Linguistics

David R. Mortensen

Language Technologies Institute
Carnegie Mellon University

May 18, 2020

Outline

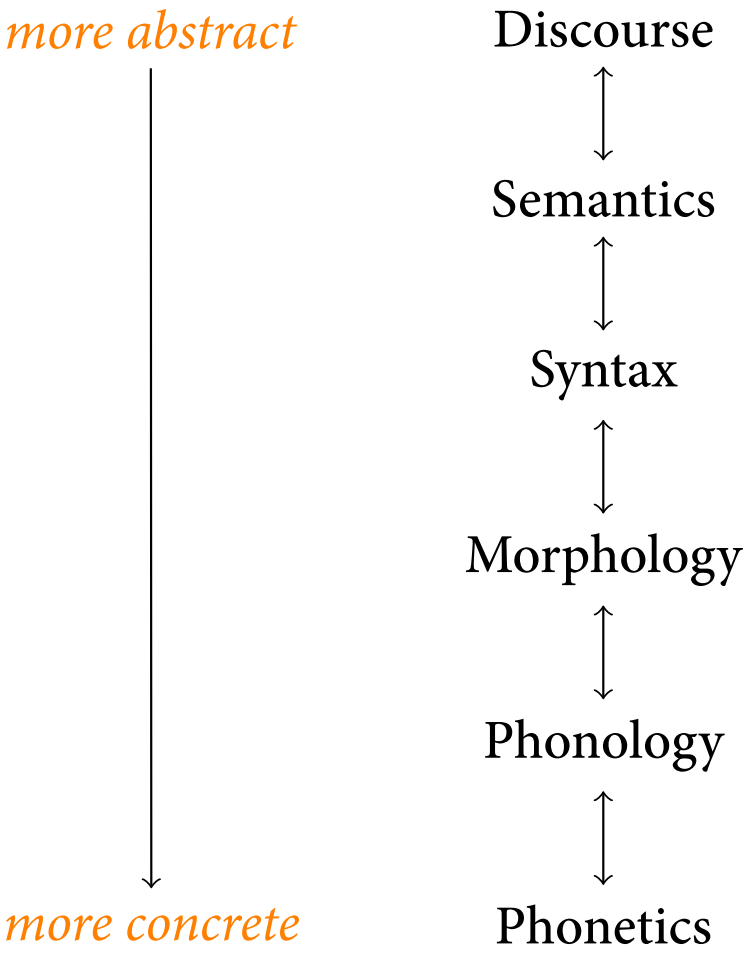
Language Structure

- People sometimes think that language is random ("Why do you park in a driveway and drive on a parkway?") but language actually has a rich and largely predictable structure.
- Native speakers of a language are generally not aware of this structure.
- However, this structure can be investigated scientifically.
- This scientific field is called LINGUISTICS. It is allied to adjacent subfields in computer science, neuroscience, cognitive science, psychology, and sociology.
- **Some patterns in linguistic structure**:
  - Language is structured at different levels—what linguists call "levels of representation." A sentence consists of a sequence of sounds (phonemes), a sequence of meaningful word-pieces (morphemes), a sequence of words, a sequence of phrases.
  - Except for phonemes, all of these units are accompanied by a meaning representation (or semantic representation).
  - In apparently all languages, there are a small finite number of phonemes; these combine to form an almost unlimited number of morphemes and words and an infinite set of sentences.

Sound over Letter

- Linguists tend to privilege spoken language over written language because spoken language is prior and written language is derived from spoken language.
- Written language is important though.
- Systems of sounds are called phonologies; systems of written language are called orthographies.
- Some of my work involves bridging the gap between these two levels of representation.

Levels of Representation (simplified)



Where Does Linguistic Knowledge Come from?

- **Introspection** (early generative linguistics, philosophy of language)
- **Field data** (documentary linguistics, sociolinguistics)
- **Analysis of linguistic databases** (typology, computational linguistics)
- **Analysis of language corpora** (lexicography, corpus linguistics, computational linguistics, documentary linguistics, philology)
- **Experiments** (psycholinguistics, neurolinguistics, sociolinguistics, phonetics, computational linguistics)

