NATURAL LANGUAGE PHONOLOGY & MORPHOLOGY

CS662: Natural Language Processing

2015-01-29

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 …?

[Source: Comprehensive Test of Phonological Processing, 2nd ed. (CTOPP-2)]

Arabic *binyanim* (K-T-B):

yu-CC*i*C-*u*	*yuktibu*	‘he dictates’
C*a*C*a*C-*a*	*kataba*	‘he wrote’
ya-CC*u*C-*u*	*yaktubu*	‘he writes’
C*aa*C*i*C	*kaatib*	‘writer’
ma-CC*a*C	*maktab*	‘letter’

Turkish agglutination:

çöp-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

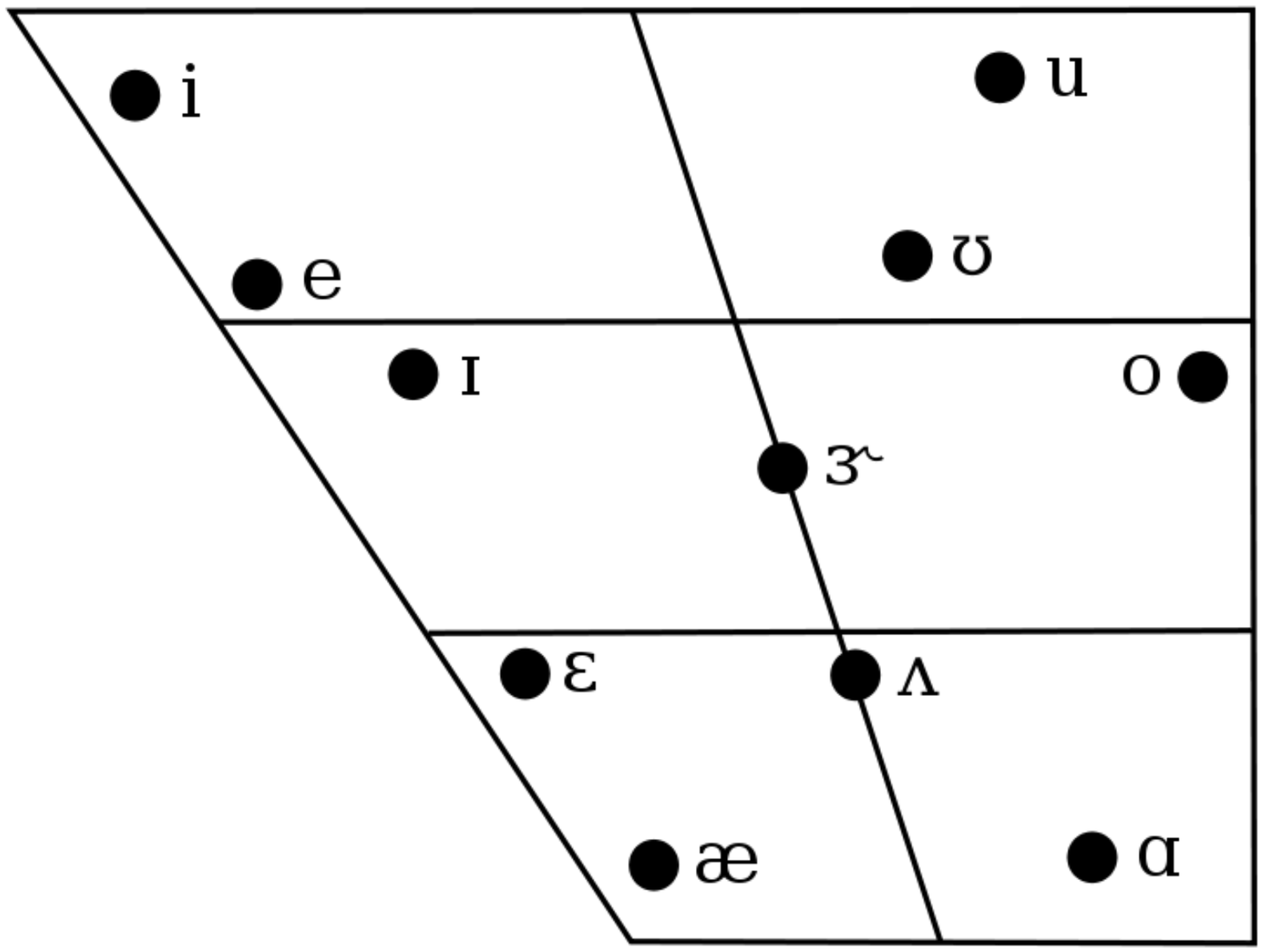
CONSONANTS

	Bilabial	Labiodental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral Fricative				ɬ ɮ							
Approximate	ʋ			ɹ		ɻ	j	ɰ			
Lateral Approximant				l		ɭ	ʎ	ʟ			

IN ENGLISH...

