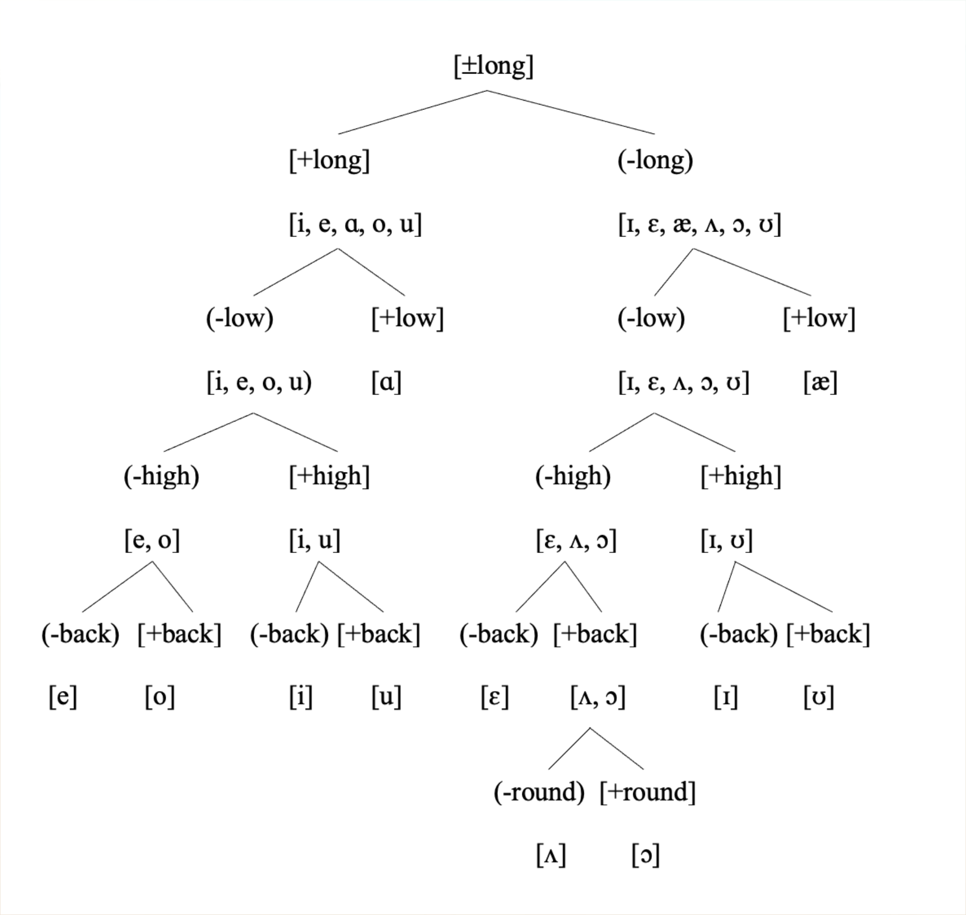


Arabic Parse of English Input



9 vowels cannot be uniquely parsed

French Parse of English Input



All vowels are successfully parsed, though, perhaps in a non-nativelike fashion.