Outline

Three subfields of phonetics

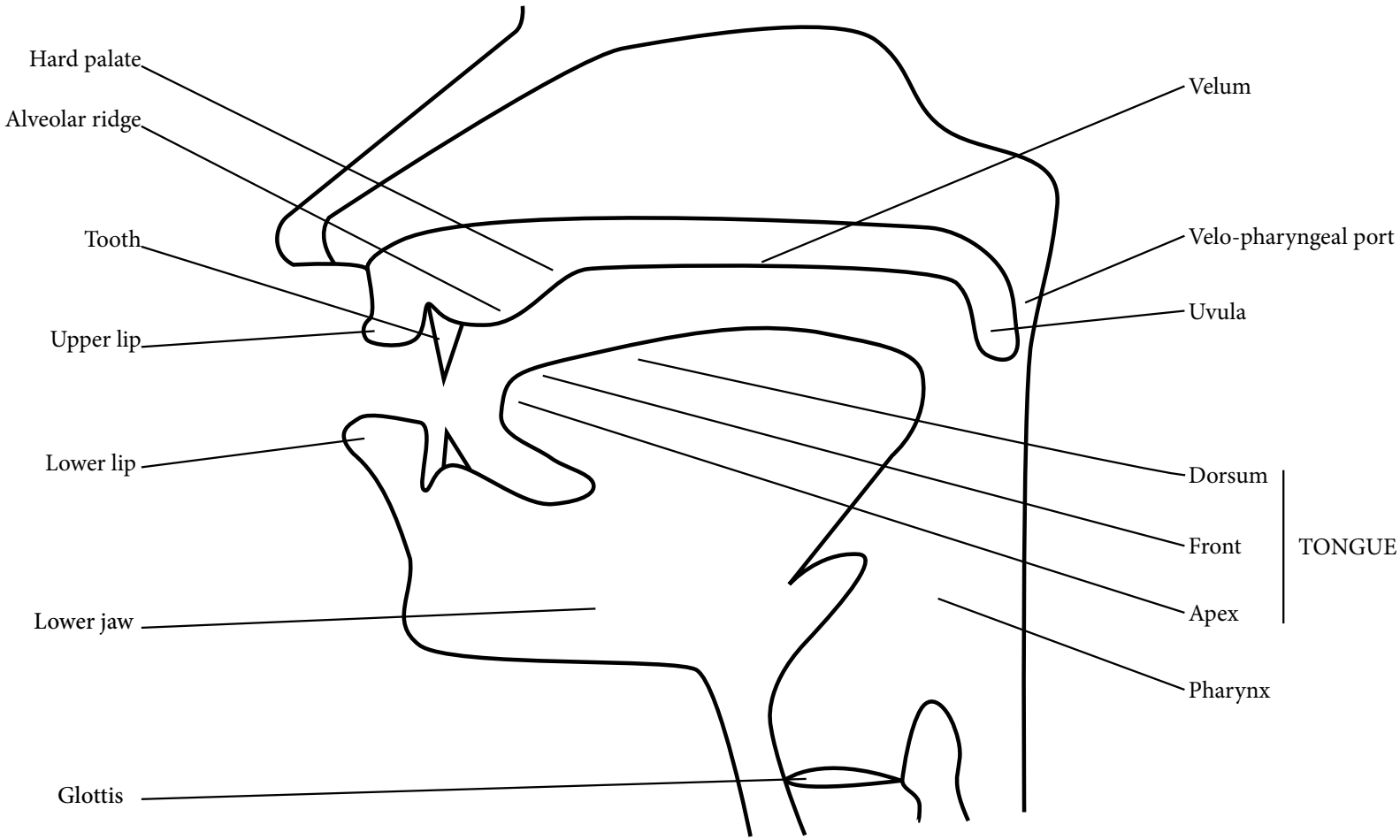
Phonetics has three subfields, each with an associated field of physics:

- Articulatory and descriptive phonetics (biophysics, aerodynamics)
- Acoustic phonetics (acoustics)
- Auditory phonetics (psychophysics)

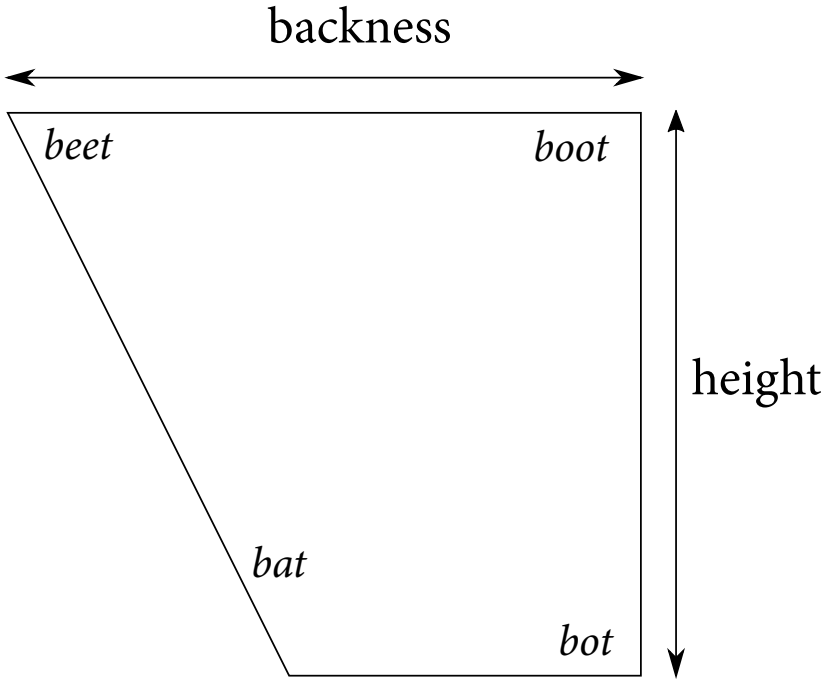
Articulatory phonetics

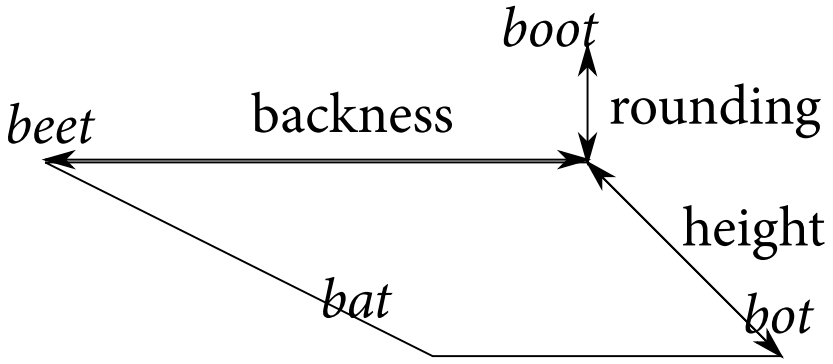
ARTICULATORY PHONETICS: the study of the mechanisms by which humans produce speech sounds.

