

Multilingual NLP



EECS 183/283a: Natural Language Processing

Adapted from slides by Graham Neubig, Aditi Chaudhary, and Xinyi Wang

Multilingual NLP



- For any task we expect out of language technologies, they should work for any language
 - Question answering, information retrieval, summarization
 - Dialogue systems and chatbots
 - Language generation
- Language technologies can also support cross-language communication
 - Machine translation
 - Language learning

Challenge: Data Modality



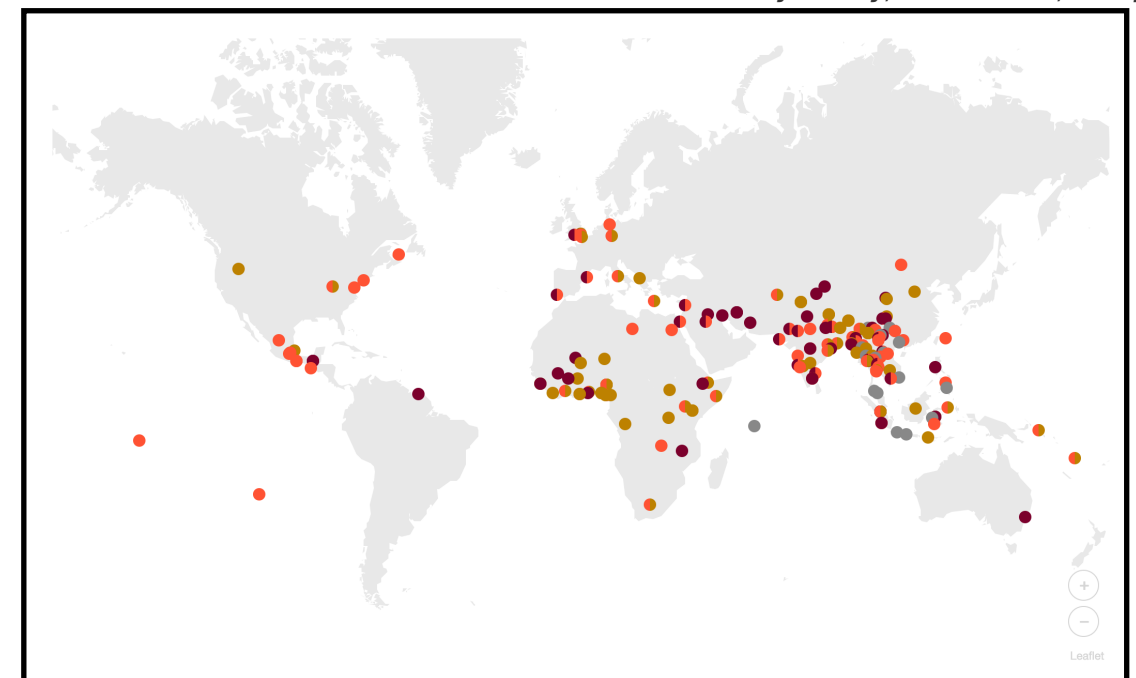
- Not all languages have writing systems, with many languages where:
 - Only audio recording is possible or available
 - Developing a consistent writing system is difficult, and finding or creating written records is extremely time-consuming relative to how much the language is used
- Some languages are rarely written or have inconsistent uses of writing systems, and only used conversationally
- Some writing systems are not-yet digitized, or all documents are handwritten

