Carnegie Mellon University

Language Technologies Institute

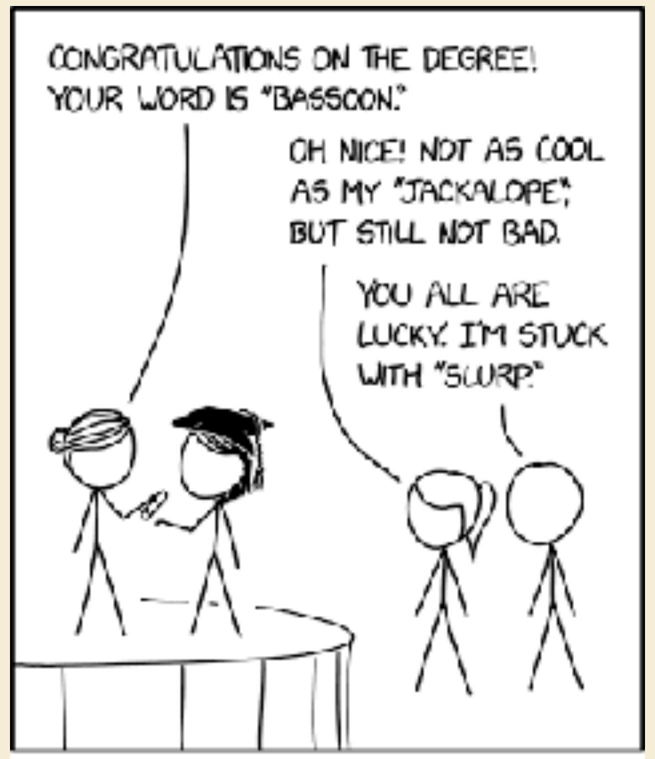
Linguistics and Computational Linguistics

A whirlwind tour 🌪️

11-711 Fall 2024

What is linguistics?

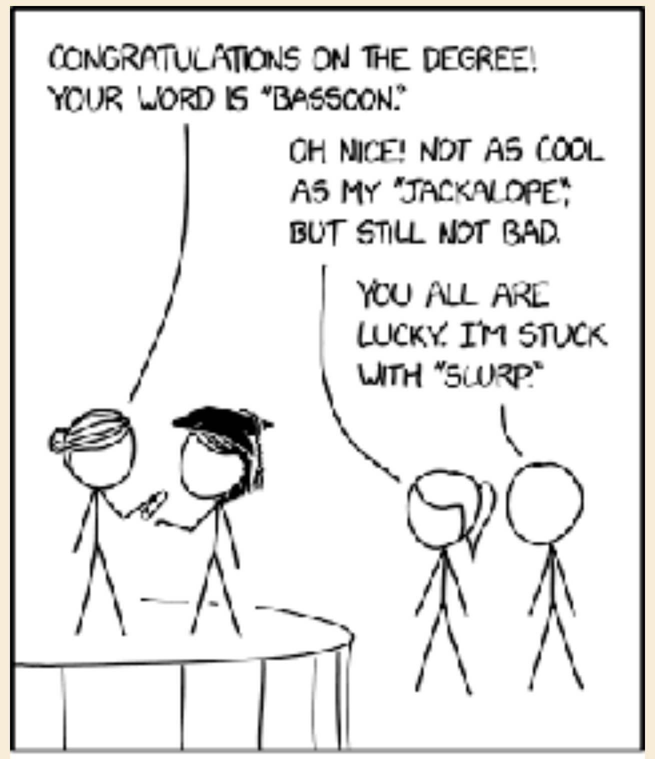
- Scientific study of language, its structure, and its use
- Theoretical linguistics tries to find a *general theory to explain the structure of language / a framework in which we can describe language*
    - While there are certain specific rules that govern the structure of individual languages, a general theory of language aims to encompass all natural languages


CONGRATULATIONS ON THE DEGREE! YOUR WORD IS "BASSOON."
OH NICE! NOT AS COOL AS MY "JACKALOPE", BUT STILL NOT BAD.
YOU ALL ARE LUCKY. I'M STUCK WITH "SLURP."

EVERY LINGUISTICS DEGREE COMES WITH ONE WORD THAT YOU'RE PUT IN CHARGE OF.

What is linguistics?

- Insights from theory can inform more applied research, e.g.:
  - What are the linguistic variations within speakers of a single language?
  - How are linguistic structures within and across languages are processed by the brain?
  - How do people acquire a new language at different stages of their life?


CONGRATULATIONS ON THE DEGREE! YOUR WORD IS "BASSOON."
OH NICE! NOT AS COOL AS MY "JACKALOPE," BUT STILL NOT BAD.
YOU ALL ARE LUCKY. I'M STUCK WITH "SLURP."

EVERY LINGUISTICS DEGREE COMES WITH ONE WORD THAT YOU'RE PUT IN CHARGE OF.

What is linguistics?

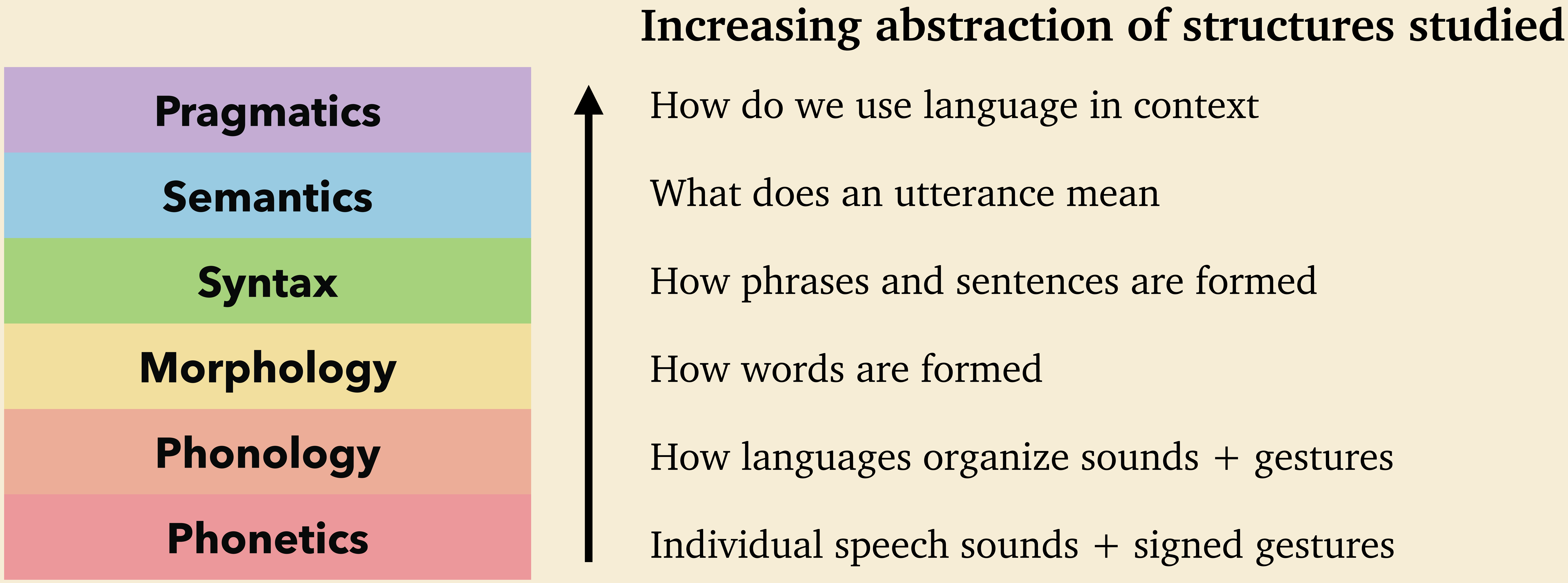
…and why should you care as an NLP practitioner?

- At minimum, allows you to **understand your data** more thoroughly
    - Especially for characterizing certain failure modes
- Gives you interesting test cases and frameworks to explore!
- Can motivate data-efficient methods (e.g. how do children learn language?)
- Linguistics posits theories for how human language is structured and processed
    - If we want to make claims about how NLP models/systems are similar to humans, being *aware* of these theories is a necessary starting point (even if you do not agree)
- It’s fun 🙂

Lecture Roadmap

- Brief overview of subfields and coverage of topics in linguistics
- For each topic group, we'll go over:
    - Main concepts and research questions
    - (Previous) computational approaches
    - Applications to NLP
- Because there's a lot in linguistics and only ~80 minutes, this might be very dense…apologies in advance

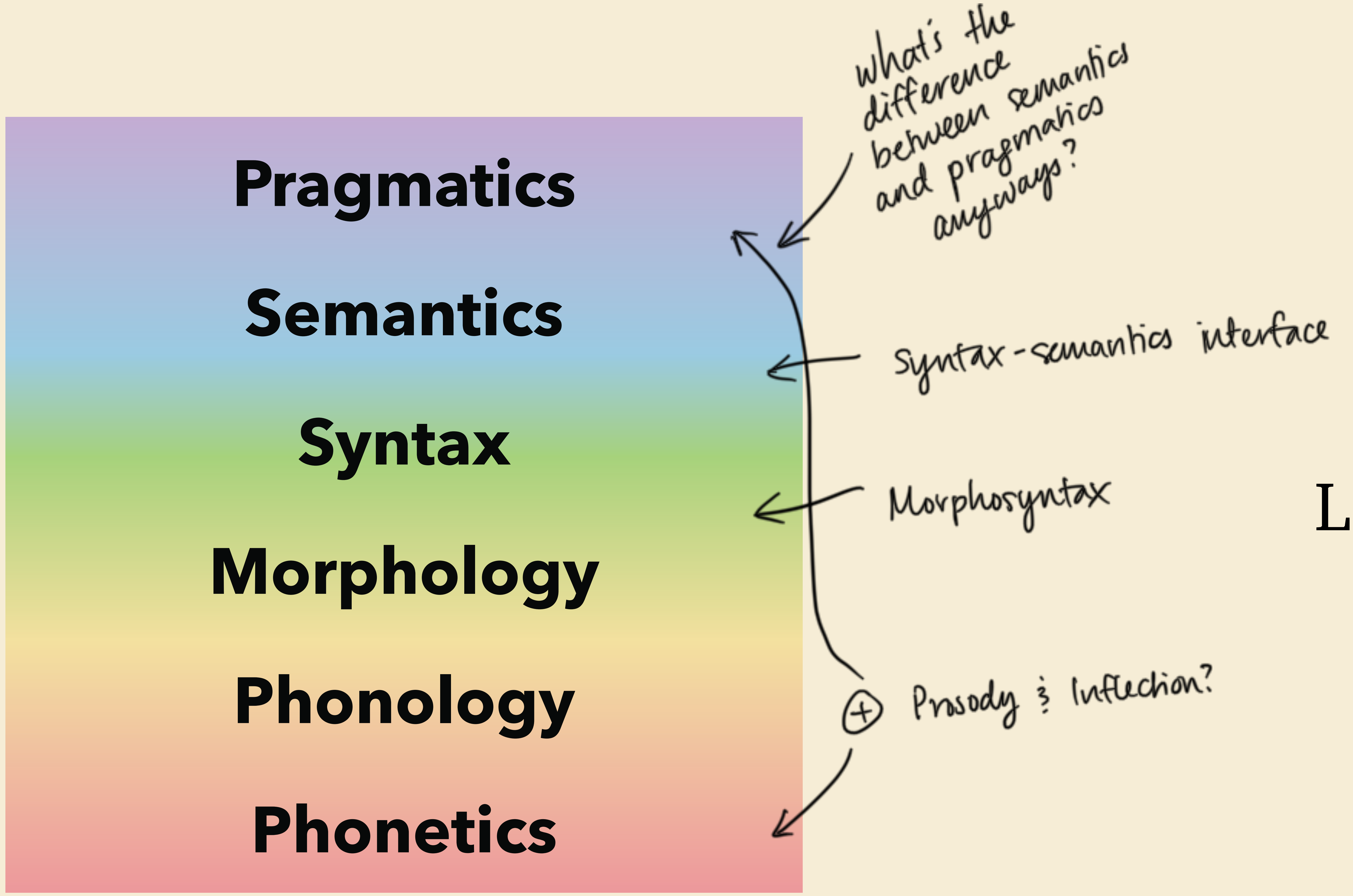
Subfields: An overview



Increasing abstraction of structures studied

Pragmatics	How do we use language in context
Semantics	What does an utterance mean
Syntax	How phrases and sentences are formed
Morphology	How words are formed
Phonology	How languages organize sounds + gestures
Phonetics	Individual speech sounds + signed gestures

Subfields: An overview



Pragmatics

Semantics

Syntax

Morphology

Phonology

Phonetics

what's the difference between semantics and pragmatics anyways?

Syntax-semantics interface

Morphosyntax

⊕ Prosody & Inflection?

Lots of interaction between levels!

Subfields: An overview

Pragmatics

Semantics

Syntax

Morphology

Phonology

Phonetics

Neurolinguistics

Psycholinguistics

Sociolinguistics

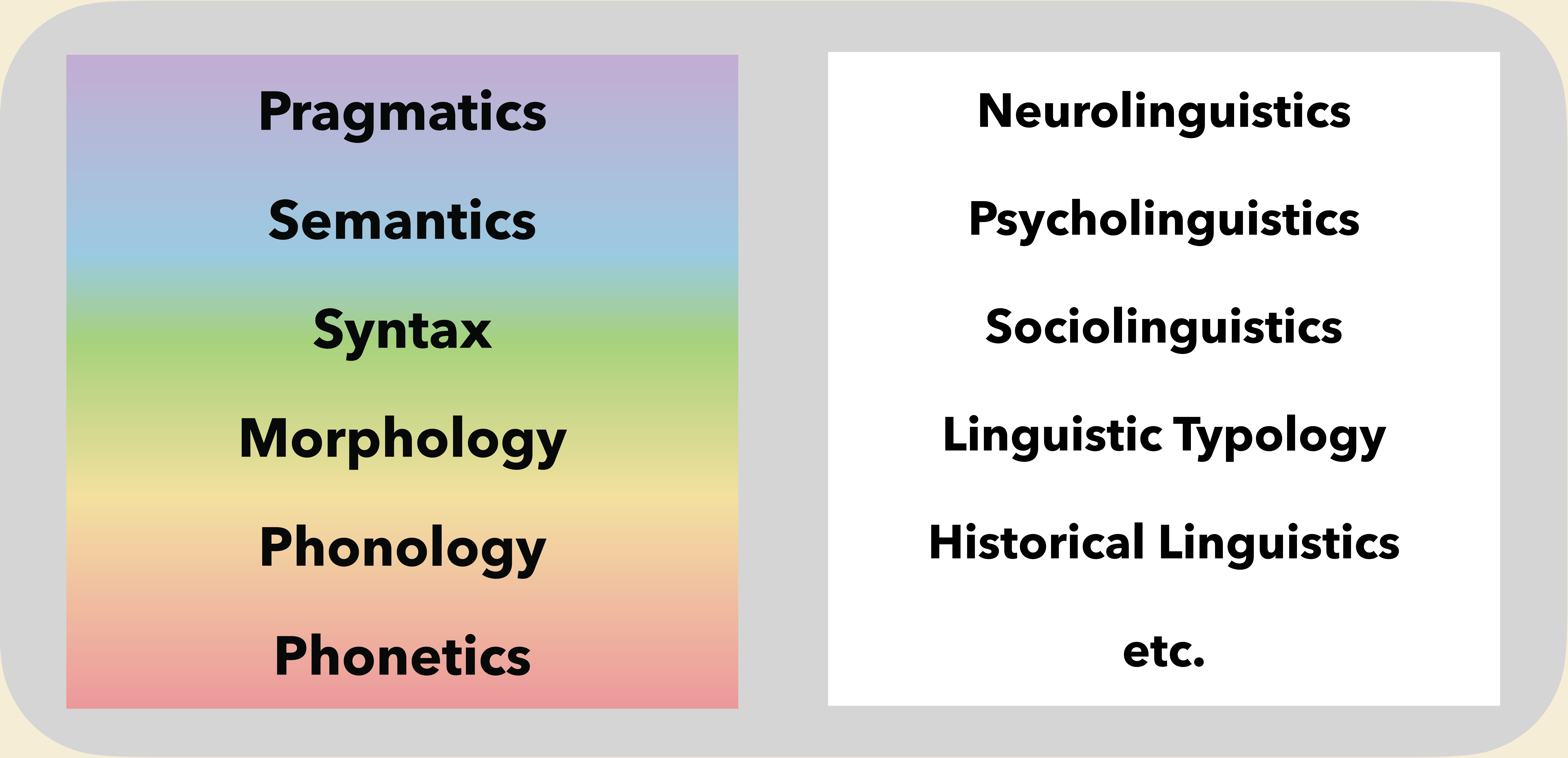
Linguistic Typology

Historical Linguistics

etc.

Subfields: An overview

We can use computational methods to explore questions within + across these subfields



Pragmatics

Semantics

Syntax

Morphology

Phonology

Phonetics

Neurolinguistics

Psycholinguistics

Sociolinguistics

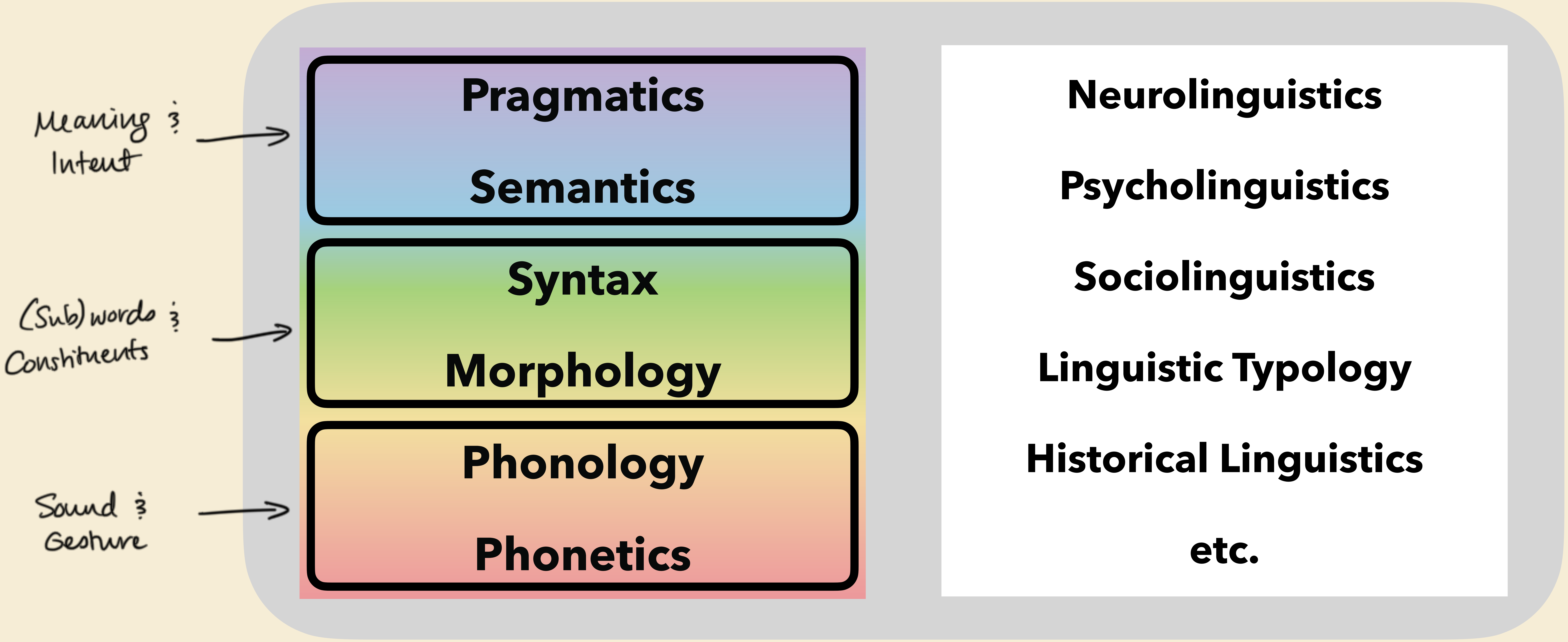
Linguistic Typology

Historical Linguistics

etc.

