Linguistics 101

Kevin Duh
Intro to NLP, Fall 2019

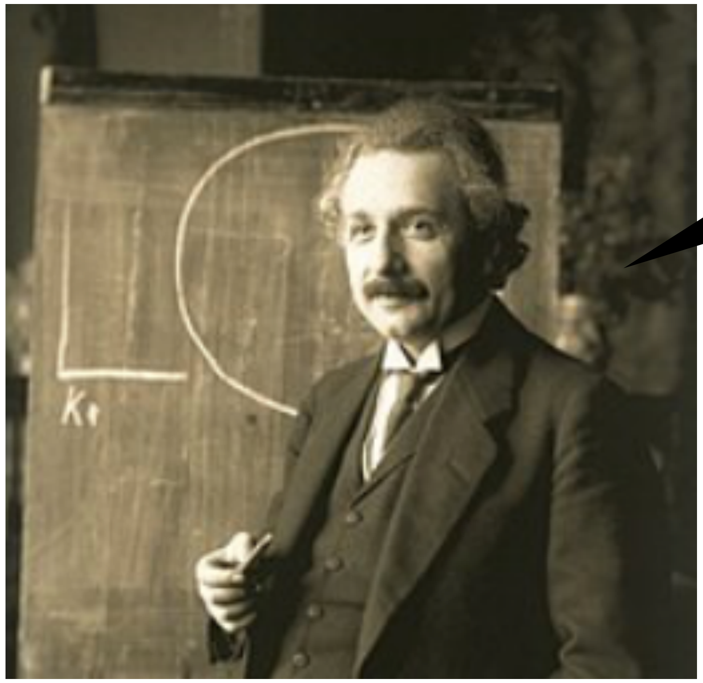
Why?

- As NLPers, we shoud know something about language!
- Studying linguistics may or may not help your NLP model, but it will give you a vocabulary to think about your data.

Outline

1. Phonetics/Phonology: the sounds of language
2. Writing Systems: transcribing language
3. Morphology: structure of words
4. Syntax: structure of sentences
5. Semantics: meaning of words/sentences
6. Pragmatics: meaning in context

Disclaimer



Everything should be made as simple as possible, but not simpler.

We're not following Eistein's advice.
These slides are probably over-simplified.
Please consult a real linguistics book for details.

Outline

1. Phonetics/Phonology: the sounds of language
2. Writing Systems: transcribing language
3. Morphology: structure of words
4. Syntax: structure of sentences
5. Semantics: meaning of words/sentences
6. Pragmatics: meaning in context

Language is not writing

- Language is a spoken phenomenon*
- Writing is a way to represent language in a physical medium
    - All kids learn to speak & listen naturally
    - Writing must be taught
    - 55% of world languages are unwritten