	a	e	i	o
ch				
k				
kw				
m				
n				
p				
s				
t				
tl				
tz				
w				
x				

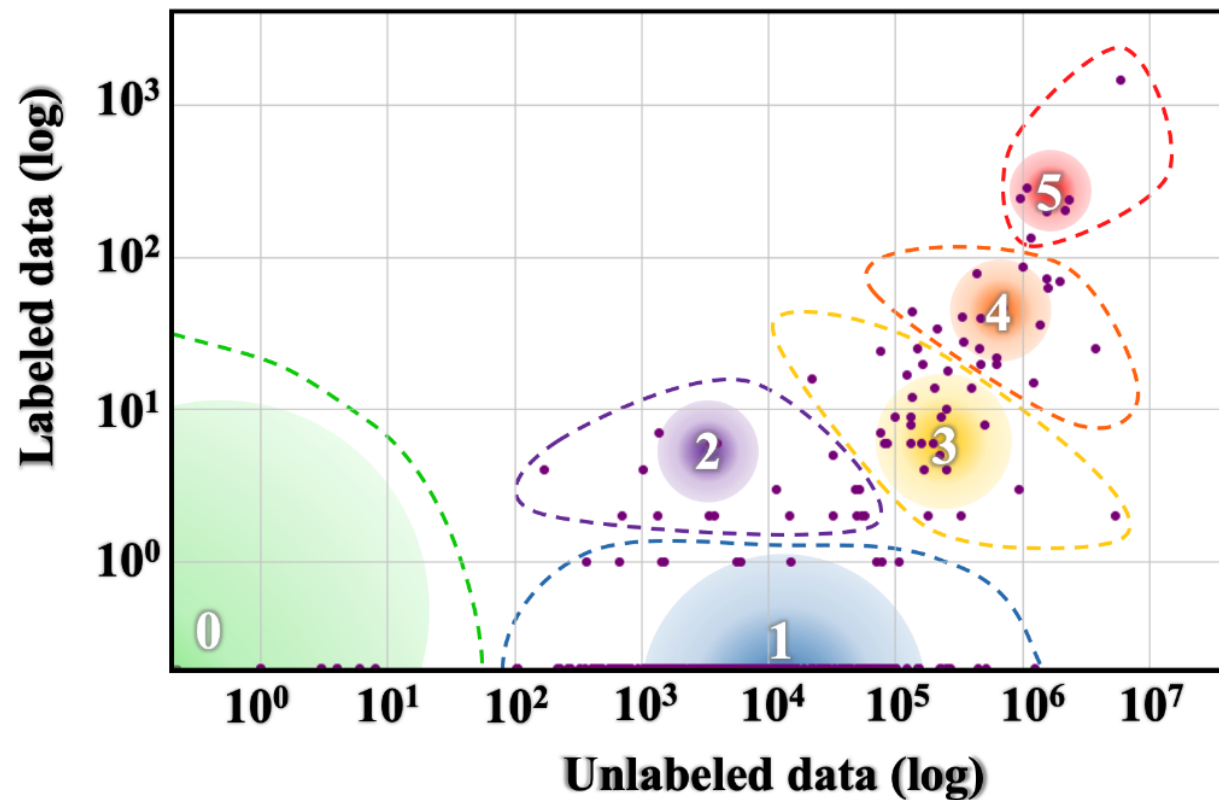
Aztec syllabary, PauloCalvo, Wikipedia

$B_a B_a^{z\sim}$ $\ddot{V}\ddot{V}\dot{a}$ $3^\perp$ $[] \sqrt{C}^\dagger \sqrt{C}^\vee$ $\} Y^\circ$ $\sqrt{G}_\Lambda <^\vee <$
 $\bar{B}_a \sqrt{B}_\Lambda^\omega$ $G^\perp$ $B_\Lambda^! B_\Lambda^\dagger$ $D \dot{A}^\otimes$ $\underline{B}_D B_D^\perp$
 $G^\triangleright$ $\wedge \dot{5}^x$ $[] \sqrt{C}^\dagger \sqrt{C}^\vee$ $X_1 X_1^\dagger$ $B_T V_D^\vee$
 $\bar{B}_a L^\#$ $X_1 X_1^\dagger$

Stokoe notation of ASL



Challenge: Data Scarcity



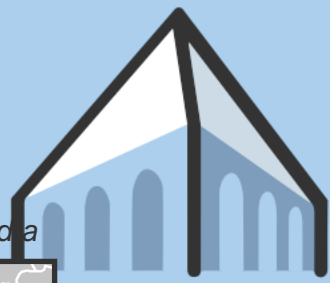
- For most languages, very little data is available for training or evaluating language technologies
- There's even less labeled or parallel data!