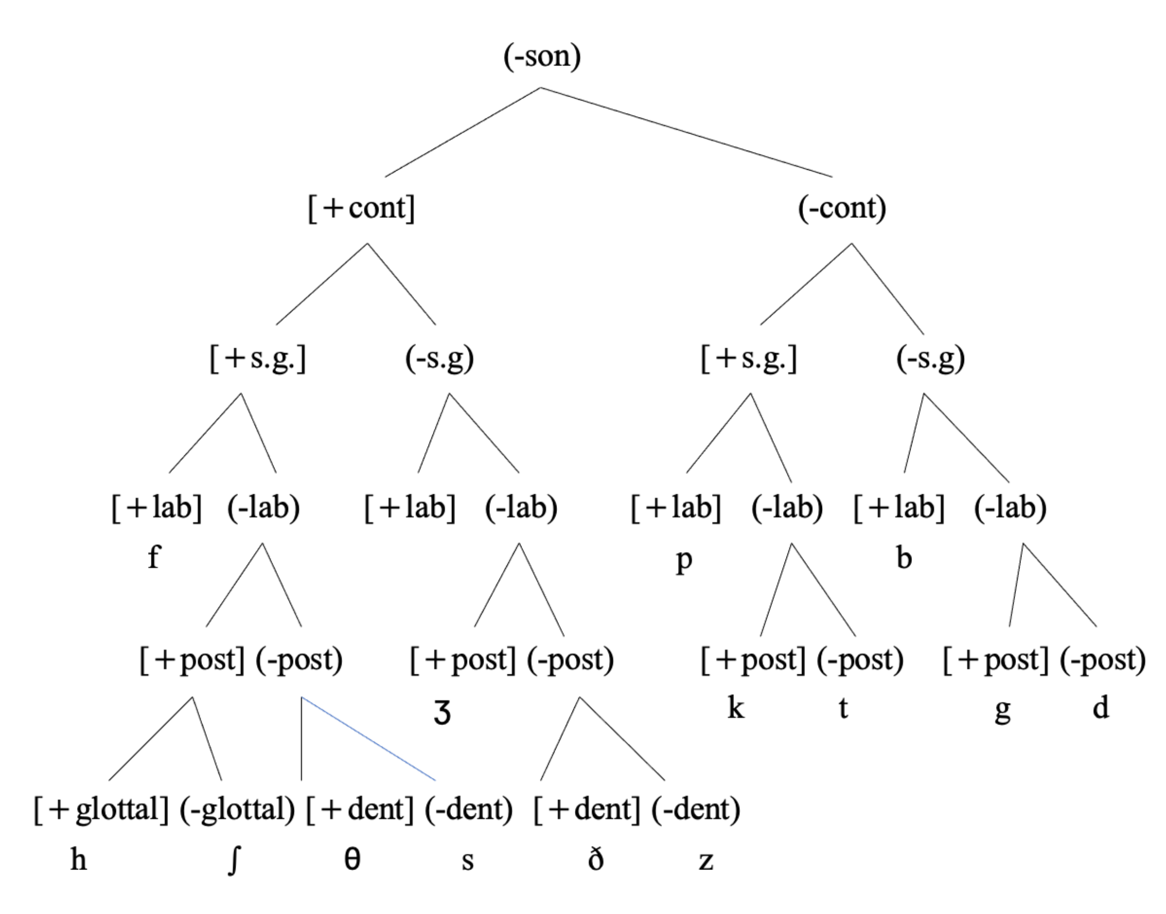
Rankings for Jaccard Distance: Vowels

French Rankings	English Rankings
nasal > long	
nasal > low	
nasal > high	
nasal > back	
nasal > round	
long > low	**long > low**
long > high	long > front
long > back	**long > high**
long > round	**long > round**
low > high	low > front
low > back	**low > high**
low > round	low > round