Subfields: An overview

We can use computational methods to explore questions within + across these subfields



Meaning & Intent →

Pragmatics

Semantics

(Sub)words & Constituents →

Syntax

Morphology

Sound & Gesture →

Phonology

Phonetics

Neurolinguistics

Psycholinguistics

Sociolinguistics

Linguistic Typology

Historical Linguistics

etc.

Sound and Gesture

Phonetics

- The study of speech sounds (spoken) / gestures (signed)
- How we:
    - Produce them (articulatory)
    - Perceive them (auditory)
    - Analyze them (acoustic)

Phonetics

Sound and Spelling

- *Phones* are individual speech sounds (or gestures)
  - E.g. the [p] sound in the English word ***pat***, [r] in ***write***

Phonetics

Sound and Spelling

- *Phones* are individual speech sounds (or gestures)
  - E.g. the [p] sound in the English word ***pat***, [r] in ***write***
- Text is often not a one-to-one mapping between characters and sounds
  - Some scripts are logographic, with little indication with how words are pronounced (e.g. Chinese)
  - Some do have consistent spellings for sounds that are one-to-one, so exact pronunciation can often be determined (e.g. Japanese *kana*, Spanish, Hindi)
  - Some have a general relationship between spelling form and sound, though it is often irregular (e.g. English, French)

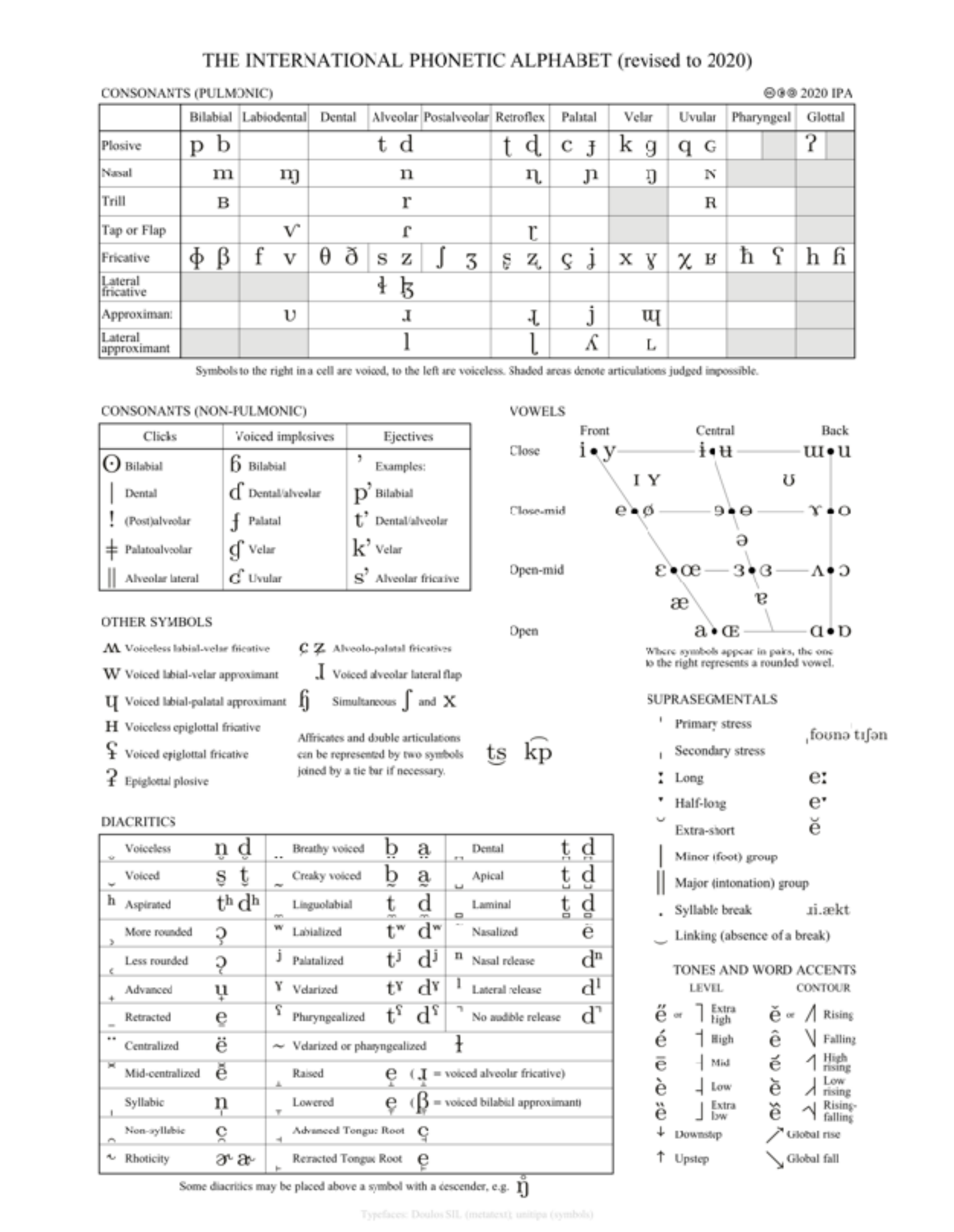
Phonetics

IPA (not the beer)

[aɪ̯ pʰiː eɪ̯]

In order to have a consistent representation of sound, linguists use the **International Phonetic Alphabet (IPA)**

- **Epitran**: library and tool for transliterating orthographic text as IPA



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

CONSONANTS (PULMONIC) ©®® 2020 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	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							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative	ɕ ʑ Alveolo-palatal fricatives
w Voiced labial-velar approximant	ɺ Voiced alveolar lateral flap
ɥ Voiced labial-palatal approximant	ɧ Simultaneous ʃ and x
ʜ Voiceless epiglottal fricative	
ʢ Voiced epiglottal fricative	
ʡ Epiglottal plosive	

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͜s k͡p

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.

DIACRITICS

̥ Voiceless n̥ d̥	̤ Breathy voiced b̤ a̤	̪ Dental t̪ d̪
̬ Voiced s̬ t̬	̰ Creaky voiced b̰ a̰	̺ Apical t̺ d̺
ʰ Aspirated tʰ dʰ	̼ Linguolabial t̼ d̼	̻ Laminal t̻ d̻
̹ More rounded ɔ̹	ʷ Labialized tʷ dʷ	̃ Nasalized ẽ
̜ Less rounded ɔ̜	ʲ Palatalized tʲ dʲ	ⁿ Nasal release dⁿ
̟ Advanced u̟	ˠ Velarized tˠ dˠ	ˡ Lateral release dˡ
̠ Retracted e̠	ˤ Pharyngealized tˤ dˤ	̚ No audible release d̚
̈ Centralized ë	̴ Velarized or pharyngealized ɫ	
̽ Mid-centralized e̽	̝ Raised e̝ (ɹ̝ = voiced alveolar fricative)	
̩ Syllabic n̩	̞ Lowered e̞ (β̞ = voiced bilabial approximant)	
̯ Non-syllabic e̯	̘ Advanced Tongue Root e̘	
˞ Rhoticity ɚ a˞	̙ Retracted Tongue Root e̙	

Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

Typefaces: Doulos SIL (metatext); unitipa (symbols)

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən
ˌ	Secondary stress	
ː	Long	eː
ˑ	Half-long	eˑ
̆	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

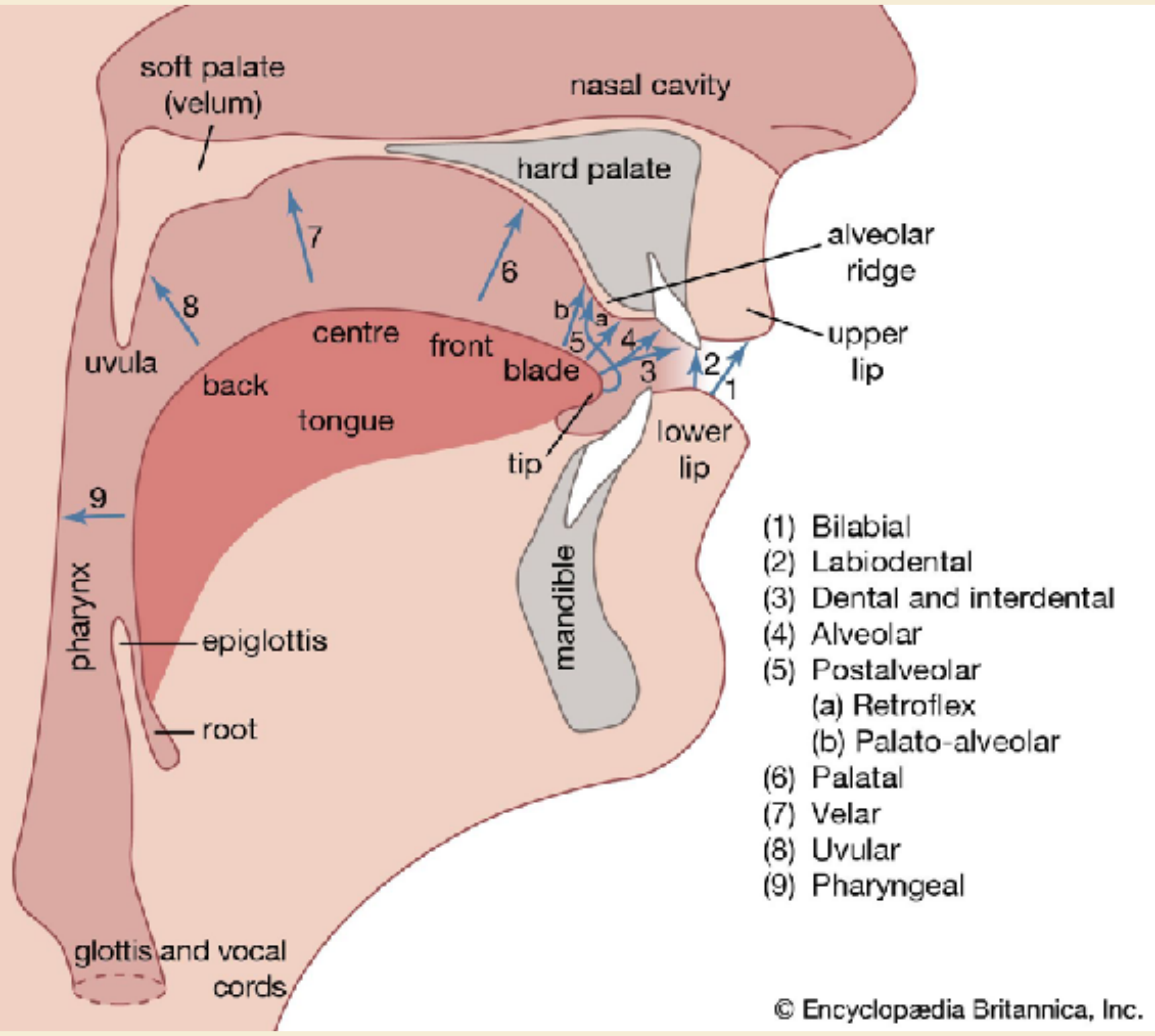
TONES AND WORD ACCENTS

LEVEL		CONTOUR	
e̋ or ˥	Extra high	ě or ˩˥	Rising
é ˦	High	ê ˥˩	Falling
ē ˧	Mid	e᷄ ˦˥	High rising
è ˨	Low	e᷅ ˩˨	Low rising
ȅ ˩	Extra low	e᷈ ˧˦˧	Rising-falling
ꜜ	Downstep	↗	Global rise
ꜛ	Upstep	↘	Global fall

Phonetics

Production of speech sounds

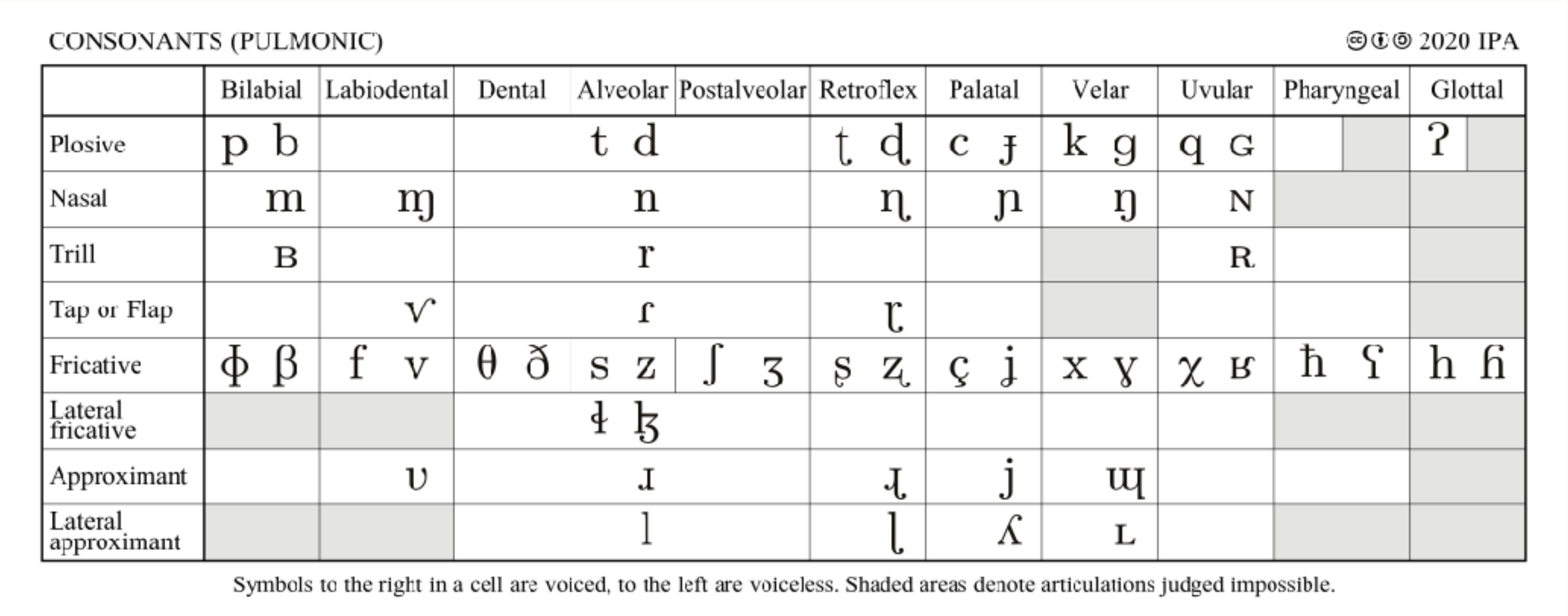
- *Articulatory phonetics* studies how speech sounds are produced
- Various organs in the mouth, nose, and throat modify airflow from the lungs
- Based on how these modifications occur, we get different kinds of sounds
  - *Vowels* are produced without any restriction
  - *Consonants* are produced with (partial or full) restriction
  - *Semi-vowels* are between a consonant and a vowel



© Encyclopædia Britannica, Inc.

Phonetics

Consonants



CONSONANTS (PULMONIC)

2020 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	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							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

We can categorize consonants based on their *place* and *manner* of articulation, as well as whether they are *voiced* or *voiceless*

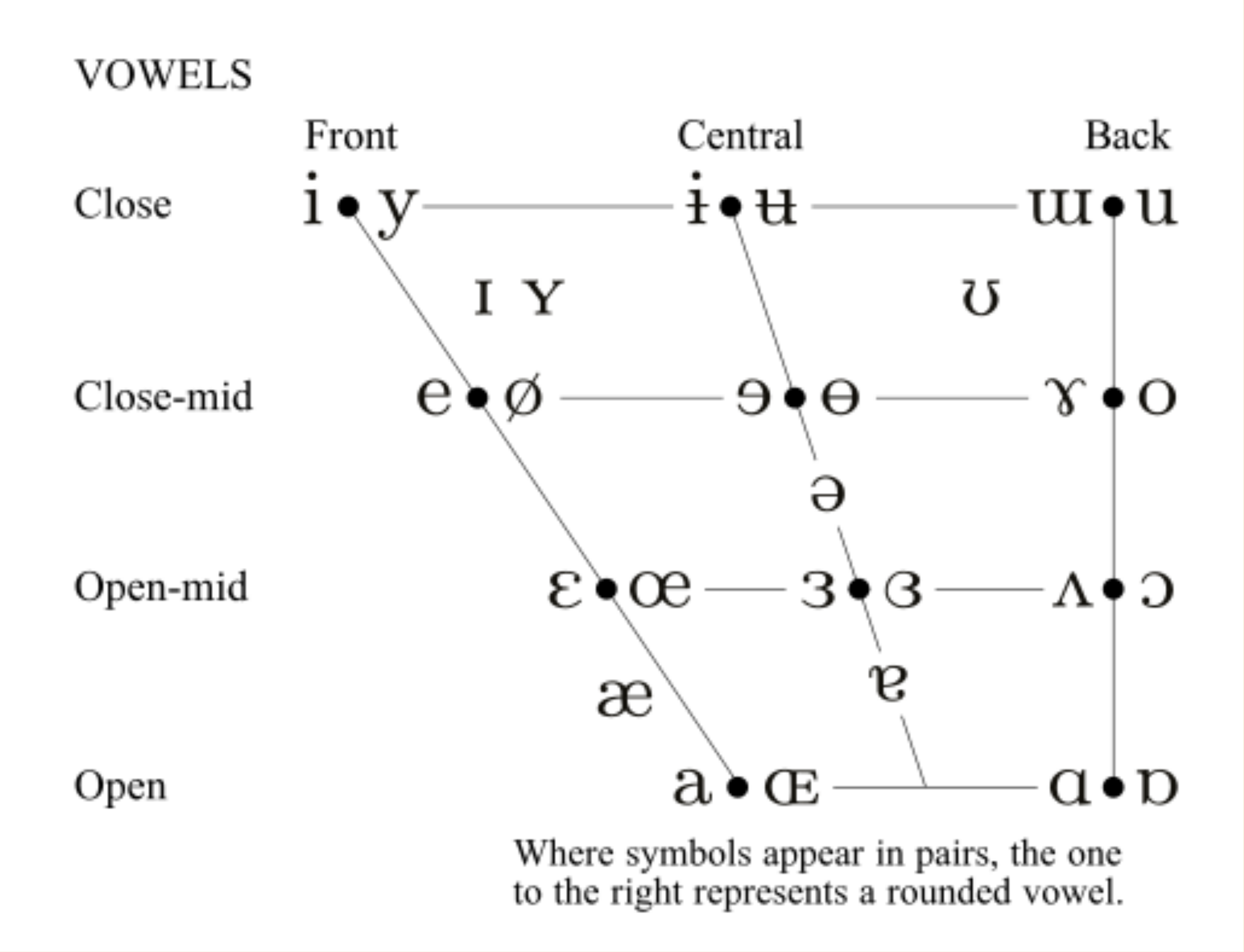
Phonetics

Vowels

Vowels can be categorized based on:

- Position of the tongue (front-back)
- How open the mouth is (close-open)
- Roundedness of the lips (rounded/unrounded)

Vowels are typically voiced, but voiceless vowels do exist!