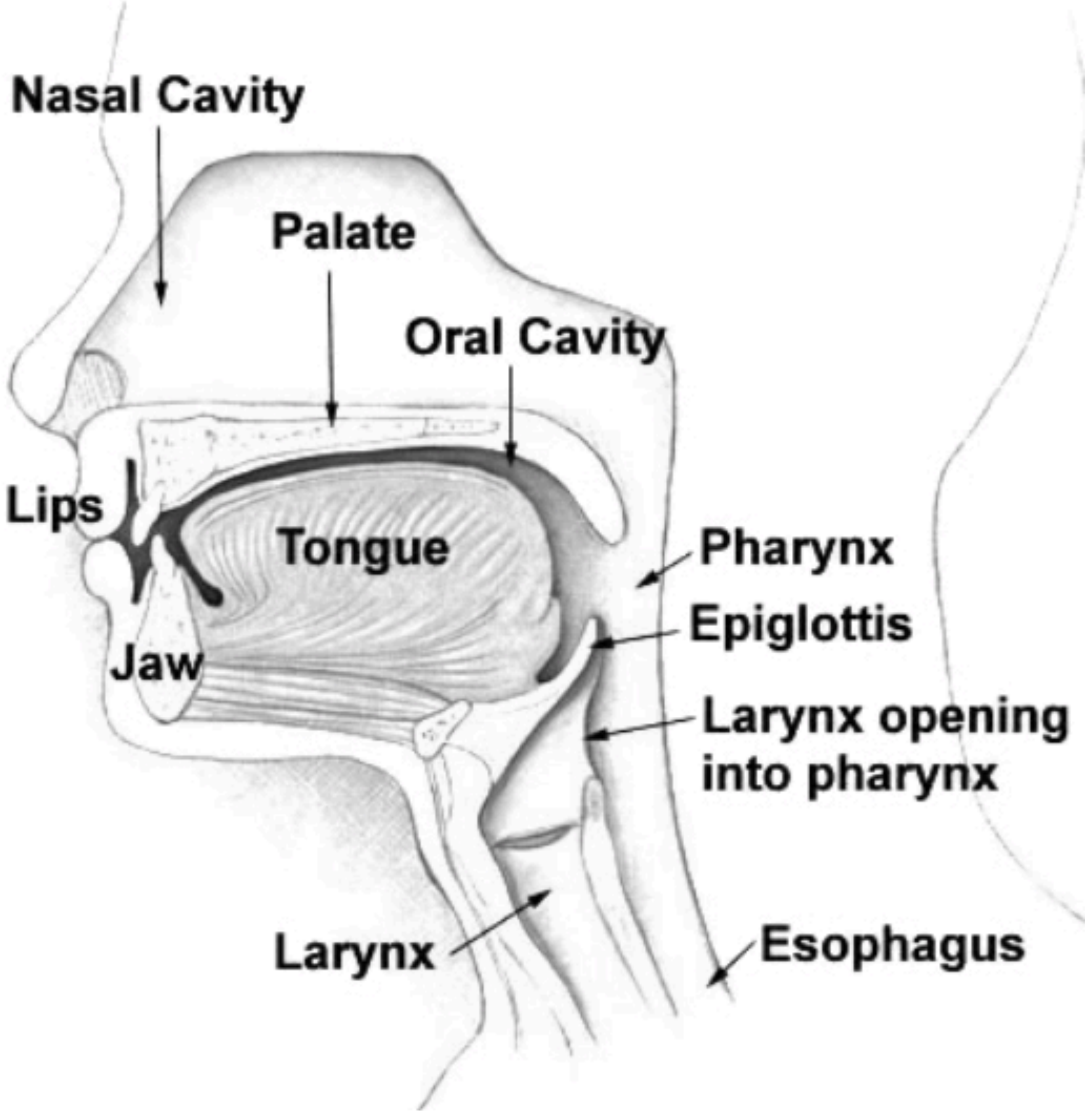
*Over-simplification: sign languages are visual, and show exhibit all the richness of spoken languages

Phonetics & Phonology

- **Phonetics**: study of the sound units
  - e.g. Vowels, Consonants, how they are produced
- **Phonology**: study of how these sound units combine

How is speech produced?

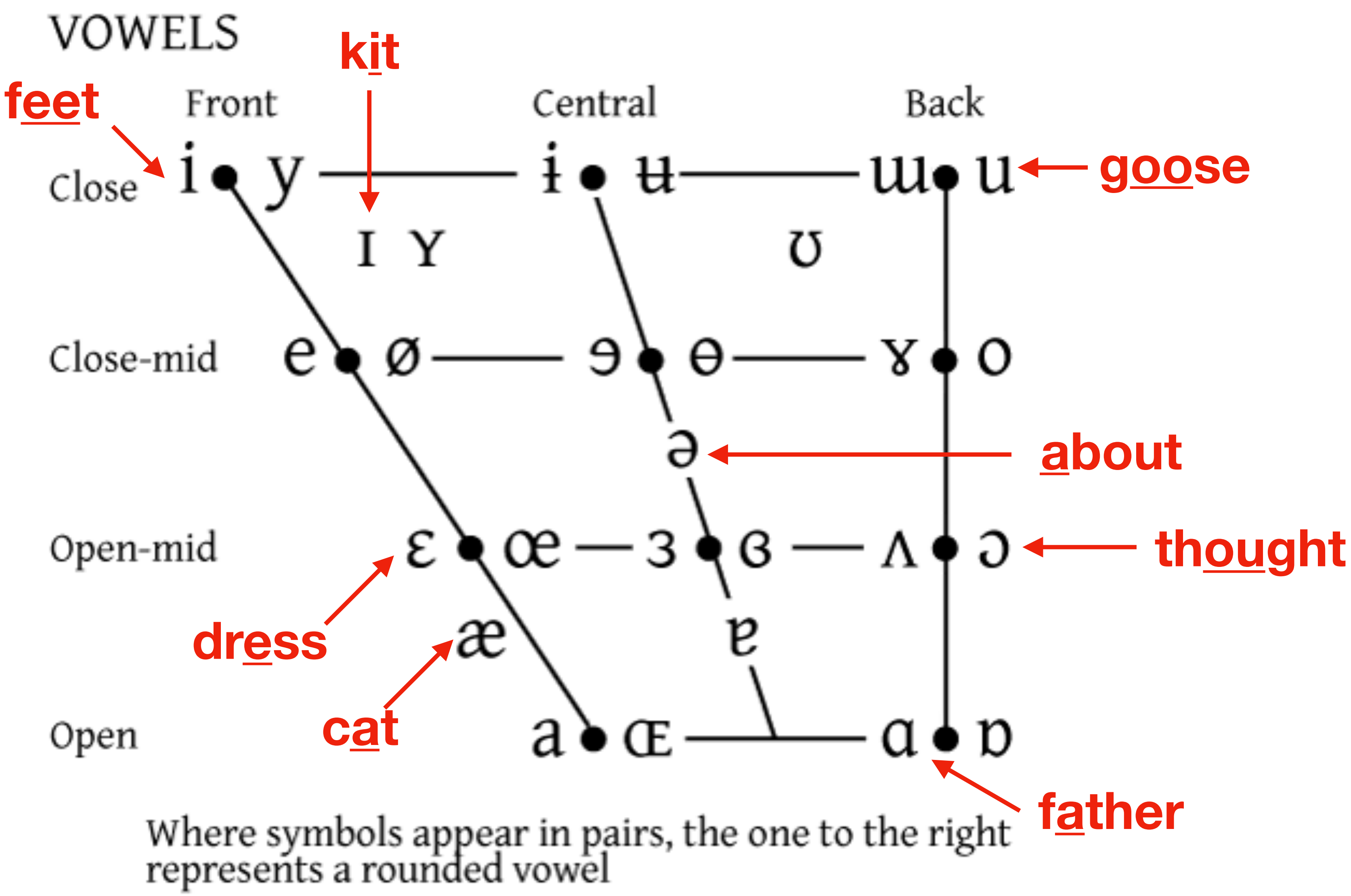
- Vocal tract: an amazing multi-purpose device
  - Breathing
  - Eating
  - Speaking
- Different sounds generated by:
  - air pushing through from lungs
  - vocal cords vibrating
  - shape formed from lips, tongue, etc.