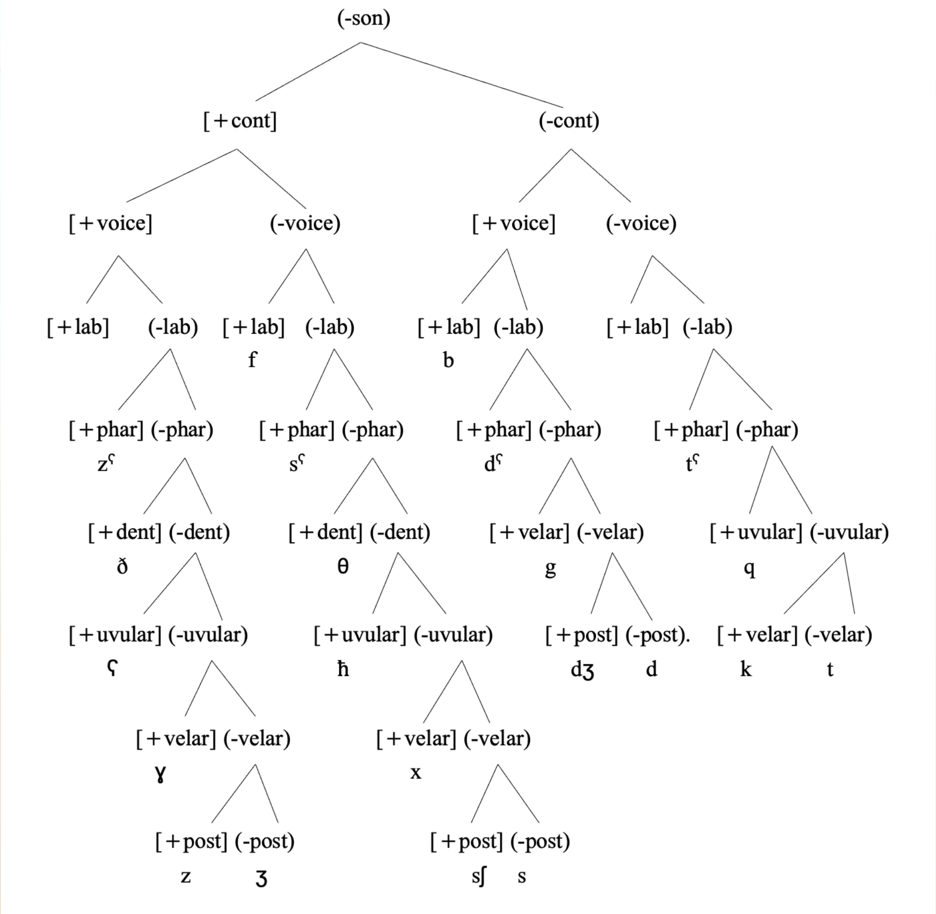
- In this case the parsing test and the Jaccard distance both point to French vowels being a better fit to English vowels
- But what about consonants?

- Ultimately I will argue that we can’t rely solely on Jaccard distance but need to supplement it with a notion of phonological parsing.

