Logistical Note

- A1 (implementing N-gram LMs) is posted. Due next Friday.
  - Lectures have covered what you need to know for parts 1–3.
- Today's lecture is self-contained. On Tuesday we will tie up loose ends on N-gram models.
  - If you want to work on the last parts of the assignment before Tuesday, you can read through Section 3.5.2 in the book.

What is Linguistics?

Nathan Schneider
ENLP | 25 January 2024

What is language?

What is language?

- Wikipedia: “Language is the ability to acquire and use complex systems of communication, particularly the human ability to do so, and a language is any specific example of such a system. The scientific study of language is called linguistics.”

What is language?

- Dictionary.com: “1. a body of words and the systems for their use common to a people who are of the same community or nation, the same geographical area, or the same cultural tradition

  “2. communication by voice in the distinctively human manner, using arbitrary sounds in conventional ways with conventional meanings; speech.”

What is language?

- Collins: “1. a system for the expression of thoughts, feelings, etc, by the use of spoken sounds or conventional symbols

“2. the faculty for the use of such systems, which is a distinguishing characteristic of man as compared with other animals”

What is language?

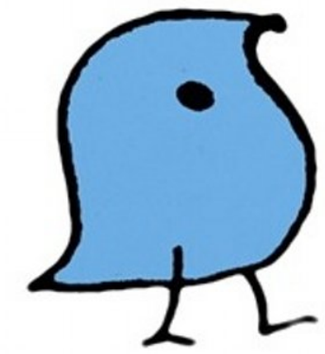
- Merriam-Webster: "**a:** the words, their pronunciation, and the methods of combining them used and understood by a community

"**b (1):** audible, articulate, meaningful sound as produced by the action of the vocal organs

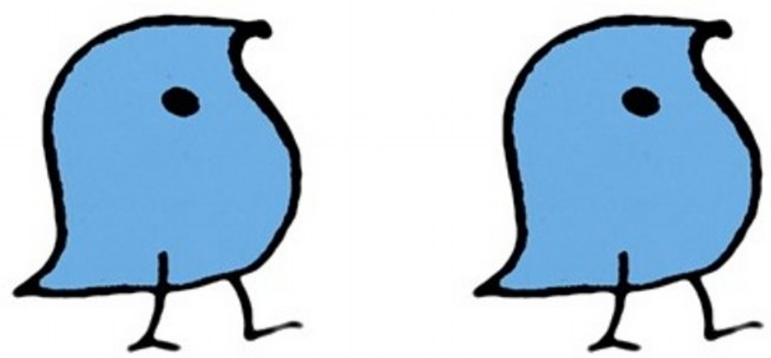
(2): a systematic means of communicating ideas or feelings by the use of conventionalized signs, sounds, gestures, or marks having understood meanings"

Productivity

- This is a **wug**:



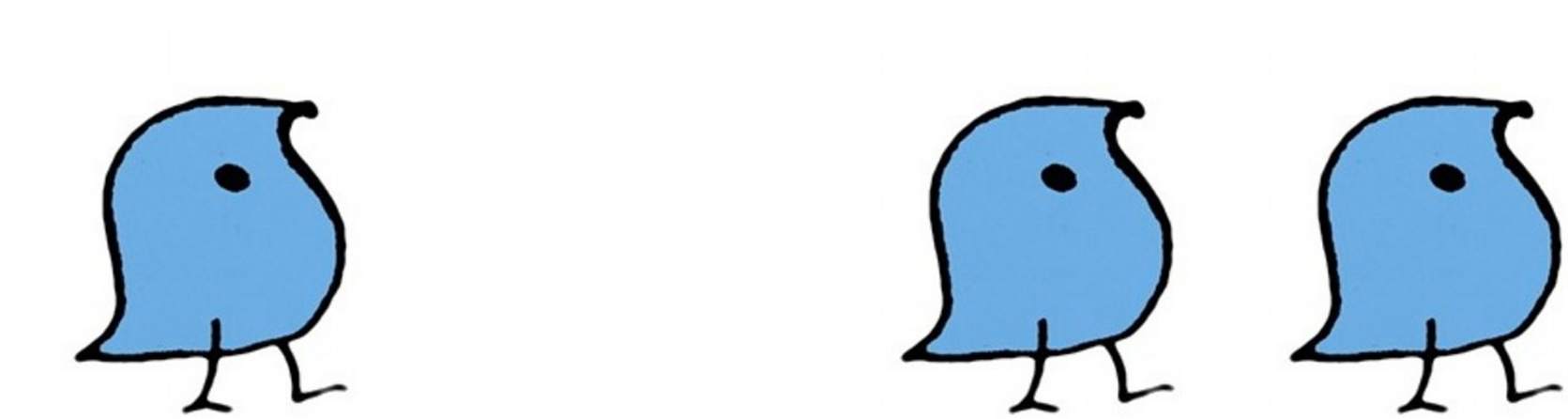
- Here there are two of them:



- There are two _____.

Productivity

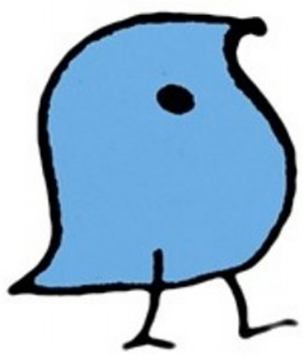
- What is happening?



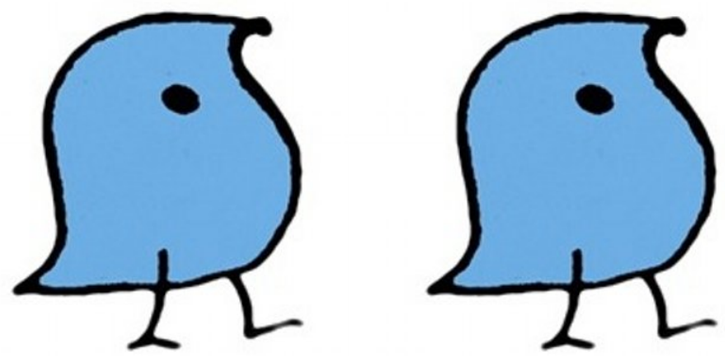
This wug is walking. _____

Productivity

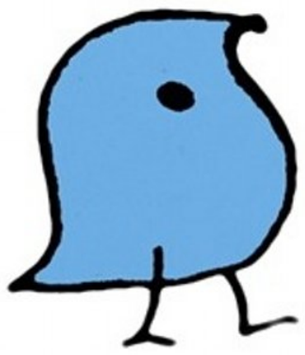
- What is happening?