Vowels

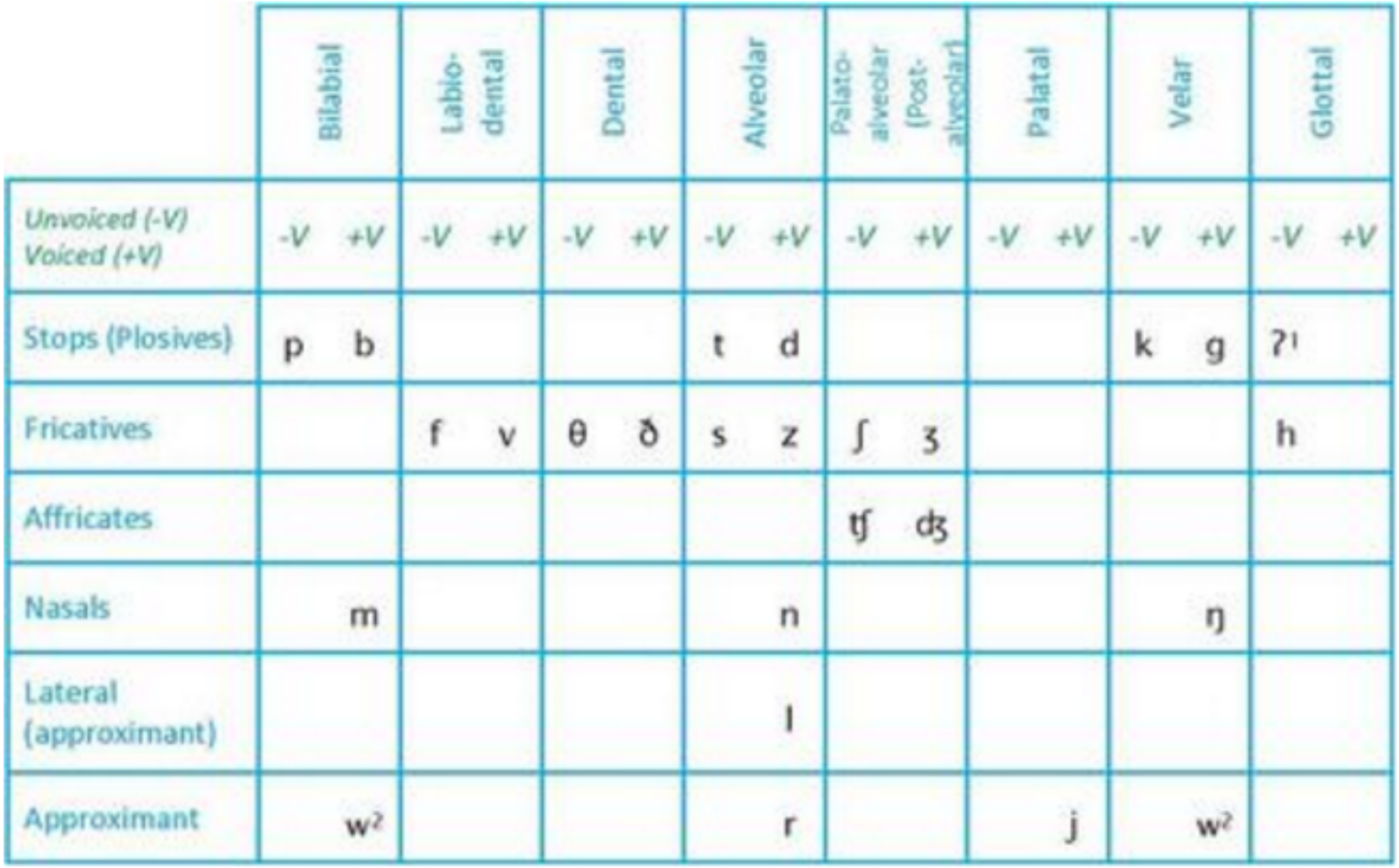
- Hold your jaw. Say **he**, **who**, **ha**.

