Elision: Say ___. Now say ___ without saying ___.

Matching: The word ___ starts with the ___ sound. Which one of these words starts with the ___ sound like ___: ___ or ___?

Isolation: The word ___ has four sounds: ___, ___, ___, and ___. What is the second sound in ___?

Blending: ___ and ___ makes …?
**Arabic binyanim (K-T-B):**

yu-CCiC-u  
_C-kultib_  
‘he dictates’

CaCaC-a  
_kaat.iba_  
‘he wrote’

ya-CCuC-u  
_yaktubu_  
‘he writes’

CaaCiC  
_kaatib_  
‘writer’

ma-CCaC  
_maktab_  
‘letter’

**Turkish agglutination:**

cöp-lük-ler-imiz-de-ki-ler-den-mi-y-di

‘Was it from those that were in our garbage cans?’

**English variable bracketing:**

We could open the door—it’s *[un-lock]able.*

We can’t keep people out—the door’s *un[lock-able].*
WHY MODEL IT?

• Most languages have richer morphology than English:
  
  • In Archi, *every single verb* has 1.5 million forms
  
  • In Turkish, there is *no known upper bound* on word length
  
  • And Zipf’s Law applies within morphological paradigms (see slides from 2015-01-20)
  
• The identities and boundaries of morphological units are aggressively obscured by phonological processes

[Source: Kibrik 1998, Hankamer 1992]
OUTLINE

• Phonology:
  • Phonemes, allophones, phones, and features
  • Phonological processes as FSTs

• Morphology:
  • Morphemes and allomorphs
  • Morphological typology
  • Morphological analysis
PHONOLOGY
**CONSONANTS**

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p b</td>
<td>t d</td>
<td>t q</td>
<td>c j</td>
<td>k g q g</td>
<td>N</td>
<td>R</td>
<td>h h</td>
<td>h h</td>
<td>h h</td>
<td>?</td>
</tr>
<tr>
<td>Nasal</td>
<td>m m̃</td>
<td>n ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>N</td>
<td>C C</td>
<td>C C</td>
<td>C C</td>
<td>C C</td>
<td>C C</td>
<td>C C</td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Tap or flap</td>
<td>v</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Fricative</td>
<td>φ β f v θ ð s z f s z ç j x y x k h q h</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
</tr>
<tr>
<td>Lateral Fricative</td>
<td></td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
<td>f b</td>
</tr>
<tr>
<td>Approximate</td>
<td>ũ</td>
<td>r</td>
<td>r</td>
<td>j w</td>
<td>j w</td>
<td>j w</td>
<td>j w</td>
<td>j w</td>
<td>j w</td>
<td>j w</td>
<td>j w</td>
</tr>
<tr>
<td>Lateral Approximant</td>
<td></td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
</tr>
</tbody>
</table>
## IN ENGLISH...

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plosive</strong></td>
<td>p b</td>
<td>t d</td>
<td>t q</td>
<td>c j</td>
<td>k q g</td>
<td>g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td><strong>Nasal</strong></td>
<td>m m̃</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trill</strong></td>
<td>B</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td></td>
</tr>
<tr>
<td><strong>Tap or flap</strong></td>
<td>v</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fricative</strong></td>
<td>φ β f v θ ð s z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lateral Fricative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Approximant</strong></td>
<td>v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lateral Approximant</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VOWELS IN ENGLISH
DEFINITIONS