Place of articulation: consonants



Place of articulation: vowels

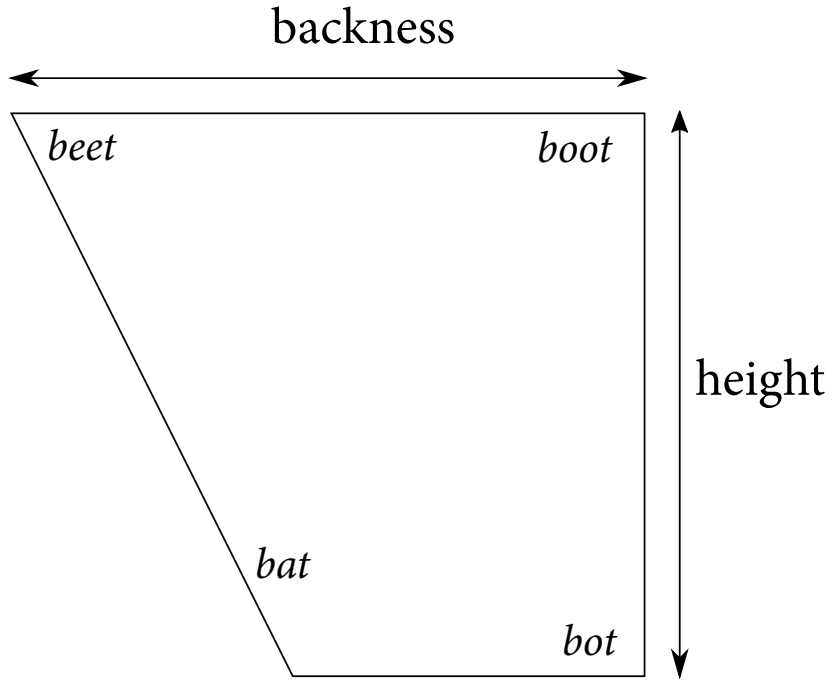




Vowel place rules of thumb

2

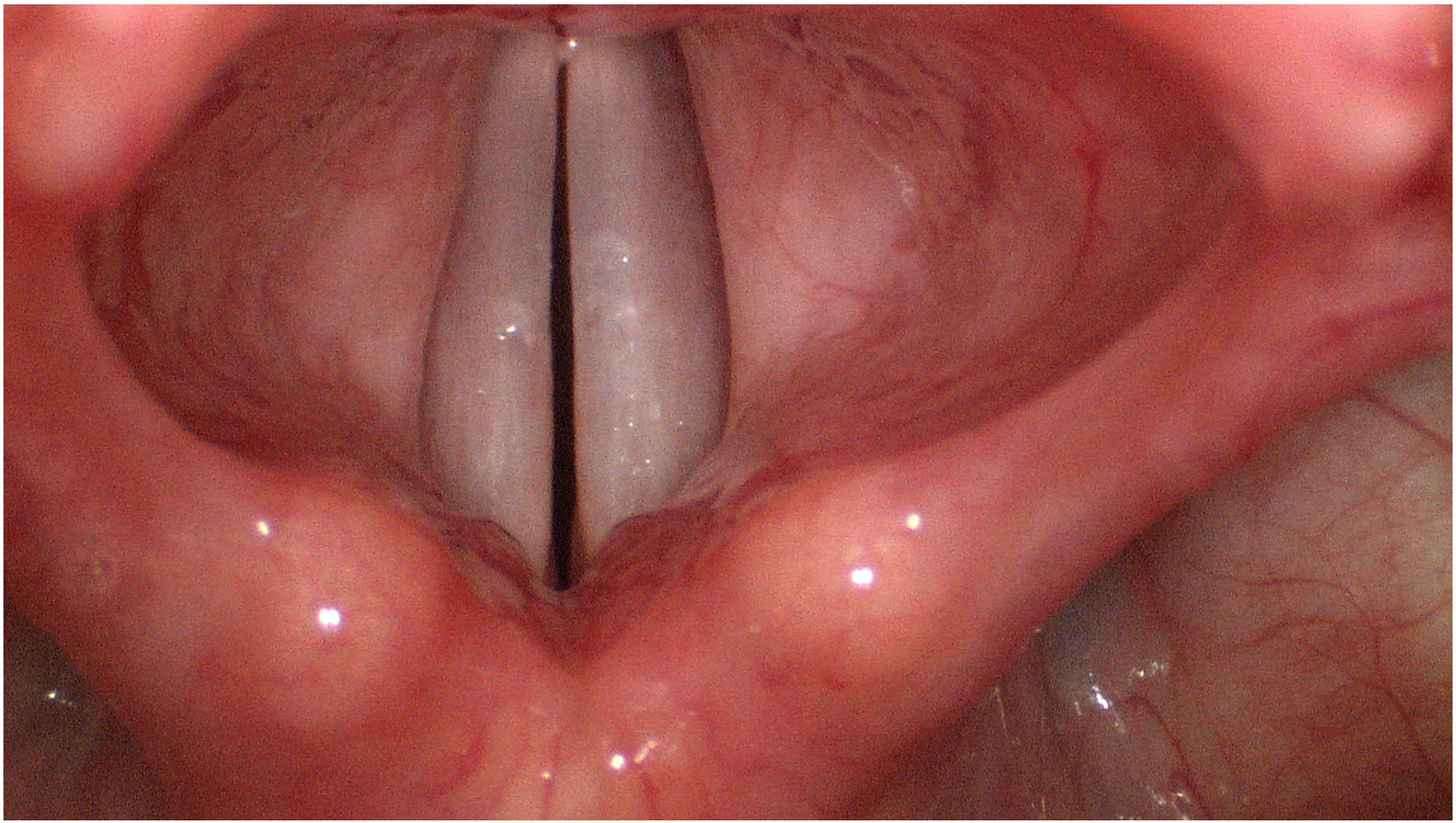
- A mid central vowel has minimal constriction
- A high front vowel typically has a PALATAL constriction
- A high back vowel typically has a VELAR constriction
- A low back vowel typically has a PHARYNGEAL constriction