- Did you feel for jaw move for **ha**?
- Different vowels are produced based on:
  - position of tongue (high vs low, front vs back)
  - rounding of lips



Where symbols appear in pairs, the one to the right represents a rounded vowel

- Vowel: made with mouth quite open
- **Consonant**: made with some part constricted
    - Place of articulation: where the vocal tract is made narrower, e.g.
        - Bilabial: pat bat mat (both lips)
        - Labial-dental: fat vat (lower lip on front teeth)
        - Inter-dental: thigh thy (tip of tongue protuding front teeth)
        - Aveolar: tab (tongue tip behind front teeth)
        - Velar: kill gill (tongue at back near velum)
    - Manner of articulation: how airstream is modified, e.g.
        - Stop: pat bat (complete obstruction of air)
        - Fricative: fat vat thigh (some air escape, turbulent noise)
    - Voiced vs Unvoiced: vat vs fat (try whispering…)



	Bilabial		Labio-dental		Dental		Alveolar		Palato-alveolar (Post-alveolar)		Palatal		Velar		Glottal	
Unvoiced (-V) *Voiced (+V)*	-V	+V	-V	+V	-V	+V	-V	+V	-V	+V	-V	+V	-V	+V	-V	+V
Stops (Plosives)	p	b					t	d					k	g	ʔ[1]	
Fricatives			f	v	θ	ð	s	z	ʃ	ʒ					h	
Affricates									tʃ	dʒ						
Nasals		m						n						ŋ		
Lateral (approximant)								l								
Approximant		w[2]						r				j		w[2]		

Note: these are IPA (International Phonetic Alphabet) symbols

Spelling (Orthography) doesn’t consistently represent sounds

- One sound, multiple spellings:
  - e.g. he, people, key
- One spelling, multiple sounds:
  - e.g. father, village
- There are 5 vowels and 21 consonants in English?
  - No, those are letters. 20 vowels and 24 consonants.

Phonemes and Phones

- Phone (Phonetic): any distinct sound produced, not specific to any language

- Phoneme (Phonemic): sound of a particular language. If swapped with another phoneme, word meaning can change

    - English: “map” with aspiration or not doesn’t make a difference in meaning

    - English: “cop” vs “keep” has slightly different [k] sounds, but doesn’t matter so one /k/ phoneme

