Lecture 26

Computational Linguistics

August 4th, 2022

Cooper Bedin

Announcements

- HW 6 due tonight
- Lab 11 due tonight
- Scheme checkpoint 2 due tomorrow; project due Tuesday
- Last discussion sections today :(
- Topical review sessions next week!
- Unfortunately, we cannot grant any extensions past Wednesday of next week (11:59PM the night before the final) because we need to get grades to the university on time
    - If there's an emergency you can email us and we'll work something out

Computational Linguistics

A brief history of linguistics

People have always been interested in studying and describing the structure of language! As far back as the 6th century BC we can find a formal description of Sanskrit grammar

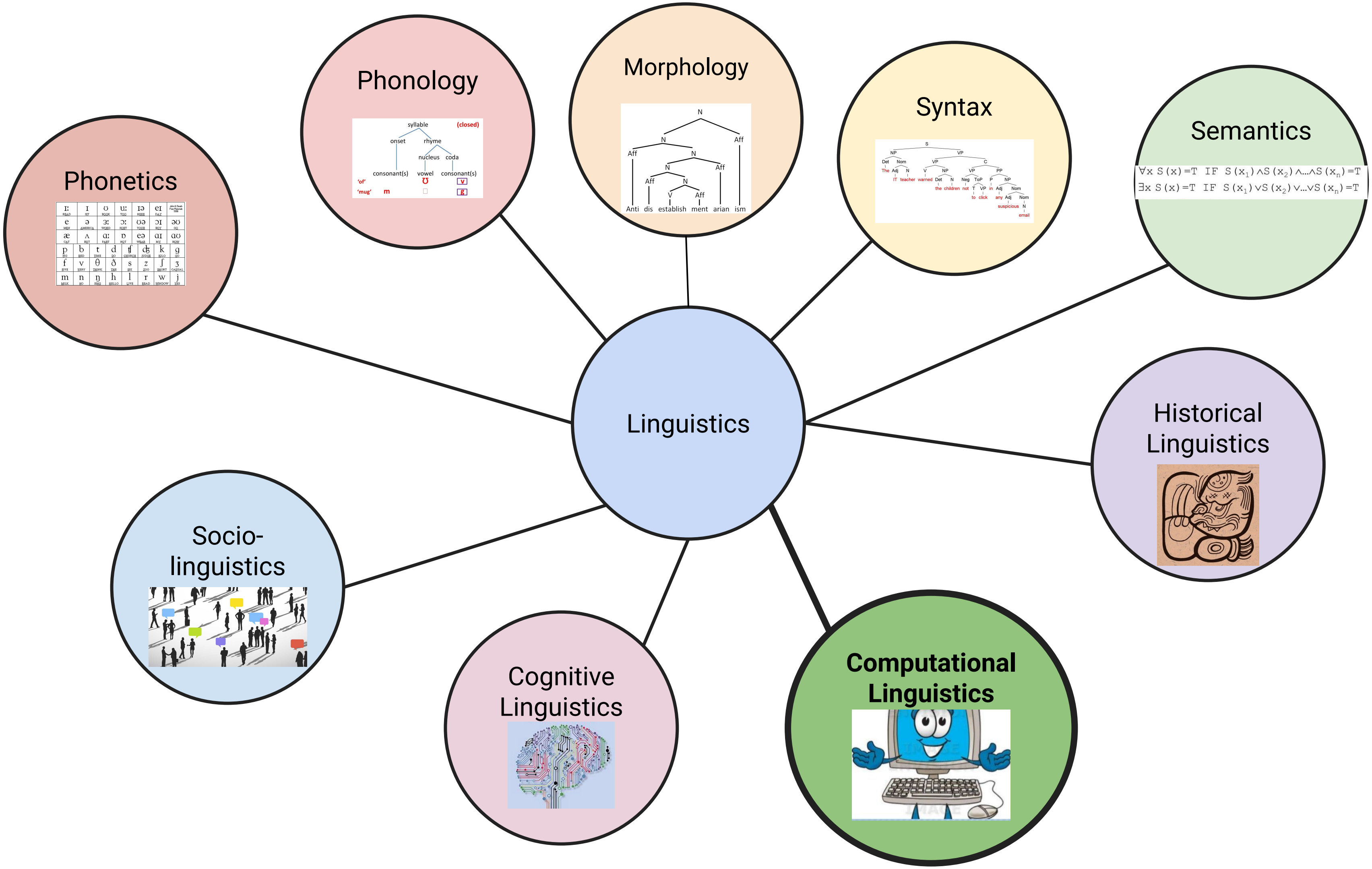
Prior to the 20th century, the study of language was often referred to as **philology**, and was predominantly focused on history, etymology, and literature

Linguistic methods as they are practiced today started to take shape in the 18th century, and in the 20th century linguistics emerged as a distinct social science

Linguistics is a field with a relatively short history, meaning that there's still a ton to learn! It's also a field with a deep history of colonialism and eurocentrism, and there's an emerging need to challenge many established research practices

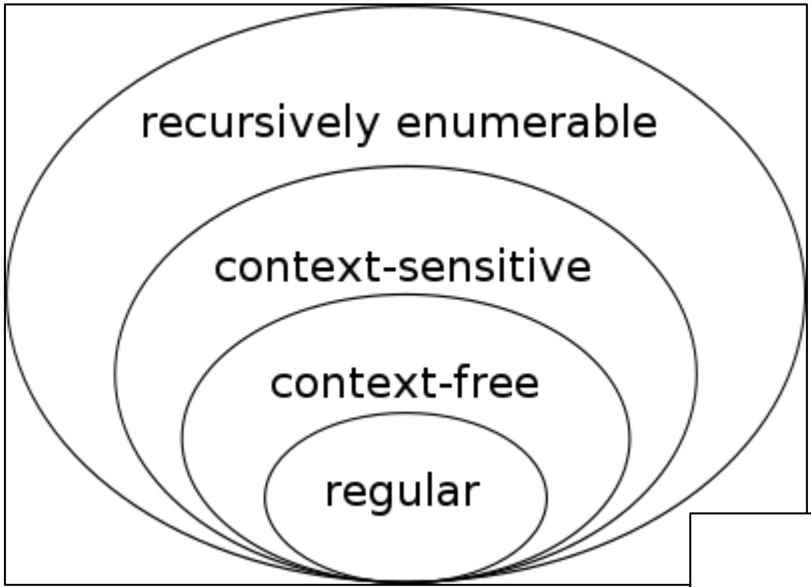
Ok, but what is linguistics?

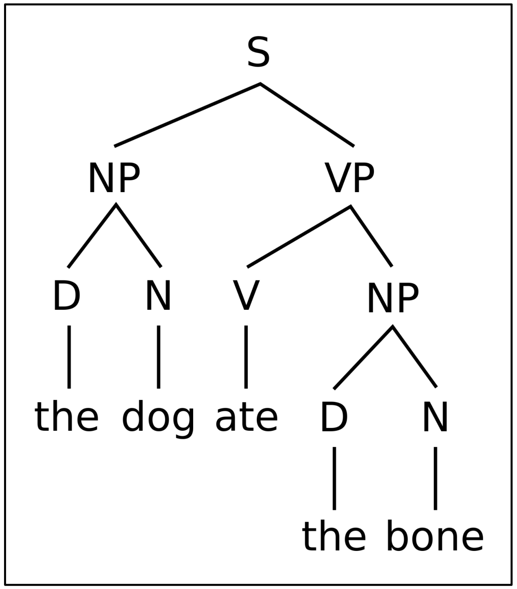
Linguistics is, at its core, a scientific approach to the study of language



Linguistics and computation

Theories about language and computation are deeply connected!





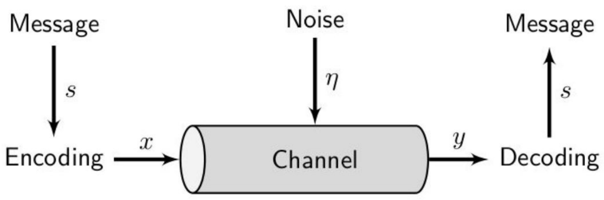


Figure 1: The communication channel. A message (data) is encoded before being used as input to a communication channel, which adds noise. The channel output is decoded by a receiver to recover the message.

What is computational linguistics?

Computational linguistics is honestly a pretty vague term—basically any time you're using computational methods or models in linguistics research, you're doing computational linguistics!

Typically, in computational linguistics we're specifically drawing on **artificial intelligence** and **machine learning** models

It turns out that we can (and do) apply computational methods to pretty much all of the subfields of linguistics from the previous slide :0

You may also have heard the term **natural language processing**—computational linguistics is related to NLP, but is also a little bit different