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

Phonology

- The study of categorical organization of speech sounds (or equivalent gestures in signed languages)
- While phonetics deals with the *physical* properties of sounds (regardless of their context in a language), phonology deals with *abstract* rules/constraints that govern interactions of sounds within a language:
  - What sounds are meaningfully distinct in a language?
  - How are sounds organized into syllables?
  - What rules govern allowable sequences of sounds?

Phonology

Phones, Phonemes, and Allophones

- *Phones* are individual speech sounds

Phonology

Phones, Phonemes, and Allophones

- *Phones* are individual speech sounds
- *Phonemes* are **perceptually distinct units of sounds** in a language
    - Can distinguish one word from another (e.g. [pit] vs. [lit] → /p/ and /l/ are separate phonemes)
    - The *phoneme inventory* of a language is the set of all such units

Phonology

Phones, Phonemes, and Allophones

- *Phones* are individual speech sounds
- *Phonemes* are **perceptually distinct units of sounds** in a language
  - Can distinguish one word from another (e.g. [pit] vs. [lit] → /p/ and /l/ are separate phonemes)
  - The *phoneme inventory* of a language is the set of all such units
- Fun fact: over time, we are conditioned to limit our mental distinction and production of sounds between those that are distinct in our native languages…but we can still re-learn!

Phonology

Phones, Phonemes, and Allophones

- [p] and [p^h] are two distinct phones that are used in English speech
  - E.g. *spat* vs *pat*

Phonology

Phones, Phonemes, and Allophones

- [p] and [p^h] are two distinct phones that are used in English speech
  - E.g. *spat* vs *pat*
- However, changing [p] for [p^h] (and vice versa) will not change the meaning of a word

Phonology

Phones, Phonemes, and Allophones

- [p] and [p^h] are two distinct phones that are used in English speech
  - E.g. *spat* vs *pat*
- However, changing [p] for [p^h] (and vice versa) will not change the meaning of a word
- [p] and [p^h] are instances of the same phoneme /p/ → they are *allophones* in English
  - Some other languages do distinguish these sounds (e.g. Thai), so their phoneme inventory would include both /p/ and /p^h/

Phonology

Phonological Rules

- Phonological rules determine how a phoneme is pronounced in context
- Whether /p/ is pronounced as [p] or [p^h] can be determined by the sounds that surround it (its *environment*)

Phonology

Phonological Rules

- Phonological rules determine how a phoneme is pronounced in context
- Whether /p/ is pronounced as [p] or [p^h] can be determined by the sounds that surround it (its *environment*)
    - **Observation**: (generally) aspiration only occurs when /p/ is at the beginning of a stressed syllable
    - This also happens with other sounds: ([t], [t^h]) and ([k], [k^h])

Phonology

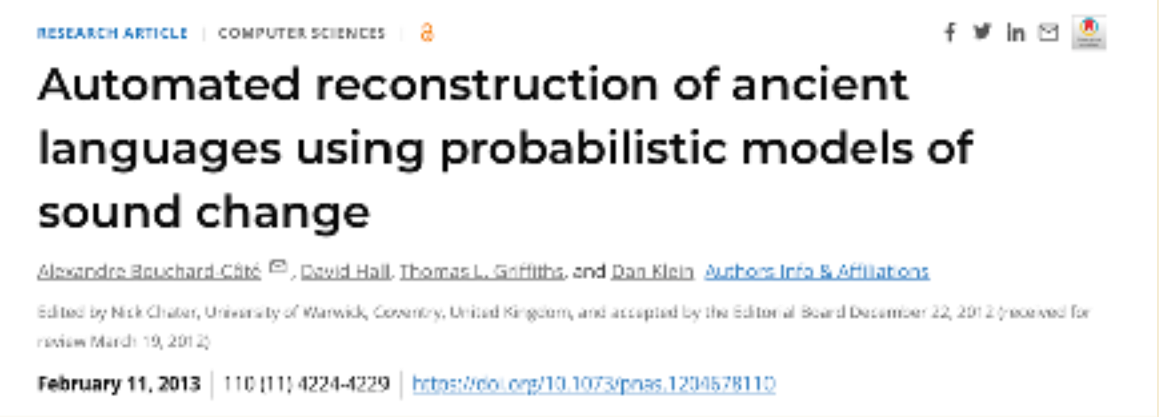
Phonological Rules

- Phonological rules determine how a phoneme is pronounced in context
- Whether /p/ is pronounced as [p] or [p^h] can be determined by the sounds that surround it (its *environment*)
  - **Observation**: (generally) aspiration only occurs when /p/ is at the beginning of a stressed syllable
  - This also happens with other sounds: ([t], [t^h]) and ([k], [k^h])
- **Rule**: these sounds (unvoiced stops) will be aspirated at the beginning of a stressed syllable, and unaspirated otherwise

Computational Phon* and Applications in NLP

A sampling

- **Automatic protolanguage reconstruction**: phonological changes over time can give us clues as to how languages have evolved over time



RESEARCH ARTICLE | COMPUTER SCIENCES

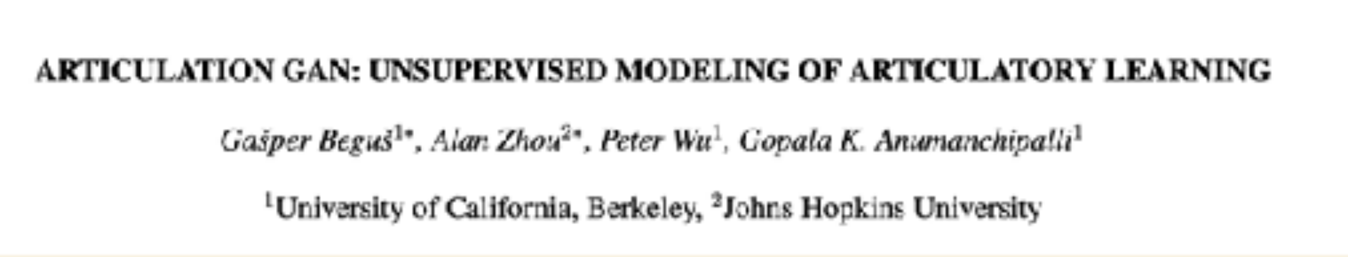
Automated reconstruction of ancient languages using probabilistic models of sound change

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein Authors Info & Affiliations

Edited by Nick Chater, University of Warwick, Coventry, United Kingdom, and accepted by the Editorial Board December 22, 2012 (received for review March 19, 2012)

February 11, 2013 | 110 (11) 4224-4229 | https://doi.org/10.1073/pnas.1204678110

- **Cognitive models of human speech production**: Training an unsupervised speech synthesis model to produce speech with human-like articulatory gestures

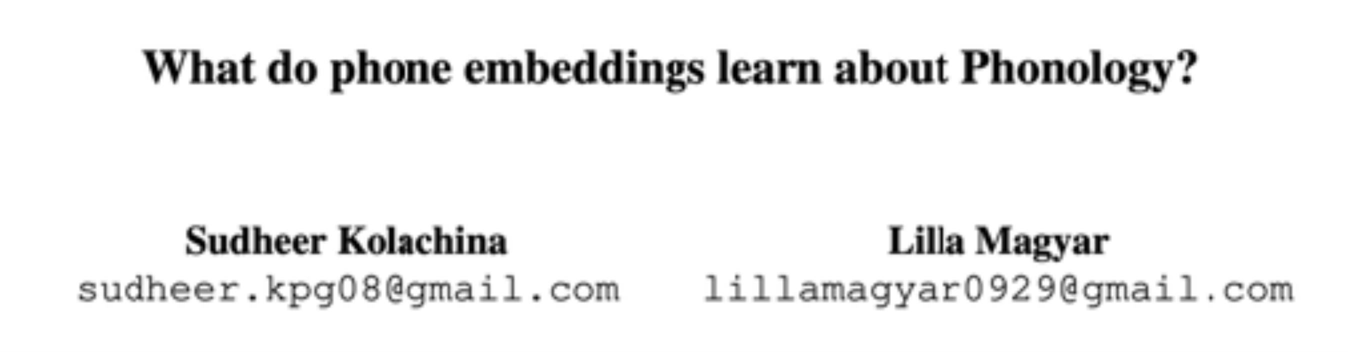


ARTICULATION GAN: UNSUPERVISED MODELING OF ARTICULATORY LEARNING

Gašper Beguš[1*], *Alan Zhou*[2*], *Peter Wu*[1], *Gopala K. Anumanchipalli*[1]

[1]University of California, Berkeley, [2]Johns Hopkins University

- **Linguistic evaluation**: do phone embeddings encode phonological relations?

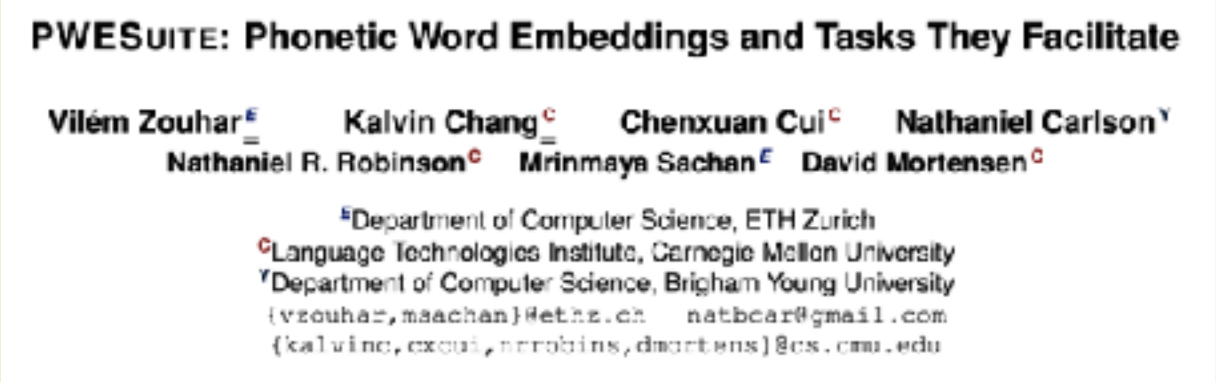


What do phone embeddings learn about Phonology?

Sudheer Kolachina
sudheer.kpg08@gmail.com

Lilla Magyar
lillamagyar0929@gmail.com

- **Incorporating phonetic information into word embeddings**: can be applied to cognate/loanword detection, multilingual NER, language identification, etc.



PWESUITE: Phonetic Word Embeddings and Tasks They Facilitate

Vilém Zouhar[E] Kalvin Chang[C] Chenxuan Cui[C] Nathaniel Carlson[Y]
Nathaniel R. Robinson[C] Mrinmaya Sachan[E] David Mortensen[C]

[E]Department of Computer Science, ETH Zurich
[C]Language Technologies Institute, Carnegie Mellon University
[Y]Department of Computer Science, Brigham Young University
{vzouhar,msachan}@ethz.ch natbcar@gmail.com
{kalvinc,cxcui,nrrobins,dmortens}@cs.cmu.edu

(Sub)words and Constituents

Morphology

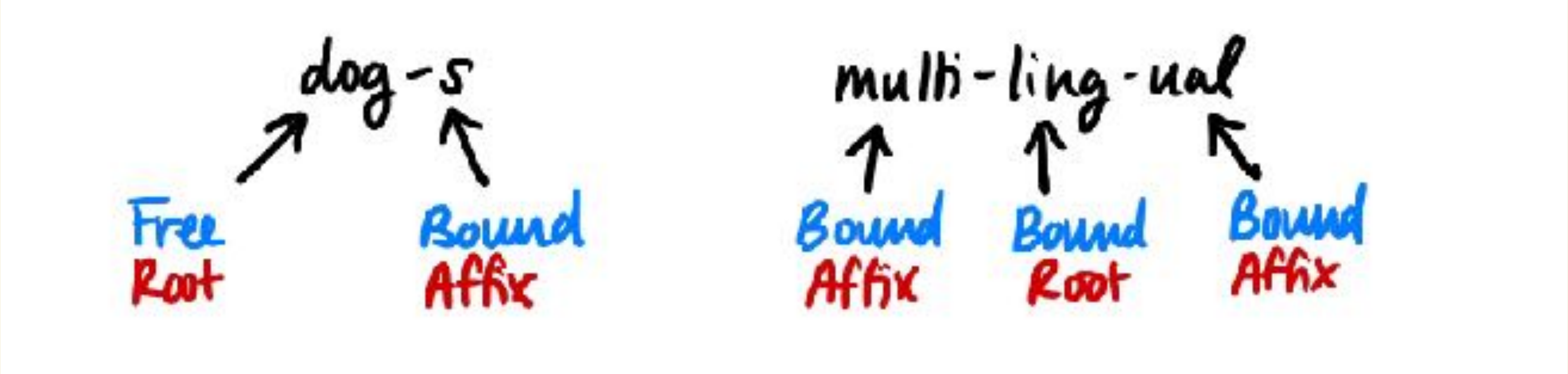
- The study of word formation and structure
    - Side note: You may be asking…What is a word? Do words actually exist? In any case, these questions are highly contested. If you ask an opinionated linguist, they can probably talk about this for a long, long time.
    - For now, let's just gloss over this and go with our intuitions
- Words are formed from linguistic units called *morphemes*
    - Smallest **meaningful** linguistic unit
    - E.g. morph (form, shape) - ology (the study of)
- Most of the examples here are in English, though English morphology is…boring. Check out some polysynthetic languages (e.g. including many indigenous American languages) for more fun!

Morphology

Morpheme Types

For the most part, we can categorize morphemes by the following properties:

- Can a morpheme occur by itself? → *Free / Bound*
  - *Also: cranberry morphemes*
- Does it comprise the "main meaning" of the word? → *Root / Affix*



Morphology

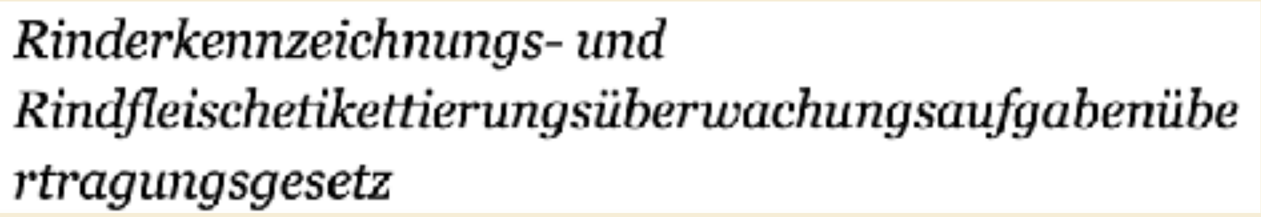
Inflection

- *Inflection* is a process that creates a **new form of the same word**
  - The main concept/meaning of the word remains the same
  - Changes a grammatical feature
    - Number: *dog* (noun, singular), *dog*-***s*** (noun, plural)
    - Person: *I run* (verb, first person), *he run*-***s*** (verb, third person)
    - Tense: *I climb* (verb, present), *I climb*-***ed*** (verb, past)
    - etc.

Morphology

Word Formation: Derivation and Compounding

- *Derivation* is a process that creates a **semantically related new word** by **operating on a base form**, often through a process like affixation
  - The main concept/meaning of the word changes
  - Part of speech often changes, though not always
    - to *teach* (verb) → a *teach*-***er*** (noun, agent)
    - *intense* (adj) → to *intens*-***ify*** (verb)
    - *easy* (adj) → *easi*-***ly*** (adv)
    - *lucky* (adj) → ***un***-*lucky* (adj)
- *Compounding* is a process that creates a **semantically related new word** by **combining words**
  - *blackbird*, *ice cream*, *skyscraper*



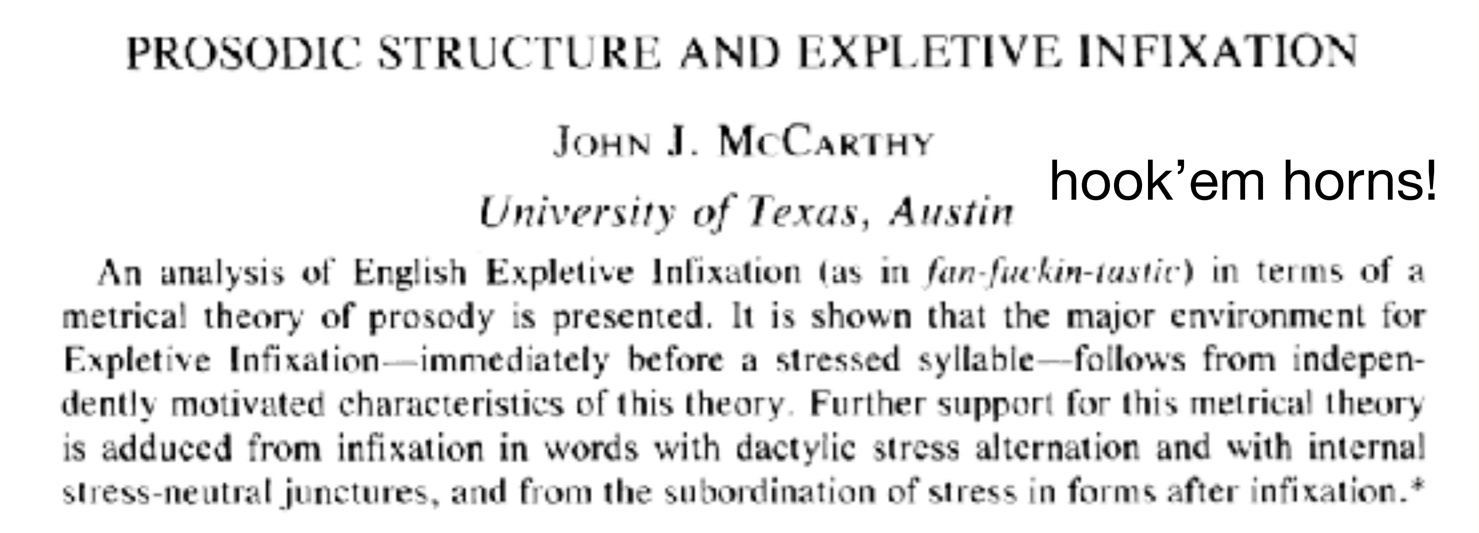
Rinderkennzeichnungs- und
Rindfleischetikettierungsüberwachungsaufgabenübe
rtragungsgesetz

"Cattle marking and beef labeling supervision duties delegation law"

Morphology

Non-Concatenative Processes

- So far, all of the examples we've looked at are formed by sequentially attaching affixes to roots
- However, not all morphological processes are this straightforward
    - Apophony (tooth → teeth)
    - Infixation (a fun example is expletive infixation)
    - Transfixation (as with Arabic and Hebrew roots)
    - Reduplication (*berjalan* [to walk] → *berjalan-jalan* [to stroll])
    - …among others!