1.2B total speakers, virtually no available data for building language technologies

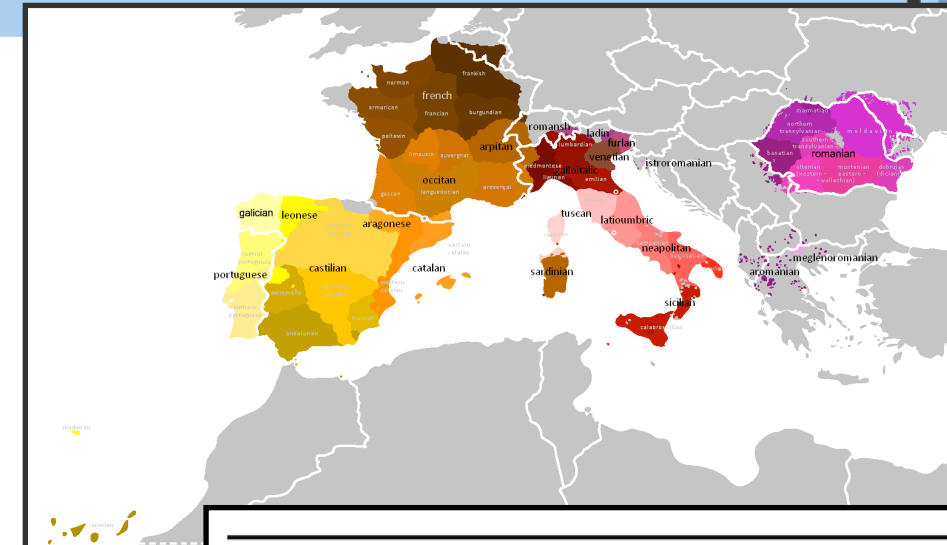
Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Challenge: Dialectical Variation

Romance languages; Simighejan, Wikipedia



- The same languages can vary significantly depending on who is speaking it where
 - Regional differences
 - Formality differences
- What can vary? Any linguistic features!
- Speakers often mix different dialects with one another in the same conversation (code-mixing)



Colloquial Indonesian	Translation
Ada yang ngetag foto lawas di FB	Someone is tagging old photos in FB
Quotenya Andrew Ng ini relevan banget	This Andrew Ng quote is very relevant
Bilo kita pergi main lagi?	When will we go play again?
Ini teh aksara jawa kenapa susah banget ?	Why is this Javanese script very difficult?

Table 5: Colloquial Indonesian code-mixing examples from social media. Color code: **English**, **Betawinese**, **Javanese**, **Minangkabau**, **Sundanese**, Indonesian.

English	Tagalog	Taglish
Could you explain it to me?	Maaaring ipaunawa mo sa akin.	Maaaring i-explain mo sa akin.
Could you shed light on it for me?	Pakipaliwanag mo sa akin.	Paki- explain mo sa akin.
Have you finished your homework ?	Natapos mo na ba ang iyong takdang-aralín ?	Finished /Natapos na ba 'yung homework mo?
Please call the driver .	Pakitawag ang tsuper .	Pakitawag ang driver .