	Bilabial	Labiodental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p͡ b͡			t͡ d͡		ʈ ɖ	c ɟ	k͡ g͡	q ɢ		ʔ
Nasal	m͡	ɱ		n͡		ɳ	ɲ	ŋ͡	ɴ		
Trill	ʙ			r					ʀ		
Tap or flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f͡ v͡	θ͡ ð͡	s͡ z͡	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h͡	ɦ
Lateral Fricative				ɬ ɮ							
Approximant	ʋ			ɹ		ɻ	j	ɰ			
Lateral Approximant				l		ɭ	ʎ	ʟ			

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* [**tʰeɪn**] has four or five phones (depending on whether you count [**eɪ**] as one or two

/-d/		/-z/	
[næp-t]	‘napped’	[læp-s]	‘laps’
[næb-d]	‘nabbed’	[læb-z]	‘labs’
[saɪt-əd]	‘sighted’	[lis-əz]	‘leases’
[saɪd-əd]	‘sided’	[tʃiz-əz]	‘cheeses’

DISTINCTIVE FEATURES & NATURAL CLASSES

- *Laryngeal* features specify the state of the glottis:

  [±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*:

  [–VOI, +CONT] ⟷ [f, θ, s, ʃ, h]

REWRITE RULE FOR ENGLISH PAST TENSE

A rewrite rule is a four-tuple $\phi \rightarrow \psi$ / λ __ ρ read as "ϕ is replaced by ψ when preceded by λ and followed by ρ"

ASSIMILATION: [+OBS] → [–VOI] / [–VOI, +OBS] ___

$\phi \rightarrow \psi$: [p, b → p], [t, d → t], [k, g → k], …
λ: [p, t, k, …]
ρ: (null)

/næp-d/ → [næpt]

EFFICIENT COMPILATION OF REWRITE RULES

- Rule $\phi \rightarrow \psi$ / λ __ ρ 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 ρ

  - *f*: insert $<_1$ and $<_2$ before ϕ followed by > (marking just those ϕ immediately before a ρ)

  - *R:* replace ϕ with ψ in the context $<_1$ __ > and delete >

  - L_1: delete $<_1$ preceded by a λ

  - L_2: delete $<_2$ not preceded by a λ

[Source: Mohri & Sproat 1996]

A → B / C __ D

C A C A D C 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 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.

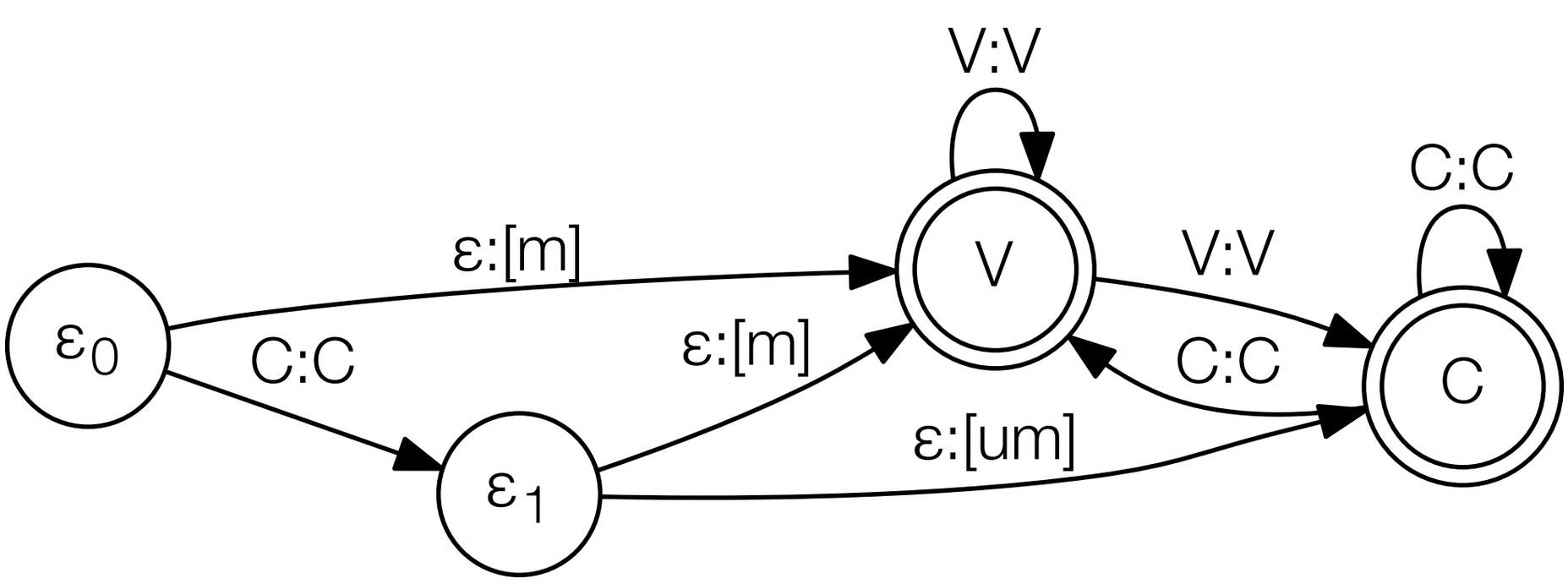
Despite the fact that FLAPPING eliminates the distinction between /**t** ~ **d**/, *writer* and *rider* still have different vowels for these speakers.

