• For a phone [tʃ] and other such complexes, how do we decide whether it is /tʃ/ or /t/+/ʃ/?
  o **Affricates**: tʃ, ts, tc, pf, kx
  o **Partially nasalized**: ʰd, ʰn, ʰnʲ, ʰn̥
  o **Labiovelars**: kp, gb, ŋmgb
  o **Labialized**: kw, gw, ŋw etc.

• "The one unifying feature of complex segments is that they are **articulatorily complex** in a way that simplex segments are not, and they can be **decomposed into parts** that can be separate segments"

• "The learning problem can be decomposed into two questions: (i) **what** language-internal **cues** do learners use to **discover complex segment representations**, and (ii) **when** does this happen."

**Criteria** for unification into a complex segment:
  o **Articulation**: "Is the duration of the sequence like that of a cluster, or like that of a segment?"
  o **Inseparability**: "Can the sequence be decomposed into parts that occur independently?"
  o **Distribution**: "Does the sequence have the same distribution as an uncontroversial singleton segment?"
  o **Syllabic structure**: "Is the sequence heterosyllabic or tautosyllabic?"
  o **Parsimony**: "Is the language’s phonemic inventory more symmetric if the sequence is analyzed as a complex segment?"

• Gouskova & Stanton:
  o "Our findings indicate that in a range of languages, **inseparability is the key** to identifying complex segments."
  o "we define inseparability as a **gradient, probabilistic measure**: the likelihood of C1 and C2 occurring together as C1C2, rather than separately or in clusters with other Cs."
  o "the **morphemic lexicon** emerges as the most appropriate model of the learning data."

• Fijian:
  o "Our Fijian corpus is from the An Crúbadán project, http://crubadan.org (26,000 words, compiled from internet texts—we cleaned the corpus to exclude English words, which left us with 17,600)."
  o [http://crubadan.org](http://crubadan.org): "The idea behind this project is to exploit the vast quantities of text freely available on the web as a way of bringing the benefits of statistical NLP to languages with small numbers of speakers and/or limited computational resources"
After the learner settles on a set of sequences to unify, it modifies the segmental inventory and its representation of the learning data.

There is no limit to the number of iterations the learner performs; it stops when it finds no more sequences that qualify for unification.
**Discussion point:**
- Must decide paradigmatic contrasts before syntagmatic constituency
- What **determines** what is the **same segment** across various contexts?

**Canadian English**
- /kat/ 'caught'
- /kat/ 'cot'  (cf. [kat])
- */kat/
- /najf/ [najf]
- /najvz/ [najvz] *[najvz]?*

**Three kinds of patterns in results**
- Clear arguments for complex vs. cluster: Ngbaka, Mbay, Turkish, Hebrew, Quechua
- Argument for complex vs. cluster *opposite* of analysts: Sundanese
- Lack of clear arguments: Latin, Russian, English

**Pattern 1:** Task is easy to do if there are few/marginal number of CC clusters (e.g. Fijian)
- What about **Hebrew, with lots of CCs**
- Only /ts/ from phonological argumentation
- "We tested our learner on the Living Lexicon of Hebrew Nouns (11,599 words, Bolozky and Becker 2006)." ⇔ **confirms only /ts/ is complex**

(12) Hebrew phonotactics (from Asherov and Bat-El 2019; we follow them in ignoring voicing assimilation)

| a. kvisa | ‘laundry’ | g. tsdaka | ‘charity’ | m. stsena | ‘scene (loan)’ |
| b. tkufa | ‘period’ | h. tʃuva | ‘answer’ | n. lantʃ | ‘lunch (loan)’ |
| c. tsfasdía | ‘frog’ | i. tzuza | ‘movement’ | o. tʃɪps | ‘chips (loan)’ |
| d. dʒima | ‘sample’ | j. tsvita | ‘pinch’ | p. dʒǐns | ‘jeans (loan)’ |
| e. psolot | ‘waste’ | k. tsnim | ‘toast’ | q. sklerozis | ‘sclerosis (loan)’ |
| f. btsalim | ‘onions’ | l. tnuva | ‘yield (n)’ | r. *ṭfn, ḏv, etc. |

**Pattern 2:**
- Making predictions: **Sundanese** as "contradicting" linguists
Discussion point: What happens when /CiCj/ contrasts with /Ci+Cj/?
  - Polish contrasts affricates and stop–fricative clusters
  - Fricatives longer in clusters than in affricates

(2) Polish (https://en.wikipedia.org/wiki/Polish_phonology)

a. czysta [ˈt͡ʂɨstə] ('clean' fem.)
  trzysta [ˈtʂɨstə] ('three hundred')
b. dżem’ [ˈdʐem] ('jam')
  drzem [ˈdʐem] ('take a nap' imper.)