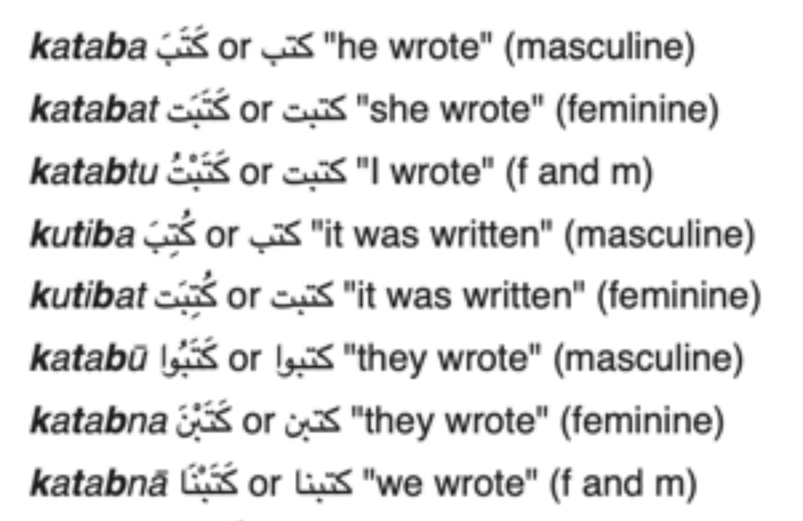
PROSODIC STRUCTURE AND EXPLETIVE INFIXATION

JOHN J. McCARTHY

University of Texas, Austin

An analysis of English Expletive Infixation (as in *fan-fuckin-tastic*) in terms of a metrical theory of prosody is presented. It is shown that the major environment for Expletive Infixation—immediately before a stressed syllable—follows from independently motivated characteristics of this theory. Further support for this metrical theory is adduced from infixation in words with dactylic stress alternation and with internal stress-neutral junctures, and from the subordination of stress in forms after infixation.*

hook'em horns!



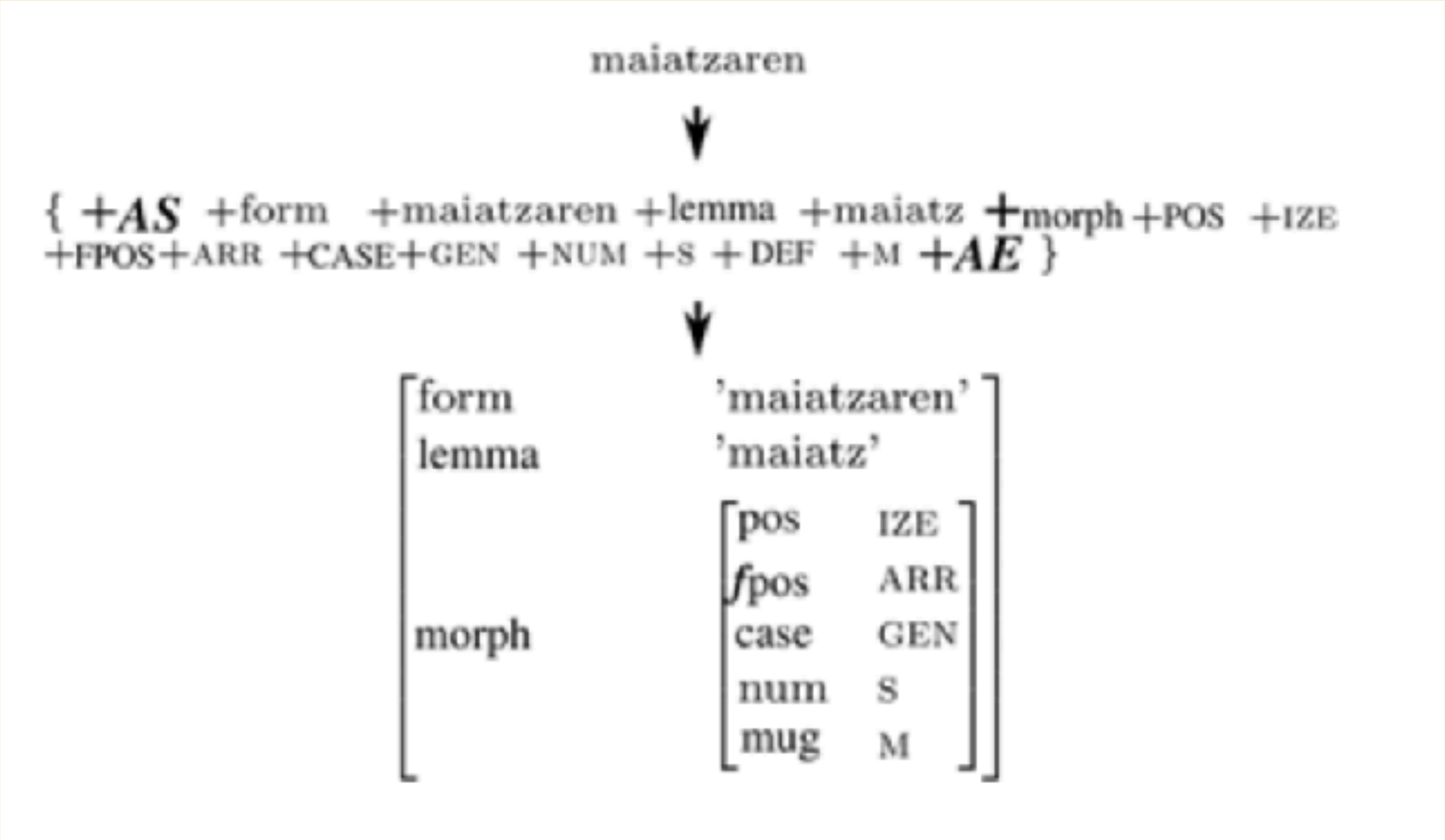
kataba كَتَبَ or كتب "he wrote" (masculine)
katabat كَتَبَتْ or كتبت "she wrote" (feminine)
katabtu كَتَبْتُ or كتبت "I wrote" (f and m)
kutiba كُتِبَ or كتب "it was written" (masculine)
kutibat كُتِبَتْ or كتبت "it was written" (feminine)
katabū كَتَبُوا or كتبوا "they wrote" (masculine)
katabna كَتَبْنَ or كتبن "they wrote" (feminine)
katabnā كَتَبْنَا or كتبنا "we wrote" (f and m)

Morphological Analyzers

Useful tool for linguistic annotation (especially understudied and endangered languages)!

Input: word form

Output: all possible morphological parses



maiatzaren

{ +*AS* +form +maiatzaren +lemma +maiatz +morph +POS +IZE +FPOS +ARR +CASE +GEN +NUM +S +DEF +M +*AE* }

$$\begin{bmatrix} \text{form} & \text{'maiatzaren'} \\ \text{lemma} & \text{'maiatz'} \\ \text{morph} & \begin{bmatrix} \text{pos} & \text{IZE} \\ \text{fpos} & \text{ARR} \\ \text{case} & \text{GEN} \\ \text{num} & \text{S} \\ \text{mug} & \text{M} \end{bmatrix} \end{bmatrix}$$

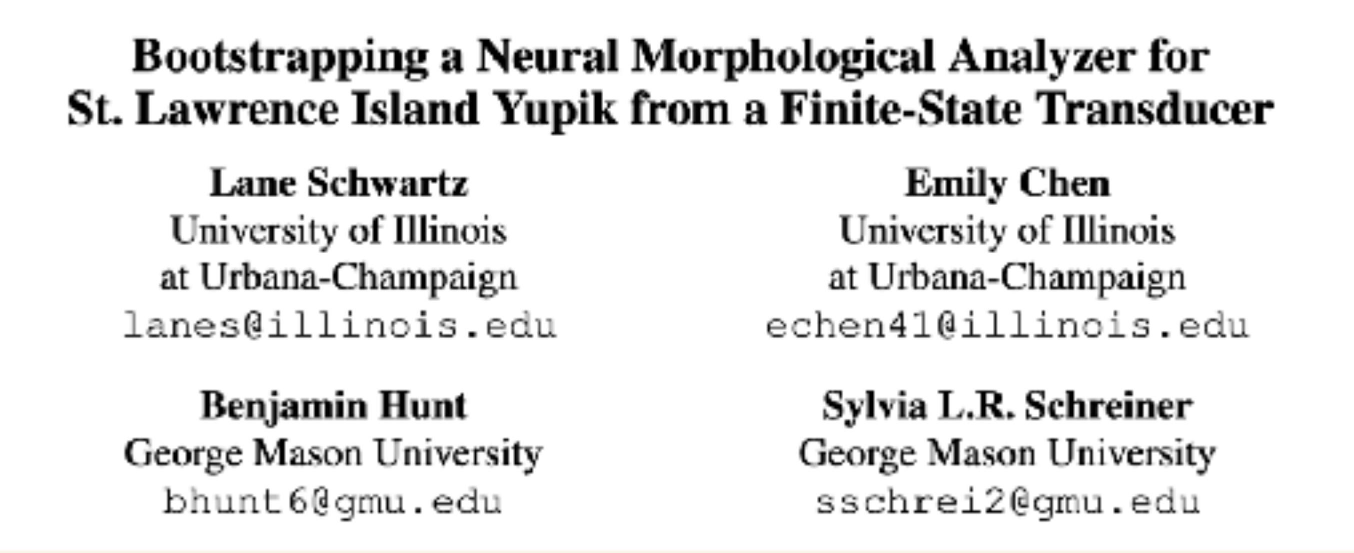
Taken from https://fomafst.github.io/morphtut.html

Morphological Analyzers

- Traditionally done with finite state transducers (FSTs)
  - Two-step creation process:
    - Map lemma+morphosyntactic description to an intermediate form that represents canonical morpheme representations (e.g. bus-PL → bus-s)
    - then map from intermediate form to surface form according to rules (bus-s → busses)
  - Can be used as a generator or an analyzer
- Tools: Foma, RustFst + OpenFst

Morphological Analyzers

- Traditionally done with finite state transducers (FSTs)
- More recently, neural models are being used
  - Can combine approaches
    - E.g. combining an FST with a neural guesser for unseen word forms
  - Can use FSTs to generate additional training data
- GlossLM (Ginn et al., 2024) continually pretrains ByT5 on ~1800 langs to generate a form of morphological annotation (interlinear gloss)



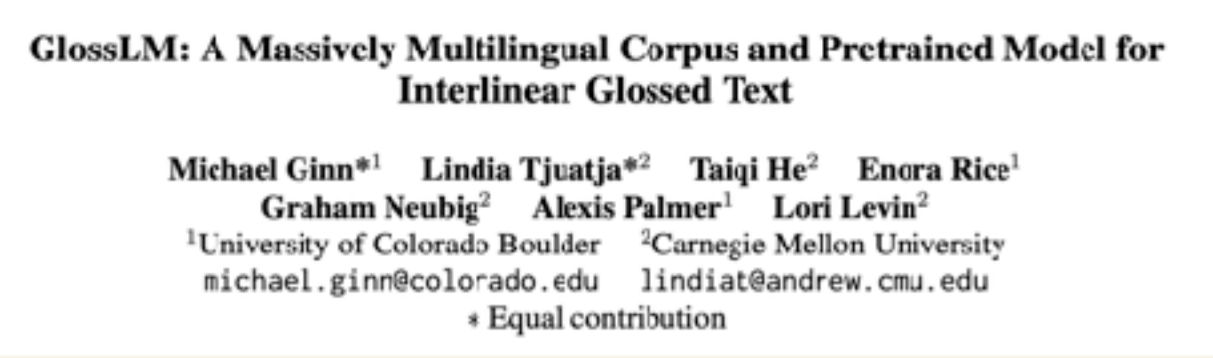
Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik from a Finite-State Transducer

Lane Schwartz
University of Illinois
at Urbana-Champaign
lanes@illinois.edu

Emily Chen
University of Illinois
at Urbana-Champaign
echen41@illinois.edu

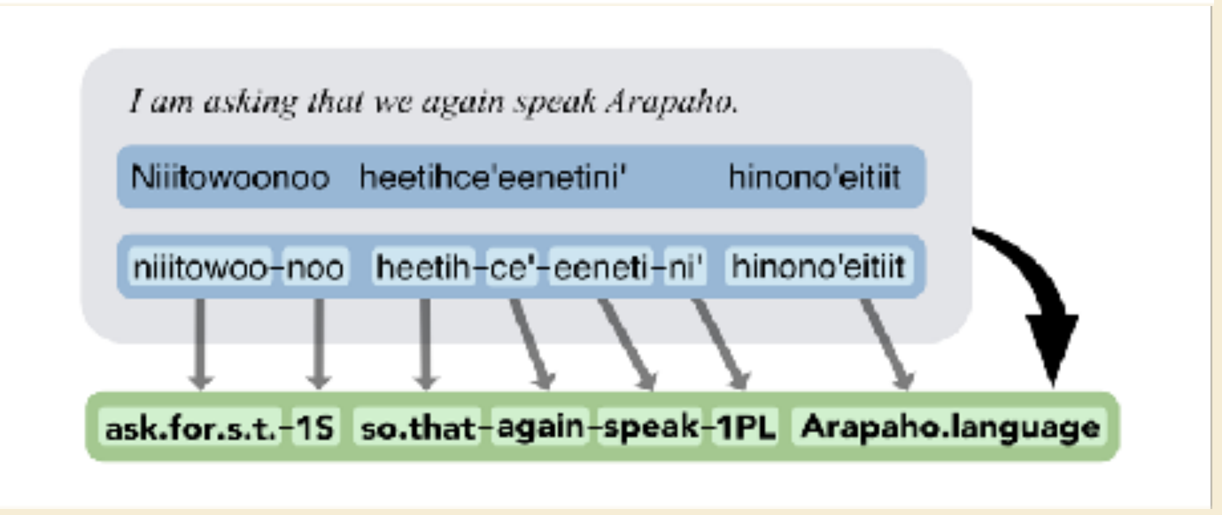
Benjamin Hunt
George Mason University
bhunt6@gmu.edu

Sylvia L.R. Schreiner
George Mason University
sschrei2@gmu.edu



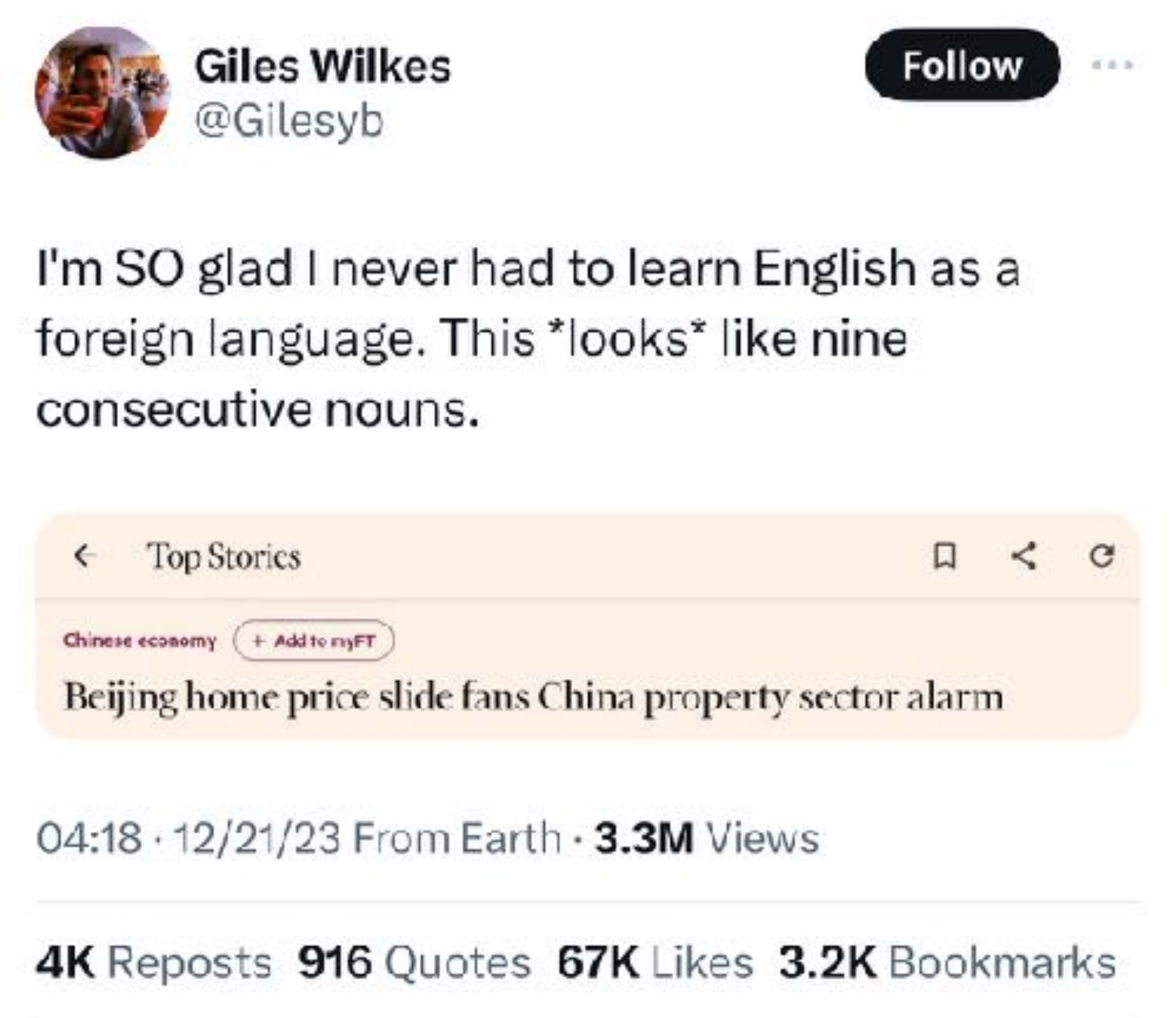
GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text

Michael Ginn*[1] **Lindia Tjuatja***[2] **Taiqi He**[2] **Enora Rice**[1]
Graham Neubig[2] **Alexis Palmer**[1] **Lori Levin**[2]
[1]University of Colorado Boulder [2]Carnegie Mellon University
michael.ginn@colorado.edu lindiat@andrew.cmu.edu
* Equal contribution


I am asking that we again speak Arapaho.
Niiitowoonoo heetihce'eenetini' hinono'eitiit
niiitowoo-noo heetih-ce'-eeneti-ni' hinono'eitiit
ask.for.s.t.-1S so.that-again-speak-1PL Arapaho.language

Syntax

- The study of how words form phrases and sentences
  - What are the principles governing phrase and sentence structure within a language and across languages?
- Aspects of syntax include:
  - Word order (e.g. SVO, SOV, etc.)
  - Agreement (e.g. subject-verb agreement)
  - Hierarchical structure (e.g. what modifies what in a sentence)
  - etc.