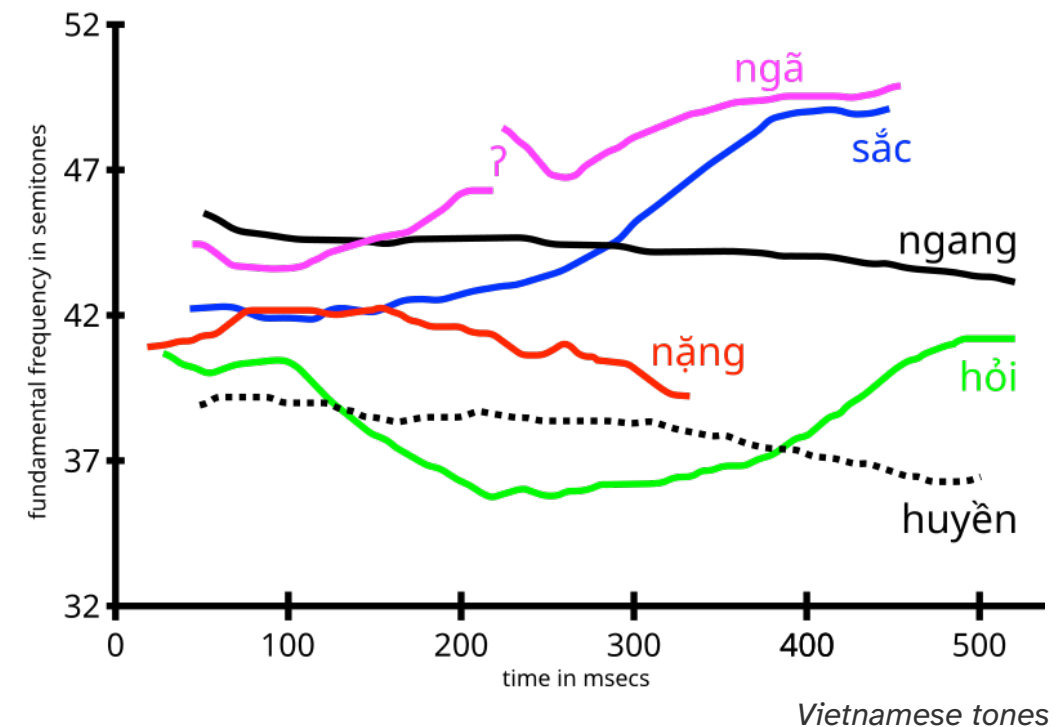
Challenge: Speech Systems



In some languages, syllables may be distinguished not only by which specific vowel is used, but also by:

- Syllable length
- Pitch contour
- Pitch height
- Phonation (breathy, creaky, etc.)

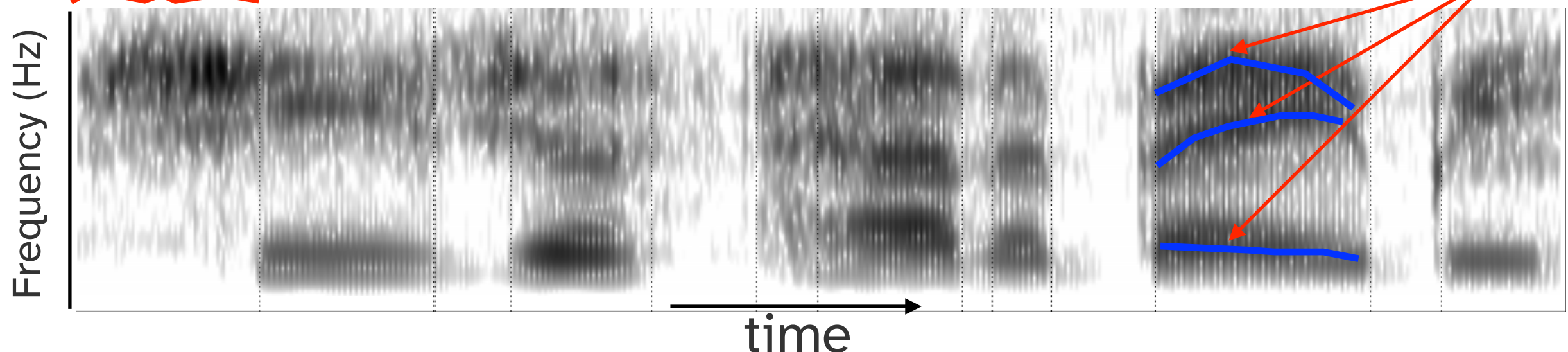
Differences between vowels may be more subtle



Unvoiced

Speech Spectrogram

Formants



Challenge: Morphology



- Synthetic languages denote syntactic relationships between words using inflection (modification of a word, e.g., conjugating a word) or agglutination (adding particles to a word)
- Issues with tokenization

English	Turkish	Formation			
I liked	sevdim	sev- "like"	-di (past tense)	-m (first person singular)	
I did not like	sevmedim	sev- "like"	-me "not"	-di (past tense)	-m (first person singular)
I like	severim	sev- "like"	-er (present tense)	-im (first person singular)	
I do not like	sevmem	sev- "like"	-me (negative present tense)	-m (first person singular)	

• Georgian

გადმოგვახტუნებინებდნენო (gadmogvakhṭunebinebdneno)

გადმო- გვ- ა- ხტუნ -ებ -ინ -ებ -დ -ნენ -ო

gadmo gv a khṭun eb in eb d nen o

"They said that they would be forced by them [the others] to make someone to jump over in this direction." (The word describes the whole sentence that incorporates tense, subject, object, relation between them, direction of the action, conditional and causative markers etc.)