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

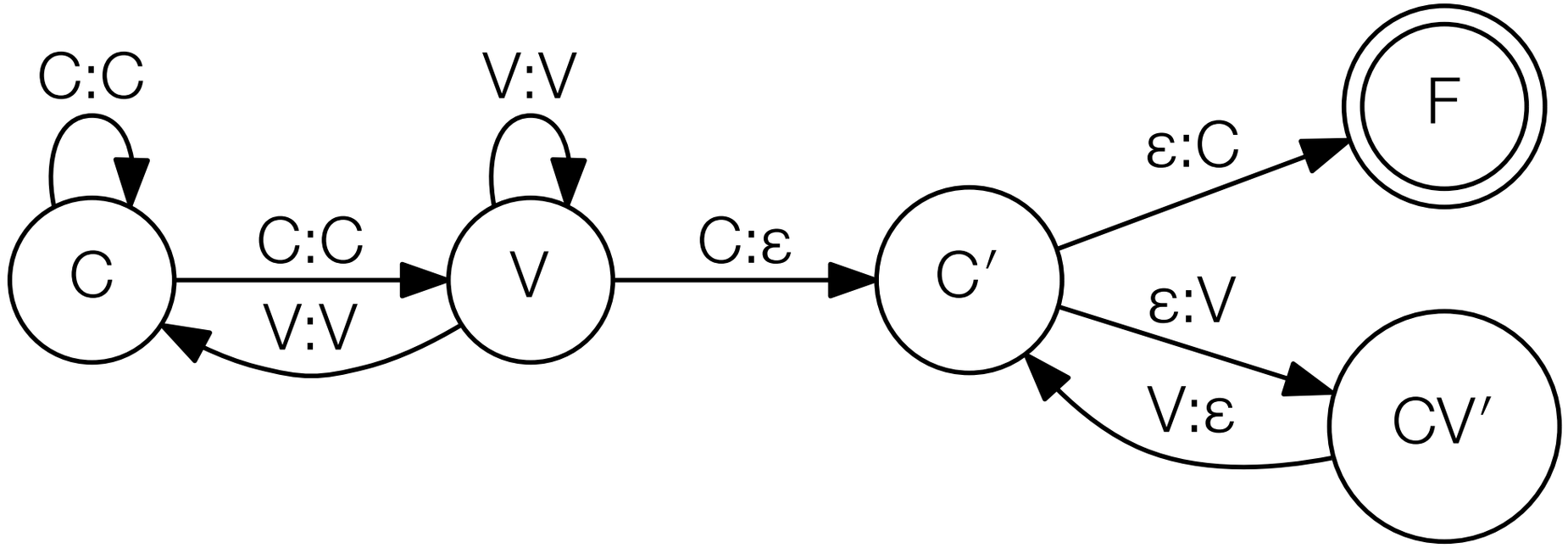
hafuj	*hmafuj*	‘chant’
tqir	*tumqir*	‘protest’
iup	*miup*	‘blow with the mouth’



[Source: Lu 2011]

METATHESIS IN ROTUMAN

hosa	*hoas*	‘flower’
pure	*puer*	‘to rule’
parofita	*parofiat*	‘prophet’



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

REDUPLICATION IN TAGALOG

ól	*olól*	‘mad’
súlat	*susúlat*	‘write’
magka-útang	*magkaka-útang*	‘owe’

áraw	*araw-aráw*	‘day’
sangpówo	*sangposangpówo*	‘ten’

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. 1POSS.SG.-MASC.
‘my sword’

Agglutinative: (Turkish)

oda-lar-ım-ız
room-PL.-1POSS.-PL.
‘our rooms’

ISOLATING (ANALYTIC) VS. SYNTHETIC LANGUAGES

Isolating: (Vietnamese)

Khi tôi dên <u>nhà ban tôi</u>, <u>chúng tôi</u> bát dâu 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 linguis-tics, a morphem-e is the small-e-s-t gramma-tical uni-t in a language.

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

  In linguistic-s[n.pl.], a morpheme is the small-est[comparative] gramma-tical[adj.] unit in a language.