Giles Wilkes
@Gilesyb

I'm SO glad I never had to learn English as a foreign language. This *looks* like nine consecutive nouns.

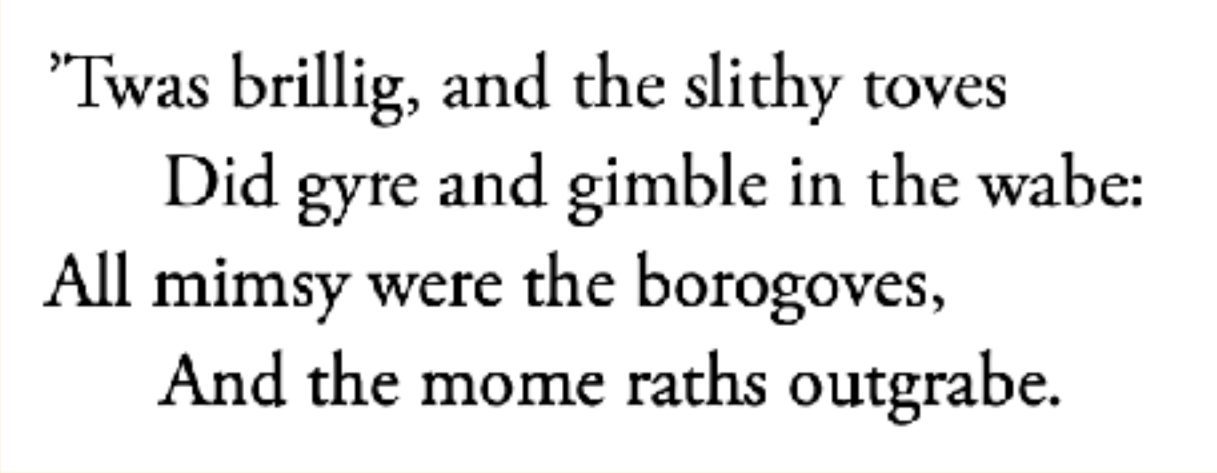
Beijing home price slide fans China property sector alarm

04:18 · 12/21/23 From Earth · 3.3M Views

4K Reposts 916 Quotes 67K Likes 3.2K Bookmarks

Syntax

Word Classes and Parts of Speech



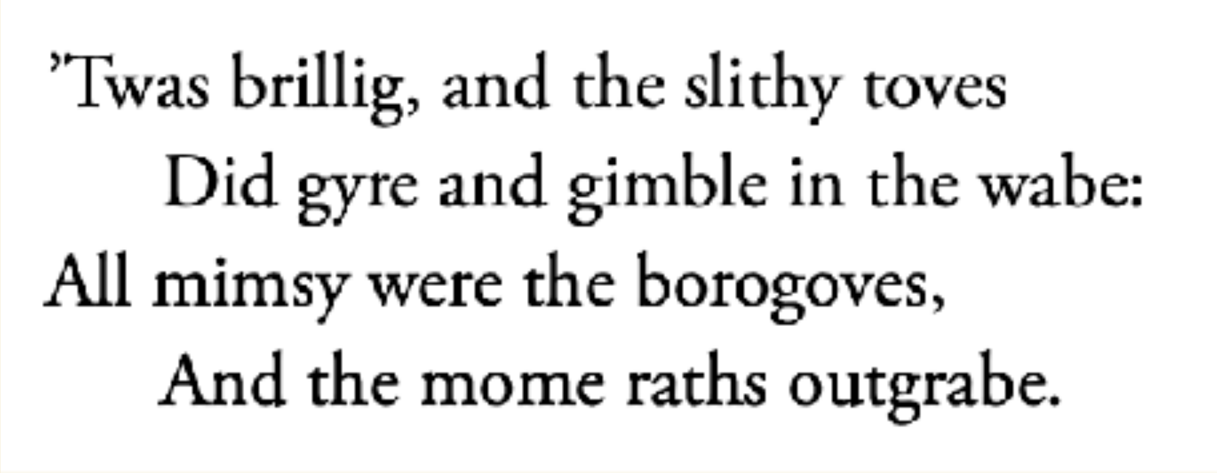
’Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

from *Jabberwocky* (aka, every linguist’s go-to POS example), by Lewis Carroll

- Words can be categorized based on their morphological, syntactic, and semantic properties
  - We refer to these categories as *parts of speech*, e.g. nouns, verbs, adjectives, etc.
  - However, this categorization is not hard-and-fast across languages, and should not be taken for granted!

Syntax

Word Classes and Parts of Speech



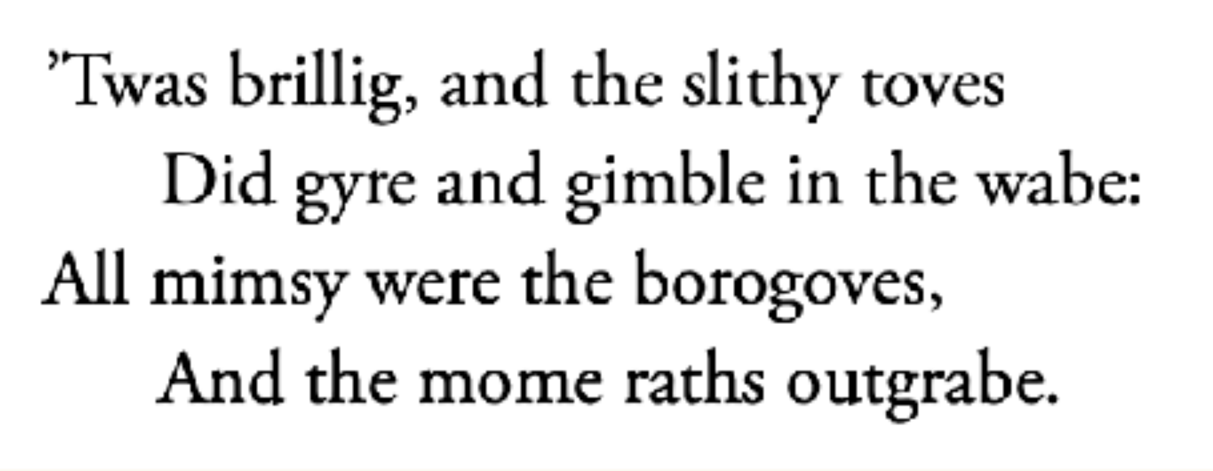
’Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

from *Jabberwocky* (aka, every linguist’s go-to POS example), by Lewis Carroll

- Words can be categorized based on their morphological, syntactic, and semantic properties
  - We refer to these categories as *parts of speech*, e.g. nouns, verbs, adjectives, etc.
  - However, this categorization is not hard-and-fast across languages, and should not be taken for granted!
- A very broad distinction we can make are between:
  - *Open class* words: new items are added over time with relative ease (e.g. *rizz*)
  - *Closed class* words: much smaller number of words, harder to add new items

Syntax

Word Classes and Parts of Speech



’Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

from *Jabberwocky* (aka, every linguist’s
go-to POS example), by Lewis Carroll

- Words can be categorized based on their morphological, syntactic, and semantic properties
  - We refer to these categories as *parts of speech*, e.g. nouns, verbs, adjectives, etc.
  - However, this categorization is not hard-and-fast across languages, and should not be taken for granted!
- A very broad distinction we can make are between:
  - *Open class* words: new items are added over time with relative ease (e.g. *rizz*)
  - *Closed class* words: much smaller number of words, harder to add new items
- Based on how a word acts in context, we can often infer its function and POS even if we’ve never seen it before, as in the Jabberwocky example

Syntax

Word Classes and Parts of Speech

Nouns

Verbs

Adjectives

Adverbs

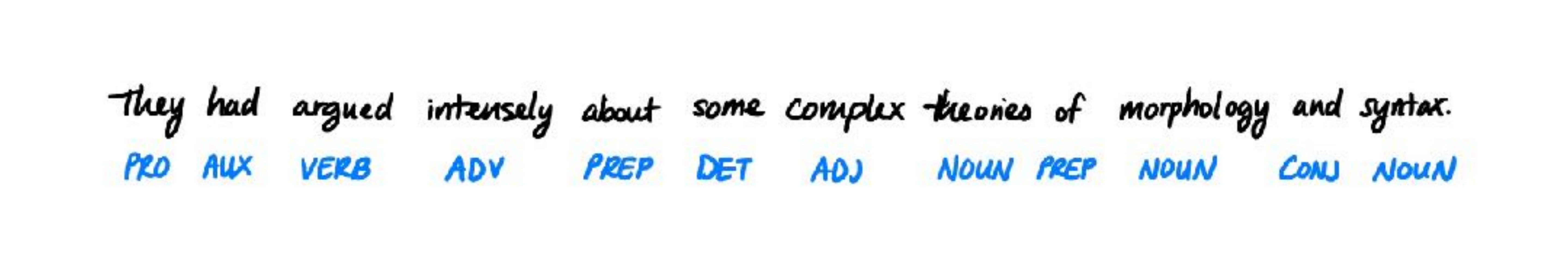
Determiners

Auxiliary Verbs

Pronouns

Prepositions

Conjunctions



They	had	argued	intensely	about	some	complex	theories	of	morphology	and	syntax.
PRO	AUX	VERB	ADV	PREP	DET	ADJ	NOUN	PREP	NOUN	CONJ	NOUN

Syntax

Phrases

Words can combine together to form different types of phrases:

- **Noun phrase** (NP): contains a **noun**, may also include a determiner and adjectival modifiers
    - [The [old [man]]]
- **Prepositional phrase** (PP): contains a **preposition** followed by a NP
    - [on the shelf]
- **Verb phrase** (VP): contains a **verb** and any NP/PP phrases that verb requires / has an slot for, as well as adverbial modifiers
    - (The old man) [sold a car to me]

Syntax

Constituents

A *constituent* consists of at least one contiguous word that **behaves as a single unit**

Anthropic released [a [new [language [model]]]].

A crucial observation is that we can replace units with smaller and smaller constituents of the same category, down to the word level.

Syntax
Constituents

A *constituent* consists of at least one contiguous word that **behaves as a single unit**

Anthropic released [a [new [agent]]].

Anthropic Wants Its AI Agent to Control Your Computer

Claude is the first major AI model to be able to take control of a computer to do useful work.

A crucial observation is that we can replace units with smaller and smaller constituents of the same category, down to the word level.

Syntax

Constituents

A *constituent* consists of at least one contiguous word that **behaves as a single unit**

Anthropic released [a [shrimp]].

A crucial observation is that we can replace units with smaller and smaller constituents of the same category, down to the word level.

Syntax

Constituents

A *constituent* consists of at least one contiguous word that **behaves as a single unit**

Anthropic released [AGI].

A crucial observation is that we can replace units with smaller and smaller constituents of the same category, down to the word level.

Syntax

Context-Free Grammars (in linguistics specifically, Phrase Structure Grammars)

- Originally introduced by Noam Chomsky
  - "A phrase-structure grammar is defined by a finite vocabulary V, and a finite set Σ of initial strings in V, and a finite set F of rules of the form: $X \rightarrow Y$, where X and Y are strings in V."
- Some example phrase structure rules for English:
  - S → NP VP [a sentence is comprised of a NP followed by a VP]
  - NP → (Det) NP_1 [a NP is comprised by an optional determiner and some NP_1]
  - NP_1 → (AP) N (PP) [NP_1 is comprised of an optional AP and a N and optional PP] …

Syntax

Context-Free Grammars (in linguistics specifically, Phrase Structure Grammars)

- Using such set of rules, we can generate lots and lots of English sentences, including those that are syntactically proper (even if semantically nonsensical)
    - E.g. *Colorless green ideas sleep furiously*
    - And amazingly, speakers have an intuition for this!

Syntax

Context-Free Grammars (in linguistics specifically, Phrase Structure Grammars)

- Using such set of rules, we can generate lots and lots of English sentences, including those that are syntactically proper (even if semantically nonsensical)
- However, if what we've seen so far seems like too simple of an approach…you're right

Syntax

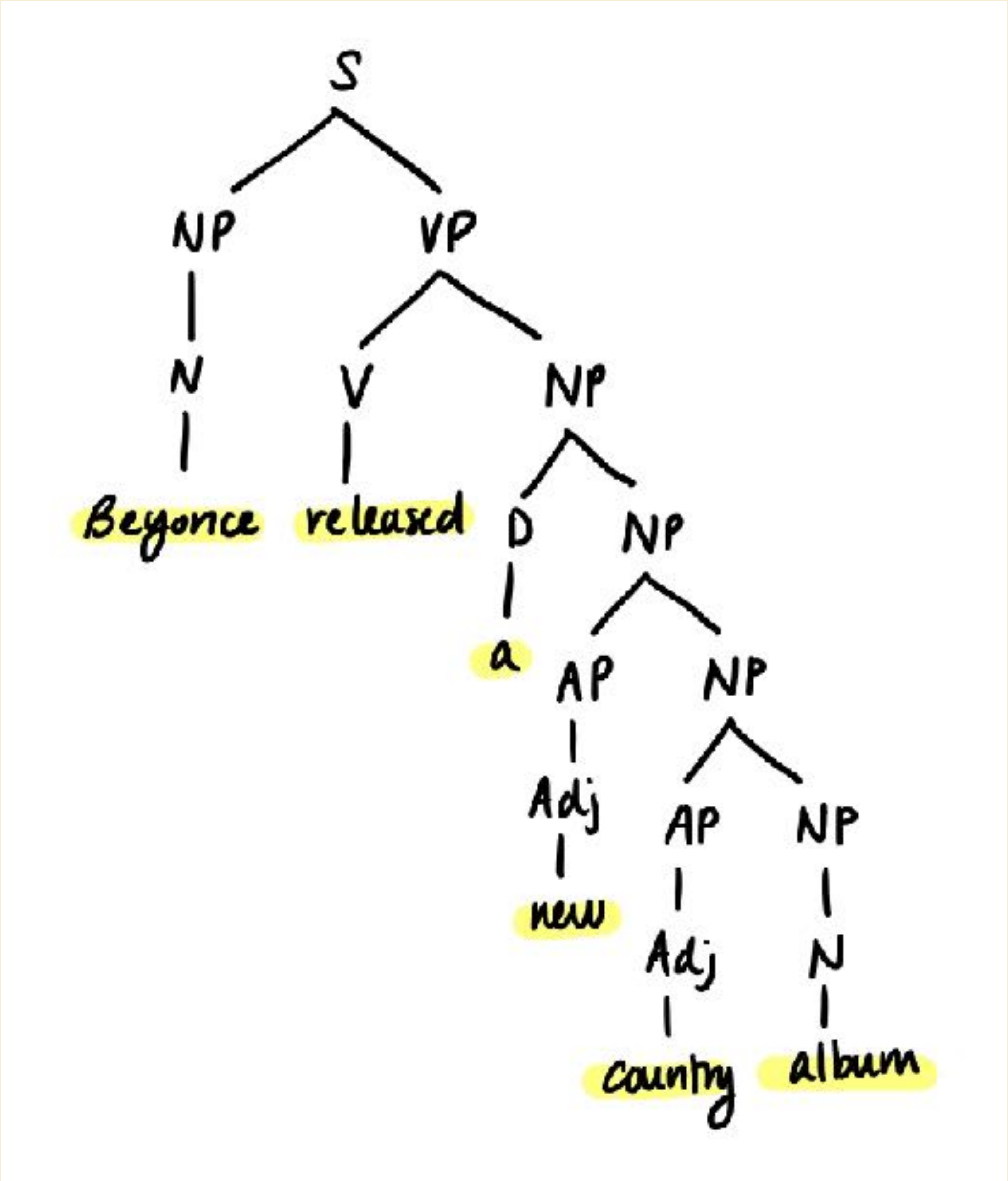
Context-Free Grammars (in linguistics specifically, Phrase Structure Grammars)

- Using such set of rules, we can generate lots and lots of English sentences, including those that are syntactically proper (even if semantically nonsensical)
- However, if what we've seen so far seems like too simple of an approach…you're right
  - Some phenomena are very difficult to model this way
  - Theoretical syntax has since expanded beyond these basic rules
    - E.g. newer generative frameworks like Minimalism, other formalisms like HPSG, cognitive linguistics approaches like construction grammar, etc.
  - Nevertheless, conceptually powerful and remains influential