This wug is walking.



These wugs are walking.



Jean Berko Gleason,
inventor of the Wug Test

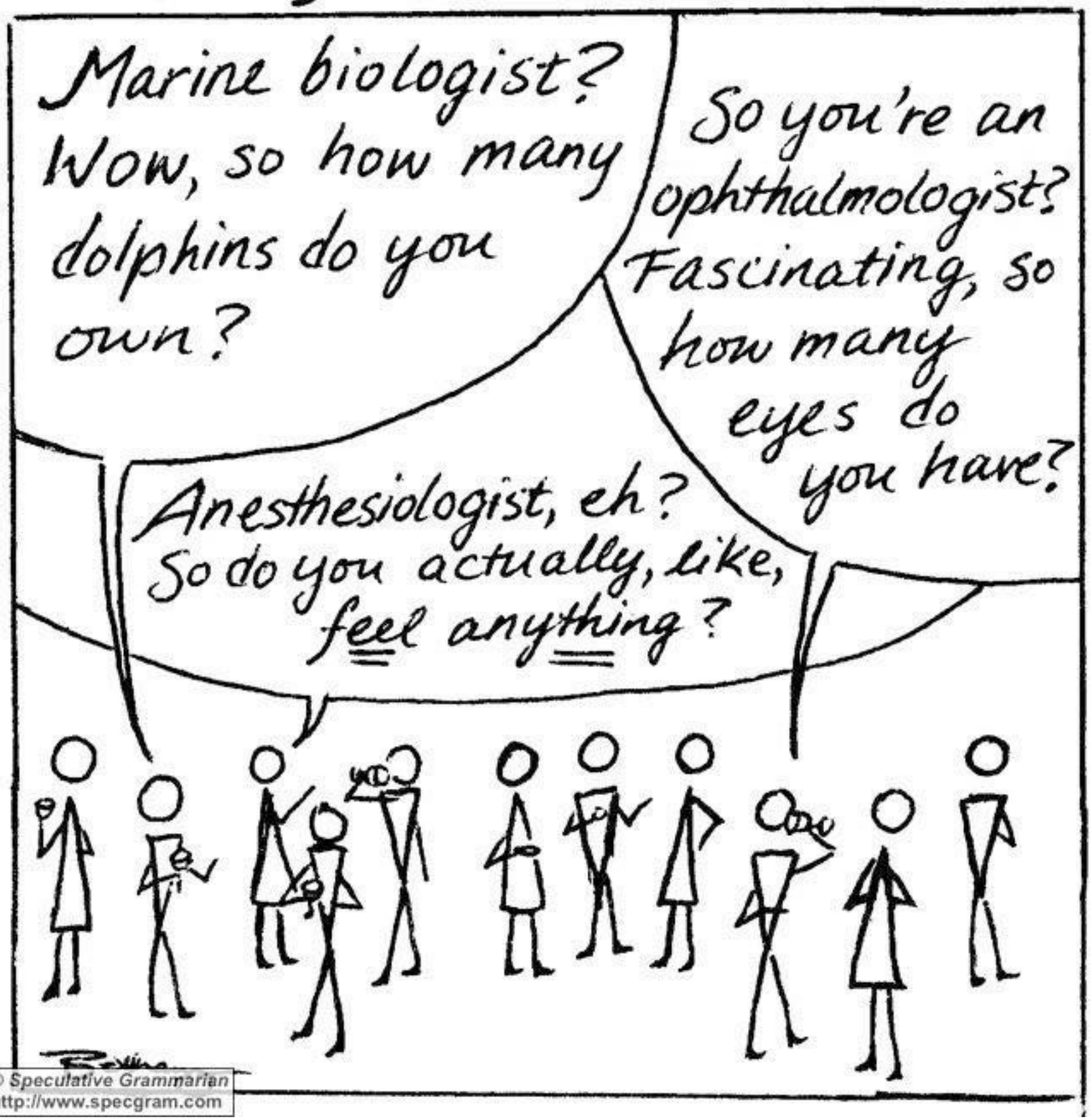
Productivity

- English-speaking school-age children can correctly infer the plural form **wugs**, though they've never heard it before!
  - "Wug" is a made-up word, a.k.a. **nonce word**, used to study linguistic ability.
  - Berko (1958): Correct production of "wugs" by 76% of preschoolers (ages 4–5), 97% of first graders (ages 5.5–7) in the study. [video]
  - Speakers can **generalize** beyond the language they've heard to produce new words and sentences that obey the grammatical patterns of the language. The ability to put together familiar pieces in new ways is called **productivity**.

Knowledge of/about language

- Every linguist gets questions like:
  - *“How many languages do you speak?”*
  - *“Which is correct in this sentence: ‘who’ or ‘whom’?”*
- These reflect misunderstandings of what linguistics is.

The linguists strike back



Knowledge of/about language

- Studying a language does not necessarily require fluency in it
  - Though it requires data, ideally from a native speaker
- Speaking a language doesn't entail understanding how it works!
  - Linguistics = studying **what speakers know, but don't know they know**. Uncovering the **implicit** knowledge behind a skill.
  - You learned your native language primarily through exposure, not being taught the rules of grammar!

Knowledge of/about language

- **Speech** is primary, **writing** is a technology
  - Most languages of the world are never or rarely written down
  - Written language can be more conservative, stylistically fixed

- Mosts linguists are **descriptivists**
  - They study what language *is* according to the practice of a speech community, not what it *should be* according to some socially accepted authority or stereotype (**prescriptivist**).
  - In linguistics, **grammar rules** describe the patterns of how people talk.

Knowledge of/about language

- Forms of evidence
  - "Thought data"/native speaker intuitions
    - *This test allows to determine whether the result is statistically significant.*
    - *Who cares about how it looks like when it tastes damn good?*
  - Use data (corpora)
  - Lab data

Sentences + glosses

(8) a. ***Kto-to*** *(/****kto-nibud'***)** *postučal v dver'.*
"Someone (/*anyone) knocked at the door."
b. *Esli* ***čto-nibud'*** *slučitsja, ja pridu srazu.*
"If anything happens, I'll come immediately."

Wolof (Niger-Congo; Northern Atlantic) [Mark 1:29]

(1)	*...génn*	*na-ñu*	***ci***	*jàngu*	*bi,*	*ñu...*	*dem*
	...exit	PERF-3SG	**PP.PROX**	church	the,	3PL	go

	ci	*kër*	*Simoŋ*	*ak*	*Andare.*
	PP.PROX	house	Simon	and	Andrew

'...when they were come **out of** the synagogue, they entered **into** the house of Simon and Andrew.'