Machine learning

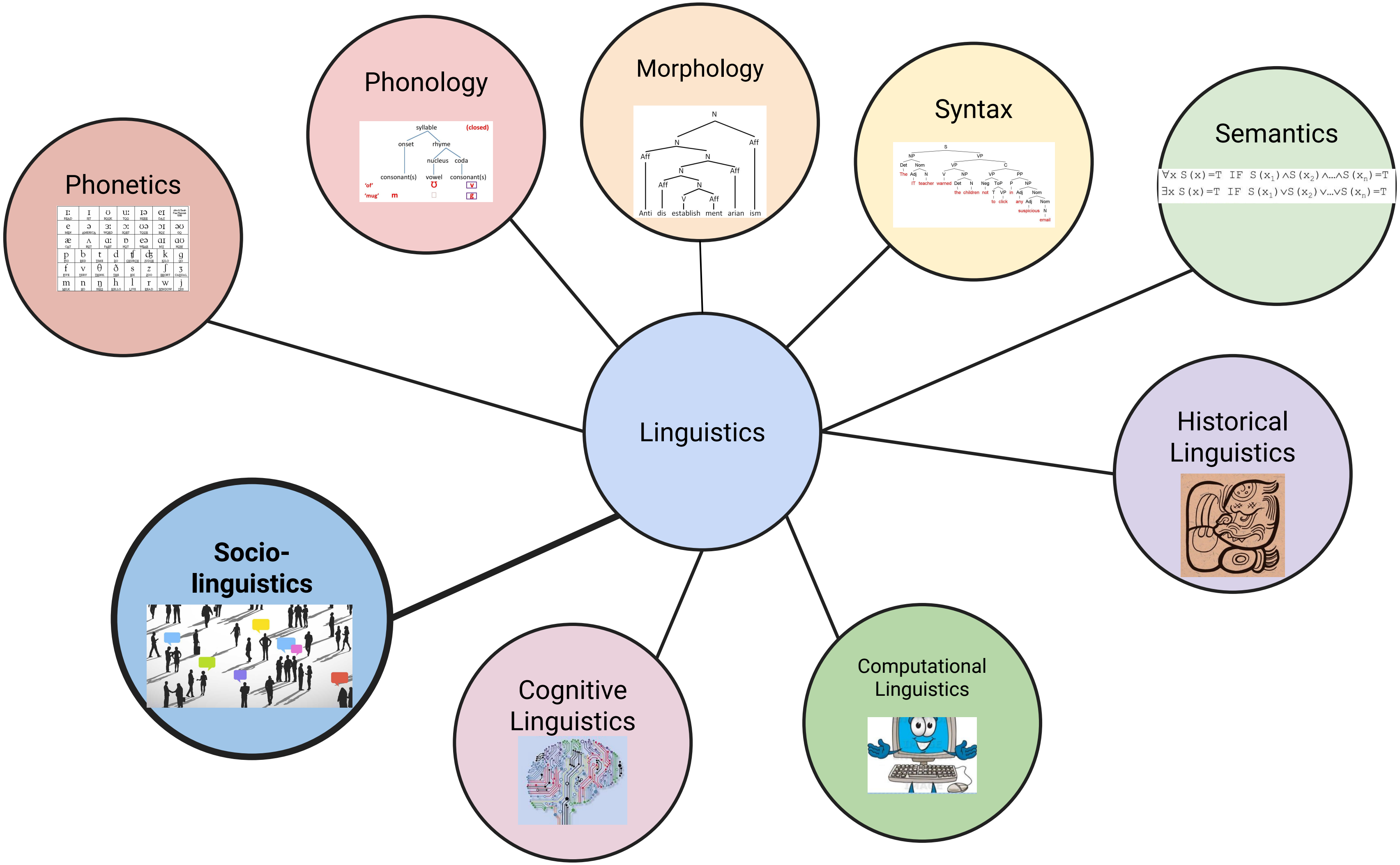
Machine learning is a field of computer science based on creating algorithms that can "learn" how to solve problems based on data

Rather than writing a narrow algorithm to solve a specific problem, ML algorithms are much more general, and they learn to solve specific problems depending on the data they're given

For example, rather than writing a program that can translate between English and French*, we could create an algorithm that learns patterns in how to translate between two languages—then we could train it on any pair of languages

**Even this is something we don't quite have a handle on yet*

My research!



Linguistics

Phonetics

Phonology

Morphology

Syntax

Semantics

$\forall x\ S(x)=T\ \text{IF}\ S(x_1)\wedge S(x_2)\wedge\ldots\wedge S(x_n)=T$

$\exists x\ S(x)=T\ \text{IF}\ S(x_1)\vee S(x_2)\vee\ldots\vee S(x_n)=T$

Historical Linguistics

Computational Linguistics

Cognitive Linguistics

Socio-linguistics

Sociolinguistics

Sociolinguistics is—as the name implies—the study of the relationship between language and society. This includes both the ways that language can shape society and culture, and the ways that society and culture shape language

A major component of sociolinguistics is looking at **variation**—the ways that speakers of the same language may make different decisions about word choice, pronunciation, etc. based on their identity and social context (often totally subconsciously)

A **sociolinguistic variable** is a specific feature of a language that can be a site of variation

Example: The fourth floor of Macy's

This is a very famous and old (1966) example of an early sociolinguistic experiment, conducted by William Labov:

Researchers went to three different department stores (Saks, Macy's, and Klein's), which are notably at different ends of the price and fashion scales

Researchers would ask an associate at each of these stores a question, the answer to which is "Fourth Floor" (e.g. "where can I find the men's shoes?")

Researchers found that, the lower the social prestige of the store, the more likely the sales associate was to drop the 'r' sound in "Fou**r**th Floo**r**"

This demonstrates that this linguistic variable is tied with a social identity! (in this case class)

Queer Speech

There's a pre-existing body of research on "gay speech"—the idea that there is a way to speak that sounds specifically "gay" or "queer," especially in reference to homosexual, cisgender men

Much of this research looks at "accuracy" on the end of the listener—whether a listener is able to determine that a speaker is either heterosexual or homosexual, based only on listening to a recording of their voice (e.g. Tracy and Satariano 2011)

In many experiments listeners are asked to read identical passages or lists of words, to allow us to look solely at differences in pronunciation as an indicator of queerness

It does turn out that listeners are better than chance at this task, which is itself super interesting

Sociophonetic variables

In the paper cited on the previous slide it was found that /s/, along with certain vowels, are particularly important indicators of queer speech—analysis showed that variations in pronunciations of these sounds correlated with listener judgements

However, other sounds such as /n/, /m/, /f/, /v/, /l/, /w/ were not. This helps us to know what **variables** to look at or not look at

This, and much other pre-existing research, is based on statistical methods—a consequence of this is that researchers have to know in advance which variables they want to look at. Additionally, most statistical methods only look at correlation for one variable at a time

My research is based on throwing machine learning at this question!

Experiment Set-Up

Original experiment

The research I did for my thesis was based on the experiment I talked about in the previous slides, which was run by Keith Johnson (UC Berkeley) and Erik Tracy (UNC Pembroke)

In the original experiment, speakers were recorded reading a list of words, and a panel of listeners was given the words to listen to

For each word, listeners rated on a scale of 1–7 how "gay" the speaker sounded to them (1=straight, 7=gay)

New experiment

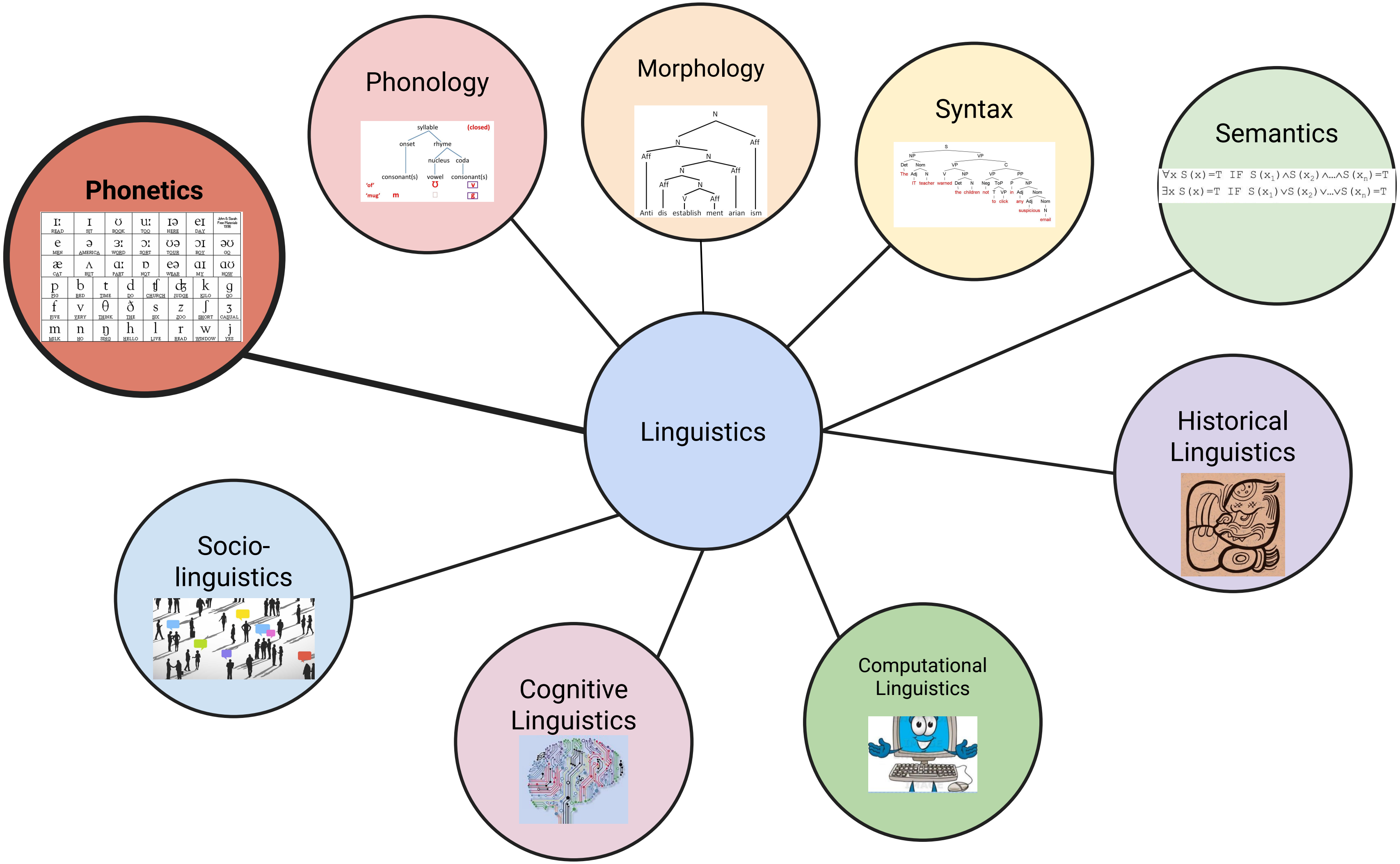
In my version, speakers were asked to record themselves reading a list of sentences aloud:

- Do you hear the sleigh bells ringing?
- Masquerade parties tax one’s imagination.
- Although always alone, we survive.
- I honor my mom.
- Jeff’s toy go-cart never worked!
- etc.