Manner of articulation

- **plosives** or **oral stops** Characterized by the complete obstruction of the vocal tract and the closure of the velopharyngeal port; like the ⟨p⟩ in *parseltongue*
- **nasal stops** or NASALS Characterized by the complete obstruction of the vocal tract but with the velopharyngeal port open; like the ⟨m⟩ in *muggle*
- **trills** Produced with a "loose" closure so that the passage of air produces an oscillation
- **flap** or **tap** essentially a momentary plosive produced when an ACTIVE ARTICULATOR strikes a PASSIVE ARTICULATOR.
- **fricatives** Characterized by a tight constriction that produced turbulence when air is blown through it; like the ⟨s⟩ in *slither*
- **lateral fricative** A special kind of fricative in which the opening is on one or both sides of the tongue; common in exotic languages like Hmong and Welsh
- **approximant** Characterized by a loose constriction; includes glides like the ⟨w⟩ in *wand* and other sounds like the ⟨r⟩ in *raven*
- **lateral approximant** A special type of approximant in which there is an opening on one or both sides of the tongue; like the ⟨l⟩ in *leprechaun*

Voicing (phonation)



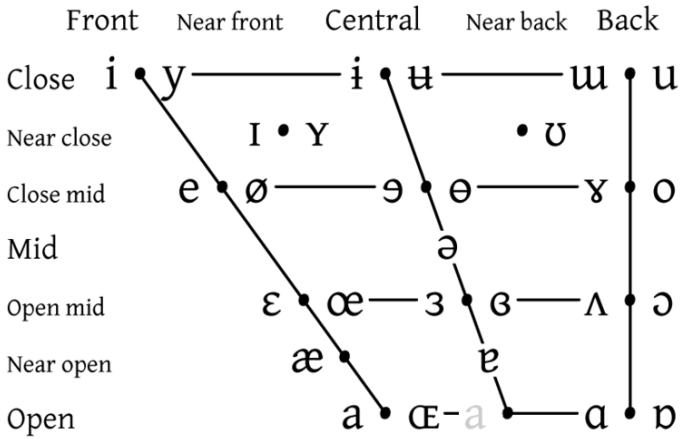
Linguists use confusion terminology

The LARYNX is a cartilaginous structure that contains the GLOTTIS. Part of the larynx protrudes from the throat of many males and is called the ADAM'S APPLE. The parts of the glottis that open, close, and vibrate are called the VOCAL FOLDS. Sometimes, they are also called the VOCAL CORDS.

The International Phonetic Alphabet

	Bilab		LabDent		Dent		Alv		PostAlv		Retr		Pal		Vel		Uvu		Phar		Glot	
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								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
Lat. Fric.							ɬ	ɮ														
Approximant				ʋ				ɹ				ɻ		j		ɰ						
Lat. Approx.								l				ɭ		ʎ		ʟ						

VOWELS



Vowels at right & left of bullets are rounded & unrounded.

Tone and other Suprasegmentals

- Many of the world's 7,000 langauges, **probably a majority**, use pitch to distinguish words. This is called TONE.
- Basically all languages use pitch in some way. Along with duration and intensity, this is called INTONATION.
- Phonetically, pitch is always controlled in the same way: by varying the tension on the vocal folds and thus modulating the rate at which they vibrate. The changes the FUNDAMENTAL FREQUENCY of the speech signal.

Why not just use orthography?

ORTHOGRAPHY refers to the conventional writing system used to write a language. Literally, it means "right-writing".

Some orthographies faithfully represent the pronunciation of a language (Spanish, Hungarian, Vietnamese); others (English, Chinese) show only a tenuous relationship between sound and symbol.

Each orthography has its own conventions. The same symbol might be pronounced in different ways and the same sound is often represented by different symbols. In IPA, in contrast, there is a near-perfect mapping between sound and symbol.

Getting IPA

For some applications, it may be necessary to have data transcribed in IPA (or some equivalent system). However, for many applications, IPA can be obtained from orthographic text. This is called grapheme-to-phoneme transduction (or G2P). Two types of G2P systems:

- **Rule-based:** Unitran, Epitran. Cover many languages where training data is not available. Only work well when orthographies are "shallow"
- **ML-based:** Based on WFSTs or seq2seq models. Require extensive training data in the form of pronouncing dictionaries, but work well for languages with "deep" orthographies like English and Arabic.

What is the IPA good for?

- **What is the IPA good for in linguistics?**
  - Documenting languages
  - Annotating speech data for many kinds of analysis
  - Providing a universal reference point for comparisons of sounds within and between languages
  - Characterizing disordered speech
- **What is the IPA good for in speech and language technologies?**
  - Textually representing pronunciation for both speech recognition and speech synthesis (X-SAMPA, which is an ASCII representation of the IPA, is also used for this)
  - Projecting data from different languages into a single representational space for transfer, identifying names, etc.
  - Others yet to be discovered.

Outline

Sound patterns

orthographic	phonetic
im-possible	[ɪm-pʰɑsəbl̩]
in-tolerable	[ɪn-tʰɑlərəbl̩]
in-conceivable	[ɪŋ-kʰənsivəbl̩]
il-legal	[ɪl-ligl̩]
ir-regular	[ɪɹ-ɹɛgjulɹ̩]

Allophones of English Plosives

aspirated	unreleased	unaspirated
p̲in	nip̲	sp̲in
t̲ick	kit̲	st̲ick
k̲in	nick̲	sk̲in

Allophany in Korean

kal	‘that’ll go’	ilkop	‘seven’	iɾɯni	‘name’
kɯnɯl	‘shade’	ipalsa	‘barber’	kiɾi	‘road’
mul	‘water’	onɯlpːam	‘tonight’	kɯɾəm	‘then’
pal	‘leg’	pulp^{h}jən	‘discomfort’	kəɾiɾo	‘to the street’
p^{h}al	‘arm’	silkwa	‘fruit’	saɾam	‘person’
səul	‘Seoul’	tɯl͡ʧaŋ	‘window’	uɾi	‘we’
tatɯl	‘all of them’	əlmana	‘how much’	yəɾɯm	‘summer’

Phonemes are contrasting units of sound.