Syntax

Constituency Trees

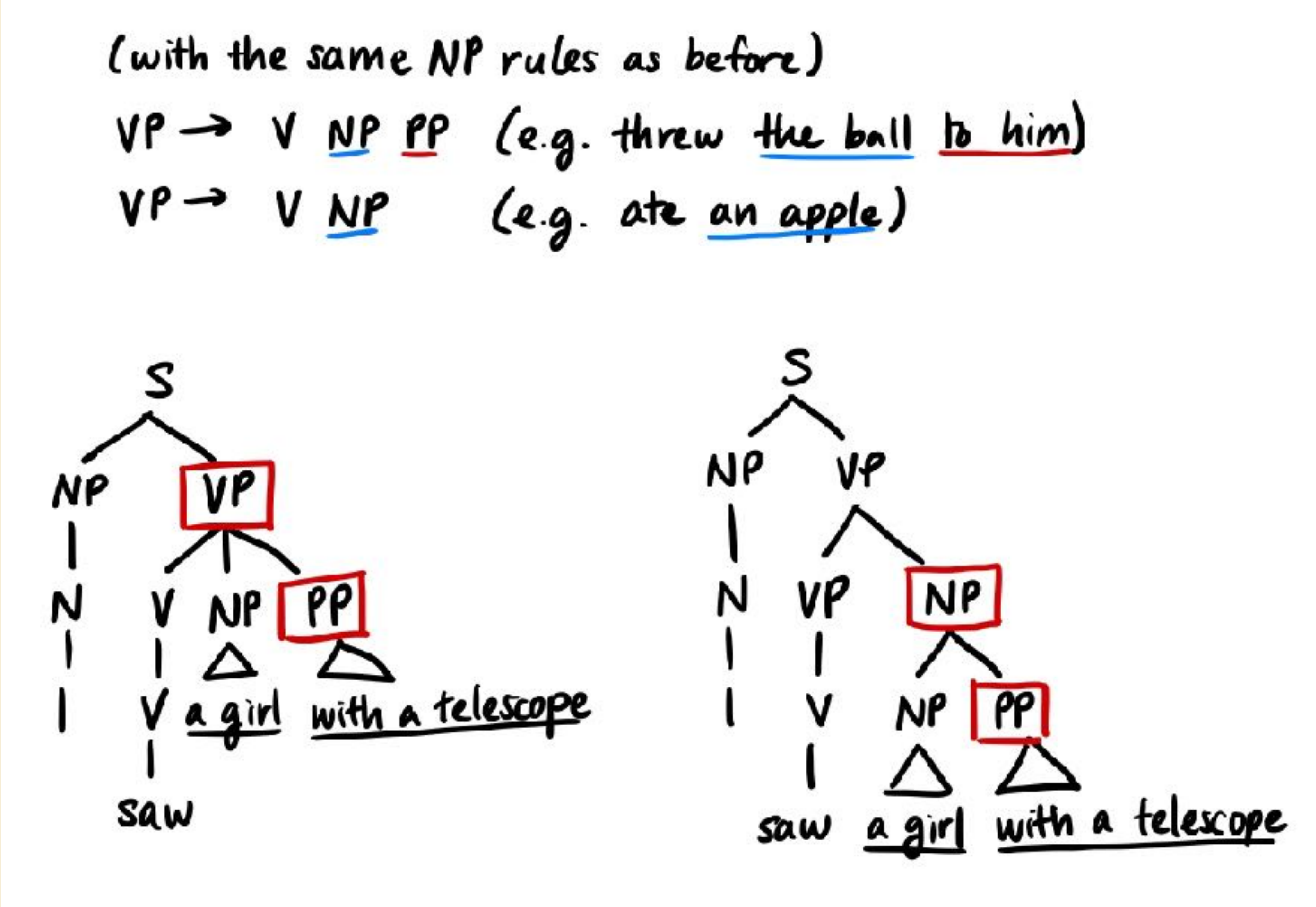
- An important aspect of this line of work (and subsequent + competing theories) is the idea of hierarchical structure in syntax
- We can represent how phrase structure rules break down sentences in a tree, with the sentence node S as the root and words as the leaves



Syntax

Ambiguity

- We can also represent *syntactic ambiguity*
- Here we have two trees for the same surface form sentence, which mean slightly different things
  - Depends on what the PP directly attaches to



(with the same NP rules as before)

VP → V NP PP (e.g. threw the ball to him)

VP → V NP (e.g. ate an apple)

Syntax

Dependency Grammars & Trees

- While constituency trees are based on constituency relations (as the name suggests), dependency trees are based on…

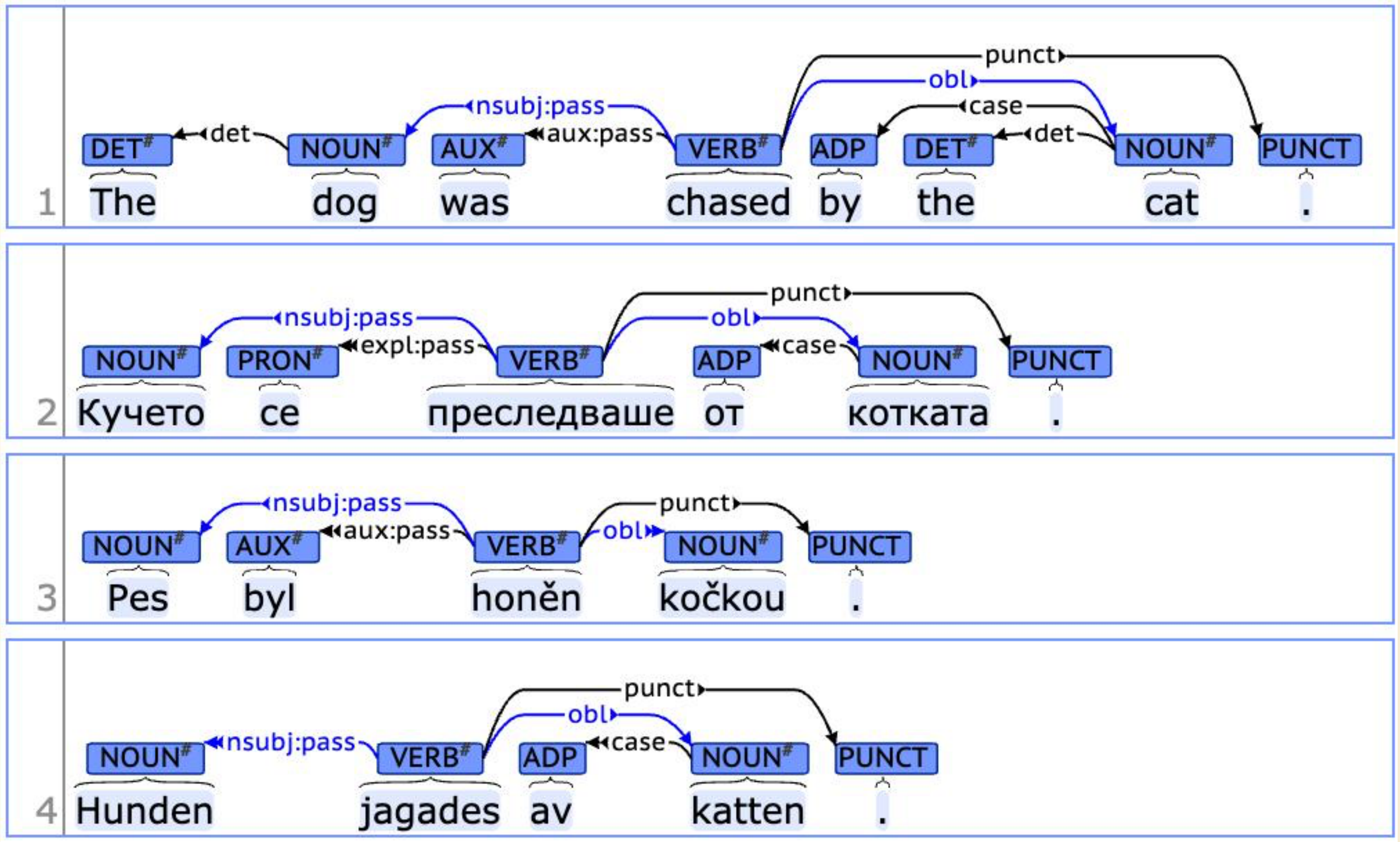
Syntax

Dependency Grammars & Trees

- While constituency trees are based on constituency relations (as the name suggests), dependency trees are based on…
- *Dependency relations* (sometimes referred to as *grammatical relations*) are binary, asymmetrical relations that connect words and phrases
  - In the relation A → B, A is the *head* and B is the *dependent*
  - The relation can be syntactic, semantic, morphological, prosodic…but most frameworks focus on syntactic relations, with the main verb serving as the root
    - Clausal relations: nominal subject, direct object, indirect object…
    - Modifier relations: nominal modifier, adjectival modifier, adverbial modifier, determiner…
    - etc.

Syntax

Dependency Grammars & Trees



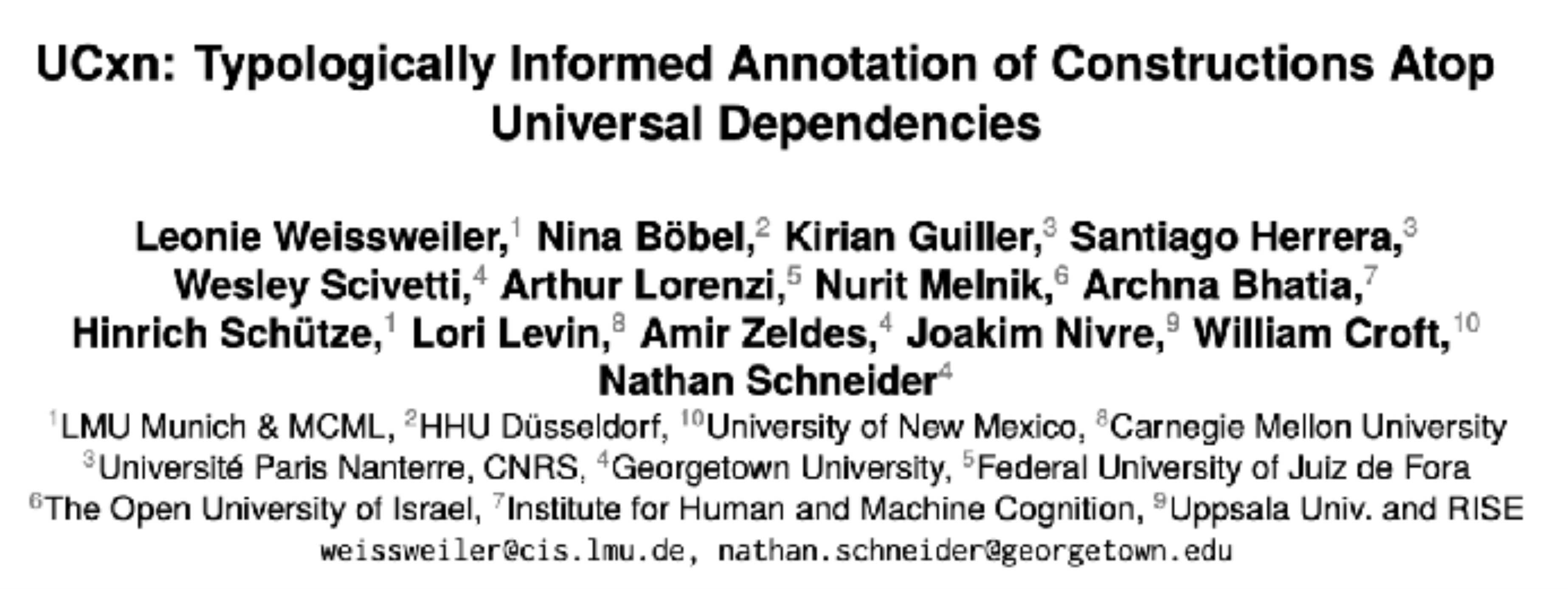
Taken from https://universaldependencies.org/introduction.html

POS Tagging, Syntactic Parsing, and Annotation

- These tasks used to be a big deal! Not as much anymore…for *high resource* languages
  - Still a valuable resource for people studying lower resource languages
  - Having linguistically annotated corpora over a wide variety of languages can enable us to do cross-linguistic studies

POS Tagging, Syntactic Parsing, and Annotation

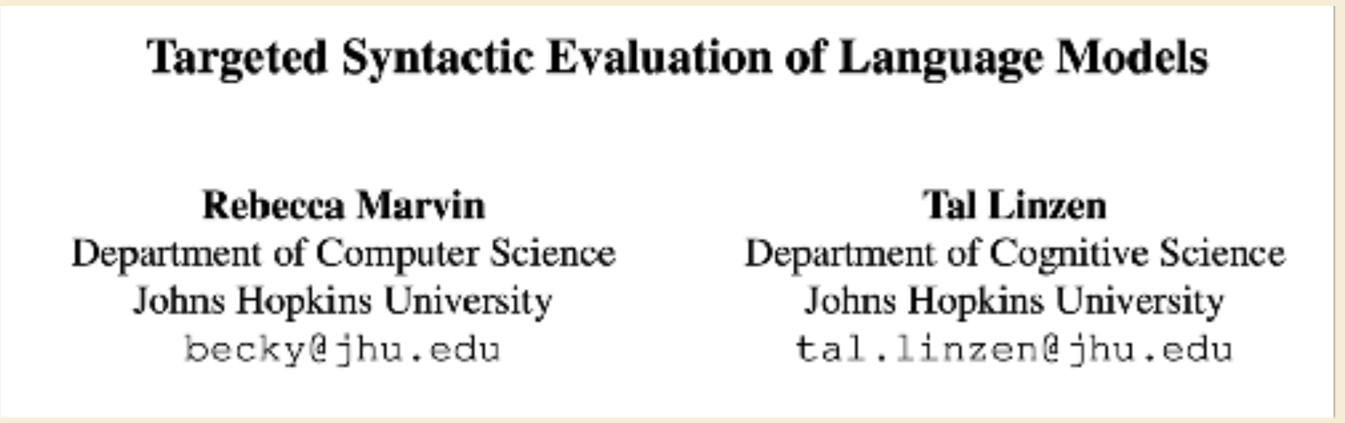
- (Eng) Brown Corpus, Corpus of Contemporary American English (COCA)
- (Eng) Penn Treebank
- (Eng) Google Syntactic N-grams
- Universal Dependencies
  - Over 140 languages
  - Still a continual effort to develop more descriptive annotations across languages



UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies

Leonie Weissweiler,[1] Nina Böbel,[2] Kirian Guiller,[3] Santiago Herrera,[3]
Wesley Scivetti,[4] Arthur Lorenzi,[5] Nurit Melnik,[6] Archna Bhatia,[7]
Hinrich Schütze,[1] Lori Levin,[8] Amir Zeldes,[4] Joakim Nivre,[9] William Croft,[10]
Nathan Schneider[4]
[1]LMU Munich & MCML, [2]HHU Düsseldorf, [10]University of New Mexico, [8]Carnegie Mellon University
[3]Université Paris Nanterre, CNRS, [4]Georgetown University, [5]Federal University of Juiz de Fora
[6]The Open University of Israel, [7]Institute for Human and Machine Cognition, [9]Uppsala Univ. and RISE
weissweiler@cis.lmu.de, nathan.schneider@georgetown.edu

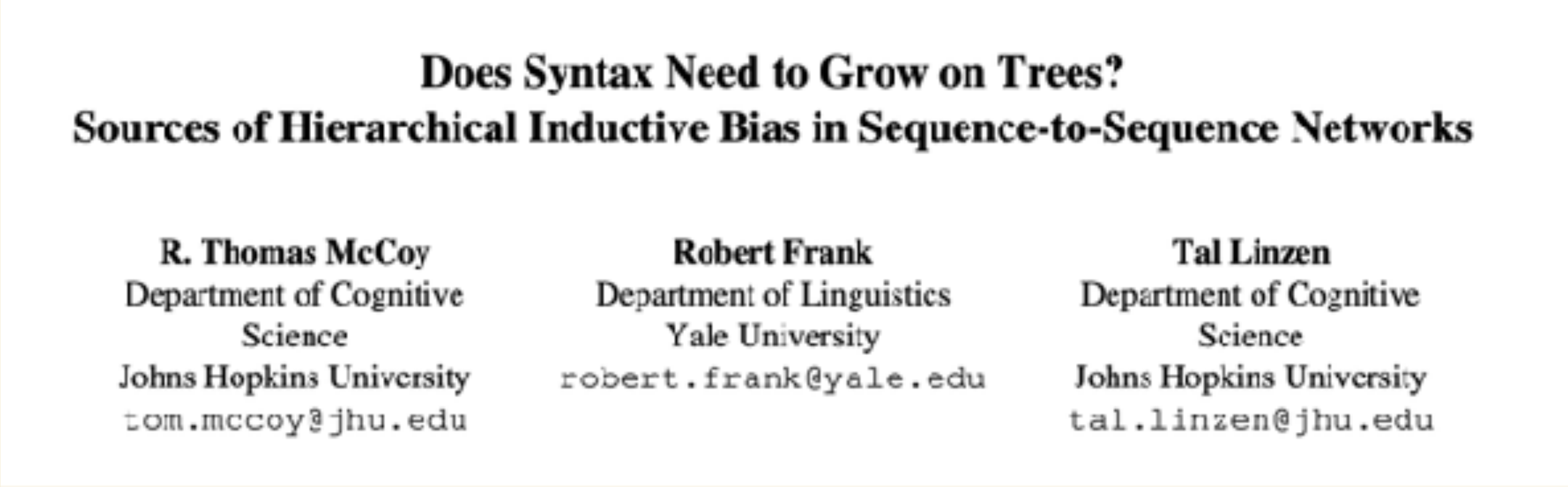
Do statistical models of language learn grammar?



Targeted Syntactic Evaluation of Language Models

Rebecca Marvin
Department of Computer Science
Johns Hopkins University
becky@jhu.edu

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

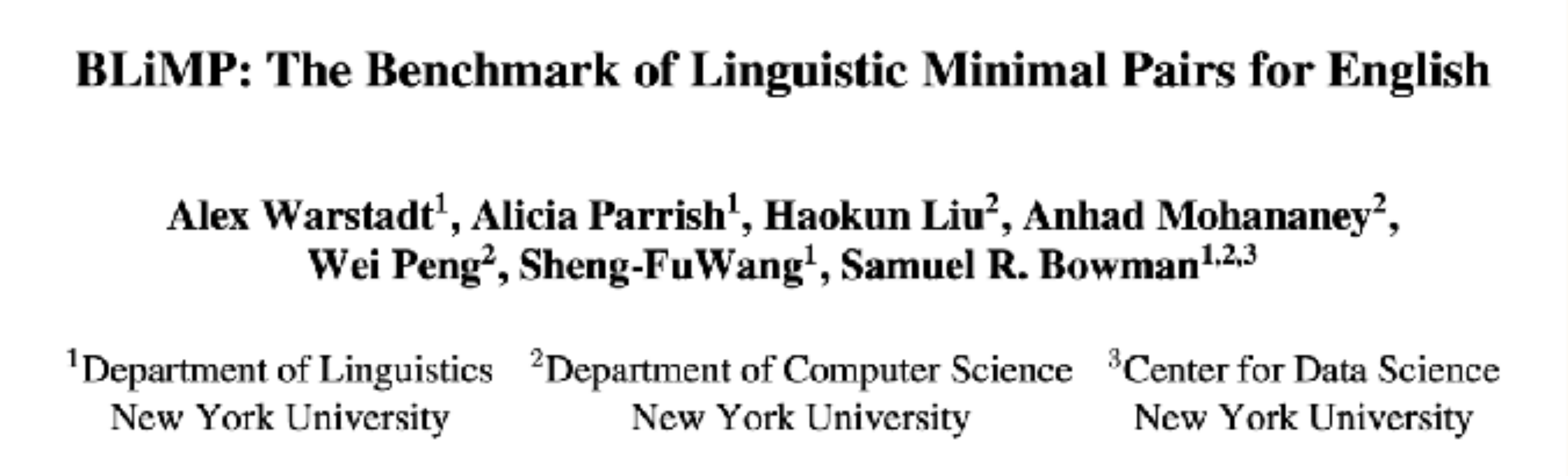


Does Syntax Need to Grow on Trees?
Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks

R. Thomas McCoy
Department of Cognitive Science
Johns Hopkins University
tom.mccoy@jhu.edu

Robert Frank
Department of Linguistics
Yale University
robert.frank@yale.edu

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu



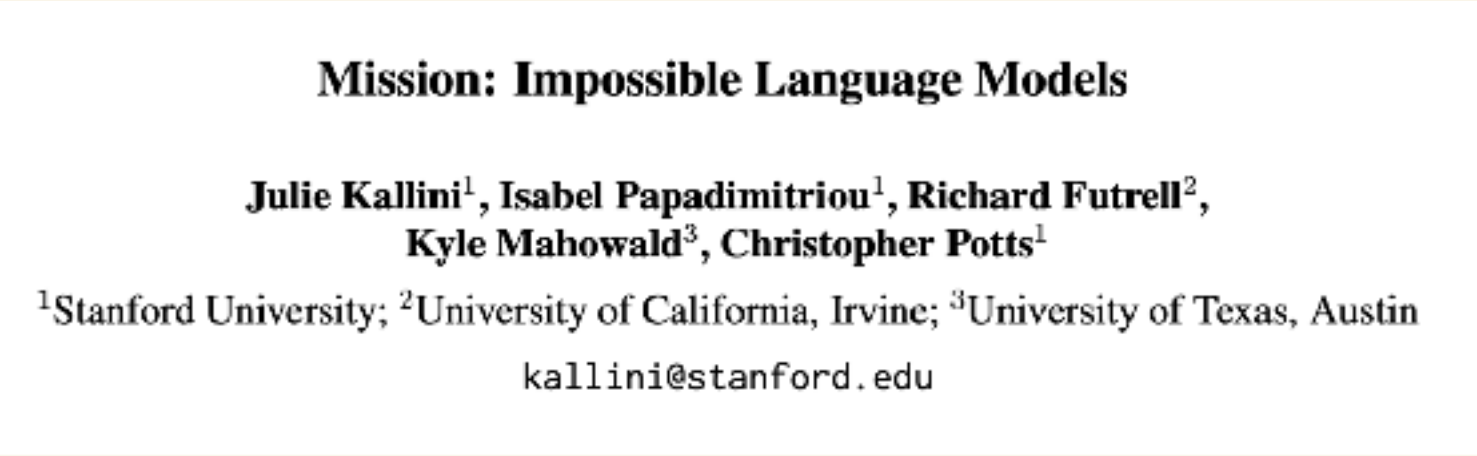
BLiMP: The Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt[1], Alicia Parrish[1], Haokun Liu[2], Anhad Mohananey[2],
Wei Peng[2], Sheng-FuWang[1], Samuel R. Bowman[1,2,3]

[1]Department of Linguistics
New York University

[2]Department of Computer Science
New York University

[3]Center for Data Science
New York University

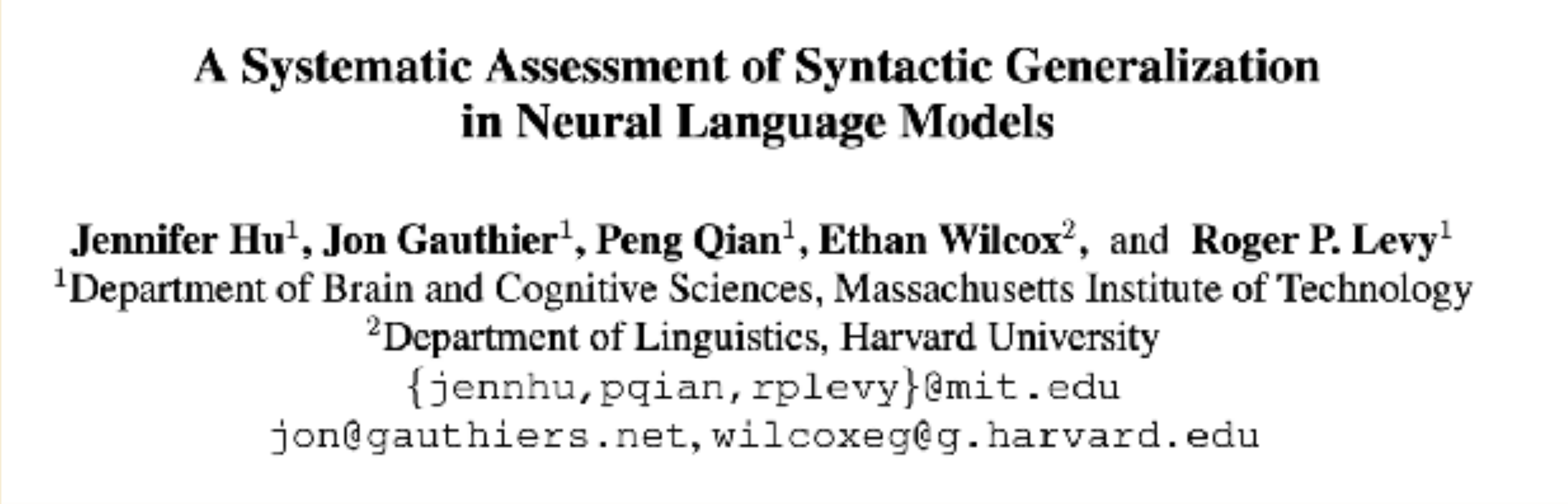


Mission: Impossible Language Models

Julie Kallini[1], Isabel Papadimitriou[1], Richard Futrell[2],
Kyle Mahowald[3], Christopher Potts[1]

[1]Stanford University; [2]University of California, Irvine; [3]University of Texas, Austin

kallini@stanford.edu



A Systematic Assessment of Syntactic Generalization
in Neural Language Models

Jennifer Hu[1], Jon Gauthier[1], Peng Qian[1], Ethan Wilcox[2], and Roger P. Levy[1]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[2]Department of Linguistics, Harvard University
{jennhu,pqian,rplevy}@mit.edu
jon@gauthiers.net, wilcoxeg@g.harvard.edu

Meaning and Intent

Semantics

- The study of linguistic meaning
- Can study this at various levels (morpheme, word, sentence)
- As we saw earlier, often interacts with morphology + syntax
    - Syntax-semantics interface: What is the relationship between syntactic form and meaning?
- Talking about meaning can veer easily into philosophy of language…we'll stick to computationally relevant topics here!
    - Even then, we have limited time, so I'll have to skip some topics that may be of interest, like propositional and first-order logic

Semantics

Lexical Semantics and Word Senses

- A *sense* of a word is a distinct meaning of a word

Semantics

Lexical Semantics and Word Senses

- A *sense* of a word is a distinct meaning of a word
- Words can have multiple, semantically related senses → word *polysemy*
  - *They **run** experiments*, *They **run** races*, *Candidates **run** for office*, *Can I **run** this idea by you?*, etc.

Semantics

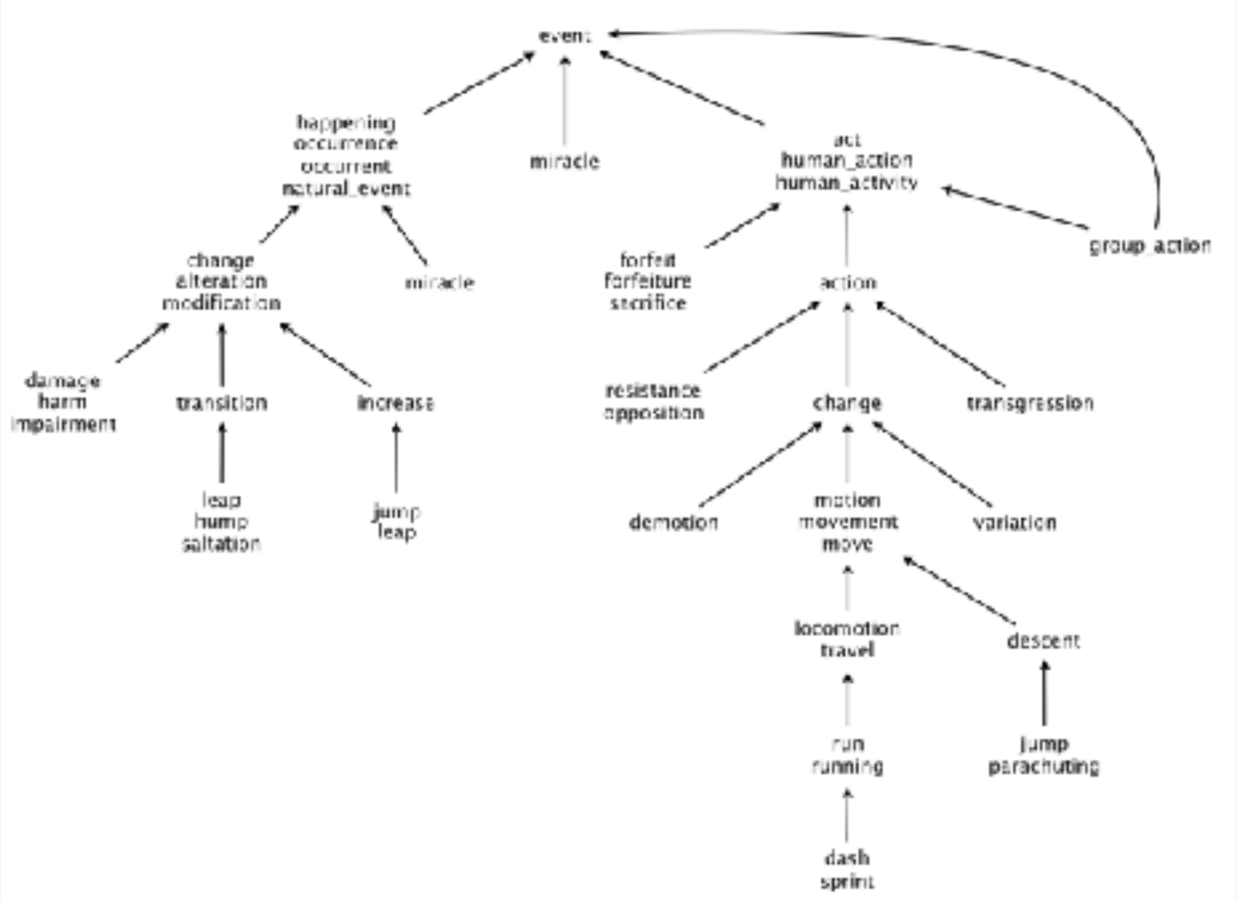
Lexical Semantics and Word Senses

- A *sense* of a word is a distinct meaning of a word
- Words can have multiple, semantically related senses → word *polysemy*
  - *They **run** experiments*, *They **run** races*, *Candidates **run** for office*, *Can I **run** this idea by you?*, etc.
- Relations between senses:
  - Synonymy-antonymy (same-opposite)
  - Hyperonymy-hyponymy (super-subordinate)
  - Meronymy-holonymy (part-whole), etc.

Semantics

Lexical Semantics and Word Senses

- **WordNet** (Fellbaum 2005): large lexical database of English words
  - Content words are grouped into sets of synonyms (synsets)
  - Synsets are linked through conceptual-semantic and lexical relations
    - Most common: super-subordinate relations (hyperonymy and hyponymy)
    - Distinguish between Types (common nouns) and Instances (proper nouns), with Instances always being terminal nodes in their hierarchies
- WordNets in different languages have since been created
- ImageNet (Deng et al. 2009) based its hierarchy according to nouns in WordNet



Taken from https://www.cs.princeton.edu/courses/archive/spring20/cos226/assignments/wordnet/specification.php

Semantics

Distributional Semantics and Word Embeddings

- *Distributional Hypothesis* (Harris 1954): linguistic items that have similar distributions have similar meanings
    - "You shall know a word by the company it keeps" (J.R. Firth)
    - This idea is the foundation for statistical approaches to (lexical) semantics

Semantics

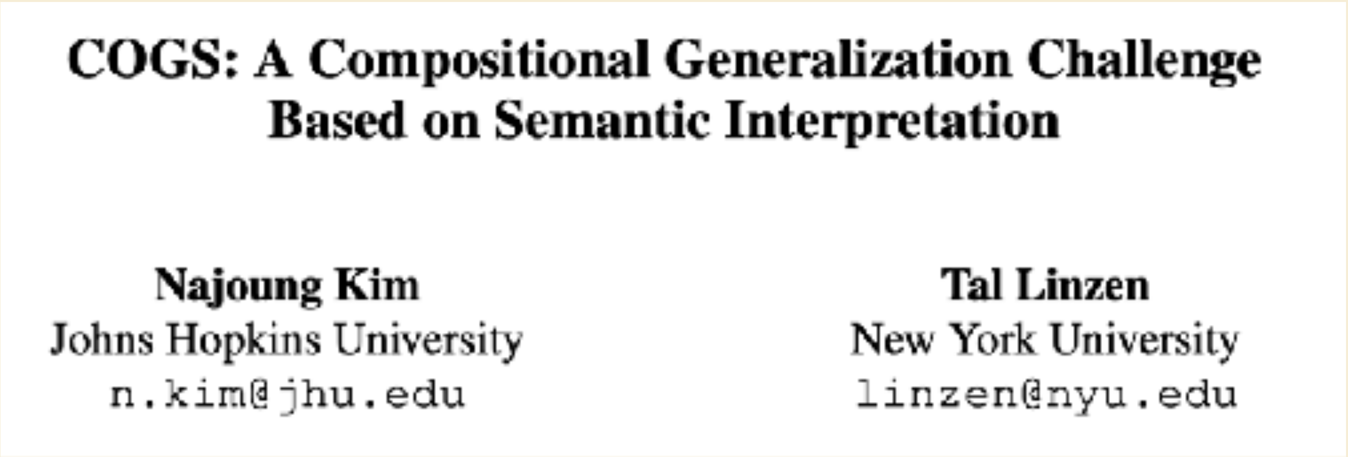
Distributional Semantics and Word Embeddings

- *Distributional Hypothesis* (Harris 1954): linguistic items that have similar distributions have similar meanings
  - “You shall know a word by the company it keeps” (J.R. Firth)
  - This idea is the foundation for statistical approaches to (lexical) semantics
- Given a large corpus, we can form vector representations of words based on statistical relationships between the words
  - Can show sense relations with cosine similarity, vector arithmetic
  - (Dense) Static Embeddings: **word2vec**, **GloVe**
  - Contextual Embeddings: **ELMo**, **BERT**

Semantics

Compositionality

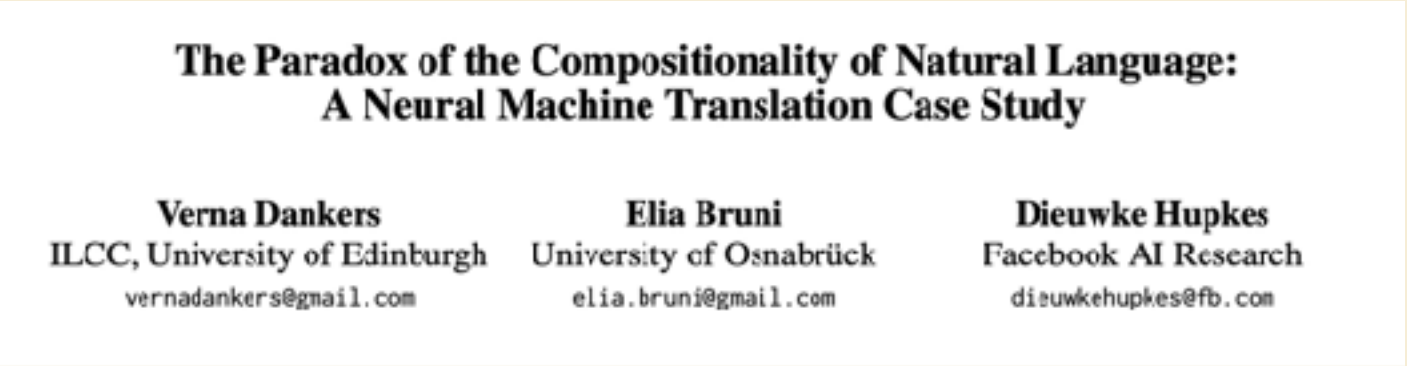
- It seems like much of natural language is *compositional*: the meaning of the whole is comprised of the structure and meaning of its parts
  - We saw this in the morphology examples!
  - In sentences, we can combine the meaning of lexical items and phrases
- We can create novel sentences and structures systematically; similarly, we can determine the meaning of novel sentences and structures
  - How well can (cognitive/language) models do this?
- There are also exceptions to compositionality, such as *idioms* and figurative language
  - A challenge in applications like MT



COGS: A Compositional Generalization Challenge Based on Semantic Interpretation

Najoung Kim
Johns Hopkins University
n.kim@jhu.edu

Tal Linzen
New York University
linzen@nyu.edu

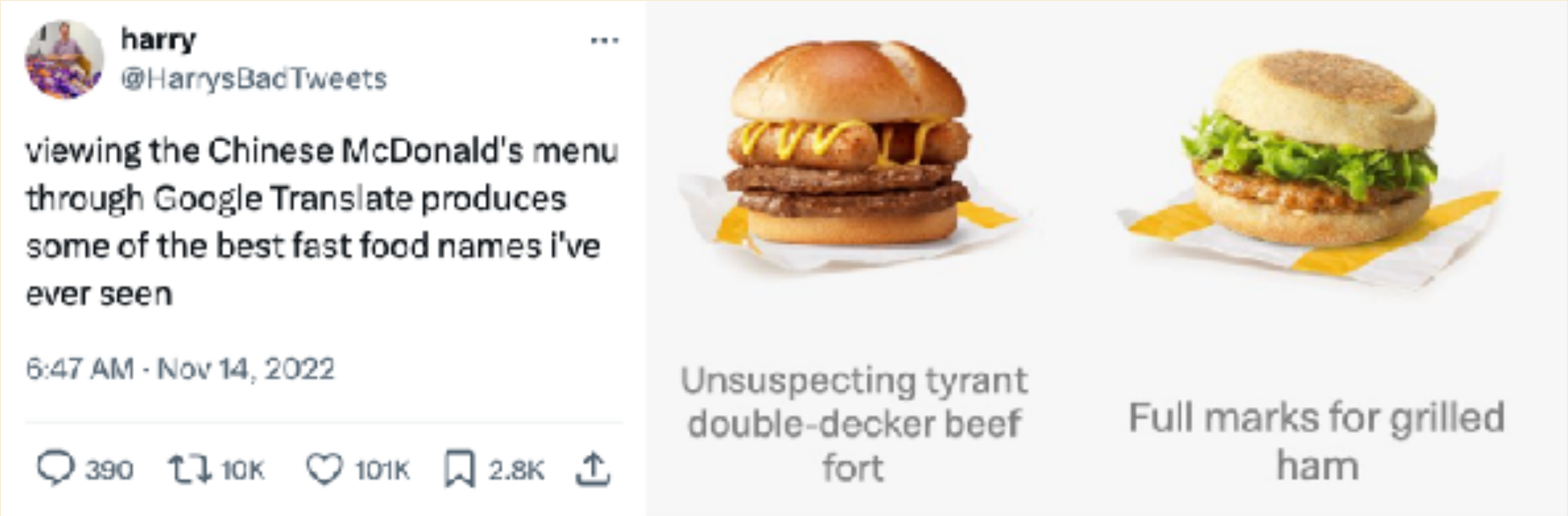


The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study

Verna Dankers
ILCC, University of Edinburgh
vernadankers@gmail.com

Elia Bruni
University of Osnabrück
elia.bruni@gmail.com

Dieuwke Hupkes
Facebook AI Research
dieuwkehupkes@fb.com



harry
@HarrysBadTweets

viewing the Chinese McDonald's menu through Google Translate produces some of the best fast food names I've ever seen