Basic verb	Reduplication	Triplication
kɔul 'to sing'	kɔukɔul 'singing'	kɔukɔukɔul 'still singing'
mejɾ 'to sleep'	mejmejɾ 'sleeping'	mejmejmejɾ 'still sleeping'

Reduplication in Pingelapese

English

GPT-3.5 & GPT-4 GPT-3 (Legacy)

OpenAI's large language models (sometimes referred to as GPT's) process text using tokens, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

Clear Show example

Tokens 58 Characters 301

OpenAI's large language models (sometimes referred to as GPT's) process text using tokens, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

Text Token IDs

Burmese/Myanmar (Google Translated)

GPT-3.5 & GPT-4 GPT-3 (Legacy)

OpenAI ၏ကြီးမားသောဘာသာစကားမော်ဒယ်များ (တစ်ခါတစ်ရံ GPT များဟုလည်းခေါ်သည်) စာသားအစုအဝေးတွင်တွေ့ရလေ့ရှိသောအက္ခရာများဖြစ်သည့် တိုက်ကပ်များကိုအသုံးပြု၍ စာသားလုပ်ဆောင်သည်။ မော်ဒယ်များသည် ဤတိုက်ကပ်များကြား ကိန်းဂဏန်းဆိုင်ရာ ဆက်နွယ်မှုများကို နားလည်ရန် သင်ယူကြပြီး တိုက်ကပ်များ၏ အတွဲလိုက် နောက်လားမည့် တိုက်ကပ်ကို ထုတ်လုပ်ရာတွင် ထူးချွန်သည်။

Clear Show example

Tokens 617 Characters 325

OpenAI ၏ကြီးမားသောဘာသာစကားမော်ဒယ်များ (တစ်ခါတစ်ရံ GPT များဟုလည်းခေါ်သည်) စာသားအစုအဝေးတွင်တွေ့ရလေ့ရှိသောအက္ခရာများဖြစ်သည့် တိုက်ကပ်များကိုအသုံးပြု၍ စာသားလုပ်ဆောင်သည်။ မော်ဒယ်များသည် ဤတိုက်ကပ်များကြား ကိန်းဂဏန်းဆိုင်ရာ ဆက်နွယ်မှုများကို နားလည်ရန် သင်ယူကြပြီး တိုက်ကပ်များ၏ အတွဲလိုက် နောက်လားမည့် တိုက်ကပ်ကို ထုတ်လုပ်ရာတွင် ထူးချွန်သည်။