Because this was during a global pandemic, speakers recorded themselves out of their homes through a virtual recorder

Listeners were then administered a virtual survey where they listened to a subset of these recordings, and rated 1–7 how "queer/gay" the speaker sounded to them

Acoustic Analysis



Linguistics

Phonetics

iː READ	ɪ SIT	ʊ BOOK	uː TOO	ɪə HERE	eɪ DAY	John & Sarah Free Materials 1996
e MEN	ə AMERICA	ɜː WORD	ɔː SORT	ʊə TOUR	ɔɪ BOY	əʊ GO
æ CAT	ʌ BUT	ɑː PART	ɒ NOT	eə WEAR	aɪ MY	aʊ HOW

p PIG	b BED	t TIME	d DO	ʧ CHURCH	ʤ JUDGE	k KILO	g GO
f FIVE	v VERY	θ THINK	ð THE	s SIX	z ZOO	ʃ SHORT	ʒ CASUAL
m MILK	n NO	ŋ SING	h HELLO	l LIVE	r READ	w WINDOW	j YES

Phonology

Morphology

Syntax

Semantics

$\forall x\ S(x)=T\ \text{IF}\ S(x_1)\wedge S(x_2)\wedge\ldots\wedge S(x_n)=T$

$\exists x\ S(x)=T\ \text{IF}\ S(x_1)\vee S(x_2)\vee\ldots\vee S(x_n)=T$

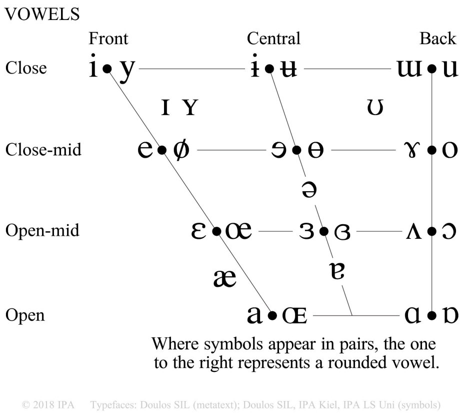
Historical Linguistics

Computational Linguistics

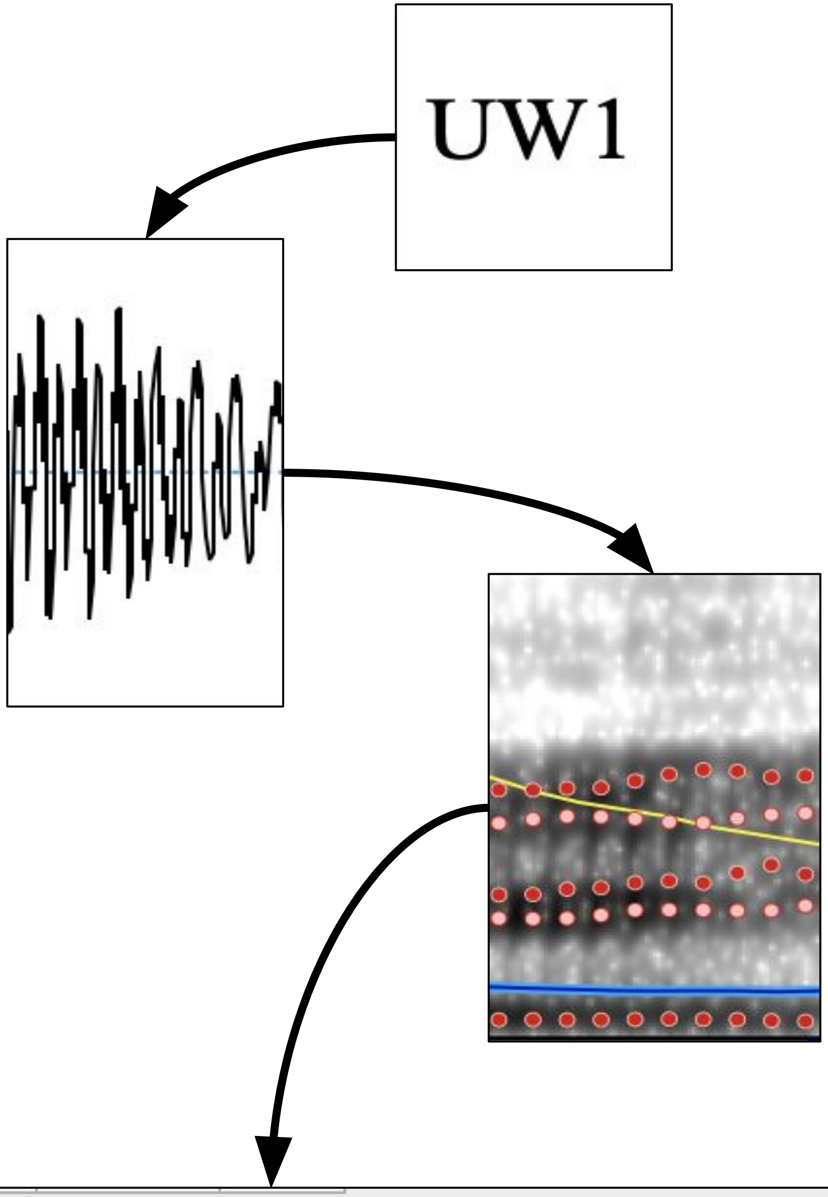
Cognitive Linguistics

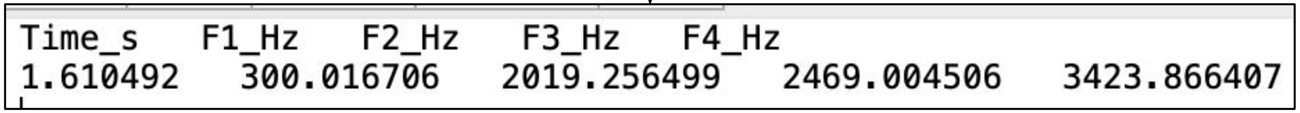
Socio-linguistics

Vowel features


VOWELS
Front
Central
Back
Close
i • y
ɨ • ʉ
ɯ • u
ɪ ʏ
ʊ
Close-mid
e • ø
ɘ • ɵ
ɤ • o
ə
Open-mid
ɛ • œ
ɜ • ɞ
ʌ • ɔ
æ
ɐ
Open
a • ɶ
ɑ • ɒ
Where symbols appear in pairs, the one to the right represents a rounded vowel.
© 2018 IPA Typefaces: Doulos SIL (metatext); Doulos SIL, IPA Kiel, IPA LS Uni (symbols)

Vowels are categorized by height and backness, which we measure quantitatively using pre-existing software

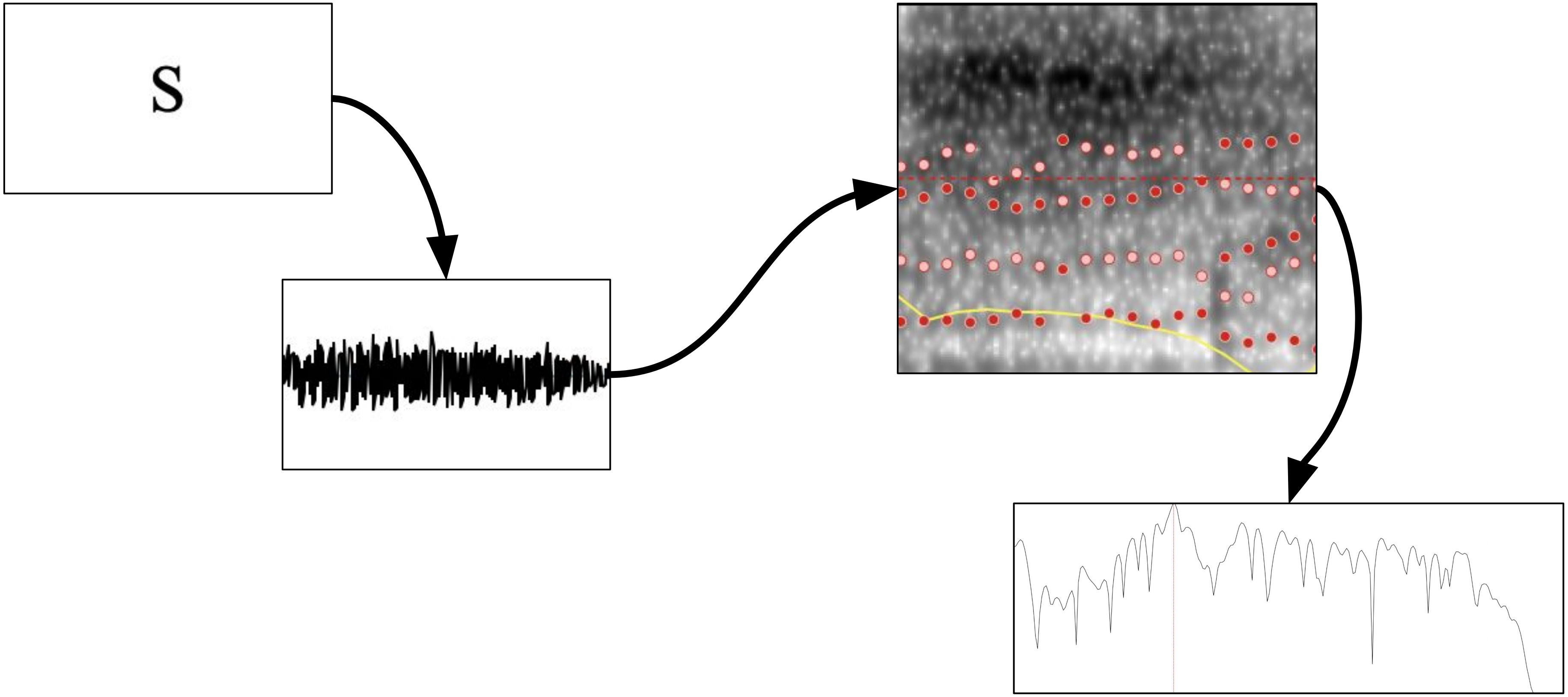




```
Time_s     F1_Hz      F2_Hz      F3_Hz      F4_Hz
1.610492    300.016706    2019.256499    2469.004506    3423.866407
```