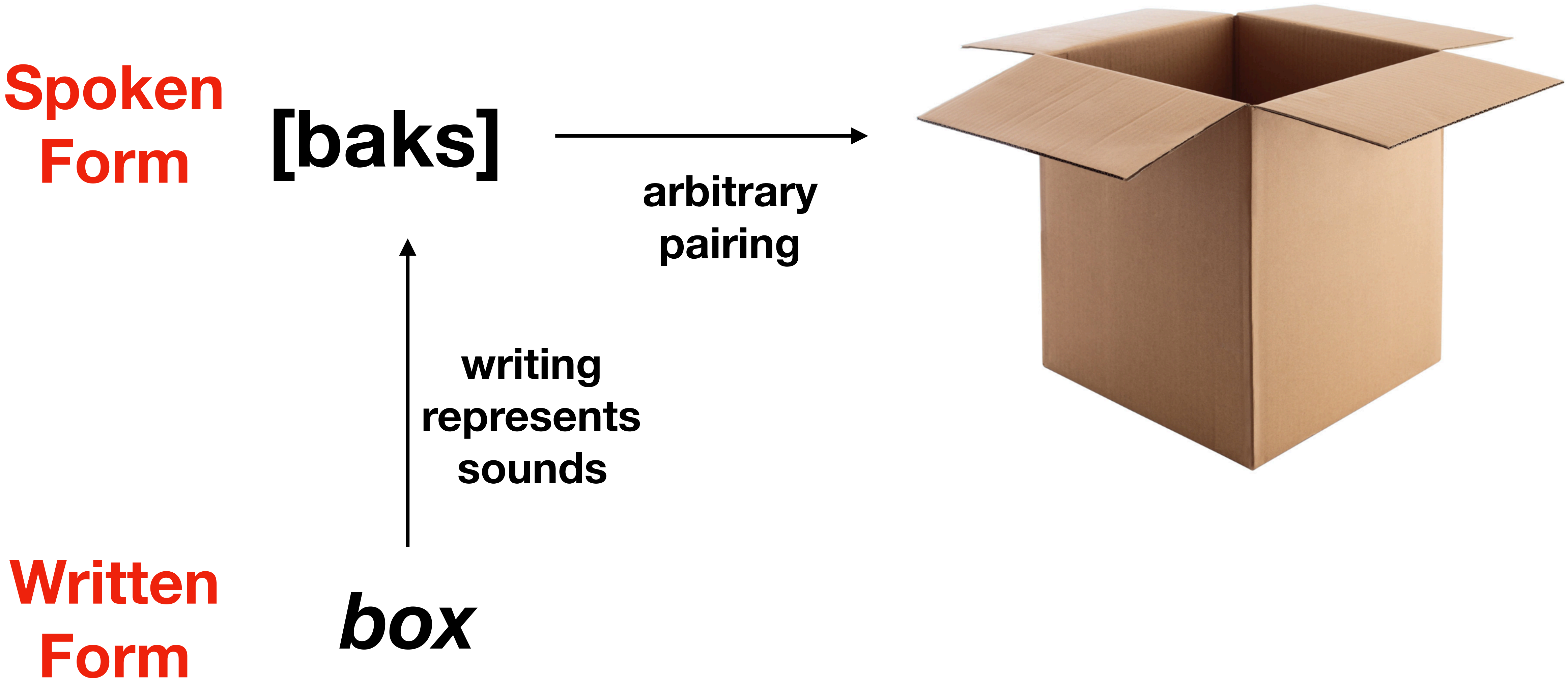
Why do we hear foreign accents?

- Phonology constraints from mother tongue, e.g.
  - English allows up to 3 consonants (C) at the beginning of a word, followed by vowel (V), i.e. CCCV “spree”
  - But not all languages allow this: Hawaiian only allows {CV, V}, Indonesian allows {CV, V, VC, CVC}

Outline

1. Phonetics/Phonology: the sounds of language
2. Writing Systems: transcribing language
3. Morphology: structure of words
4. Syntax: structure of sentences
5. Semantics: meaning of words/sentences
6. Pragmatics: meaning in context

Linguistic Sign
= Form + Meaning



Spoken Form

[baks]