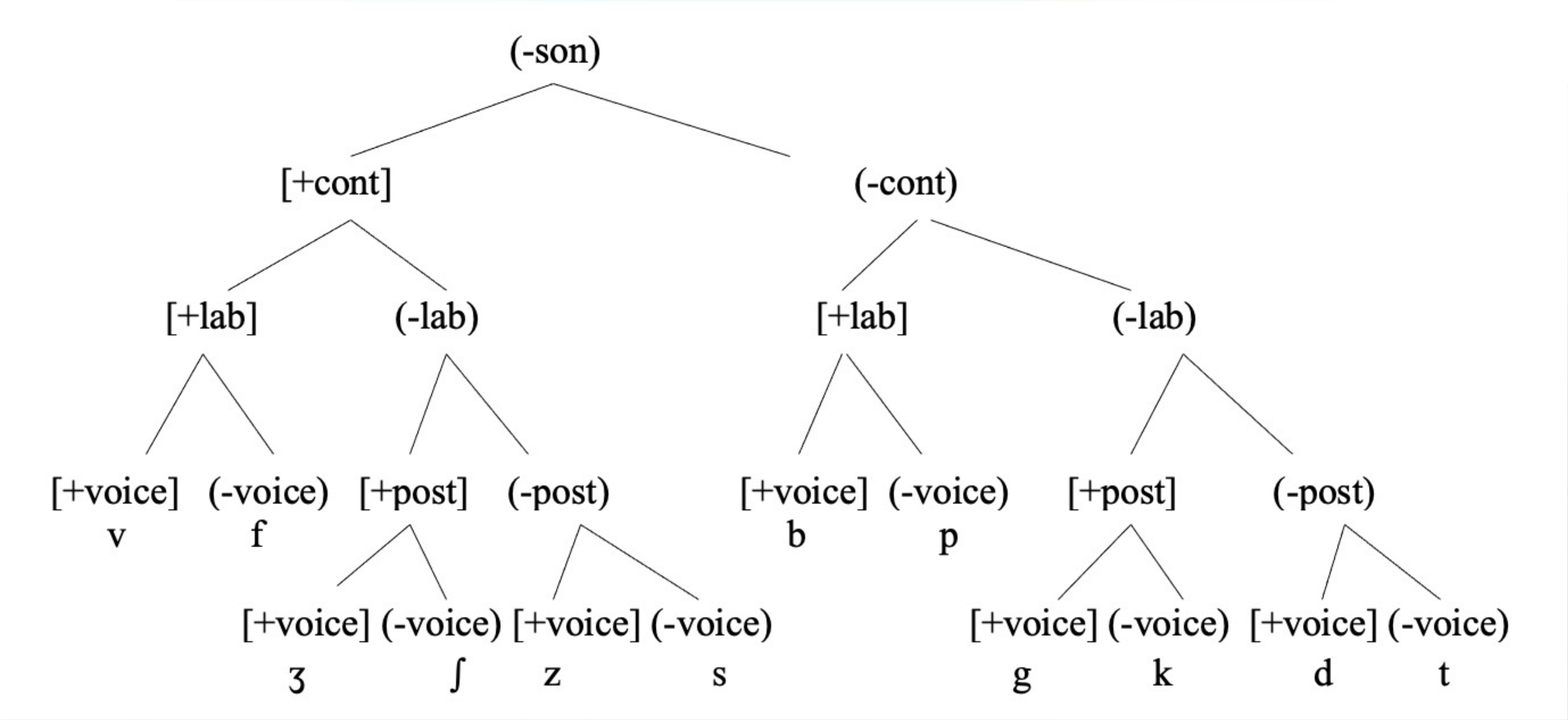
English Obstruents



Arabic Consonants



French Consonants



Rankings for Jaccard Distance: Consonants

French Rankings	English Rankings	Arabic Rankings
continuant> labial	continuant > spread glottis	continuant > voice
continuant > posterior	**continuant > labial**	**continuant > labial**
continuant > voice	**continuant > posterior**	continuant > pharyngeal
labial > posterior	spread glottis > labial	continuant > dental
labial > voice	spread glottis > posterior	continuant > uvular
posterior > voice	**labial > posterior**	continuant > velar
		continuant > posterior
		voice > labial
		voice > pharyngeal

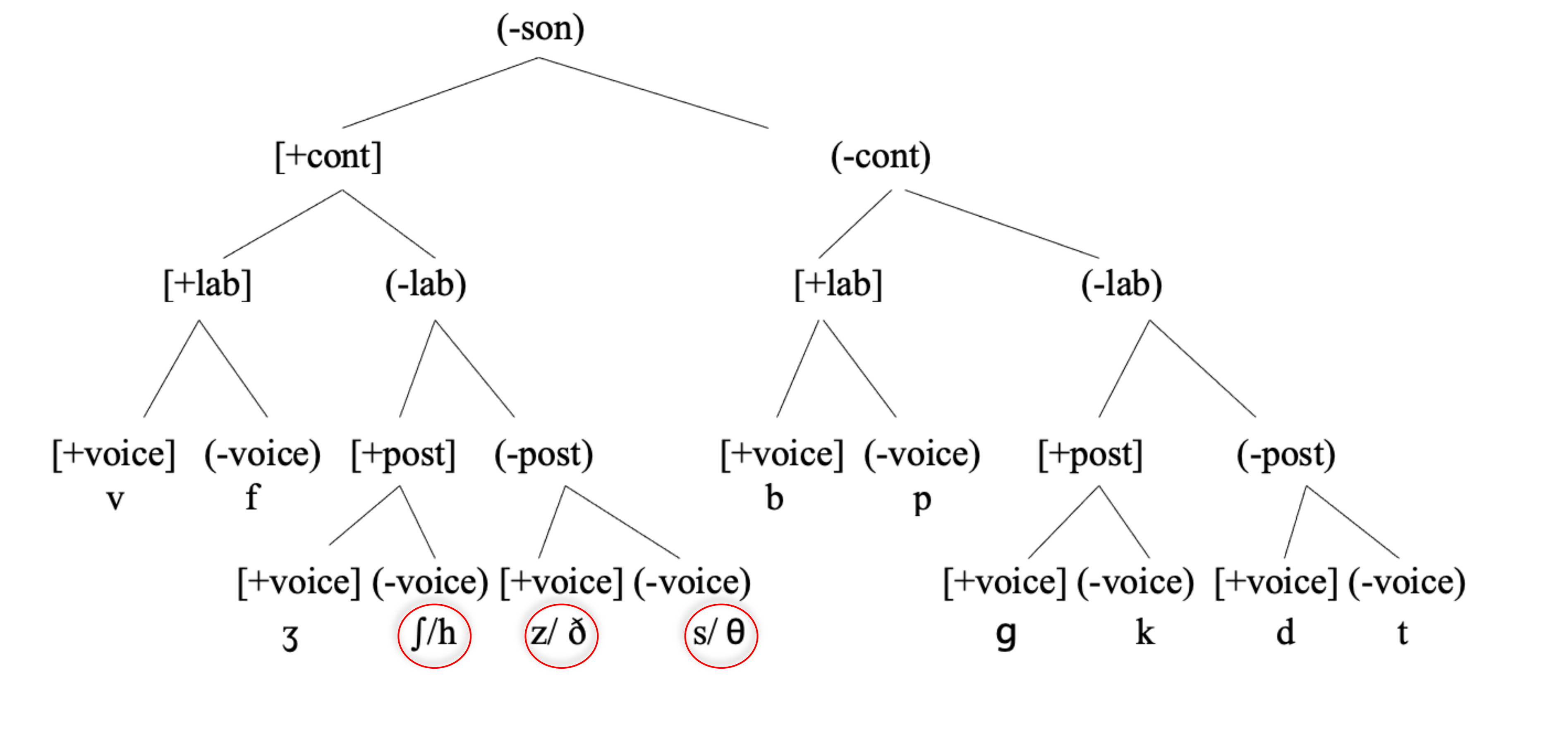
Etc.