/S/

If we try to apply the same algorithms to /s/, it won't work because /s/ has different acoustic properties, so we have to take a different kind of measurement



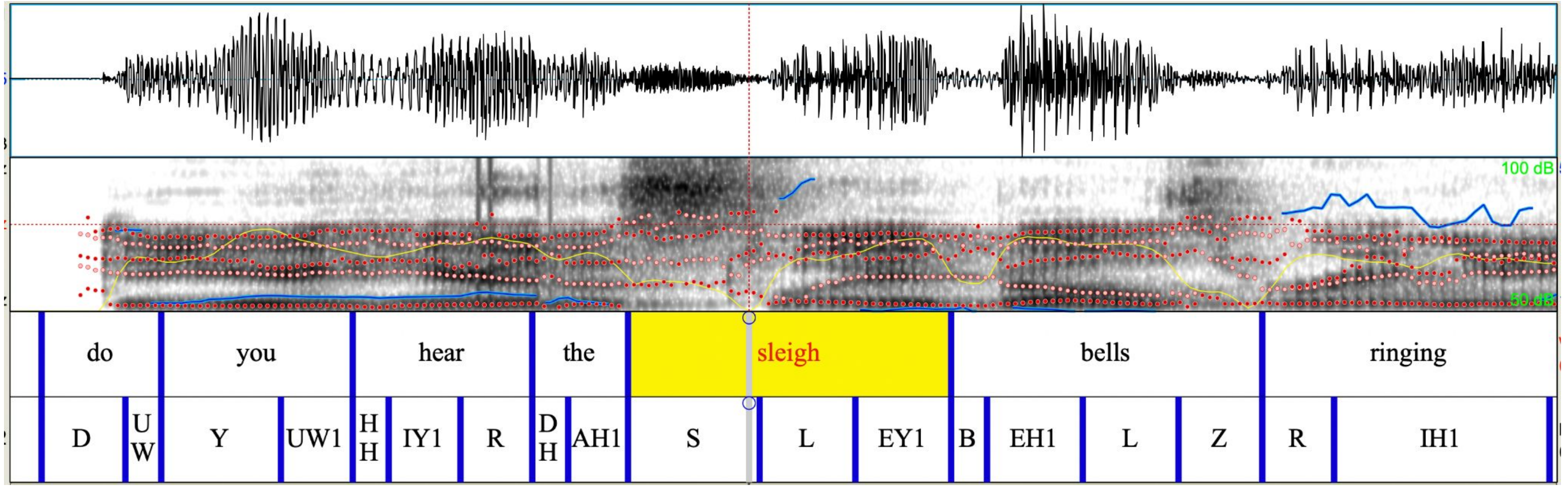
S

Segmentation

The algorithms from the previous slides allow us to extract numerical measurements of natural speech, but we also just saw that different kinds of sounds need different kinds of measurements

This means that we also need to do something called **segmentation**, where we chunk up a long recording into all the little sounds inside

The algorithm to do this is based on machine learning!!



Analysis

Where's the computational linguistics?

So far we've already seen several instances of computation:

- Speakers used a recorder embedded in a survey to record their sentences, which was written in JavaScript
- Listeners were administered their task through a survey written in PHP
- Recordings were segmented using the Montreal Forced Aligner, which is a pre-packaged aligning algorithm

All of these used instances of computation, but many folks who don't call themselves "computational linguists" also use these resources—computation is helping us a lot with data collection, but we have yet to apply any computational models to our analysis

I also didn't write any of these things myself—nothing we've seen so far is super new

Machine learning (again!)

Acoustic analysis allows us to take recordings of speakers and generate numerical data from these recordings—this is hugely helpful for when we want to do real, quantitative analysis

Because speakers were all given exactly the same list of sentences, any difference in data between two speakers will be due to differences in those speakers' pronunciations

The original experiment's analysis was primarily based at looking at correlations between specific acoustic measurements and the ratings given by listeners

My analysis is based on turning that idea around: what if we train a model to take in acoustic measurements as inputs, and give back ratings as outputs? Will it be able to learn meaningful trends in the data?

Modules

Now we use a lot of Python! We'll leverage a lot of libraries here:

- `parselmouth` implements algorithms for extracting acoustic measurements
- numpy allows us to arrange those measurements into **matrices** so that we can operate on them at a large scale
- `sklearn` (Scikit-Learn) implements a huge host of machine-learning algorithms
- `matplotlib` converts data into visualizations

Models

Imagine we have a recording of a speaker reading this sentence

"Do you hear the sleigh bells ringing?"

This sentence contains 21 different sounds, from which I was able to extract 192 acoustic measurements using Parselmouth

For my analysis, I trained **linear models** on this data to find patterns. Linear models are based on a formula like

$$\mathtt{RATING} = a_0 \cdot \mathtt{MEAS}_0 + a_1 \cdot \mathtt{MEAS}_1 + \ldots$$

Where, for every speaker we multiply each measurement $\mathtt{MEAS}_i$ by some constant number a_i, and add them all together to get a rating for that speaker

The goal of the model is to "learn" the best a's (called **weights**) so that for every speaker, we get as close as possible to the actual RATING through this algorithm

Choosing good models

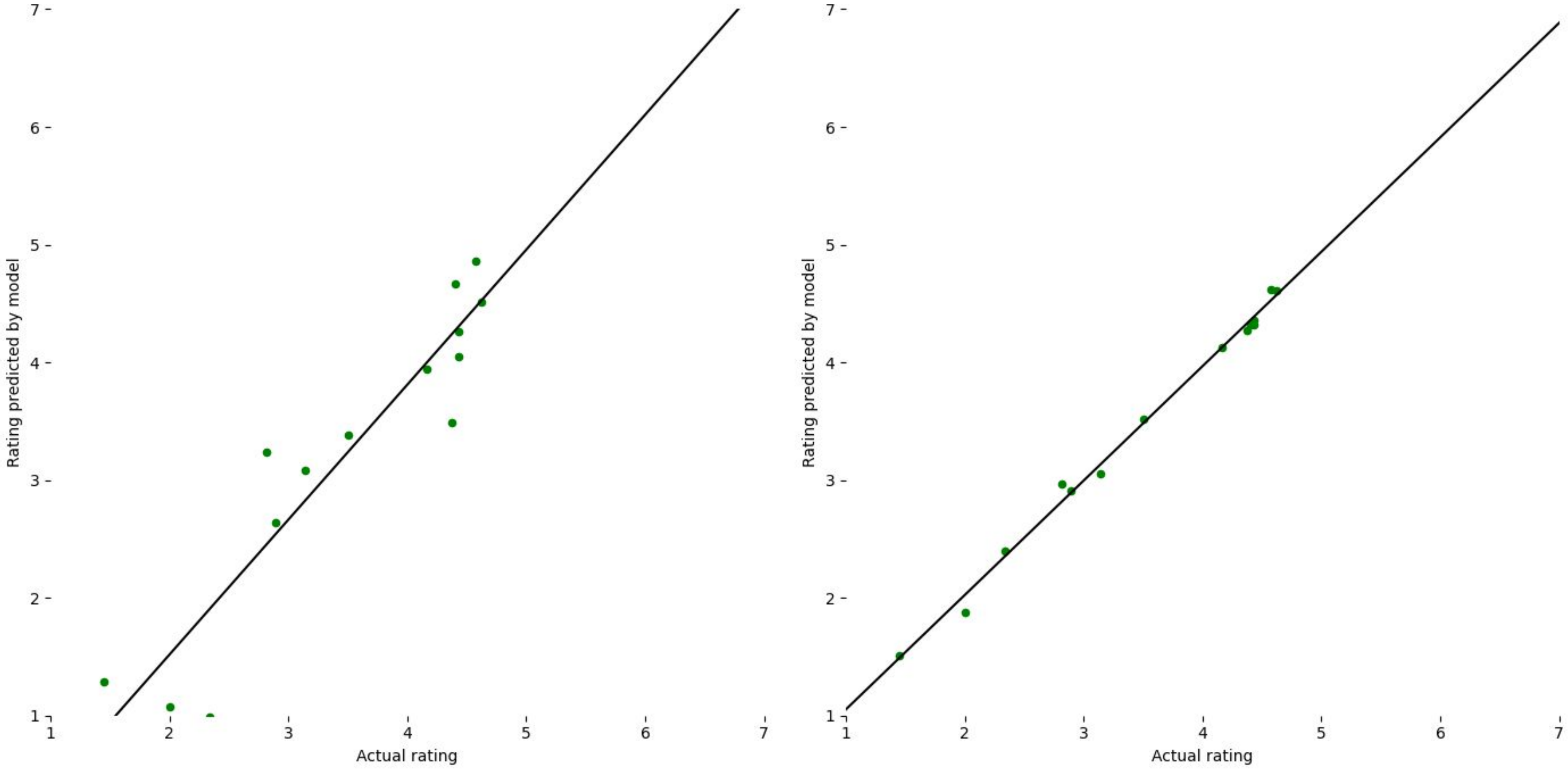
Choice of model is super important when doing machine learning!

In this project, the goal is to train a model that can identify which acoustic measurements correlate well with queerness ratings, and which ones don't seem to matter

With that in mind, I used two models—LASSO and "greedy" feature selection. I chose these models in particular because they both learn as few weights as possible—by their design, they'll only focus on the measurements they find the most helpful, and ignore the rest

Results

Good fits!



Model fits for "Do you hear the sleigh bells ringing?"

What did the models learn?

These are the features and weights that the "greedy" model learned for this sentence

Sound	**Meas**	**Weight**
TH	CoG	0.13
ER	F2	0.44
N	F1	0.47
Er	F0	0.74
SH	CoG	0.81
W	F3	1.48
SH	CoG	1.64

This /SH/ is super interesting! Lots of currently existing research looks at /S/ (I talked about some a bit earlier), but there's not a lot of folks looking at /SH/

Let's listen!