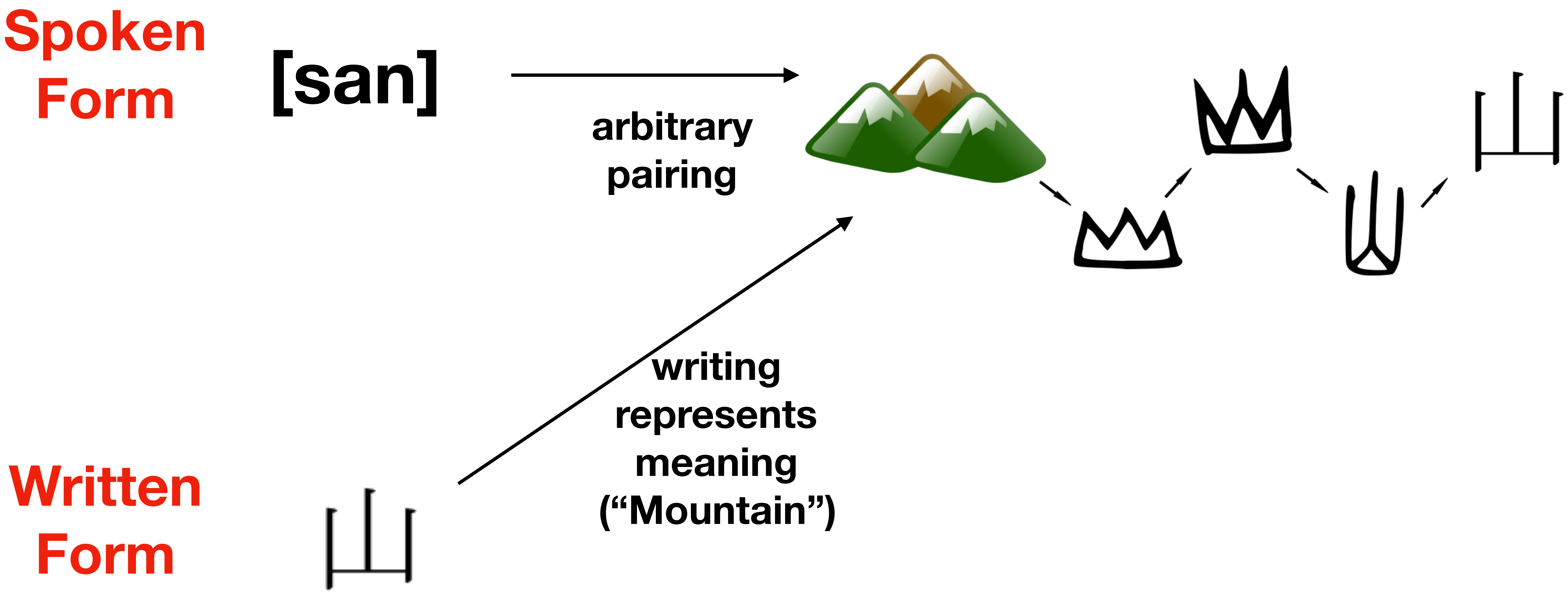
arbitrary pairing

writing represents sounds

Written Form

box

Linguistic Sign
= Form + Meaning



Spoken Form

[san]

arbitrary pairing

Written Form

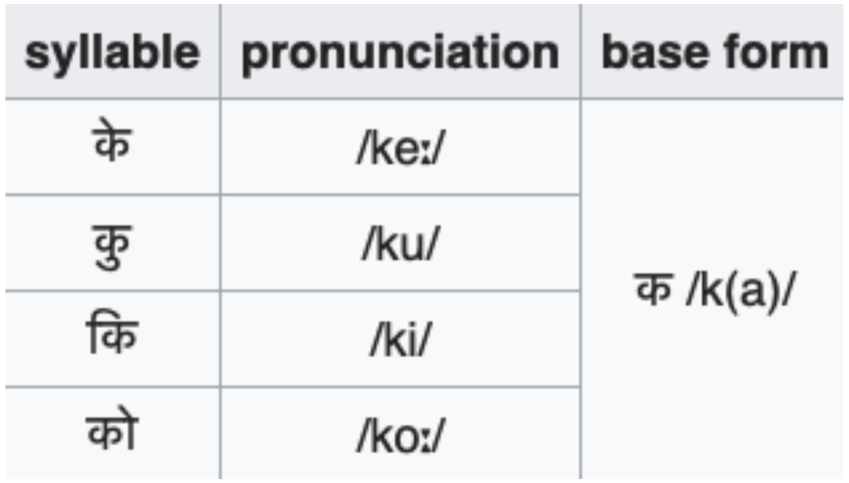
山

writing represents meaning (“Mountain”)

Note: Very few languages use logograms (Chinese, Hieroglyphs).
Even those that do contain many sound-based gylphs

Types of Writing Systems

- Logographic: symbols correspond to meaning/morpheme
- Phonographic: symbols correspond to sounds
  - Syllabary: symbol => syllable, e.g. Japanese Kana
  - Alphabet: represents both consonant & vowel, e.g. Roman
  - Abugida: represent consonants with full symbol and vowel with extra marks, e.g. Devanagari
  - Abjad: only consonant, e.g. Hebrew



syllable	pronunciation	base form
के	/keː/	क /k(a)/
कु	/ku/	
कि	/ki/	
को	/koː/	

Outline

1. Phonetics/Phonology: the sounds of language
2. Writing Systems: transcribing language
3. Morphology: structure of words
4. Syntax: structure of sentences
5. Semantics: meaning of words/sentences
6. Pragmatics: meaning in context

What are words?