Jaccard Scores

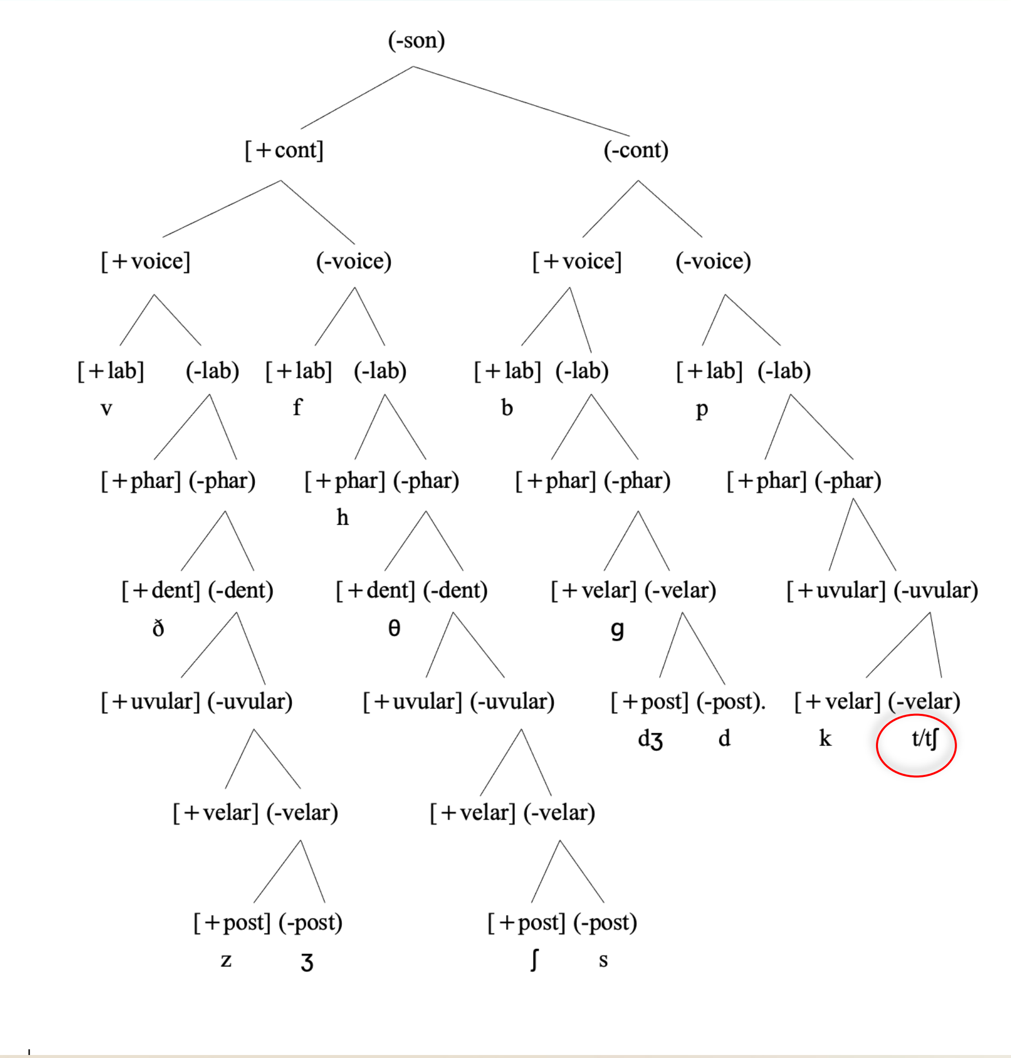
- French/English: .2
- Arabic/English: .2