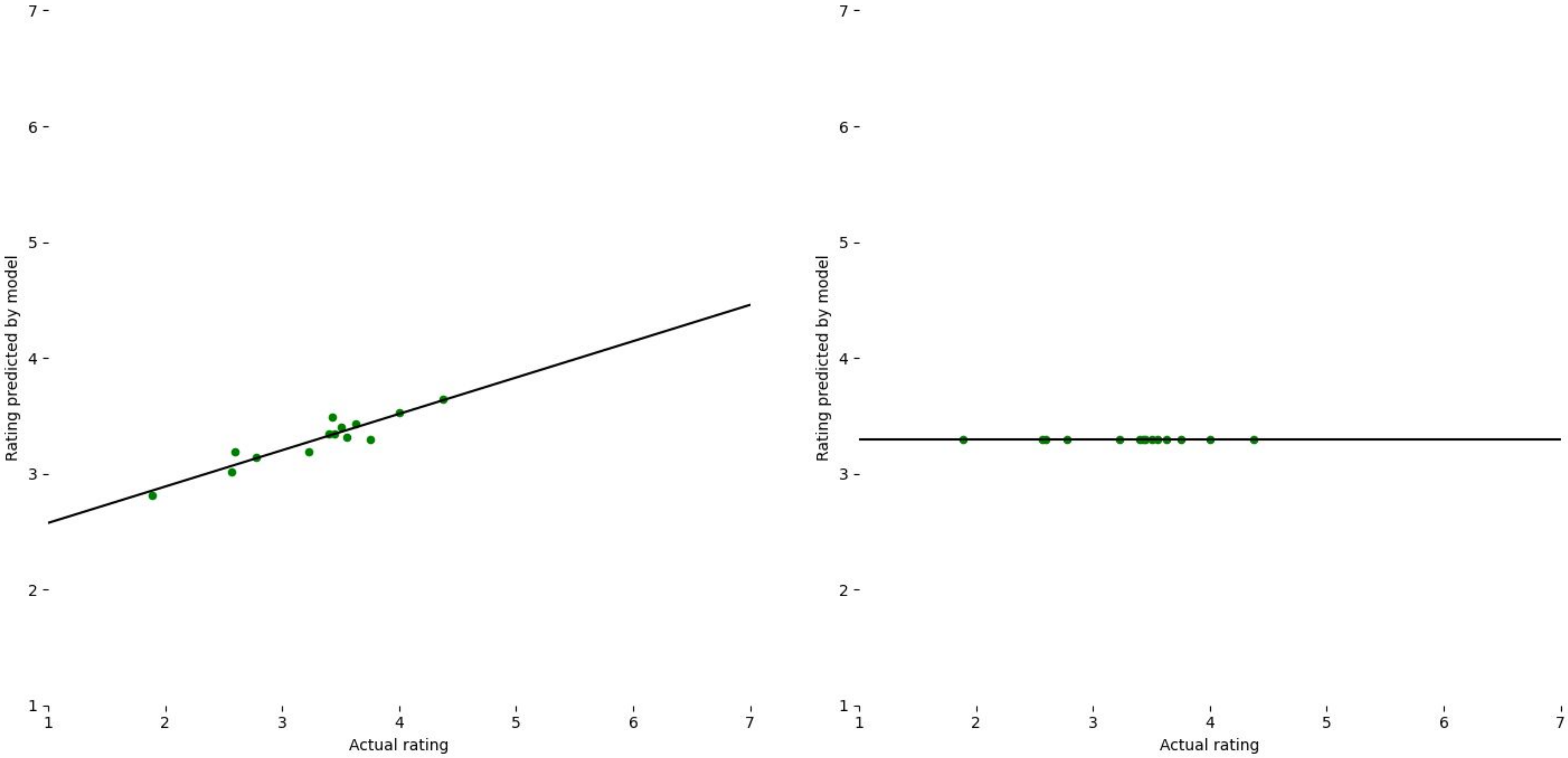
🔊

Speaker that was rated to sound most "straight" by the model

🔊

Speaker that was rated to sound most "queer/gay" by the model

Bad fits :(



Model fits for "Allow leeway here."

Major takeaways

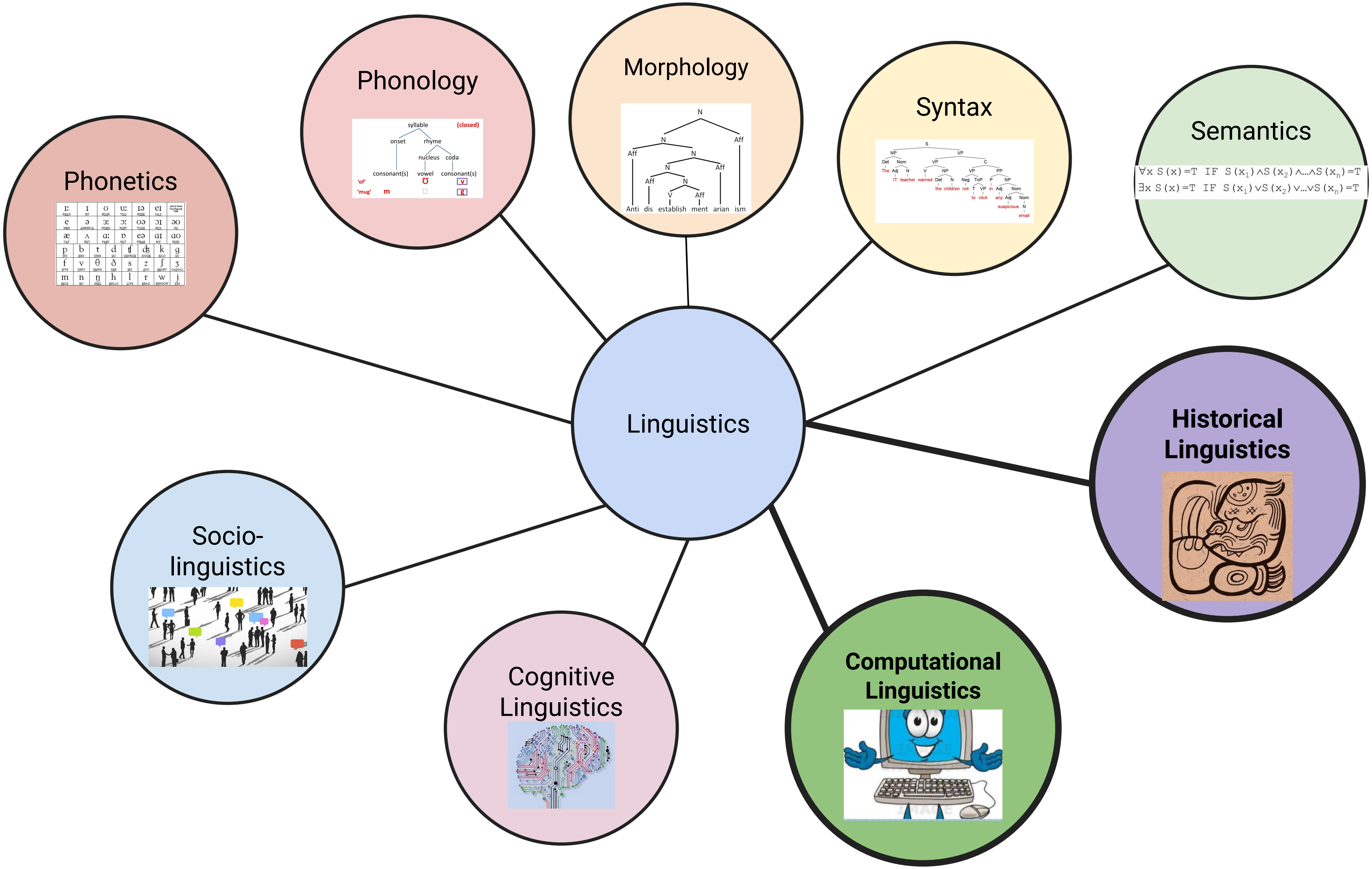
The model identified a feature to look at that pre-existing research hasn't spent a lot of time on yet!

They also were able to process data really efficiently—I was able to crunch a huge amount of information (14 speakers x 25 sentences x 23 listeners) in a matter of weeks

The models also confirmed something that we already know—if listeners don't have access to the sounds that "carry" the most information about queerness, it's really hard to actually find a pattern in the ratings they're giving

There's a lot to be said about improving the data and models (this entire paper was written in less than a year), but it seems like there are some cool directions we can go!

Other cool computational linguistics research!