Some language myths

- Kids today are ruining the previously pure form of our language.
  - Commentary of this nature goes back over the centuries. In fact, language is constantly evolving. It is an organic system, which means it complex and “messy” but adapts to the needs of speakers.

Some language myths

- When <low-prestige group members> talk they are being lazy/using bad grammar.
  - Relative to Standard American English, dialects like African-American English have some differences in **vocabulary** and **grammar** (including pronunciation and syntax).
  - Scientifically, is nothing better or worse about any dialect; there is just social prestige and acceptance.

Some language myths

- It's easy to define the boundaries of a language.
  - Roughly speaking, if two dialects are **mutually intelligible**, they are said to be from the same language. In practice, there can be a lot of gray area—e.g., Arabic has many dialects, some of which are quite different from each other.
  - **Geopolitical considerations** often interfere as well: colloquially we call Chinese a language, but Mandarin and Cantonese are not mutually intelligible. Conversely, by linguistic criteria, Hindi and Urdu are considered dialects of the same language.

Some language myths

- Sign language is less systematic than spoken language.
  - There are actually many sign languages: **American Sign Language** and **British Sign Language** are quite different, for example. This is because all languages develop subject to a community of speakers.
  - Sign languages also have grammar, with patterns and structure in how hands are shaped, how they are positioned and moved, facial expressions, etc.

Some language myths

- People are hereditarily predisposed to have an easier time learning some languages.
  - Fact: Children are capable of natively acquiring any language given sufficient exposure at the right age. Inability to do so is attributed to a mental or communicative deficit or disability.

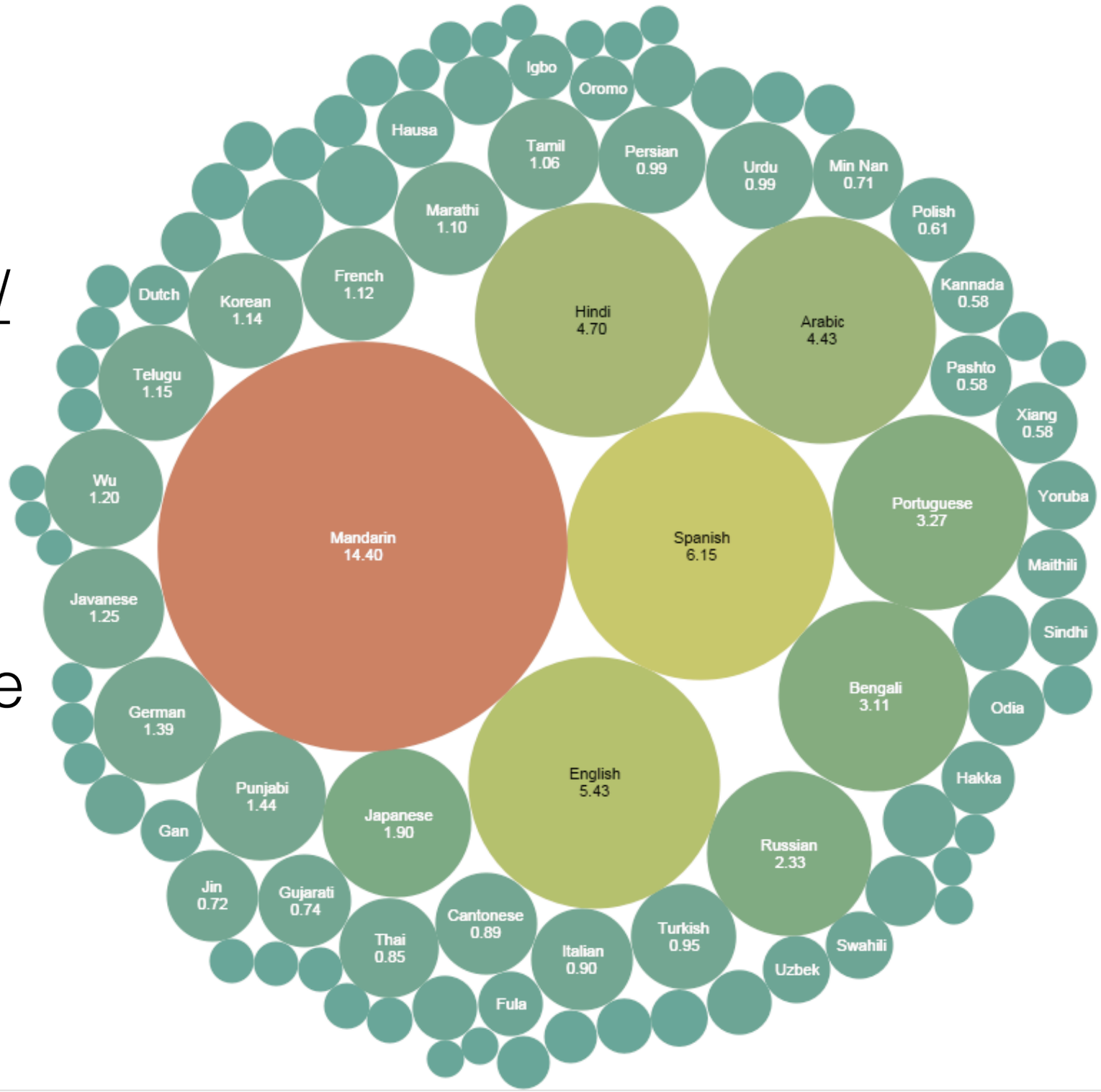
Some language myths

- Most languages have millions of speakers.
  - Fact: There are approximately 6000–7000 languages spoken today. About **a third** have small native speaker populations and are in danger of extinction.