- Are these same word or different words?
    - ***cat*** vs ***dog***
    - ***cat*** vs ***cats***
    - ***cat*** vs ***catalog***
- There’s some structure in ***cat*** vs ***cats*** that tells us they’re related

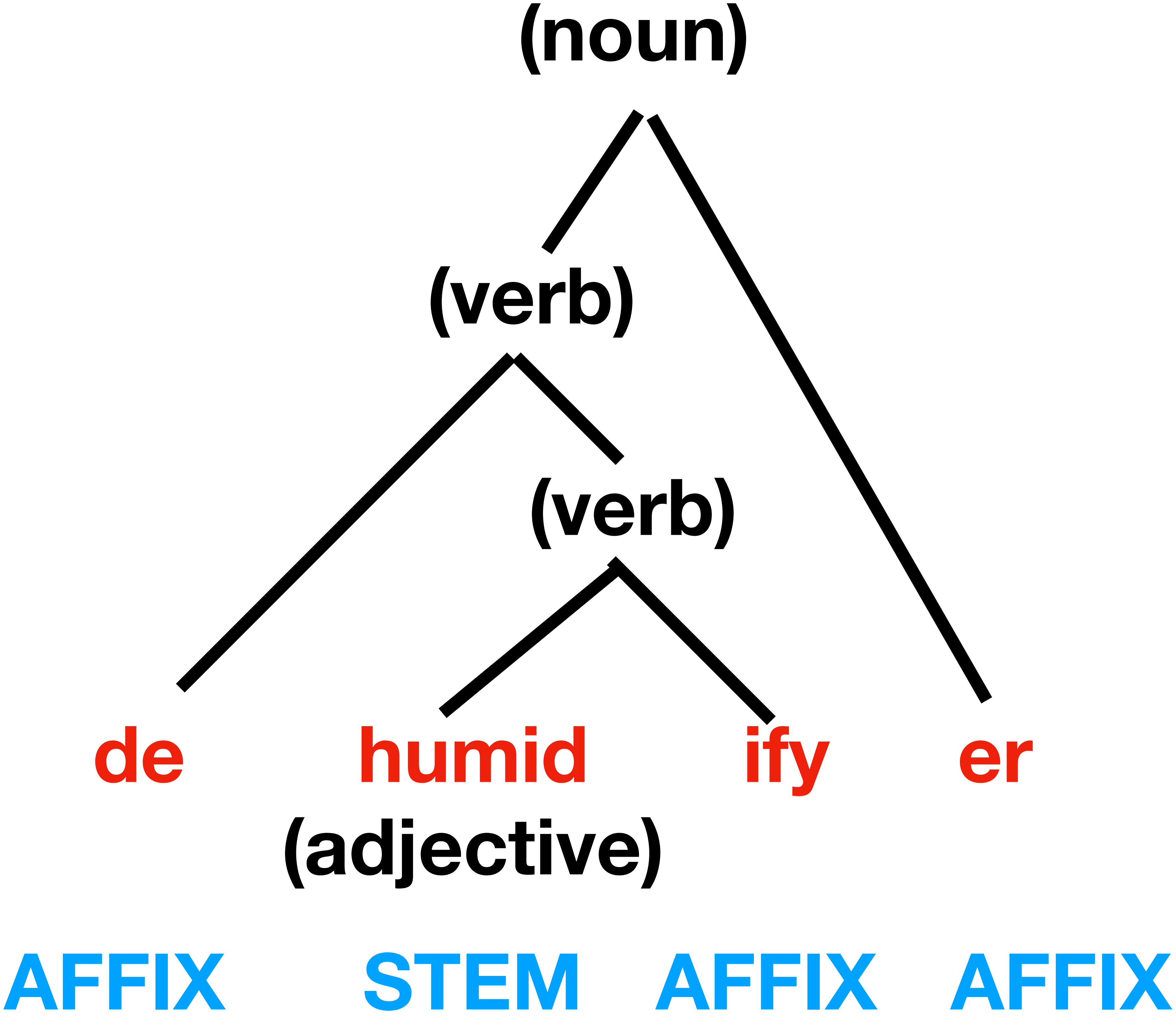
Morpheme

- Morpheme = smallest linguist unit with meaning or grammatical function
    - e.g. ***cats*** = ***cat*** + <***s:plural***>
- Types of morphemes:
    - Free morpheme: can be used as words by themselves
    - Bound morphemes: e.g. affix, suffix

- Inflection: create variants of the main word, e.g.
    - ***cats*** = ***cat*** + <***s:plural***>
    - ***walked*** = ***walk*** + <***ed:past-tense***>
    - ***taller*** = ***tall*** + <***er:comparison***>
- Derivation: create new word, changing meaning or part-of-speech
    - ***establishment*** (noun) = ***establish*** (verb) + <***ment***>
    - ***happiness*** (noun) = ***happy*** (adjective) + <***ness***>
    - ***undo*** = ***un*** + ***do***