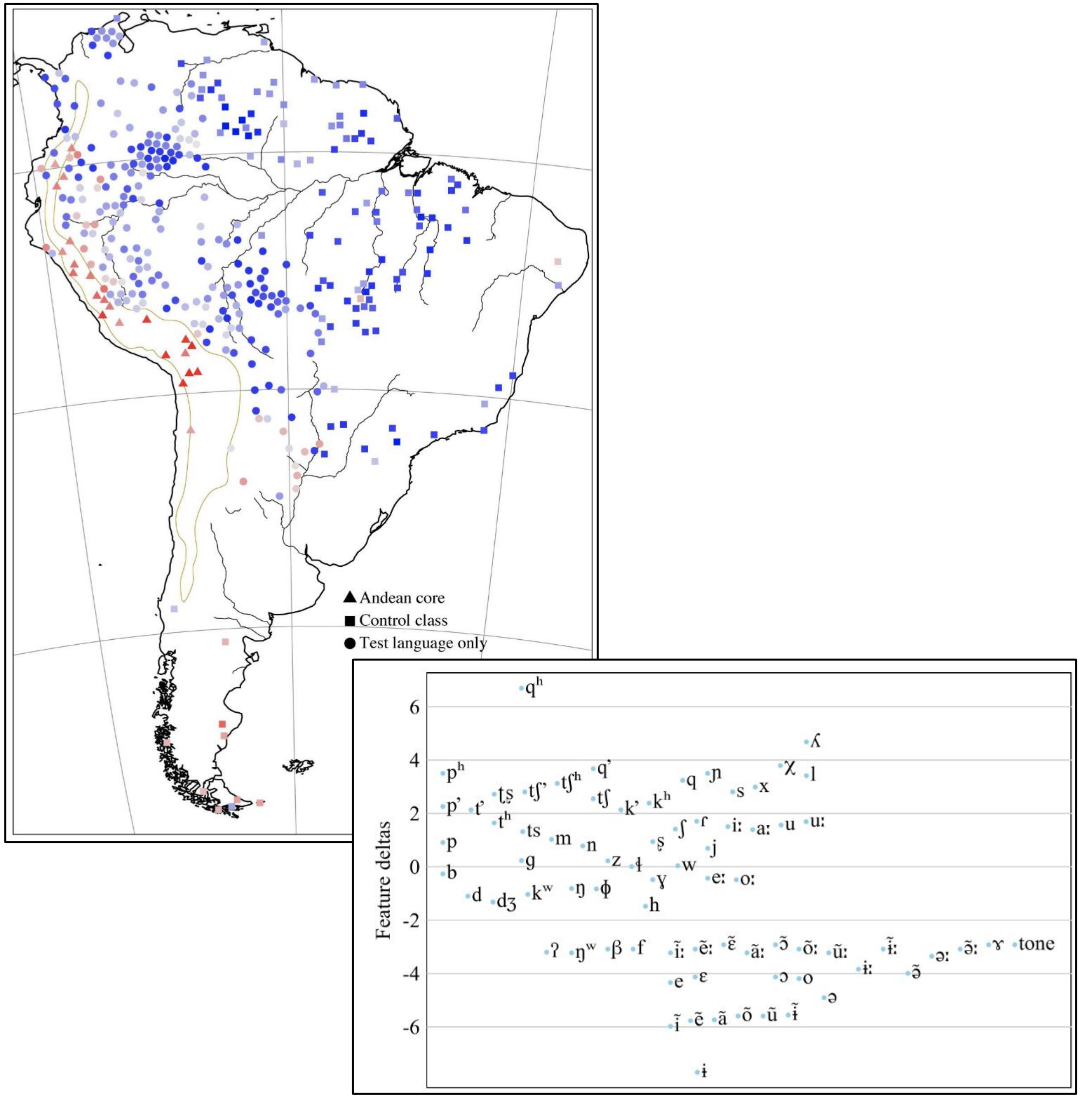

Phonology
Morphology
Syntax
Semantics
Phonetics
Linguistics
Historical Linguistics
Socio-linguistics
Cognitive Linguistics
Computational Linguistics
syllable
(closed)
onset
rhyme
nucleus
coda
consonant(s)
vowel
consonant(s)
'of'
'mug'
m
N
Aff
Anti
dis
establish
ment
arian
ism
S
NP
VP
The
teacher
warned
the
children
not
to
click
any
suspicious
email
∀x S(x)=T IF S(x₁)∧S(x₂)∧…∧S(xₙ)=T
∃x S(x)=T IF S(x₁)∨S(x₂)∨…∨S(xₙ)=T

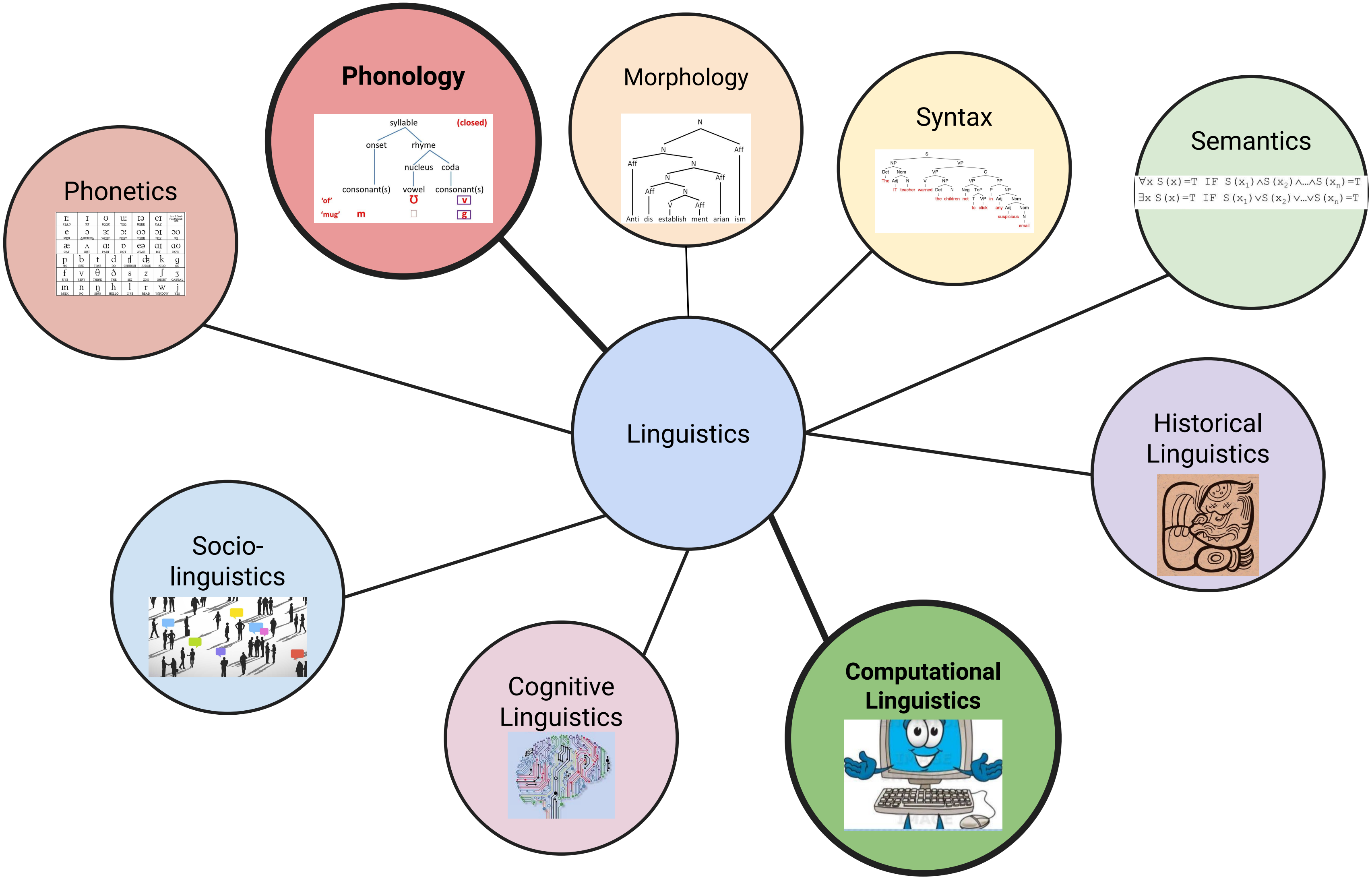
Phonological areality with naive Bayes

Naive Bayes is a kind of **classification model** that attempts to learn how to sort data points into categories

Michael et al. 2014 looks at applying this classifier to the problem of **phonological areality**

Basically, the model is given the **phonological inventories** of languages (literally just the sounds those languages have) and attempts to learn what family or language group those languages belong to





Linguistics

Phonetics

Phonology

Morphology

Syntax

Semantics

$\forall x\ S(x)=T$ IF $S(x_1) \wedge S(x_2) \wedge \ldots \wedge S(x_n)=T$

$\exists x\ S(x)=T$ IF $S(x_1) \vee S(x_2) \vee \ldots \vee S(x_n)=T$

Historical Linguistics

Computational Linguistics

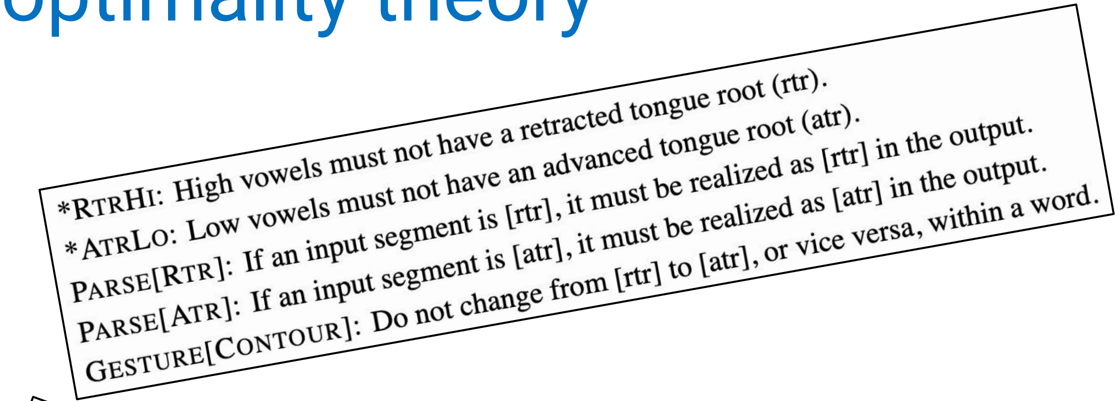
Cognitive Linguistics

Socio-linguistics

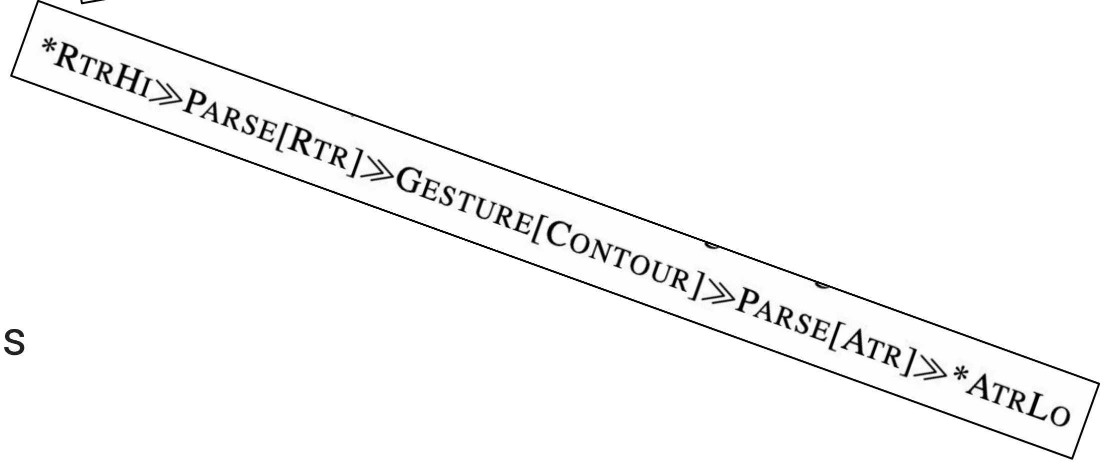
Maximum entropy optimality theory

Optimality theory is a linguistic theory in phonology that attempts to describe the sound patterns of languages by **constraints** on pronunciation and syllable structure