    http://www.ethnologue.com/world

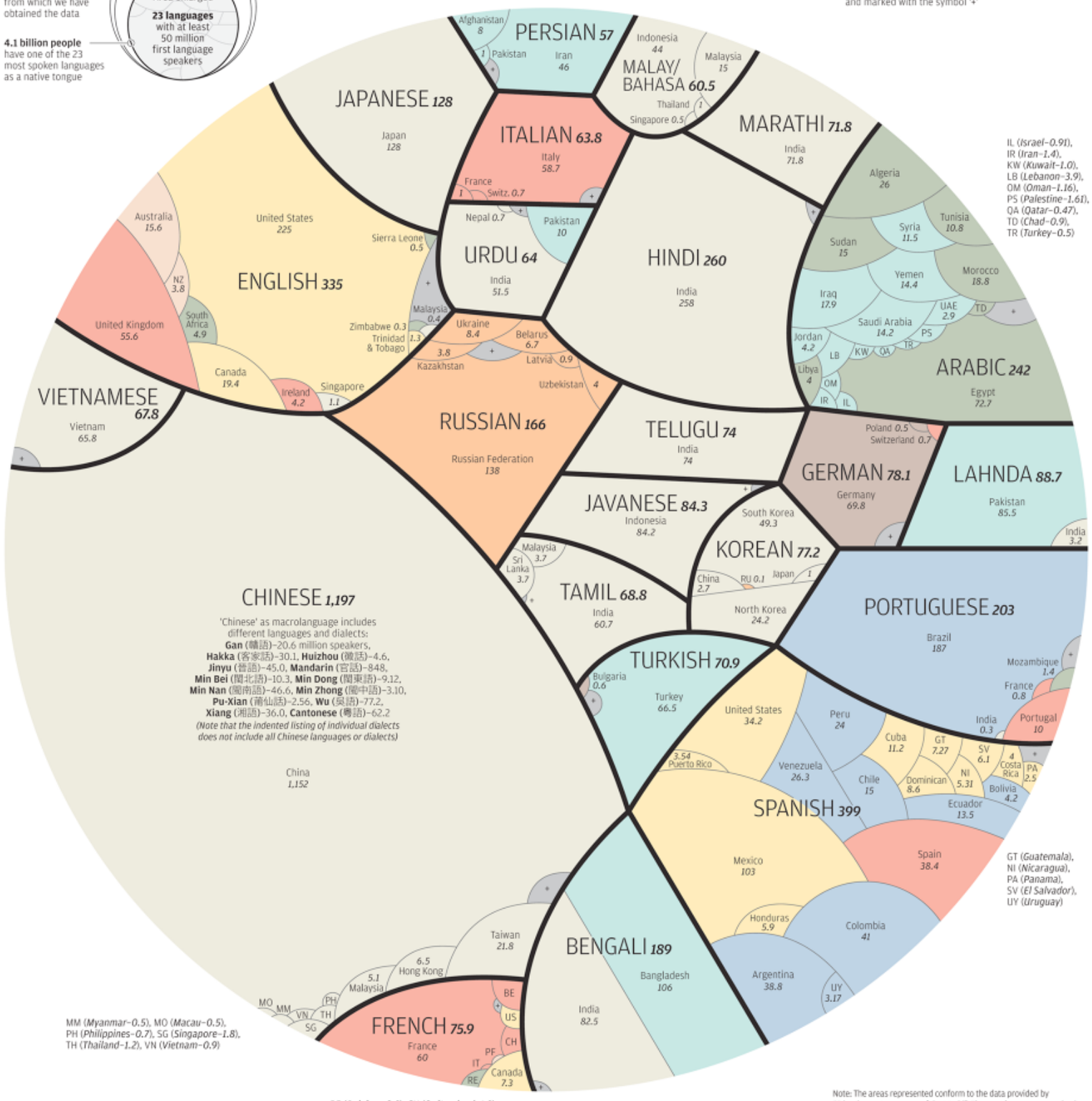
Language populations are Zipfian

According to www.ethnologue.com/statistics/size, only **5.6%** of languages have ≥1M native speakers—but these account for **94%** of the world's population.



https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

23 languages with ⩾50M native speakers



Administrivia

- Office hour times
  - **Nathan:** Mondays 2pm in STM 315H
  - **TA:** Thursdays 2:45-3:45pm?

Areas of study

Structure / Grammar		Language in the world	Methods/ Applications
Form	**Function**		
Phonetics	Semantics	Sociolinguistics / within-lang. variation	**Computational, Corpus**
Phonology	Pragmatics	Typology / between-lang. variation	Psycholinguistics, Neurolinguistics
Orthography	Discourse	Language acquisition (L1, L2)	Fieldwork, documentation
Morphology		Language change / historical	“Applied Linguistics”: teaching, policy, forensics, …
Syntax		Linguistic anthropology	

