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
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
• 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.

<table>
<thead>
<tr>
<th>Languages</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic: English</td>
<td>(11-1)/11= .9</td>
</tr>
<tr>
<td>French: English</td>
<td>(24-9)/24= .6</td>
</tr>
<tr>
<td>Mandarin: English</td>
<td>(17-3)/17= .8</td>
</tr>
</tbody>
</table>
• 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 L-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

[\[\pm\text{low}\]]

(-low)  \quad (+\text{low})

(/i/) \quad (/\text{a}/)

(-\text{back}) \quad (+\text{back})

(/u/)
French Hierarchy

[±long]

[±long] (±long)

[+low] (-low) [-low] [+low]

[+high] (-high) [-high] [+high]

[+back] (-back) [-back] [+back]
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

<table>
<thead>
<tr>
<th>French Rankings</th>
<th>English Rankings</th>
</tr>
</thead>
<tbody>
<tr>
<td>nasal &gt; long</td>
<td></td>
</tr>
<tr>
<td>nasal &gt; low</td>
<td></td>
</tr>
<tr>
<td>nasal &gt; high</td>
<td></td>
</tr>
<tr>
<td>nasal &gt; back</td>
<td></td>
</tr>
<tr>
<td>nasal &gt; round</td>
<td></td>
</tr>
<tr>
<td>long &gt; low</td>
<td>long &gt; low</td>
</tr>
<tr>
<td>long &gt; high</td>
<td>long &gt; front</td>
</tr>
<tr>
<td>long &gt; back</td>
<td>long &gt; high</td>
</tr>
<tr>
<td>long &gt; round</td>
<td>long &gt; round</td>
</tr>
<tr>
<td>low &gt; high</td>
<td>low &gt; front</td>
</tr>
<tr>
<td>low &gt; back</td>
<td>low &gt; high</td>
</tr>
<tr>
<td>low &gt; round</td>
<td>low &gt; round</td>
</tr>
</tbody>
</table>
• 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

(-son)

[ + cont] [ -cont]

[ + s.g.] [ -s.g.]

[ + lab] [ -lab]

f

[ + post] [ -post]

3

[ + glottal] [ -glottal]

h

[ + dent] [ -dent]

s
d

[ + s.g.] [ -s.g.]

[ + lab] [ -lab]

p

[ + post] [ -post]

k

t

g

d
Arabic Consonants
French Consonants

[Diagram showing the classification of French consonants based on their features such as [+cont], [+lab], and voice.]
# Rankings for Jaccard Distance: Consonants

<table>
<thead>
<tr>
<th>French Rankings</th>
<th>English Rankings</th>
<th>Arabic Rankings</th>
</tr>
</thead>
<tbody>
<tr>
<td>continuant &gt; labial</td>
<td>continuant &gt; spread glottis</td>
<td>continuant &gt; voice</td>
</tr>
<tr>
<td>continuant &gt; posterior</td>
<td>continuant &gt; labial</td>
<td>continuant &gt; labial</td>
</tr>
<tr>
<td>continuant &gt; voice</td>
<td>continuant &gt; posterior</td>
<td>continuant &gt; pharyngeal</td>
</tr>
<tr>
<td>labial &gt; posterior</td>
<td>spread glottis &gt; labial</td>
<td>continuant &gt; dental</td>
</tr>
<tr>
<td>labial &gt; voice</td>
<td>spread glottis &gt; posterior</td>
<td>continuant &gt; uvular</td>
</tr>
<tr>
<td>posterior &gt; voice</td>
<td>labial &gt; posterior</td>
<td>continuant &gt; velar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>continuant &gt; posterior</td>
</tr>
</tbody>
</table>

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.