*Phoneme*: a contrastive unit of sound

e.g., *pat* /pæt/ vs. *bat* /bæt/

*Allophone*: a surface contextual variant of a phoneme

e.g., regular past tense *-ed* and regular plural *-s*

*Phone*: a single sound unit/segment (contrastive or not)

e.g., *train* [tɹeɪn] has four or five phones (depending on whether you count [eɪ] as one or two
<table>
<thead>
<tr>
<th></th>
<th>/-d/</th>
<th></th>
<th>/-z/</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[næ̆p-t]</td>
<td>'napped'</td>
<td>[læ̆p-s]</td>
</tr>
<tr>
<td></td>
<td>[næ̆b-d]</td>
<td>'nabbed'</td>
<td>[læ̆b-z]</td>
</tr>
<tr>
<td></td>
<td>[saɪt-əd]</td>
<td>'sighted'</td>
<td>[lis-əz]</td>
</tr>
<tr>
<td></td>
<td>[saɪd-əd]</td>
<td>'sided'</td>
<td>[tʃɪz-əz]</td>
</tr>
</tbody>
</table>
DISTINCTIVE FEATURES & NATURAL CLASSES

• Laryngeal features specify the state of the glottis:

  \([\pm \text{VOI}]\) distinguishes \([b, d, g]\) from \([p, t, k]\)

• Place features specify place of articulation:

  \([p]\) is [LAB], \([t]\) is [COR]

• Manner features specify manner of articulation:

  \([s]\) is [+CONT], \([d]\) is [−CONT]

• Conjunctions of features define natural classes:

  \([−\text{VOI}, +\text{CONT}] \leftrightarrow [f, θ, s, ʃ, h]\)
REWRITE RULE FOR ENGLISH PAST TENSE

A rewrite rule is a four-tuple $\phi \rightarrow \psi / \lambda \models \rho$ read as “$\phi$ is replaced by $\psi$ when preceded by $\lambda$ and followed by $\rho$”

ASSIMILATION: $[+\text{OBS}] \rightarrow [-\text{VOI}] / [-\text{VOI}, +\text{OBS}] \models$

$\phi \rightarrow \psi$: $[p, b \rightarrow p], [t, d \rightarrow t], [k, g \rightarrow k], \ldots$

$\lambda$: $[p, t, k, \ldots]$

$\rho$: (null)

$/næp-d/ \rightarrow [næpt]$
EFFICIENT COMPILATION OF REWRITE RULES

• Rule $\phi \rightarrow \psi / \lambda \_ \_ \rho$ can be represented as the composition of 5 finite state transducers $r \circ f \circ R \circ L_1 \circ L_2$ such that

  • $r$: insert $>$ before every $\rho$

  • $f$: insert $<_1$ and $<_2$ before $\phi$ followed by $>$ (marking just those $\phi$ immediately before a $\rho$)

  • $R$: replace $\phi$ with $\psi$ in the context $<_1 \_ \_ >$ and delete $>$

  • $L_1$: delete $<_1$ preceded by a $\lambda$

  • $L_2$: delete $<_2$ not preceded by a $\lambda$

[Source: Mohri & Sproat 1996]
A → B / C __ D

C A C A D C D D D C A A C A D

insert > before every D:
C A C A> D C> D> D C A A C A> D

insert 21 before A>:
C A C21A> D C> D> D C A A C21A> D

replace 1A> with 1B:
C A C21B D C> D> D C A A C21B D

clean up:
C A C B D C D D D C A A C B D
RULE INTERACTIONS

In most dialects of English, the consonants /t, d/ undergo FLAPPING when preceded by a stressed vowel; as a result the word coder thus sounds the same as coater, though code and coat do not.

For Canadian speakers (a.o.), the stressed vowel in words like price undergoes RAISING before /t/ but not before /d/; as a result, write and ride have different vowels.

Despite the fact that FLAPPING eliminates the distinction between /t ~ d/, writer and rider still have different vowels for these speakers.

This shows that RAISING applies before FLAPPING; i.e., the output of FLAPPING is the input to RAISING.