Areas of study

Structure / Form

Phonetics

Phonology

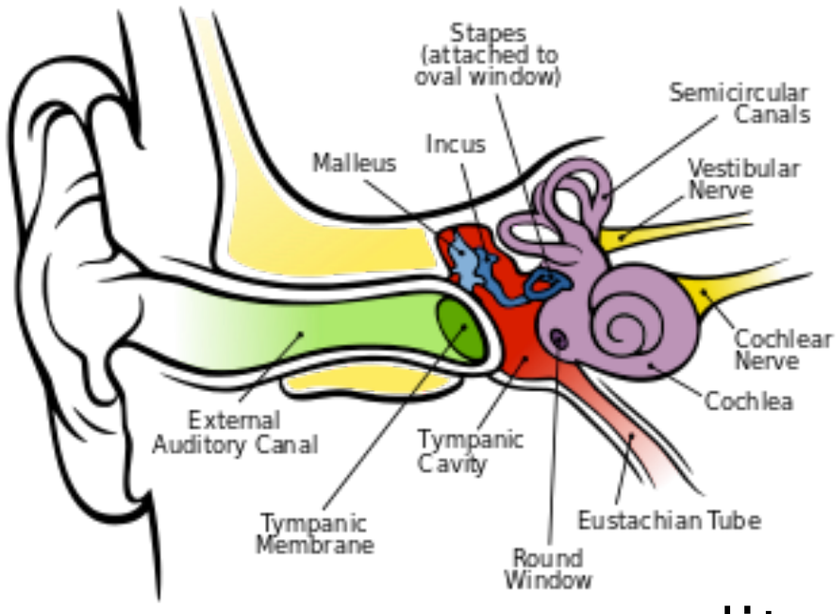
Orthography

Morphology

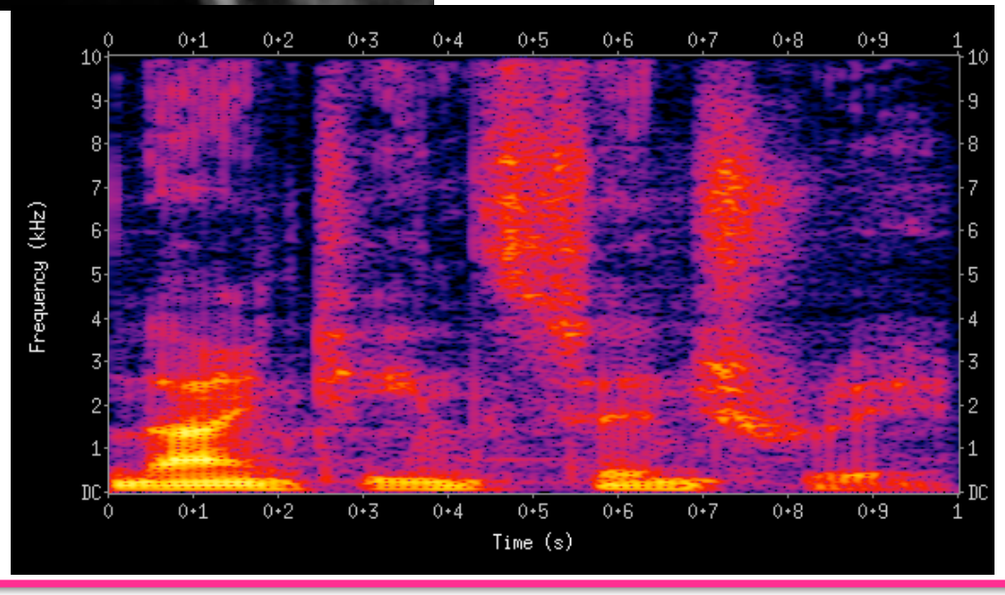
Syntax

Phonetics: the **sounds** of language

articulatory



auditory



acoustic

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

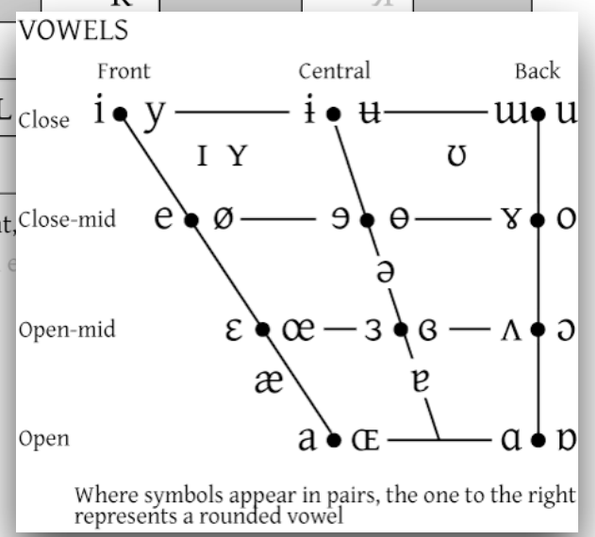
Syntax

Phonetics: the **sounds** of language

the international phonetic alphabet (2005)

consonants (pulmonic)	LABIAL		CORONAL				DORSAL				RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ		ɲ	ŋ	ɴ			
Plosive	p b			t d		ʈ ɖ		c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ɕ ʑ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ		j	ɰ				
Tap, flap		ⱱ		ɾ		ɽ							
Trill	ʙ			r						ʀ		ʡ	
Lateral fricative				ɬ ɮ									
Lateral approximant				l		ɭ		ʎ		ʟ			
Lateral flap				ɺ									

Where symbols appear in pairs, the one to the right represents a modally voiced consonant.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial e



International Phonetic Alphabet (IPA)

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Phonology: how sounds work in **particular languages**

blick sounds like a possible word of English, but not **bnick*

Why the first sound of *pit* is different from the second sound of *spit*

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Orthography: how a language is **written down**

th at the beginning of an English word corresponds to a single sound (/θ/ or /ð/)

Instead of alphabets (≈1 symbol per sound), some languages are written with **abjads** (unwritten vowels), **abugidas**, **syllabaries**, or **logograms**. The character-set of a language is called a **script**.

漢 汉 ኢትዮጵያ አፍሪቃ ውስጥ ናት 한국어 조선말

字 字 Minh là giáo viên. العربية

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Morphology: how **words** are formed

Inflection: systematic alternation in gender, number, case, tense, person, etc.
horse/horses, man/men;
decide/decides/decided, *eat/eats/ate/eaten*

Derivation or **compounding:** affects the meaning of the word more fundamentally

Why the negation of *advisable* is ***in****advisable*, but the negation of *possible* is ***im****possible*

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Morphology: how **words** are formed

A **morpheme** is a minimal unit of meaning:
in- (prefix), *advise* (stem), *-able* (suffix)

Some morphemes combine in predictable (rule-governed) patterns. Such a pattern is said to be **productive** if it can give rise to new words. Other patterns only apply to specific words, e.g., *man* (sg)/*men* (pl).

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Morphology: how **words** are formed

English is **morphologically impoverished** compared to most languages (except Chinese, which has even less morphology).

German has some famously long **compounds**:
rindfleischetikettierungsüberwachungsaufgabenübertragungs-gesetz
'the law for the delegation of monitoring beef labeling'

In Turkish, an **agglutinative** language, a "word" can be an entire sentence:
İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz
'You were (evidentially) one of those who we may not be able to convert to an Istanbulite'

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Syntax: how **sentences** are formed from words

Why in English we don't say **I happy*—
we say *I am happy*: with a **copula** (be-verb)

How questions are formed:
Why are you crying?
**Why you cry?*
**You are crying why?*

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Syntax: how **sentences** are formed from words

Linguistic categories help us to describe syntactic patterns.

Part of speech (POS): the grammatical category of a word
noun, pronoun, verb, adjective, adverb, determiner, preposition, …

Grammatical relation: how a word functions relative to other words in the sentence
subject, predicate, object, modifier, …