Korean T-Charts

l	
a	#
ɯ	#
u	#
a	#
a	#
u	#
ɯ	#
i	k
a	s
ɯ	pː
u	p^h
i	k
ɯ	$\widehat{tʃ}$
ə	m

ɾ	
i	ɯ
ɯ	ə
ə	i
a	a
u	i
ə	ɯ

AlloVera: An Allophone Database

AlloVera `http://www.github.com/dmort27/allovera` is a database of phonemes and their allophones. One use case is universal automatic speech recognition (ASR):

- Training data sets come with orthographic transcriptions
- Existing tools can convert these to phonemic representations
- However, /p/ does not mean the same thing in English as it does in Chinese. In general, phonemes are language specific representations
- Our ASR system, Allosaurus, uses AlloVera to know what phonemes have what allophones in what languages. It can then learn phonetic, rather than phonemic, representations
- The result is an ASR system that can recognize (allo)phones in arbitrary languages.

Different Forms of the Same Morpheme

Singular	Phonemic	Plural	Phonemic
dog	/dɑg/	dogs	/dɑg-z/
cat	/kæt/	cats	/kæt-s/
horse	/hɔɹs/	horses	/hɔɹs-əz/

Different Forms of the Same Morpheme

Infinitive	Phonemic	3sg	Phonemic
take	/tejk/	takes	/tejks/
give	/gɪv/	gives	/gɪvz/
watch	/wɑt͡ʃ/	watches	/wɑt͡ʃəz/

Maori I

Active	Passive	Gerund	Gloss
hopu	hopukia	hopukaŋa	‘to catch’
aru	arumia	arumaŋa	‘to follow’
tohu	tohuŋia	tohuŋaŋa	‘to point out’
maatu	maaturia	maaturaŋa	‘to know’

Maori II

Active	Passive	Gerund	Gloss
hopu	hopuk-ia	hopuk-aŋa	‘to catch’
aru	arum-ia	arum-aŋa	‘to follow’
tohu	tohuŋ-ia	tohuŋ-aŋa	‘to point out’
maatu	maatur-ia	maatur-aŋa	‘to know’

Maori II

UR of Root	Active	Passive	Gerund	Gloss
/hopuk/	hopu	hopuk-ia	hopuk-aŋa	‘to catch’
/arum/	aru	arum-ia	arum-aŋa	‘to follow’
/tohuŋ/	tohu	tohuŋ-ia	tohuŋ-aŋa	‘to point out’
/maatur/	maatu	maatur-ia	maatur-aŋa	‘to know’

Maori Apocope

$$\left\{\begin{array}{c} \text{p} \\ \text{t} \\ \text{k} \\ \text{m} \\ \text{n} \\ \text{ŋ} \\ \text{r} \\ \dots \end{array}\right\} \rightarrow 0 \; / \; \_ \; \#$$

Voicing Assimilation

$$z \rightarrow s \,/\, \left\{ \begin{array}{c} p \\ t \\ k \\ f \\ \theta \\ s \\ \int \\ \widehat{t\int} \end{array} \right\} \_ \#$$

Schwa Epenthesis

$$0 \rightarrow \text{ə} \,/\, \left\{ \begin{matrix} \text{s} \\ \text{z} \\ \text{ʃ} \\ \text{ʒ} \\ \widehat{\text{tʃ}} \\ \widehat{\text{dʒ}} \end{matrix} \right\} \_ \left\{ \begin{matrix} \text{s} \\ \text{z} \\ \text{ʃ} \\ \text{ʒ} \\ \widehat{\text{tʃ}} \\ \widehat{\text{dʒ}} \end{matrix} \right\}$$

How Are They Ordered?

- **Voicing assimilation**

$$z \rightarrow s \,/\, \left\{ \begin{array}{c} p \\ t \\ k \\ f \\ \theta \\ s \\ \int \\ \widehat{t\int} \end{array} \right\} \_ \#$$

- **Epenthesis**

$$0 \rightarrow ə \,/\, \left\{ \begin{array}{c} s \\ z \\ \int \\ ʒ \\ \widehat{t\int} \\ \widehat{dʒ} \end{array} \right\} \_ \left\{ \begin{array}{c} s \\ z \\ \int \\ ʒ \\ \widehat{t\int} \\ \widehat{dʒ} \end{array} \right\}$$

Vowel Place Features

	i	y	ɨ	u	e	ø	ʌ	o	æ	œ	ɑ	ɒ
high	+	+	+	+	−	−	−	−	−	−	−	−
low	−	−	−	−	−	−	−	−	+	+	+	+
back	−	−	+	+	−	−	+	+	−	−	+	+
round	−	+	−	+	−	+	−	+	−	+	−	+

Some Consonant Place Features

	p	t̪	t	t͡ʃ	ʈ	c	k	q	ʕ	ʔ
anterior	+	+	+	−	−	−	−	−	−	−
coronal	−	+	+	+	+	−	−	−	−	−
distributed		+	−	+	−					
high	−	−	−	−	−	+	+	−	−	−
back	−	−	−	−	−	−	+	+	+	−
low	−	−	−	−	−	−	−	−	+	−