6:47 AM · Nov 14, 2022

390 10K 101K 2.8K

Unsuspecting tyrant double-decker beef fort

Full marks for grilled ham

Semantics

Entailment and Natural Language Inference

- One aspect of an expression's meaning is its *truth condition(s)*, or the condition(s) under which the expression would be true
    - *Emmy is a cute cat* is `True` if *Emmy is a cat*, but `False` if *Emmy is a dog*

Semantics

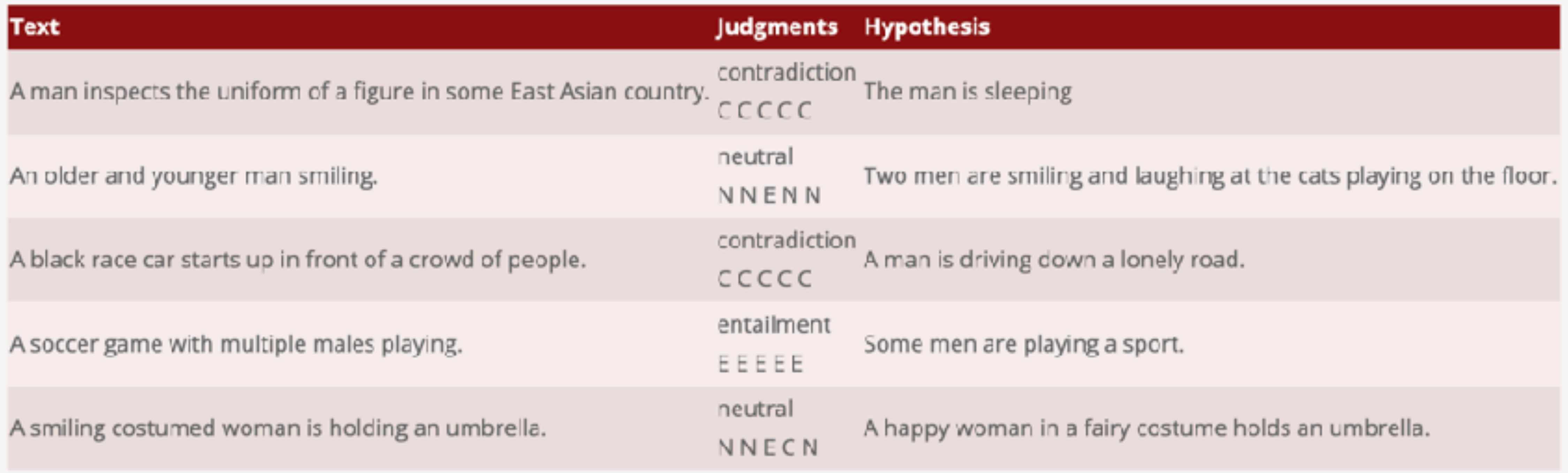
Entailment and Natural Language Inference

- One aspect of an expression's meaning is its *truth condition(s)*, or the condition(s) under which the expression would be true
    - *Emmy is a cute cat* is `True` if *Emmy is a cat*, but `False` if *Emmy is a dog*
- *Entailment* is a relationship between expressions
    - If A entails B, then B must be `True` if A is `True`
        - In other words, B is a *truth condition* of A
    - *Emmy is my adorable, little orange cat* entails:
        - *Emmy is a cat*
        - *Emmy is little*
        - *Emmy is adorable*

Semantics

Entailment and Natural Language Inference

- *Natural Language Inference* is an NLP task where given a *premise*, determine if a *hypothesis* is **entailed** or **contradicted** by that premise
  - Datasets: SNLI, Multi-NLI, SciTail, XNLI



Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Examples from the SNLI dataset

Taken from https://nlp.stanford.edu/projects/snli/

- Entailment models can be useful for factuality checking in generation, checking if two sources agree, etc.!

Pragmatics

- The study of language use in context
  - How is language used in social interactions?
  - How does context (linguistic or otherwise) influence language use?
  - What do we intend to mean when we say something, and how does this influence its interpretation?
- *Speech act theory* — the meaning of an utterance is comprised of not just the statement itself, **but also the intended effect of the utterance on the listener**
  - “Can you pass me the salt?”
  - “Do you mind if I sit next to you?” → {“Yes (go ahead)”, “No, I don’t mind”}

Pragmatics

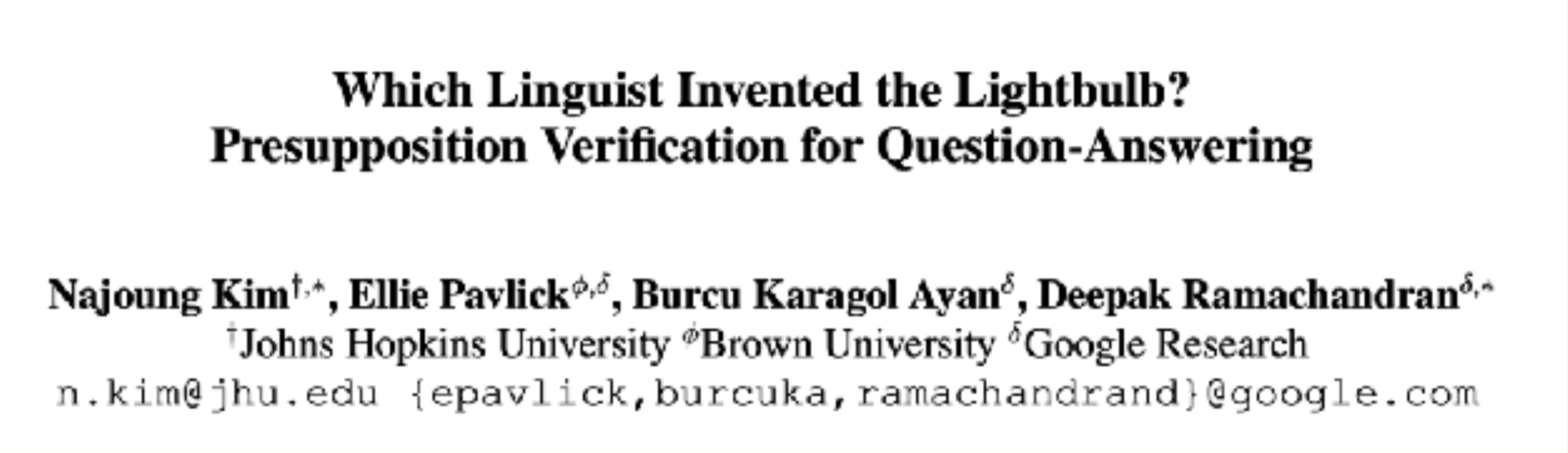
Presupposition

- *Presuppositions* are implicit assumptions about the world that are used in discourse
  - *Everyone thinks **my cat** is cute* presupposes that I have a cat…it would be super strange for me to say this otherwise

Pragmatics

Presupposition

- *Presuppositions* are implicit assumptions about the world that are used in discourse
- Presuppositions can be "triggered" by certain lexical items or constructions
  - Definite descriptions: *The king of France* ("the X" presupposes you're referring to one thing, and that such a thing exists)
  - Factives: *I **regret** staying up all night to watch the election* (presupposes I did in fact stay up to watch the election)
  - Questions: ***Which** linguist invented the lightbulb?* (Presupposes some linguist invented the lightbulb)
  - Etc.



Which Linguist Invented the Lightbulb?
Presupposition Verification for Question-Answering

Najoung Kim[†,*], Ellie Pavlick[φ,δ], Burcu Karagol Ayan[δ], Deepak Ramachandran[δ,*]
[†]Johns Hopkins University [φ]Brown University [δ]Google Research
n.kim@jhu.edu {epavlick,burcuka,ramachandrand}@google.com

Pragmatics

Implicature

- *Implicatures* are things that are suggested by an utterance, though not necessarily literally expressed
    - [It's lightly raining outside] *Today's weather is **the worst***.
        - Not literally the worst, but quite bad and I don't it
    - Q: Did you vote?

      A: I was sick on Tuesday.

Pragmatics

Implicature

- *Implicatures* are things that are suggested by an utterance, though not necessarily literally expressed
  - [It's lightly raining outside] *Today's weather is* ***the worst****.*
    - Not literally the worst, but quite bad and I don't it
  - Q: Did you vote?

    A: I was sick on Tuesday.

    A': I was sick on Tuesday, but I voted anyways.

Unlike entailments, implicatures are *defeasible*

Pragmatics

Gricean Maxims

How do people conduct conversations and achieve effective communication?

Pragmatics

Gricean Maxims

For the most part, people are rational speakers and expect+follow certain conversational conventions (maxims):

1. *Quantity* (don't undershare, don't overshare)
2. *Truth* (don't lie)
3. *Relation* (be relevant)
4. *Manner* (be clear)

Pragmatics

Gricean Maxims

For the most part, people are rational speakers and expect+follow certain conversational conventions (maxims):

1. *Quantity* (don’t undershare, don’t overshare)
2. *Truth* (don’t lie)
3. *Relation* (be relevant)
4. *Manner* (be clear)

- Speakers can *flout* maxims (e.g. sarcasm, irony, hyperbole), usually with the intent that the listener understands the underlying implicature

- Breaking maxims covertly: *violating* a maxim (e.g. lying, half truths, overcomplicating)

Pragmatics

Information Structure

- There are oftentimes multiple ways of saying what we mean…how do we choose which of these options is the best?
  - Can pick between different grammatical structures, intonation and stress patterns, words and constructions, etc.

Pragmatics

Information Structure

- There are oftentimes multiple ways of saying what we mean…how do we choose which of these options is the best?
- This in large part depends on the speaker’s knowledge of *common ground*, their *communicative goals*, and what is *desired by the listener*
  - *We can launch a bunch of small Llamas* probably doesn’t make sense to listeners that aren’t familiar with the current state of NLP (lack of common ground)
  - *Salt!* vs. *Could you please pass me the salt?* (Urgent command vs. request)
  - ***I*** *train Llamas* vs. *I* ***train*** *Llamas* vs. *I train* ***Llamas*** (focus changes depending on info requested)

Pragmatics

Rational Speech Acts (Frank and Goodman 2012)

- Bayesian model of communication

- Views communication (about a world state *w*) as a **recursive reasoning process** between a speaker *S* and a listener *L*

- Inference over the other person's mental state is very closely tied to another concept from psychology…

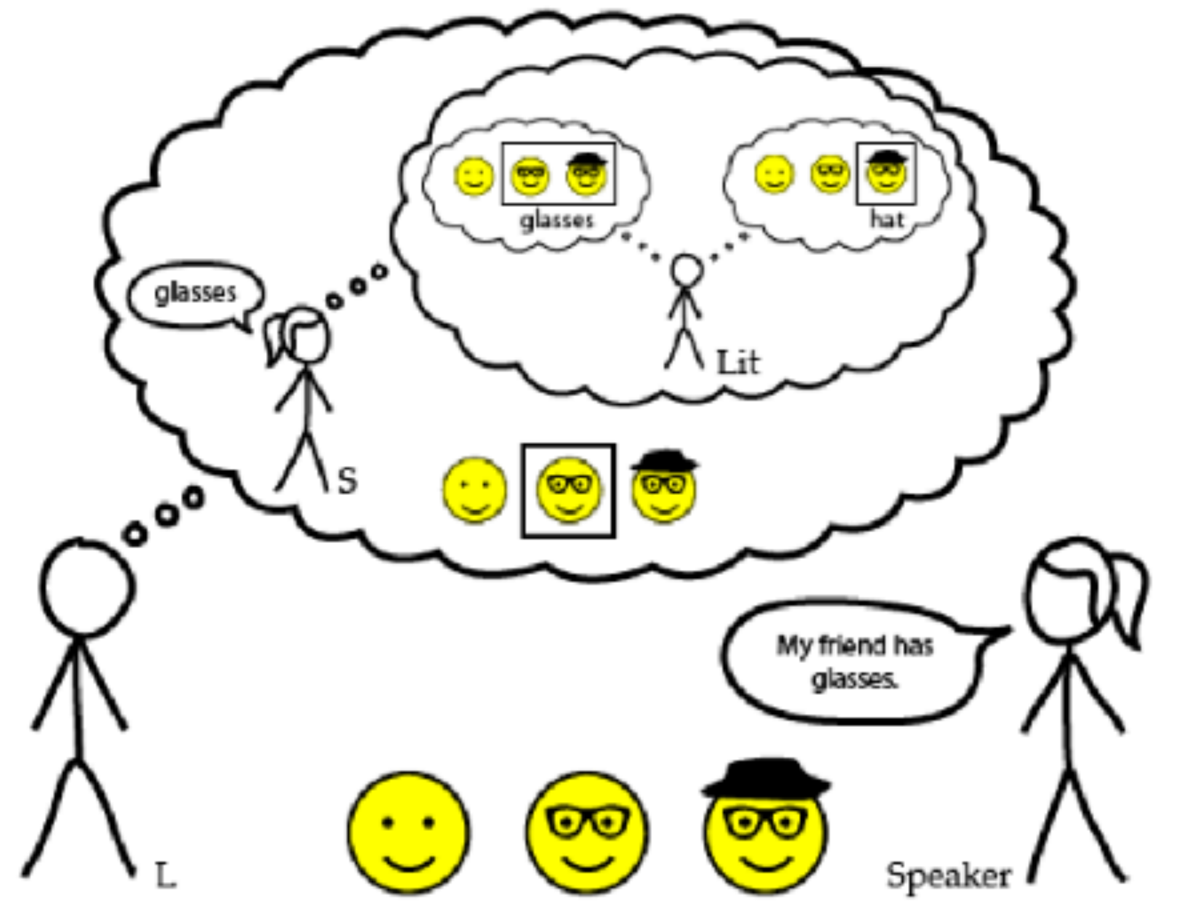


Figure 1: Application of RSA-style reasoning to a signaling game (shown by the three faces along the bottom). Agents are depicted as reasoning recursively about one another's beliefs: listener L reasons about an internal representation of a speaker S, who in turn is modeled as reasoning about a simplified literal listener, Lit. Boxes around targets in the reference game denote interpretations available to a particular agent.

From Goodman and Frank (2016)

Pragmatics

Rational Speech Acts (Frank and Goodman 2012)

For simplicity, we can consider the setting of a reference game, with a fixed set of possible world states (*w*) and utterances (*u*):

- The base case is a literal listener that selects *w* only considering *u*
- The speaker reasons about potential interpretations by *L*, and chooses *u* such that *L* is most likely to infer *w* given *u*
- The listener reasons about potential states of *w* given an utterance *u* by *S*, assuming *S* is attempting to be (maximally) informative
- Can iterate over this process however many times

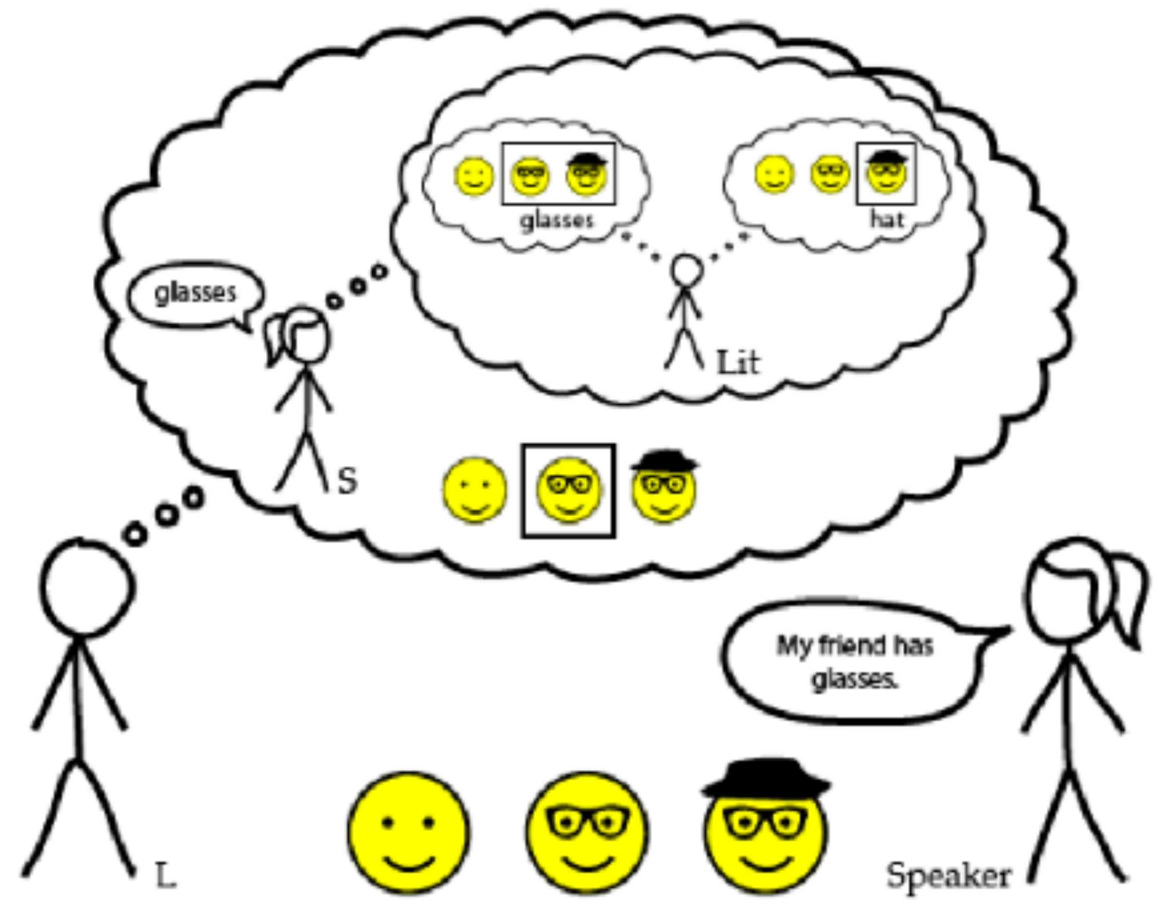


Figure 1: Application of RSA-style reasoning to a signaling game (shown by the three faces along the bottom). Agents are depicted as reasoning recursively about one another's beliefs: listener L reasons about an internal representation of a speaker S, who in turn is modeled as reasoning about a simplified literal listener, Lit. Boxes around targets in the reference game denote interpretations available to a particular agent.

From Goodman and Frank (2016)

"am I thinking what you're thinking I'm thinking that you're thinking I'm thinking…."

Interesting things I did not have time for

…and things that remain to be studied!

- Lots of overlap between questions in applied fields and current NLP, like neurolinguistics, psycholinguistics, sociolinguistics, typology, etc.
- Humans seem to be really data efficient (in terms of linguistic input)…how can we imbue that in models?
  - How do we learn to generalize from linguistic exemplars?
- How can we design fair comparisons between human and model language competence?
- How can we make NLP systems that work better for everyone, including people who speak non-standard dialects and marginalized languages?
  - Who do current NLP systems leave behind, and why?