Phrasal category:
noun phrase, prepositional phrase, clause

Areas of study

Levels of structure

Structure /

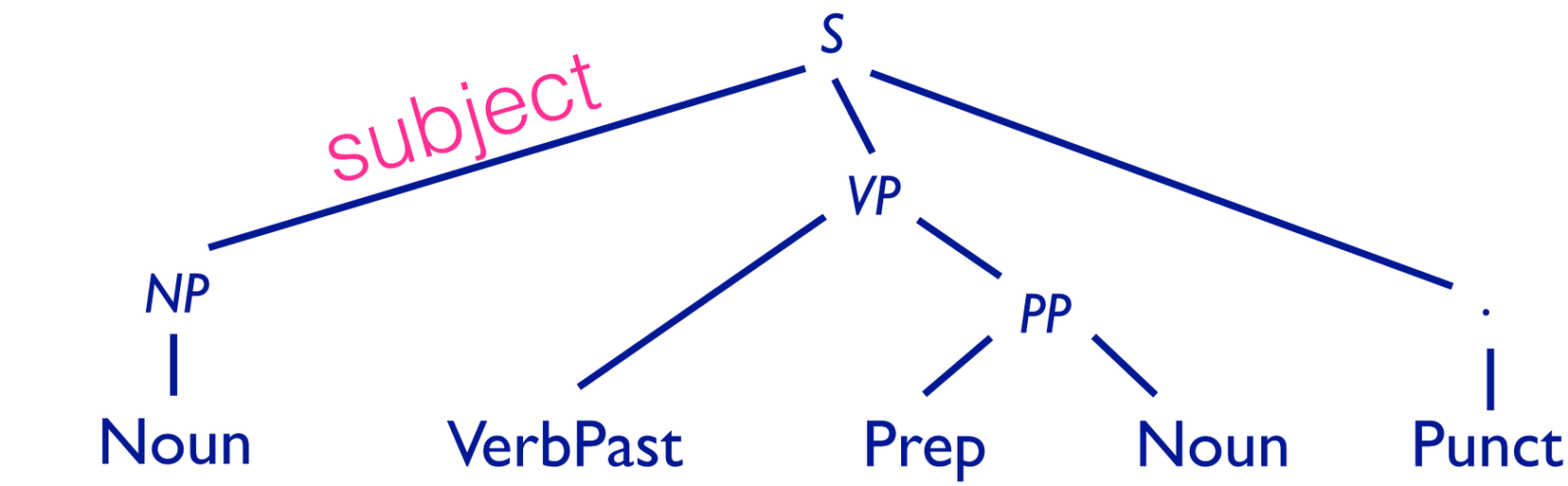
Syntax: Constituents

Syntax: Part of Speech

Words

Morphology

Characters



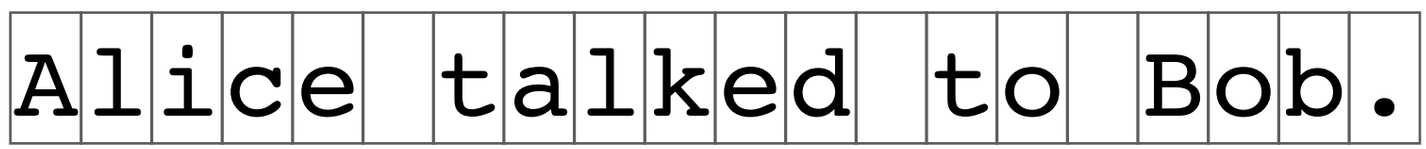


figure from Brendan O'Connor

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Syntax vs. Morphology: a tradeoff

English is called an **analytic** language because it mainly relies on word order/syntax to indicate sentence structure:

The cat ate the fish ≠ The fish ate the cat

Synthetic languages make heavier use of morphology to indicate how words function in a sentence.

synthetic ⟷ **analytic**

Cree Turkish Japanese Finnish German Russian French Spanish English Chinese

Areas of study

Structure / Form

Phonetics

Phonology

Orthography

Morphology

Syntax

Syntax vs. Morphology: a tradeoff

A **case marker** signals whether a verb's argument is the subject, object, etc.
Remnants of case in English pronouns:
She loves him / He loves her

English is strict about word order (*Him loves she), but synthetic languages with case are more flexible.

Areas of study

Structure / Grammar	
Form	**Function**
Phonetics	Semantics
Phonology	Pragmatics
Orthography	Discourse
Morphology	
Syntax	