PanPhon

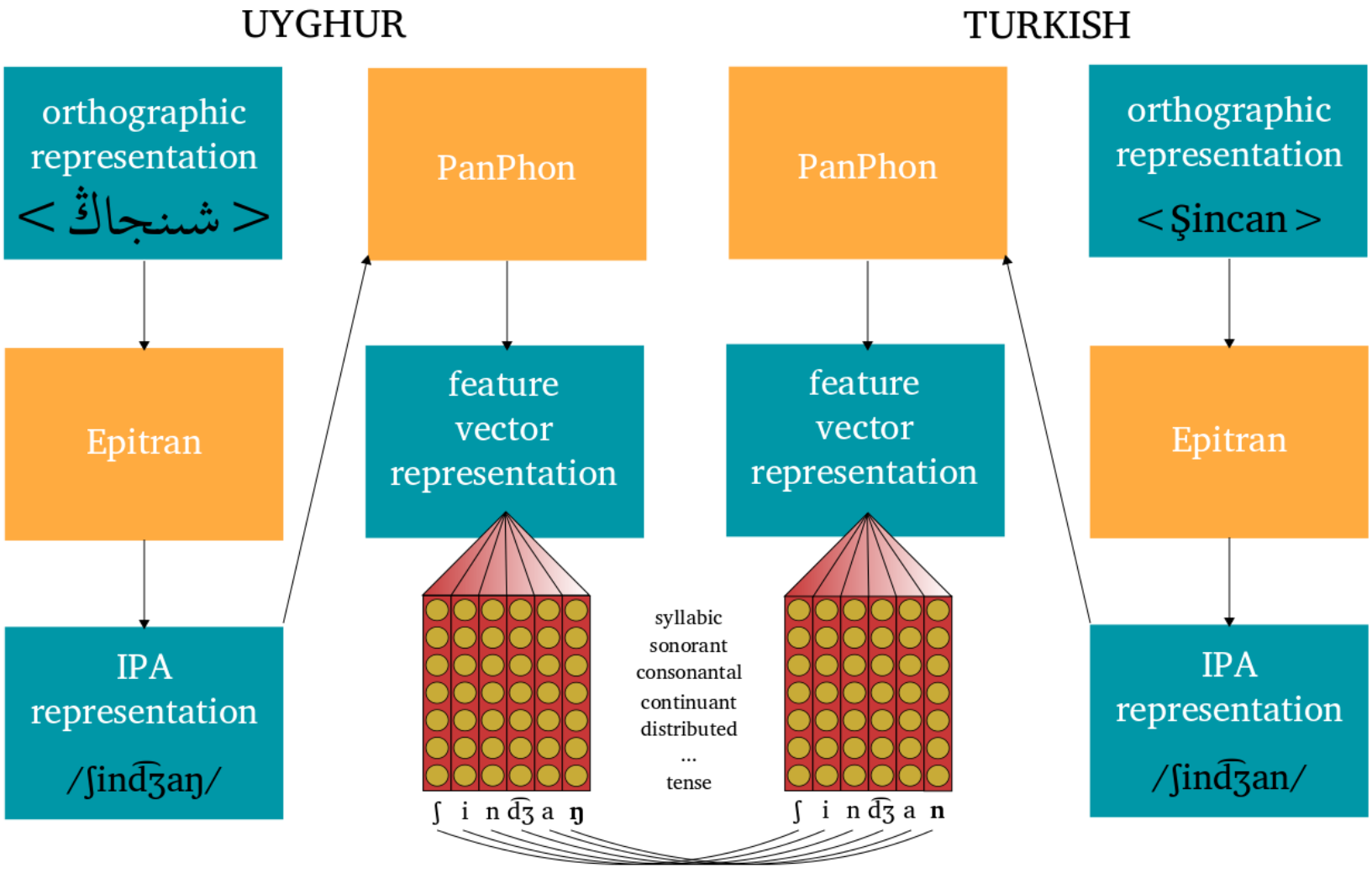
PanPhon is an ontology and associated Python library for dealing with phonological feature representations.

- `https://github.com/dmort27/panphon`
- Feature tables at `https://github.com/dmort27/panphon/blob/master/panphon/data/ipa_bases.csv` and `https://github.com/dmort27/panphon/blob/master/panphon/data/ipa_all.csv`
- `pip install panphon`
- Python 2 and 3
- Compute subsumption relations among features
- Compute edit distance between features
- Compatible with **Epitran** G2P

Similarity in Names

Language	Orthography	IPA
English	Xinjiang	/ʃɪnd͡ʒæŋ/
Chinese	新疆	/ɕint͡ɕjaŋ/
Uyghur	شىنجاڭ	/ʃind͡ʒaŋ/
Uzbek	Sinszyan	/sinszjan/
Turkish	Sincan	/sind͡ʒan/
Kazakh	Шыңжаң	/ʃəŋʒɑŋ/