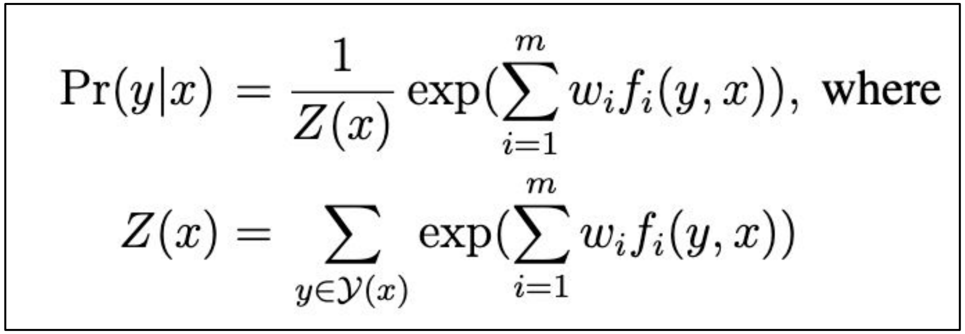
Maximum entropy optimality theory (MaxEnt OT), which can been seen for example in Goldwater and Johnson 2003, is a version where we allow a model to learn the rankings of the constraints, based on which ranking will give us the actual behavior of a language with maximum probability



*RtrHi: High vowels must not have a retracted tongue root (rtr).
*AtrLo: Low vowels must not have an advanced tongue root (atr).
Parse[Rtr]: If an input segment is [rtr], it must be realized as [rtr] in the output.
Parse[Atr]: If an input segment is [atr], it must be realized as [atr] in the output.
Gesture[Contour]: Do not change from [rtr] to [atr], or vice versa, within a word.

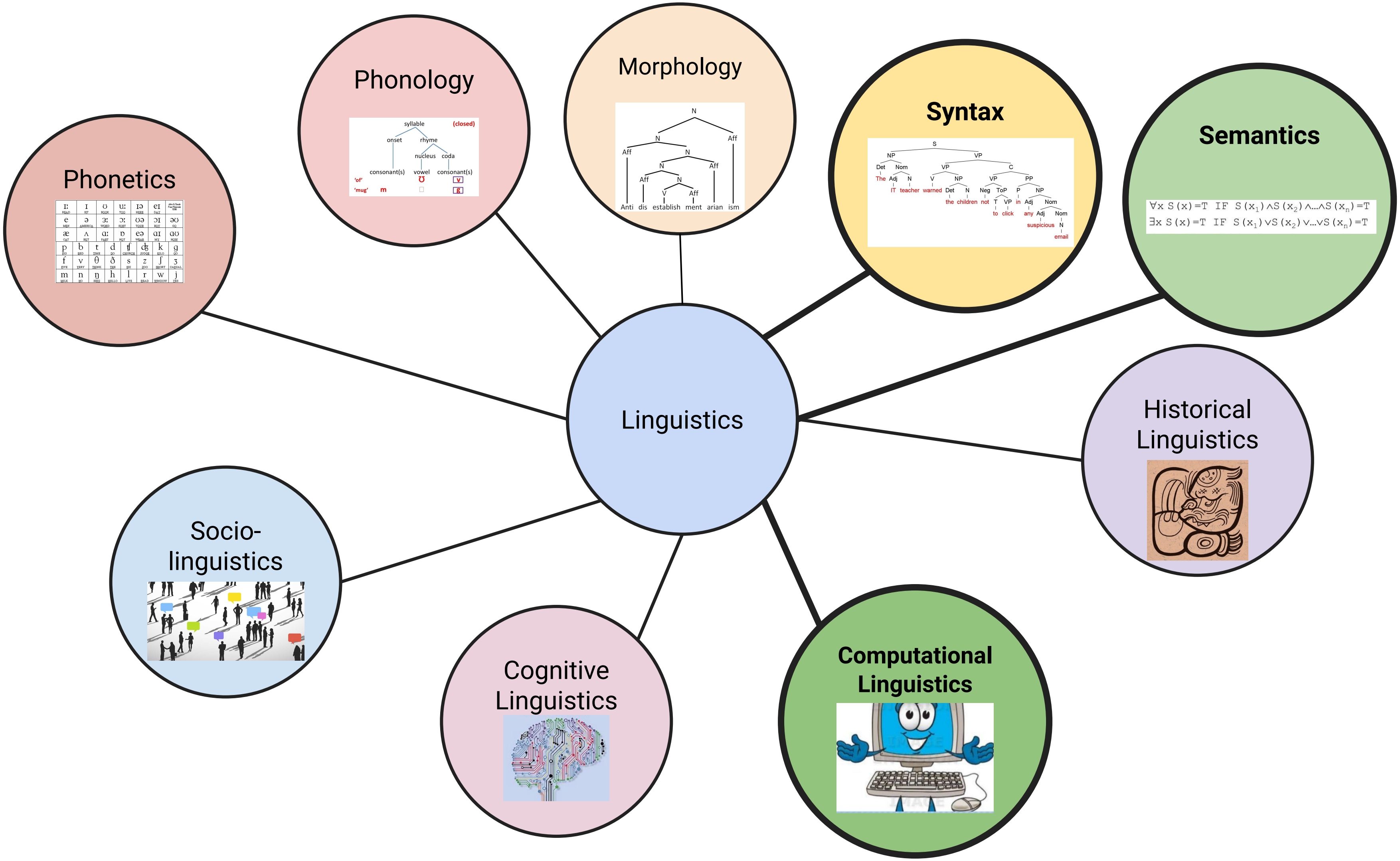


*RtrHi $\gg$ Parse[Rtr] $\gg$ Gesture[Contour] $\gg$ Parse[Atr] $\gg$ *AtrLo



$$\Pr(y|x) = \frac{1}{Z(x)} \exp\Big(\sum_{i=1}^{m} w_i f_i(y,x)\Big), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp\Big(\sum_{i=1}^{m} w_i f_i(y,x)\Big)$$



Linguistics

Phonetics

ɪː READ	ɪ SIT	ʊ BOOK	uː TOO	ɪə HERE	eɪ DAY	John & Sarah Free Materials 1996
e MEN	ə AMERICA	ɜː WORD	ɔː SORT	ʊə TOUR	ɔɪ BOY	əʊ GO
æ CAT	ʌ BUT	ɑː PART	ɒ NOT	eə WEAR	ɑɪ MY	ɑʊ HOW

p PIG	b BED	t TIME	d DO	ʧ CHURCH	ʤ JUDGE	k KILO	g GO
f FIVE	v VERY	θ THINK	ð THE	s SIX	z ZOO	ʃ SHORT	ʒ CASUAL
m MILK	n NO	ŋ SING	h HELLO	l LIVE	r READ	w WINDOW	j YES

Phonology

Morphology

Syntax

Semantics

$$\forall x\ S(x)=T\ \ \text{IF}\ \ S(x_1)\wedge S(x_2)\wedge\ldots\wedge S(x_n)=T$$
$$\exists x\ S(x)=T\ \ \text{IF}\ \ S(x_1)\vee S(x_2)\vee\ldots\vee S(x_n)=T$$

Historical Linguistics

Computational Linguistics

Cognitive Linguistics

Socio-linguistics

Clustering and categorization of attitude verbs

Clustering is a kind of computational model that attempts to sort data points into natural groups

White et al. 2014 applies clustering models to syntactic and semantic features of **attitude verbs** to see (1) what subcategories of these verbs exist and (2) how strongly syntax and semantics can pattern with each other for these verbs