Semantics: the **meaning** of a word or sentence

Areas of study

Structure / Grammar	
Form	**Function**
Phonetics	Semantics
Phonology	Pragmatics

Pragmatics: how meaning can depend on **conversational context**

"Can you pass the salt?" is usually a request, not a literal question

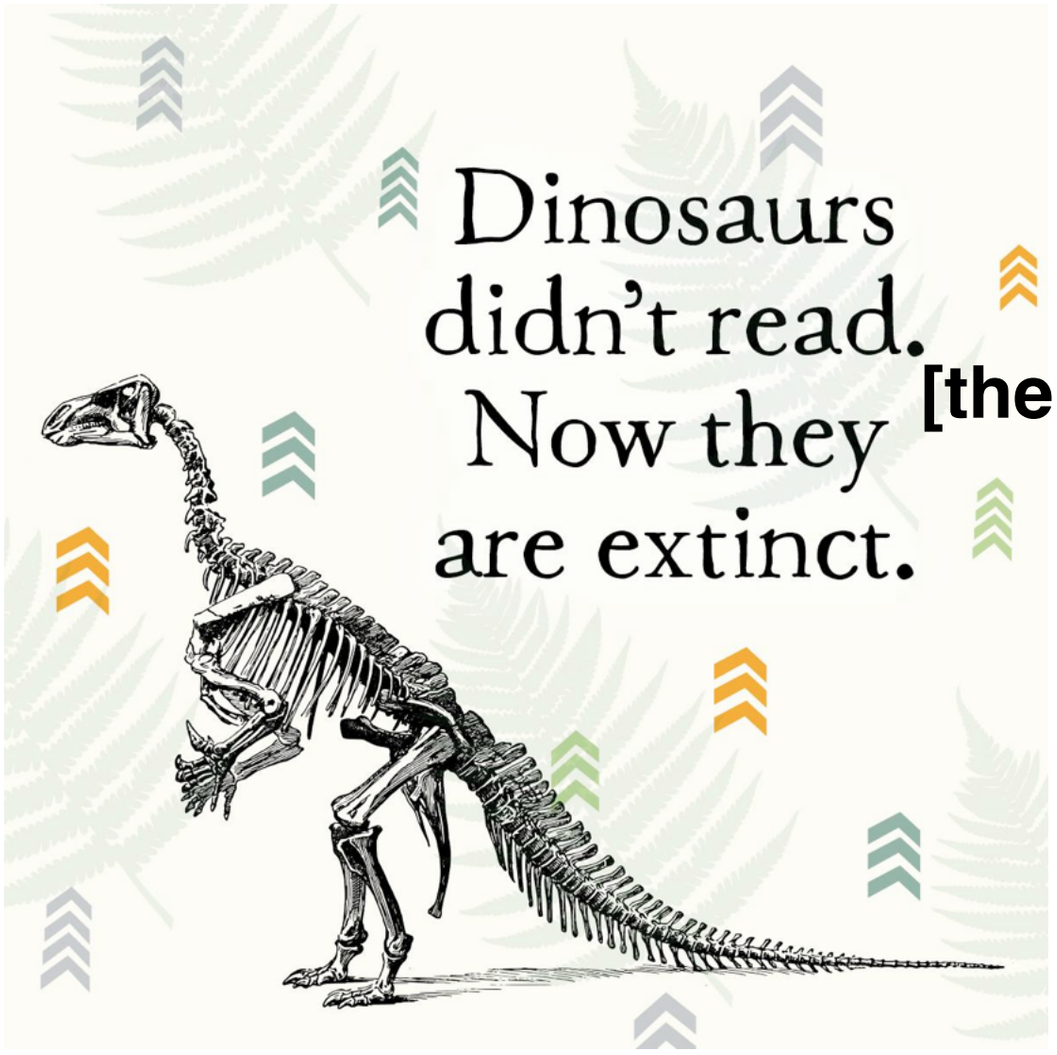
Ort

Mc

Areas of study

Structure / Grammar	
Form	**Function**
Phonetics	Semantics
Phonology	Pragmatics
Orthography	Discourse
Morphology	
Syntax	

Discourse: how sentences fit together in **texts** or **conversations**

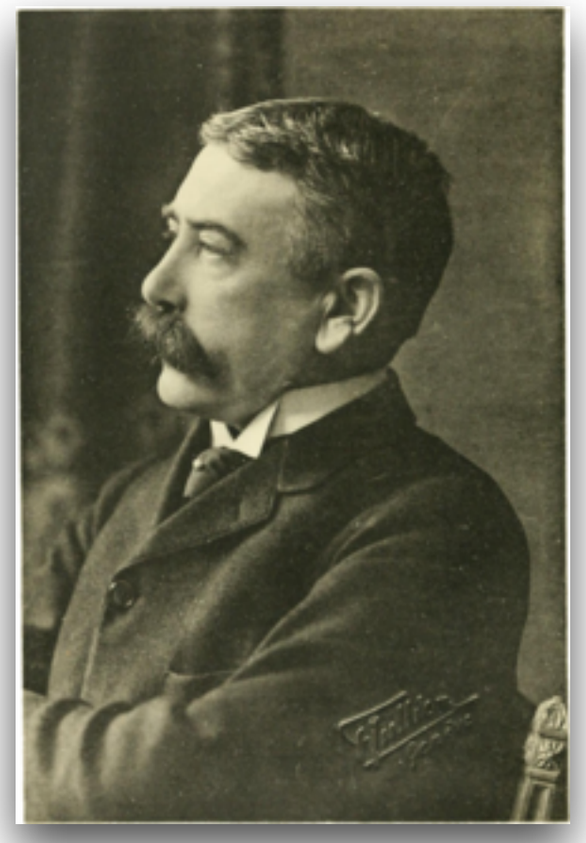


[therefore]

A sampling of figures and ideas in linguistics…

Early Linguists

Pāṇini (4th century BCE): systematic study of Sanskrit grammar; “father of linguistics”

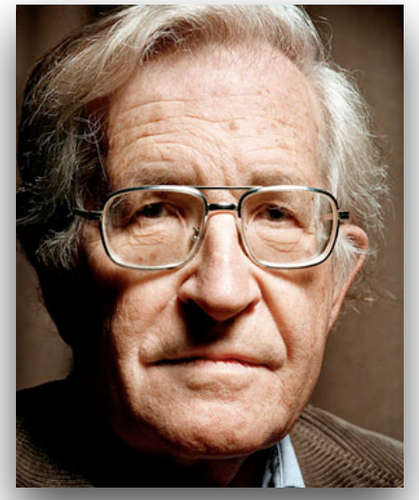


Saussure (1870s–1910s): “arbitrariness of the sign”—there is nothing intrinsically doglike about the sound of the word *dog* (or *chien*, *gǒu*, or *skýlos*).

Bloomfield, Wittgenstein, ...

Noam Chomsky

1950s – present



- Language is a complex cognitive system, not a set of simple reactive behaviors. Crucially, we can produce/comprehend utterances we have never heard before.

- "Colorless green ideas sleep furiously"
  - Claim: Perfectly *grammatical*, though *meaningless*.
  - Grammatical **competence** is the ability to decide whether a sentence has a valid *form* according to one's intuitions as a native speaker. Thus, theories that explain formal patterns should be the focus of linguistics.

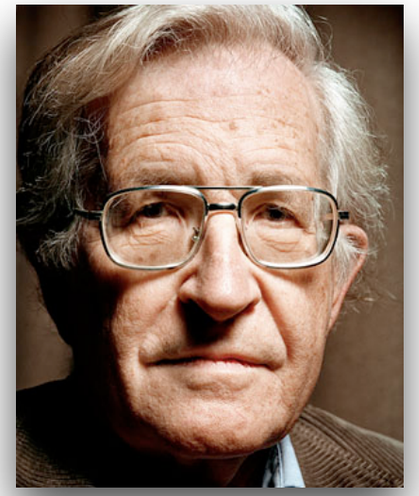
- A finite symbolic system can characterize an infinite set of strings. Different classes of formal languages require different levels of complexity to describe and parse (**Chomsky Hierarchy**).
  - Regular languages (described by regular expressions) are the simplest, with a finite number of states.
  - Context-free, context-sensitive, ...

- **Recursion** is the key property that distinguishes human language from animal language.

- "Poverty of the Stimulus" claim: It is impossible that children are exposed to enough language input that would allow them to learn all the intricacies of grammar. Much of it must be **innate** and, because any child can learn any language with the right exposure, **universal**.

Noam Chomsky

1950s – present



- Language is a complex cognitive system, not a set of simple reactive behaviors. Crucially, we can produce/comprehend utterances we have never heard before.

- "Colorless green ideas sleep furiously"
  - Claim: Perfectly *grammatical*, though *meaningless*.
  - Grammatical **competence** is the ability to decide whether a sentence has a valid *form* according to one's intuitions as a native speaker. Thus, theories that explain formal patterns should be the focus of linguistics.

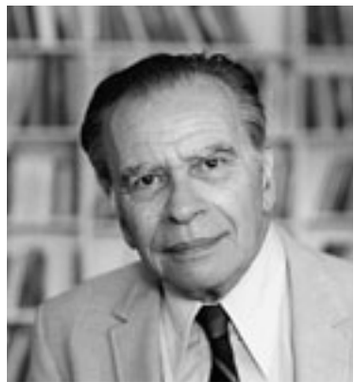
✦ Over the decades, Chomsky exerted tremendous influence on the field of linguistics from MIT. Today the "formalist" view of grammar is the dominant one in most U.S. linguistics departments. But many aspects of Chomsky's theories remain controversial.