Epitran and PanPhon in Action



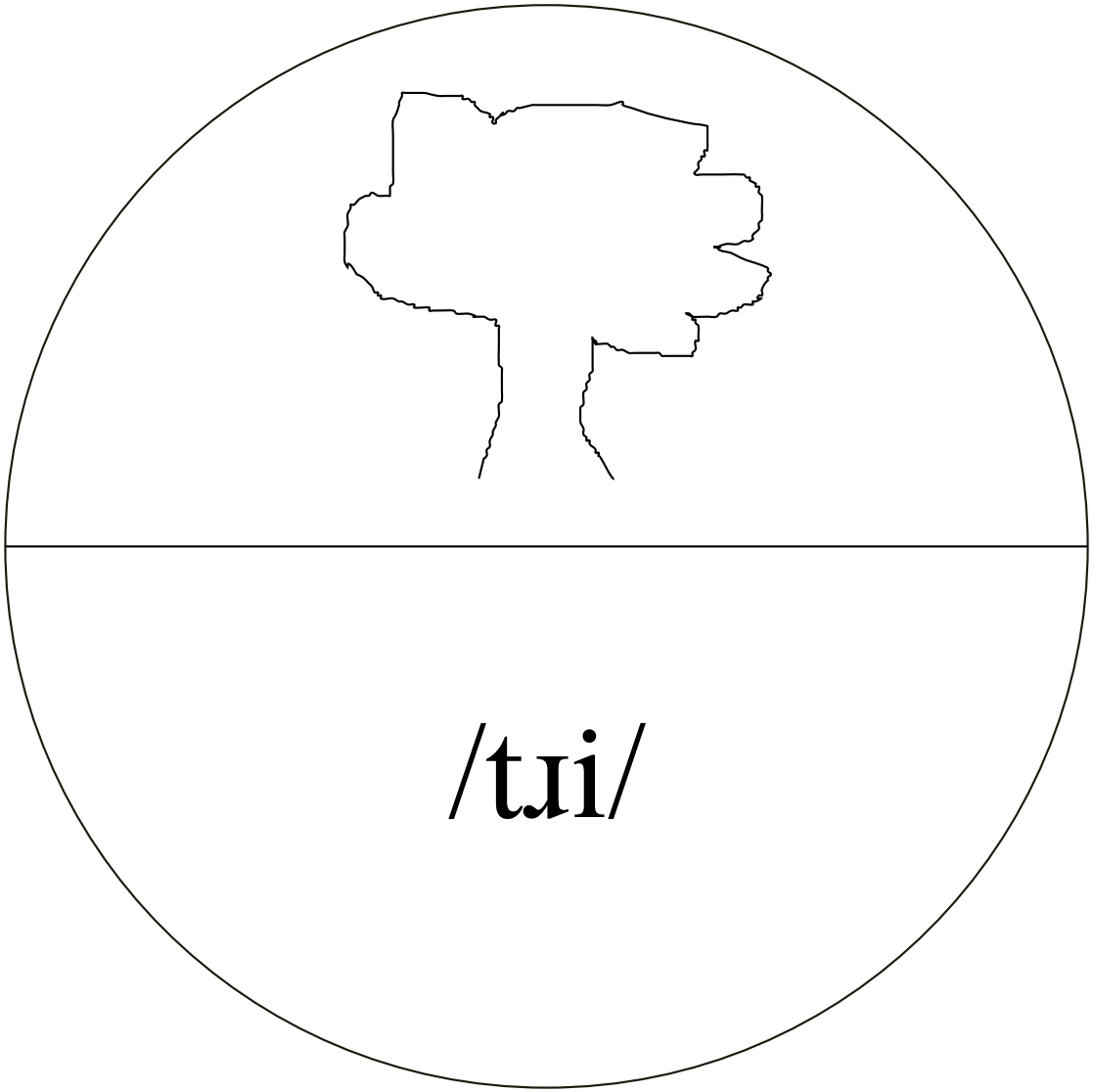
PHOIBLE

PHOIBLE is a database of segment inventories for a large number (2186) of languages.

- `https://phoible.org/`
- `https://github.com/phoible/dev`
- Phonological feature for each segment in each inventory
- Somewhat different system that **PanPhon**

Outline

Signs



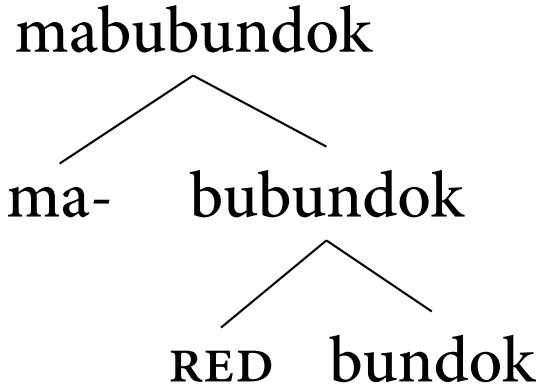
Words, Lexemes, and Listemes

- Listemes
- Lexemes
- Words

Tagalog Adjectives I

Stem	Singular	Plural	Gloss
laki	malaki	malalaki	‘big’
ganda	maganda	magaganda	‘beautiful’
bundok	mabundok	mabubundok	‘mountainous’

Tagalog Adjectives II



Tagalog Verbs

Stem	Perfective	Contemplative	Imperfective	Gloss
tapos	tinapos	tatapusin	tinatapos	‘finish’
kain	kumain	kakain	kumakain	‘eat’
sulat	sumulat	susulat	sumusulat	‘write’
hanap	humanap	hahanap	humahanap	‘seek’

German Verbs

	Present	Perfect	Preterit
1SG	mache	gemacht	machte
2SG	machst	gemacht	machtest
3SG	macht	gemacht	machte
1PL	machen	gemacht	machten
2PL	macht	gemacht	machtet
3PL	machen	gemacht	machten

Arabic Morphology

	Perfect		Imperfect		Participle	
	Active	Passive	Active	Passive	Active	Passive
I	katab	kutib	ktub	ktab	kaatib	ktuub
II	kattab	kuttib	kattib	kattab	kattib	kattab
III	kaatab	kuutib	kaatib	kaatab	kaatib	kaatab
IV	ʔaktab	ʔuktib	ktib	ktab	ktib	ktab
V	takattab	tukuttib	takattab	takattab	takattib	takattab
VI	takaatab	tukuutib	takaatab	takaatab	takaatib	takaatab
VII	nkatab	nkutib	nkatib	nkatab	nkatib	nkatab
VIII	ktatab	ktutib	ktatib	ktatab	ktatib	ktatab
IX	ktab(a)b	ktab(i)b	ktab(i)b			
X	staktab	stuktib	staktib	staktab	staktib	staktab

Mandarin Personal Pronouns

我	*wo*	1SG	我们	*women*	1PL
你	*ni*	2SG	你们	*nimen*	2PL
他	*ta*	3SG	他们	*tamen*	3PL

Mandarin Compounding

客厅	‘living room’	沙发	‘sofa’	‘living room sofa’
眼	‘eye’	药	‘medicine’	‘eye medicine’
马	‘horse’	房	‘house’	‘manger’
雨	‘rain’	帽	‘hat’	‘rain hat’

Internal Change in English

- Ablaut affects verbs
  - *sing* : *sang* : *sung*
  - *begin* : *began* : *begun*
  - *bleed* : *bled* : *bled*
- Umlaut affects nouns
  - *foot* $\rightarrow$ *feet*
  - *tooth* $\rightarrow$ *teeth*
  - *goose* $\rightarrow$ *geese*

Derivational morphology creates new lexemes.

It always changes meaning, part of speech, or both.