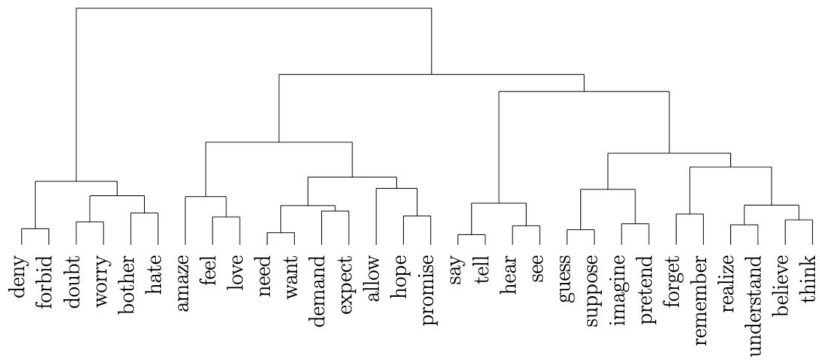


Figure 1: Semantic clusters

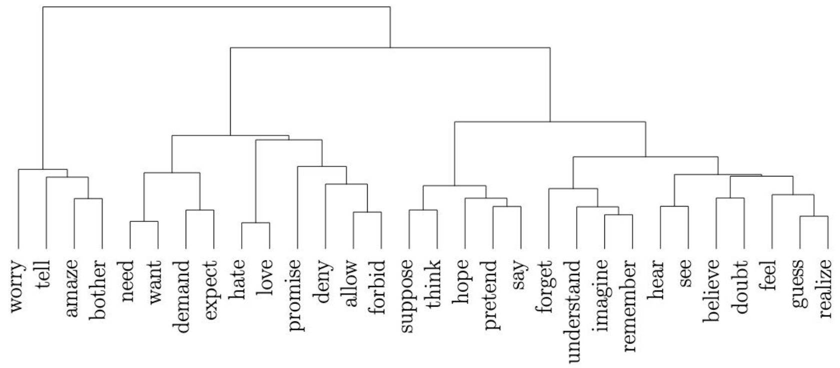
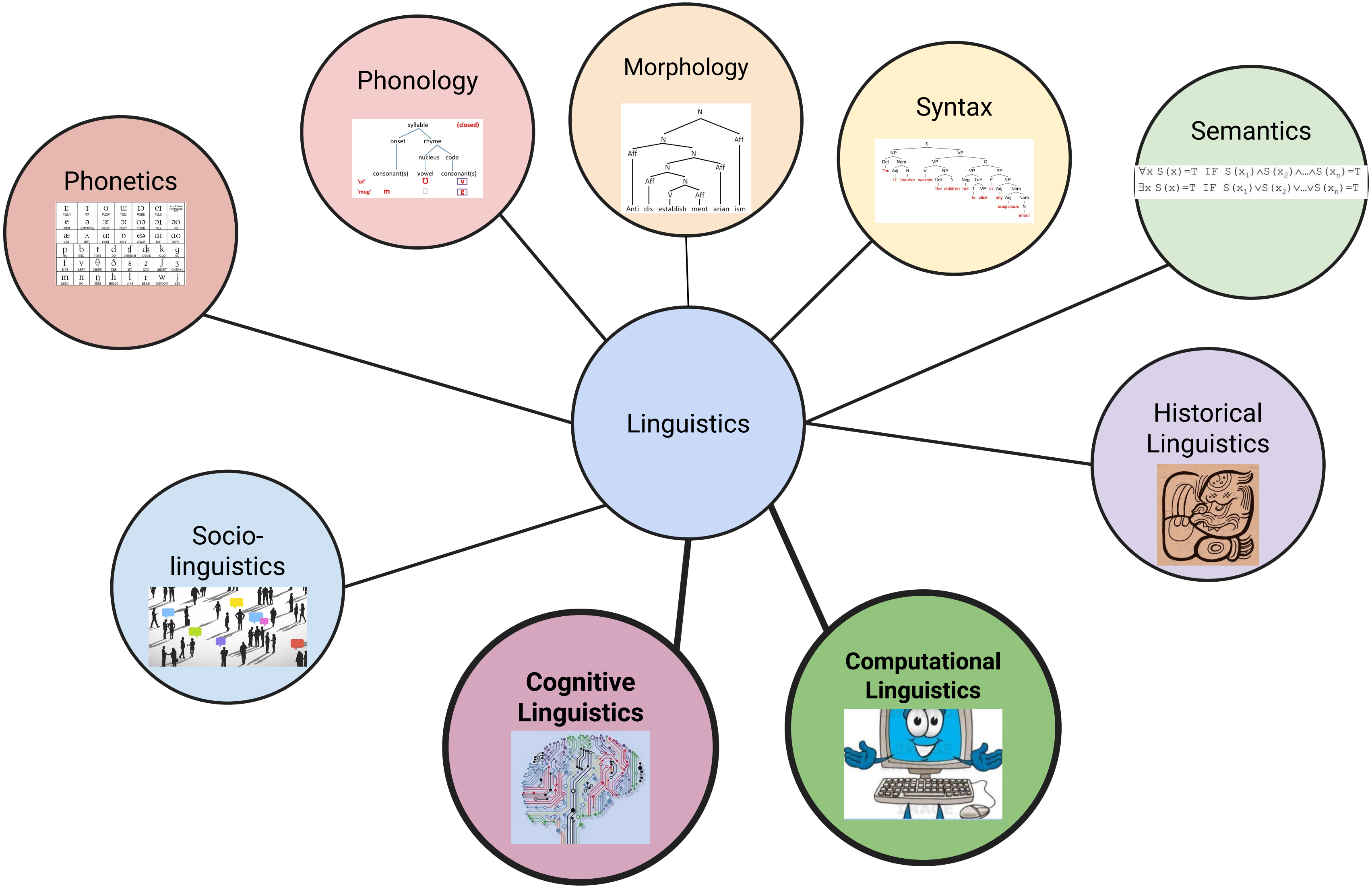


Figure 2: Syntactic clusters



Linguistics

Phonetics

ɪː READ	ɪ SIT	ʊ BOOK	uː TOO	ɪə HERE	eɪ DAY	John & Sarah Free Materials 1996
e MEN	ə AMERICA	ɜː WORD	ɔː SORT	ʊə TOUR	ɔɪ BOY	əʊ GO
æ CAT	ʌ BUT	ɑː PART	ɒ NOT	eə WEAR	ɑɪ MY	ɑʊ HOW

p PIG	b BED	t TIME	d DO	ʧ CHURCH	ʤ JUDGE	k KILO	g GO
f FIVE	v VERY	θ THINK	ð THE	s SIX	z ZOO	ʃ SHORT	ʒ CASUAL
m MILK	n NO	ŋ SING	h HELLO	l LIVE	r READ	w WINDOW	j YES

Phonology

Morphology

Syntax

Semantics

$\forall x\ S(x)=T\ \ \mathrm{IF}\ \ S(x_1)\wedge S(x_2)\wedge\ldots\wedge S(x_n)=T$

$\exists x\ S(x)=T\ \ \mathrm{IF}\ \ S(x_1)\vee S(x_2)\vee\ldots\vee S(x_n)=T$

Historical Linguistics

Computational Linguistics

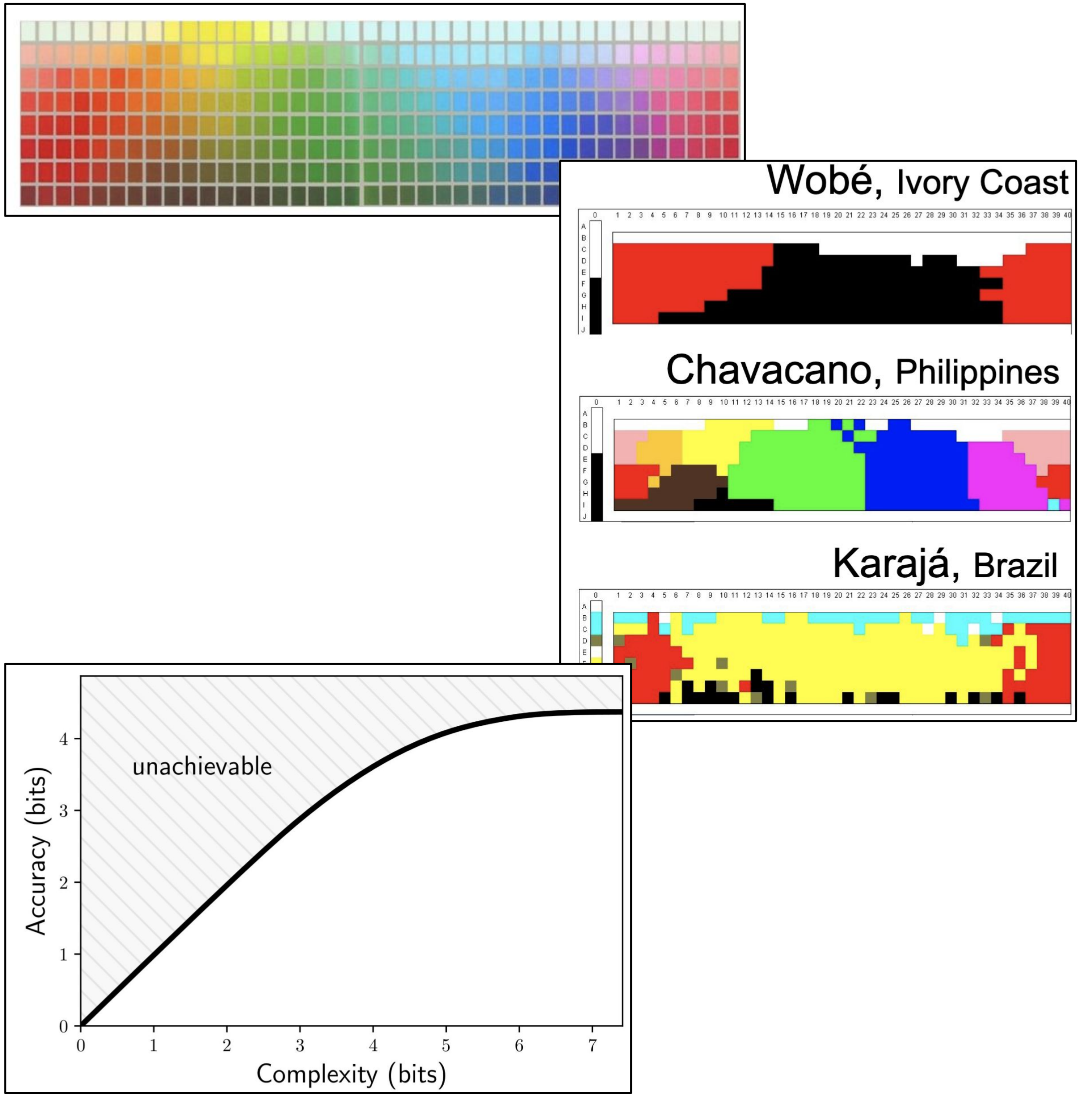
Cognitive Linguistics

Socio-linguistics

Information theory and color naming

Different languages have different color systems! In English we have a handful of basic color words, but many languages have fewer than English, and some have more

Regier et al. 2005 looks at analyzing this variation, but one option we can use is to apply **information theory** to see if the ways languages categorize colors are optimal, or close to optimal, from an information-theoretic standpoint (many of them are!)



Classes YOU can take!

Linguistics classes:

- Linguistics 100 (introduction to linguistic science)
- Linguistics 110 (phonetics)
- Linguistics 120 (syntax)
- Linguistics 121 (formal semantics)
- Linguistics 252 (computational linguistics)

CS classes:

- CS 189 (introduction to machine learning)
- CS 288 (natural language processing)
- INFO 159 (natural language processing)