[Source: Joos 1942, Chomsky 1957]
INFIXATION IN THAO

\[ \begin{align*}
hafuj & \quad h_mafuj & \quad \text{‘chant’} \\
tqir & \quad tumqir & \quad \text{‘protest’} \\
iup & \quad miup & \quad \text{‘blow with the mouth’} \\
\end{align*} \]

[Source: Lu 2011]
METATHESIS IN ROTUMAN

\[ \text{hosa} \quad \text{hoas} \quad \text{‘flower’} \]
\[ \text{pure} \quad \text{puer} \quad \text{‘to rule’} \]
\[ \text{parofita} \quad \text{parofiat} \quad \text{‘prophet’} \]

[Source: Churchward 1940, Chandlee et al. 2012]
REDUPLICATION IN TAGALOG

<table>
<thead>
<tr>
<th>Tagalog Word</th>
<th>Reduplicated Form</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ól</td>
<td>olól</td>
<td>‘mad’</td>
</tr>
<tr>
<td>súlat</td>
<td>susúlat</td>
<td>‘write’</td>
</tr>
<tr>
<td>magka-útang</td>
<td>magkaka-útang</td>
<td>‘owe’</td>
</tr>
<tr>
<td>áraw</td>
<td>araw-aráw</td>
<td>‘day’</td>
</tr>
<tr>
<td>sangpówo</td>
<td>sangposangpówo</td>
<td>‘ten’</td>
</tr>
</tbody>
</table>

Total reduplication is very difficult to model with FSTs.

[Source: Blake 1917]
Many NLP systems really work with written (orthographic) forms. That’s an interesting subject, but beyond the scope of this class.
MORPHOLOGY
Morphological differences distinguishing *H. neaderthalensis* from *H. sapiens*:

- Projecting mid-face
- Low, flat, elongated skull
- Lack of a protruding chin
- No groove on canine teeth
- Barrel-shaped rib cage
- Large kneecaps
- Long collar bones, wider shoulders

[Source: Wikipedia]
DEFINITIONS

• **Morpheme**: a unit of meaning defined such that each word contains at least one
  
  e.g., *dog* has one morpheme, *dogs* has two

• **Allomorph**: a surface contextual variant of a morpheme
  
  e.g., the English regular past tense /-d/ has three allomorphs: [-d] as in *nabbed*, [-t] as in *napped*, and [-əd] as in *sighted*
TYPES OF AFFIXATION

- Prefixation: un-lock
- Suffixation: excite-ment
- Infixation: saxo-ma-phone, Minne-bloody-sota
- Circumfixation: in-vigor-ate, in-toxic-ate
- Templatic affixation: Arabic yuktibu vs. kaatib

[Source: Yu 2004, Ruszkiewicz 2003]
FUSIONAL (INFLECTING) VS. AGGLUTINATIVE LANGUAGES

Fusional: (Latin)

gladi-us me-us
sword-MASC.NOM.SG. 1 POSS.SG.-MASC.
‘my sword’

Agglutinative: (Turkish)

oda-lar-im-iz
room-PL.-1 POSS.-PL.
‘our rooms’
ISOLATING (ANALYTIC) VS. SYNTHETIC LANGUAGES

Isolating: (Vietnamese)

*Khi tôi đến nhà bạn tôi, chúng tôi bắt đầu làm bài.*
‘When I came to my friend’s house, we began to do the lesson.’

Synthetic: (Turkish)

*Çöp-lük-ler-imiz-de-ki-ler-den-mi-y-di?*  
‘Was it from those that were in our garbage cans?’

[Source: Comrie 1989, Hankamer 1986]
INFLECTION VS. DERIVATION

*Inflection*: affixation that preserves the major part of speech, adding syntactically required information

e.g., *dog* vs. *dogs*

*Derivation*: affixation that (usually) changes the part of speech, adding additional information

e.g., *destroy* vs. *destruction*
AUTOMATIC ANALYSIS OF MORPHOLOGY

• Stemming (e.g., Porter II):

  In linguistics, a morpheme is the smallest grammatical unit in a language.

• Morphological segmentation (e.g., Morfessor):

  In linguistics, a morpheme is the smallest grammatical unit in a language.

• Morphological analysis (e.g., KIMMO):

  In linguistics, a morpheme is the smallest grammatical unit in a language.