Text Token IDs

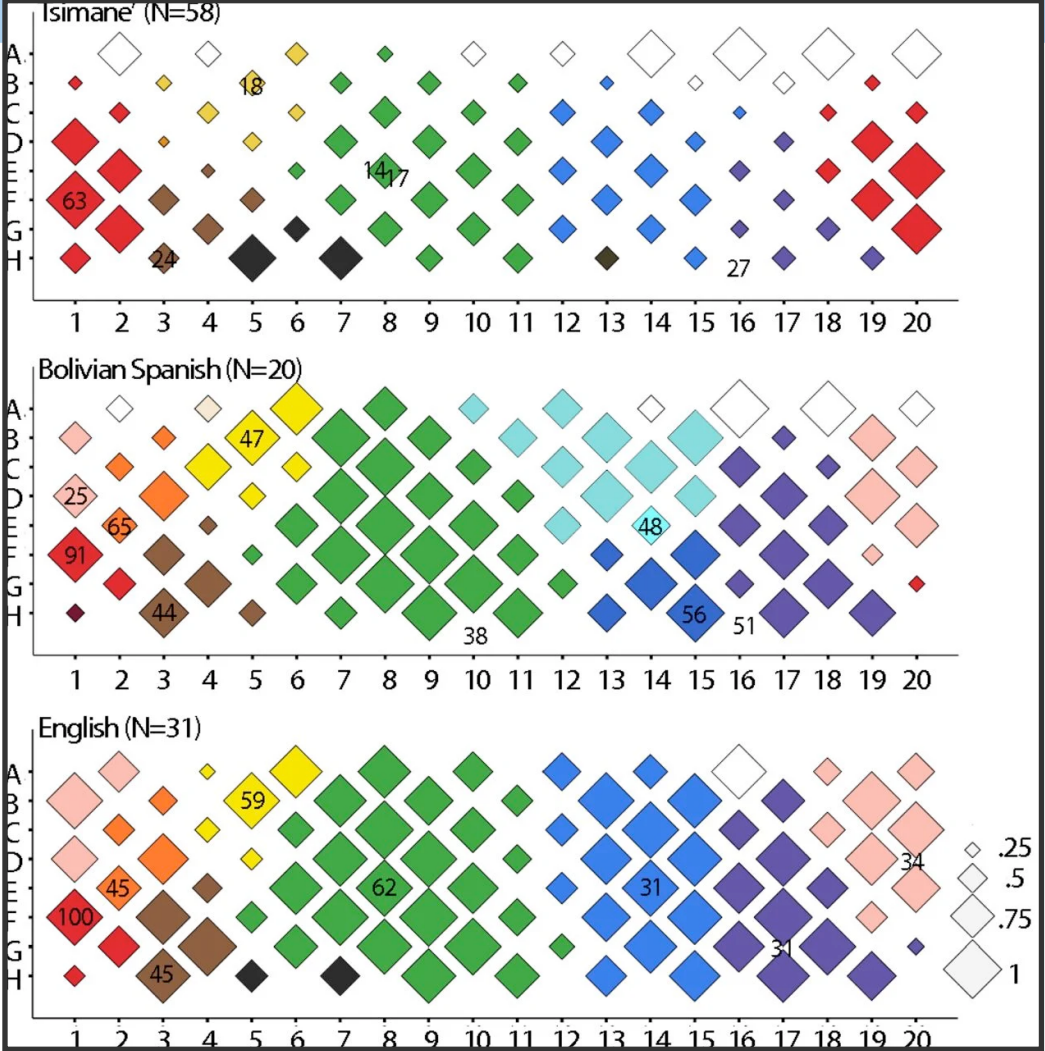
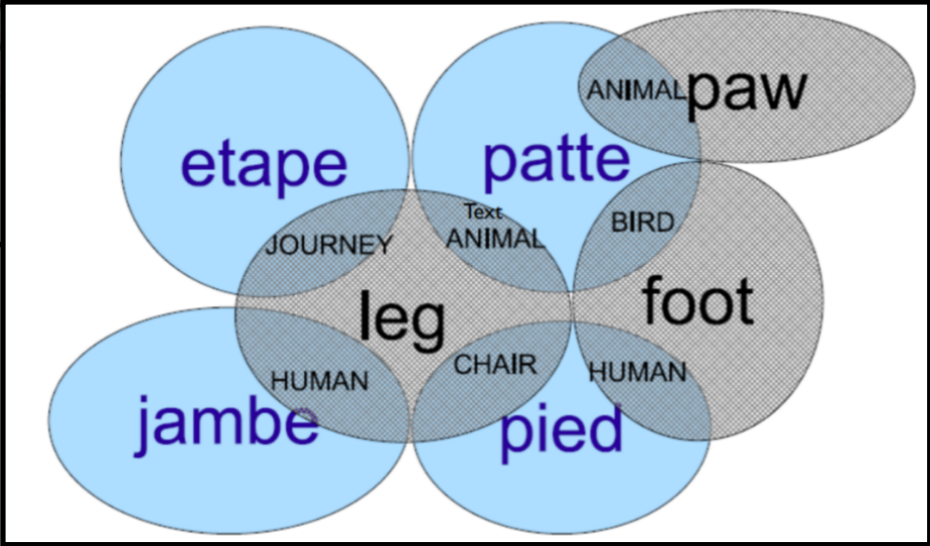
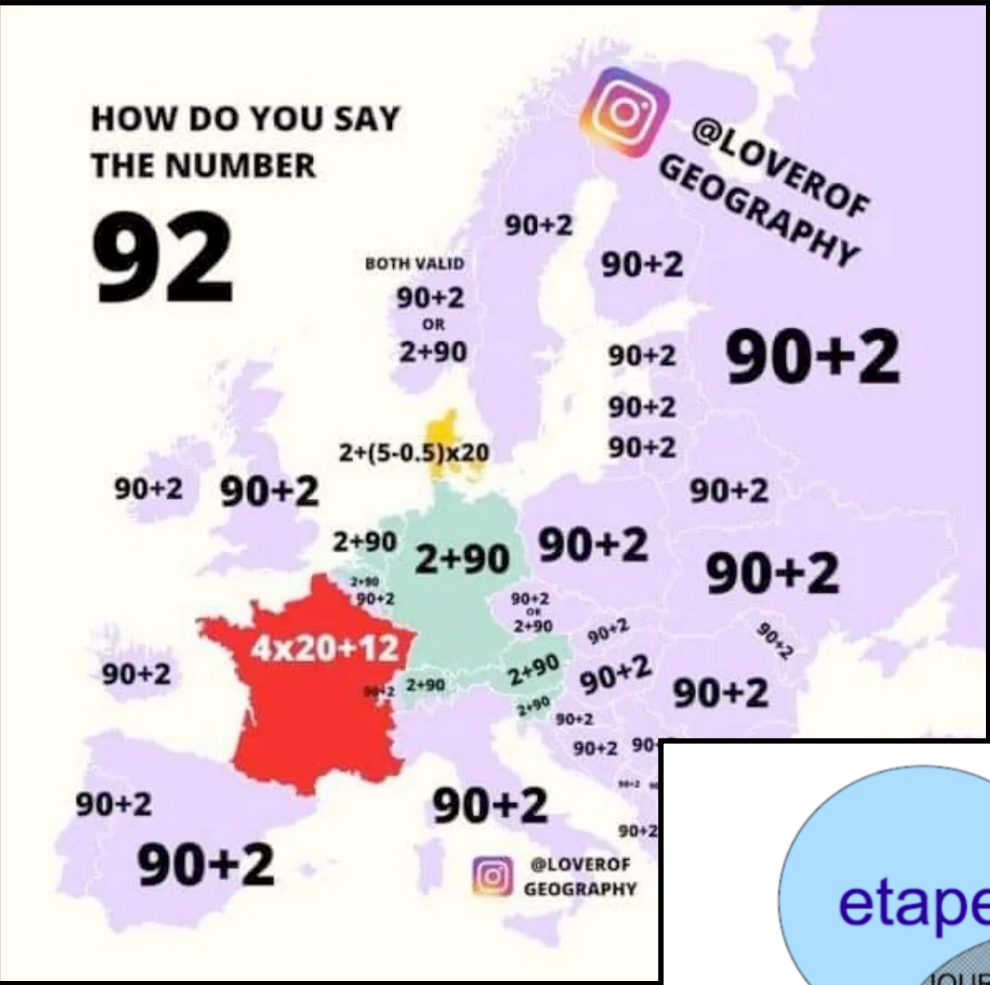
Similar content, 10.6x the tokens!