Parsing Differences

French Parsing of English Input



Arabic Parsing of English Input



Parsing vs Jaccard

- When we compare English/Arabic and English/French, the Jaccard scores were equal
- Yet the parsing capabilities of the two contrastive hierarchies were quite different
  - Arabic hierarchy: 1 English pair undifferentiated ([t/tʃ])
  - French hierarchy: 3 pairs undifferentiated ([ʃ/h]; [z/ð]; [s/θ])

Subcomponents & Jaccard

- Vowels
  - French/English (.6) < Arabic/English (.9)

- Consonants
  - French/English (.2) = Arabic/English (.2)

Subcomponents & Parsing Failures

- Vowels
  - Arabic/English (7) > French/English (3)
- Consonants
  - Arabic/English (1) < French/English (3)

Conclusion

- Jaccard Distance has the potential of assessing the difference between two sets (in this case, sets of feature rankings)
- While it may be useful for the linguist, I am less sure of its utility for the learner (not necessarily *feasible* in the sense of Chomsky, 1965)
- Sometimes identical Jaccard scores can lead to different parsing failures
- ∴ monitoring parsing failures appears to be the preferred metric for both learner and linguist in this domain.

- Archibald, J. (2022). Segmental and prosodic evidence for property-by-property phonological transfer in L3 English in northern Africa. *languages*.
- Bohn, G & R. Santos (2018). The acquisition of pre-tonic vowels in Brazilian Portuguese. *Alfa* 62(1): 191-221.
- Cowper, E., & Hall, D. (2019). Scope variation in contrastive hierarchies of morphosyntactic features. In D. Lightfoot & J. Havenhill (Eds.) *Variable properties in language: Their nature and acquisition* (pp. 27–41). Georgetown University Press.
- Dresher, B. E. 2009. *The Contrastive Hierarchy in Phonology*. Cambridge: Cambridge University Press.
- Matthe, T., R. De Caluwe, G. de Tré, A. Hallez, J. Verstraete, M. Leman, O. Cornelis, D. Moelants, and J. Gansemans. 2006. Similarity between multi-valued thesaurus attributes: Theory and application in multimedia systems. *Flexible Query Answering Systems*: (Lecture notes in computer science 4027), 331–42. Heidelberg: Springer.
- Oxford, W. (2015). Patterns of contrast in phonological change: Evidence form Algonquian vowel systems. *Language* 91: 308-357.
- Purnell, Raimy & Salmons (2019). Old English vowels: Diachrony, privativity, and phonological representations. *Language* 95(4): 447-473.
- Rothman, J. (2015). Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: Timing of acquisition and proficiency considered. *Bilingualism: Language and Cognition, 18*(2), 179-190.
- Westergaard, M. (2021). Microvariation in multilingual situations: the importance of property-by-property acquisition. *Second Language Research, 37(*3), 397-407.
- Wu, Junyu. (2021). A contrastive hierarchy analysis of the Mandarin vowel system. Canadian Linguistics Association Conference.