G&S Footnote 12:
  - "Note that we did not give the learner a chance to consider Russian palatalized Cs for unification. This is because it is not clear to us how to transcribe the distinction between Cʲ and Cj in the initial state"
  - [lʲot] 'ice' vs. [lʲot] 'pours', [abjom] 'volume' vs. [grʲib'am] 'we row'
  - Transcribing all as [C j] would neutralize this distinction
  - One possibility would be to transcribe the [j]s with different lengths [b j] vs. [b jː]

Pattern 3: English
  - "quantitative support for the affricate analysis of English [tʃ] is not strong, … because [tʃ] is not frequent enough to counterbalance the individual frequencies of [t] and [ʃ], so the learner fails to unify it without being given more detailed phonetic information"

<table>
<thead>
<tr>
<th>inseparability</th>
<th>N(C1C2)</th>
<th>N(C1)</th>
<th>N(C2)</th>
<th>p(C1C2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>d tʃ</td>
<td>1.64</td>
<td>4002</td>
<td>25859</td>
<td>4332</td>
</tr>
<tr>
<td>tʃ</td>
<td>0.25</td>
<td>2765</td>
<td>39042</td>
<td>9162</td>
</tr>
</tbody>
</table>

Table 14: English inseparability calculations for Iteration 1 under broad transcriptions

Discussion point: Can we use more sophisticated phonotactics to probe internal structure?
Important take-away points of G&S

**Point:** Lack of 3+ part complex segments tied to rarity of long consonant clusters in general

**Point:** Phonetic difference does not entail phonological difference
  - Bura-Margi languages (Chadic, NE Nigeria)
  - "They are claimed to have many complex segments, most controversial of which are the labiocoronal\s [pt, bd, mnpt, mnbd, ?bd, pts, ptʃ] (Maddieson 1983:287)."
  - "appear to counterexemplify the claim that there is a link between duration and segmenthood: they pattern like single segments, yet are longer than single segments."

**Lots of cross-linguistic variability in clusters:**
"Homorganic [s t] is shorter than heterorganic [s p] and [s k] in Greek, but not in English (Arvaniti 2007:21–22)"

Cf. Karee Garvin, Myriam Lapierre & Sharon Inkelas "A Q-theoretic approach to distinctive subsegmental timing" (also Inkelas & Shih 2016)

"Pycha (2009, 2010) discusses relative timing differences between the closure and fricative portions of [ts] and [tʃ] in Hungarian, where [t] has a longer closure in [tʃ] than in [ts]."

\[(13) \text{a. } [ts] \quad \text{b. } [tʃ] \]
\[
\begin{array}{c}
(t^1 s^2 s^3) \\
\downarrow \\
(t^1 t^2 f^3)
\end{array}
\]

**Point:** it crucially matters what kind of data is fed to the learner
  - Type vs. token
  - Underlying vs. Intermediate rep vs. Surface
  - Morphemes vs. phonological words vs. orthographic words
  - Internal boundaries indicated

"With the exception of Quebecois French and Russian, all these languages point to the same conclusion: the more informative distributions are in morphemes/roots, not phonological words or connected speech."

**Quechua:**
  - "Quechua has mostly templatic roots, CV(C)CV, but its suffixes are atemplatic and often begin with consonant clusters (e.g., -sqa ‘nominalizer’, -jku ‘1pl. excl.’, -rqa ‘past’)."
  - "All of this results in [tʃ] being common (3,494 occurrences) and inseparable (5.32), but its ejective and aspirated counterparts are less common and less inseparable than certain clusters that occur in common suffixes." …
  - "Training the learner on morphologically complex words in such a language makes it inevitable that it will unify the wrong things."

Some potentially complex segments and complex topics which were not discussed:
  - Diphthongs
  - Nasalized vowels
  - Vowels with various laryngeal states, e.g. Udihe
    *laka-čan > l’a-sa  [ləʔ-sa]~[ləʔ-sa]  'bullhead (fish)'
• Going about our business as phonologists without the certainty of complex segmenthood status

• For example, patterning of prenasalized stops in Komo: nt vs. nd (Cahill 2008)

(5) (b) Kimasa word game with prenasalized voiced stops

| mbáŋga   | ṭgámba   | ‘fast’ |
| mbókó    | kómbó    | ‘horn’ |
| ndúmbú   | mbúndú   | ‘nude’ |
| ṭgána    | nánga    | ‘refuse’ |
| ɓongbiŋbi | ɓγbiŋbiŋo | ‘length’ |

(6) Kimasa word game with syllabic nasals

| mpáti   | tipám ~ tipá?m | ‘track’ |
| núpo    | póm ~ pó?m    | ‘then’ |
| mφáse   | seφám ~ seφá?m | ‘twin’ |
| µčáŋŋá | jŋačáŋ ~ jŋačá?ŋ | ‘shallows’ |
| ntíndí | ndítíŋ ~ ndítí?ŋ | ‘civet’ |
| ṭkpá   | kpaŋ ~ kpaŋ?ŋ | ‘person’ |