Challenge: Lexical Semantics



kàmpwóò ñkwuu sicyɛɛré ná bée-tàànre ná ké ná báár-ìcyèèrè
four hundred eighty four and twenty-three and ten and five-four
799 [i.e. $400 + (4 \times 80) + (3 \times 20) + \{10 + (5 + 4)\}$]

Base 80 numerical system (Supyire)



English → Farsi

she brings me dried dates and tells me old stories

khorma rotab kharak

Challenge: Syntax



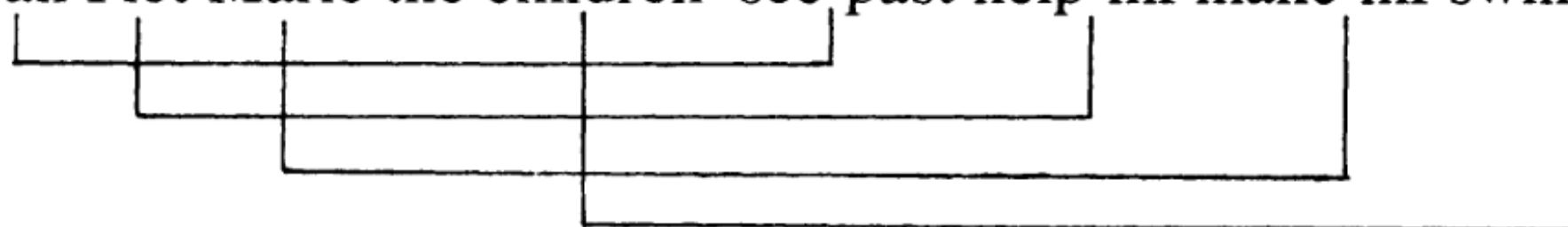
Type	Languages	%	Families	%
SOV	2,275	43.3%	239	65.3%
SVO	2,117	40.3%	55	15%
VSO	503	9.5%	27	7.4%
VOS	174	3.3%	15	4.1%
NODOM	124	2.3%	26	7.1%
OVS	40	0.7%	3	0.8%
OSV	19	0.3%	1	0.3%

The development of artificial intelligence is a really big deal.
 El desarrollo de la inteligencia artificial es un asunto realmente importante.
 The development of artificial intelligence is a really big deal.
 人工知能の発展は本当にすごいことです。

Maria non vuole mangiare.
 Maria not want [to-]eat
 "Maria does not want to eat."