Word formation processes

- Affixation: free morpheme + suffix, prefix, or infix
- Compounding: combines free morphemes
    - e.g. ***textbook*** = ***text*** + ***book***
- Reduplication: doubling of morphemes
    - Indonesian: ***rumah*** = house , ***rumahrumah*** = houses
- Alternation: morpheme-internal modifications
    - ***goose — geese***, ***foot — feet***, ***drink — drank***



(noun)

(verb)

(verb)

de

humid

ify

er

(adjective)

AFFIX

STEM

AFFIX

AFFIX

- Analytic language: each word is a single morpheme
- Synthetic language: each word is free + bound morpheme
    - Agglutinative: morphemes joined loosely, e.g. Swahili
        - [ni-na-soma] = <I>-<present>-<read> = I am reading
        - [u-na-soma] = <you>-<present>-<read> = You are reading
    - Fusional: morpheme boundaries fused, e.g. Spanish
        - [ablo] = I am speaking
        - [abla] = She/He is speaking
        - [-o], [-a] seem to suggest [abl-] means speak but it never occurs as a free morpheme
- Polysynthetic language: multiple stems and affixes in a word

Outline

1. Phonetics/Phonology: the sounds of language
2. Writing Systems: transcribing language
3. Morphology: structure of words
4. Syntax: structure of sentences
5. Semantics: meaning of words/sentences
6. Pragmatics: meaning in context

Grammaticality

- Some sentences are grammatical and some are not.
- What are general syntactic properties that determine this?
  - Word Order
  - Argument Structure
  - Agreement

Word Order

- Is there a strict order for Subject (S), Verb (V), Object (O)?
    - Grammatical: John (S) drank (V) coffee (O)
    - Ungrammatical: drank (V) John (S) coffee (O)
- In languages of the world:
    - 35% SVO, 44% SOV, 19% VSO. Other patterns rare.
    - Note, not all sentences in SVO language have to be SVO
    - Some languages allow more free word order

Argument Structure

- Why are some of these grammatical and some not*?

I run marathons.	I like it.	*I sneezed it.
I run.	*I like.	I sneezed.

- Different types of verbs expect different # of arguments

- Not just verbs. May be strict about form of an argument

It rained.   He relied on her.

Agreement

- In English, must have subject-verb agreement on number

He likes it.	*They likes it.
*He like it.	They like it.

- In German, determiner-noun agreement on gender

Der Salat
Das Krokodil
Die Kartoffel

- Things expressed via syntax in one language might be expressed via morphology in another

- e.g. Subject, Direct Object, Indirect Object are indicated by word order in English, but case markers in Japanese

I	**gave**	**Mike**	**the book**
(S)		**(IO)**	**(DO)**

*** I gave the book Mike**

私が	**マイクに**	**本を**	**あげた**
I-(S)	**Mike-(IO)**	**book-(DO)**	**gave**
私が	**本を**	**マイクに**	**あげた**
I-(S)	**book-(DO)**	**Mike-(IO)**	**gave**

Outline

1. Phonetics/Phonology: the sounds of language
2. Writing Systems: transcribing language
3. Morphology: structure of words
4. Syntax: structure of sentences
5. Semantics: meaning of words/sentences
6. Pragmatics: meaning in context

There are many ways to study semantics

- Lexical semantics:
  - Word meaning and its relationships
  - When we say “Time flies” — what does “flies” mean?
- Compositional semantics:
  - How do sentence meaning arise from word meaning?
  - e.g. What’s the meaning +? 3? 2? How about (3+2)?

Outline

1. Phonetics/Phonology: the sounds of language
2. Writing Systems: transcribing language
3. Morphology: structure of words
4. Syntax: structure of sentences
5. Semantics: meaning of words/sentences
6. Pragmatics: meaning in context

Sentence meaning depends on the context in which it’s uttered

- **Question: “Do you know the time?”**
- Answer 1: “Yes”
- Answer 2: “It’s 11:30am”

- **Question: “Can you take out the trash”?**
- Interpretation 1: Physically-speaking, do you have the ability?
- Interpretation 2: Do it!!

Some lessons for NLPers

1. Phonetics/Phonology
2. Writing Systems

The training data we observe is a result of complex processes involving the written representation of some spoken phenomena

3. Morphology
4. Syntax

Words and sentences are very productive, but follow their own rules depending on language. There is a diversity on how languages code information in morphology and syntax.

5. Semantics
6. Pragmatics

Meaning is challenging to pin down. This might be the holy grail, but there are lots of open questions.