- A finite-state grammar is a finite set of rules. Different kinds of formal languages require different levels of complexity to describe and parse (**Chomsky Hierarchy**).
  - Regular grammar (described by regular expressions) are the simplest, with a finite number of states.
  - Context-free, context-sensitive, …
- **Recursion** is the key property that distinguishes human language from animal language.

- "Poverty of the Stimulus" claim: It is impossible that children are exposed to enough language input that would allow them to learn all the intricacies of grammar. Much of it must be **innate** and, because any child can learn any language with the right exposure, **universal**.

Functionalists

A counterweight to the formalist camp anchored by Chomsky, functionalists argue that language is primarily a tool for **communicating meaning** and for **social interaction**.



Joseph Greenberg/Typologists: We can compare/categorize the languages of the world and discover **universals**.

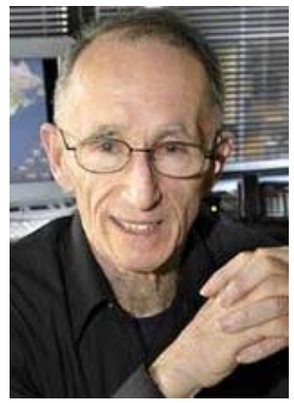
Joan Bybee: Frequency matters: Words and patterns that are frequent behave differently from those that are infrequent.

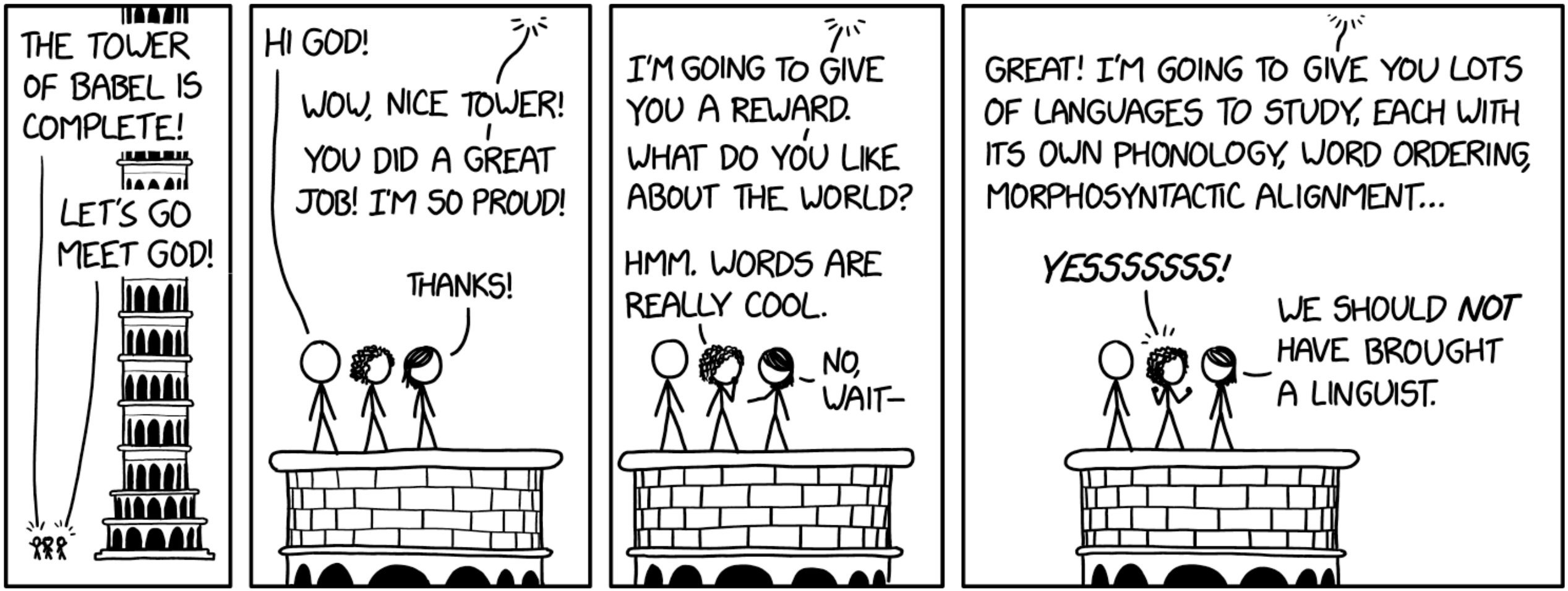
Benjamin Lee Whorf (1920s–1940s), **Lera Boroditsky**: Different languages influence how we perceive the world.

George Lakoff/Cognitive Linguists: Language is deeply connected to nonlinguistic cognition. Meaning is embodied and involves metaphor.

Leanne Hinton/Documentary linguists: We can help indigenous communities rescue their languages from extinction.

William Labov/Sociolinguists: We can trace how individuals and groups express their identity and build relationships using language.





https://xkcd.com/2421/

Language Spotlight Lighting Presentations

- As a practical measure, most of the lectures will focus on English. But other languages raise other challenges for NLP/language technologies.

- From now on, we'll start class with a 5-minute presentation from one of you that describes a different language. This will showcase the diversity of the world's languages.

Language Spotlight Lighting Presentations

- Ground rules:
  - 1 presentation per enrolled student. Sign up for a slot after class today. Indicate your choice of language at least a week in advance.
  - You must choose a language that (a) is not English and (b) has not been presented yet.
  - The style of presentation is up to you: you may use slides, handouts, multimedia, etc.
  - 5 minutes. PRACTICE WITH A TIMER. We WILL cut you off if you go over.

Language Spotlight Lighting Presentations

- Your presentation should cover:

  1. **Typological overview:** how many speakers, where spoken, what language family/related languages; synthetic vs. analytic, SVO/VSO/etc., what kinds of inflectional morphology on nouns and verbs, what kinds of agreement

     - http://ethnologue.com/, http://wals.info/

  2. A couple of **interesting phenomena** in the language (probably: different from English). Give examples (with IPA or romanized transliteration if a non-roman script). E.g., German compounds.

  3. What about this language would be especially **difficult** for NLP/language technologies?

Homework

- Sign up for a slot (see assignment on Canvas)
-