Examples of Derivation

(1) a. Pittsburgh-er
b. re-surface
c. black-ish
d. dov-ish
e. acquitt-al
f. efficient-ly

Inflectional morphology does not create new lexemes.

It adds information based on the syntactic context in which a word occurs.

Some Kinds of Inflection

- Case
- Number
- Clusivity
- Grammatical gender
- TAM
  - Tense
  - Aspect
  - Modality
  - Evidentiality
- Formality
- Voice (?)

Examples of Inflection

(2) dog-s

(3) a. walk-s

b. walk-ed

c. walk-ing

Cumulative Exponence in Latin

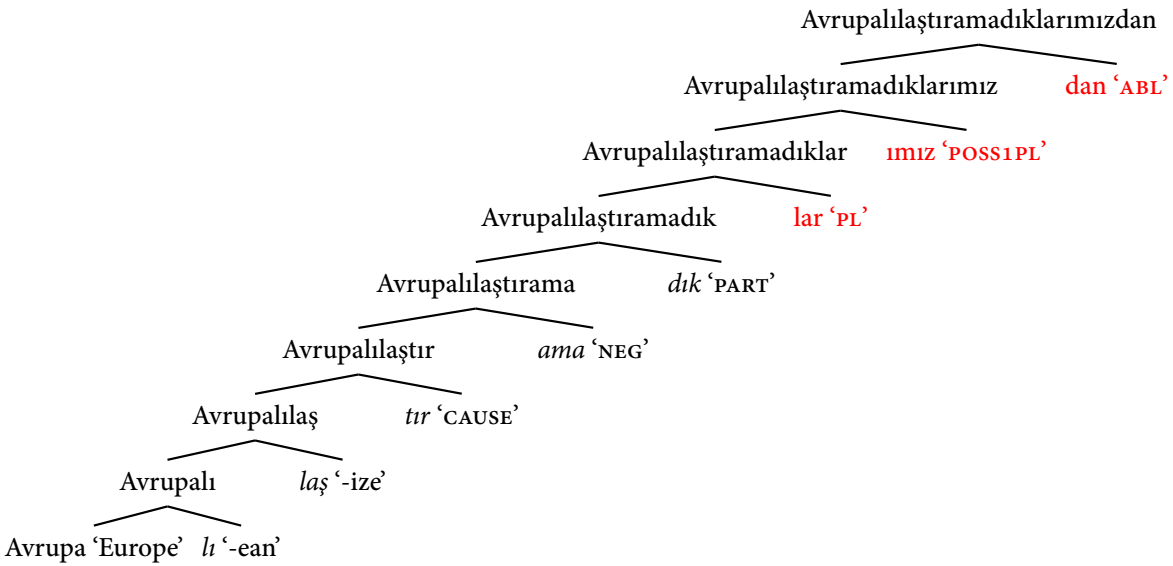
	SG	PL
NOM	amīca	amīcae
VOC	amīca	amīcae
ACC	amīcam	amīcās
GEN	amīcae	amīcārum
DAT	amīcae	amīcīs
ABL	amīcā	amīcīs

Isolating Languages

(4) 里斯 对 这 个 案件 的 调查
Lisi dui zhei ge anjian de diaocha
Lisi to this CLF case DE investigation
进行 了 一 个 小时
jinxing le yi ge xiaoshi
last ASP one CLF hour
'Lisi's investiagation of the case lasted an hour.'

Agglutinative Languages

An example from Turkish:



'of ours that were unable to be Europeanized'

Fusional/Flexional Languages

	Present	Perfect	Preterit
1SG	mache	gemacht	machte
2SG	machst	gemacht	machtest
3SG	macht	gemacht	machte
1PL	machen	gemacht	machten
2PL	macht	gemacht	machtet
3PL	machen	gemacht	machten

Templatic Languages

	Perfect		Imperfect		Participle	
	Active	Passive	Active	Passive	Active	Passive
I	katab	kutib	ktub	ktab	kaatib	ktuub
II	kattab	kuttib	kattib	kattab	kattib	kattab
III	kaatab	kuutib	kaatib	kaatab	kaatib	kaatab
IV	ʔaktab	ʔuktib	ktib	ktab	ktib	ktab
V	takattab	tukuttib	takattab	takattab	takattib	takattab
VI	takaatab	tukuutib	takaatab	takaatab	takaatib	takaatab
VII	nkatab	nkutib	nkatib	nkatab	nkatib	nkatab
VIII	ktatab	ktutib	ktatib	ktatab	ktatib	ktatab
IX	ktab(a)b	ktab(i)b	ktab(i)b			
X	staktab	stuktib	staktib	staktab	staktib	staktab

Polysynthetic Languages

Polysynthetic languages are languages in which noun arguments like objects can be expressed as part of a verb, meaning that full sentences can be expressed as a verb alone (not just through agreement with person and number, but through the "incorporation" of the noun into the verb). Take the following example from Nahuatl:

- *ni-c-qua in nacatl*
  I-it-eat the flesh
  'I eat the flesh.'
- *ni-naca-qua*
  I-flesh-eat
  'I eat flesh.'

Sapir's Two-Dimensional Typology

- **Degree of synthesis.** How many properties are there per word?
- **Degree of fusion.** How many properties are there per operation?

Outline

Syntax

- Syntax is about the structure of phrases and sentences—structure above the level of words.
- Two kinds of structure:
  - Constituency structure/phrase structure
  - Dependency structure
- Syntax will be covered in a later lecture.

Outline

Semantics

- Semantics is about the (literal) meaning of words, phrases, and sentences.
- As a field, it overlaps with both linguistics and philosophy of language.

Pragmatics and Discourse

- Pragmatics is about language use in context.
- Discourse is specifically about language use in it's **linguistic** context above the level of the sentence.

Outline

Sociolinguistics

- Language is a social phenomenon
- Sociolinguistics explores how language interacts with social factors
  - How it reflects social identity
  - How it constructs social identity