Non vuole mangiare.
 Subject not want [to-]eat
 "[(S) he] does not want to eat."

. . . dat Jan Piet Marie de kinderen zag helpen laten zwemmen
 that Jan Piet Marie the children see-past help-inf make-inf swim-inf



‘. . . that Jan saw Piet help Marie make the children swim’

Challenge: Semantics



Wide variety of possible language features:

- Noun classifiers and grammatical genders
- What must be specified via declension (e.g., tense, aspect, number, gender, evidentiality, etc.)

三位 学生 (三位學生)
sān wèi xuéshēng
 three CL[human] student
 "three students"

三 棵 树 (三棵樹)
sān kē shù
 three CL[tree] tree
 "three trees"

三 只 鸟 (三隻鳥)
sān zhī niǎo
 three CL[animal] bird
 "three birds"

三 条 河 (三條河)
sān tiáo hé
 three CL[long-wavy] river
 "three rivers"

Aspectual Marking in AAVE

Aspect/Tense	Prototypical	Stressed / Emphatic Affirmative	Negative
Habitual	'be eating' (see Habitual be)	'DO be eating'	'don('t) be eating'
Remote Past	'BEEN eating' (see [17])	'HAVE BEEN eating'	'ain('t)/haven't BEEN eating'
Remote Past <i>Completive</i>	'BEEN ate'	'HAD BEEN ate'	'ain('t)/haven't BEEN ate'
Remote Past Perfect	'had BEEN ate'	'HAD BEEN ate'	'hadn't BEEN ate'
Resultant State	'done ate'	'HAVE done ate'	'ain('t) done ate'
Past Perfect Resultant State	'had done ate'	'HAD done ate'	'hadn't done ate'
Modal Resultant State	'should'a done ate'	--	--
Remote Past Resultant State	'BEEN done ate'	'HAVE BEEN done ate'	'ain('t)/haven't BEEN done ate'
Remote Past Perfect Resultant State	'had BEEN done ate'	--	--
Future Resultant State/ <i>Conditional</i>	'a be done ate'	'WILL be done ate'	'won't be done ate'
Modal Resultant State	'might/may be done ate'	'MIGHT/MAY be done ate'	'might/may not be done ate'

Evidentials in Eastern Pomo^[7]

Evidential type	Example verb	Gloss
nonvisual sensory	<i>p^ha·bék^h-ink'e</i>	"burned" [speaker felt the sensation]
inferential	<i>p^ha·bék-ine</i>	"must have burned" [speaker saw circumstantial evidence]
hearsay (reportative)	<i>p^ha·bék^h-le</i>	"burned, they say" [speaker is reporting what was told]
direct knowledge	<i>p^ha·bék-a</i>	"burned" [speaker has direct evidence, probably visual]

Challenge: Idioms and Figurative Speech



ወሃቸን ከለቀ (ehel wehachen aleke)

literally: *we ran out of grain and water*

actual meaning: *our time [usually romantic] has come to an end*

armar un pancho

literally: *to make/create a Pancho*

actual meaning: *to cause a big scene*



Pancho Villa

能ある鷹は爪を隠す (nō aru taka wa tsume o kakusu)

literally: *a skilled hawk hides its talons*

actual meaning: *truly talented people don't show off their skills*



Soga Nichokuan's "Eagle on a rock", 1624—44

馬馬虎虎 (mǎ mǎ hǔ hǔ)

literally: *horse-horse-tiger-tiger*

actual meaning: *sloppy*



Challenge: Differences in Language Use



(2) *An éisteann Seán lena mháthair riamh?*

Q listen.PRES Seán to his mother ever

Does Seán ever listen to his mother?"

(2.1) *Éisteann.*

listen.PRES

Yes, he does.

(2.2) **Éisteann sé.*

listen.PRES 3S.M.CNJV

(2.3) *Ní éisteann.*

not listen.PRES

No, he does not.

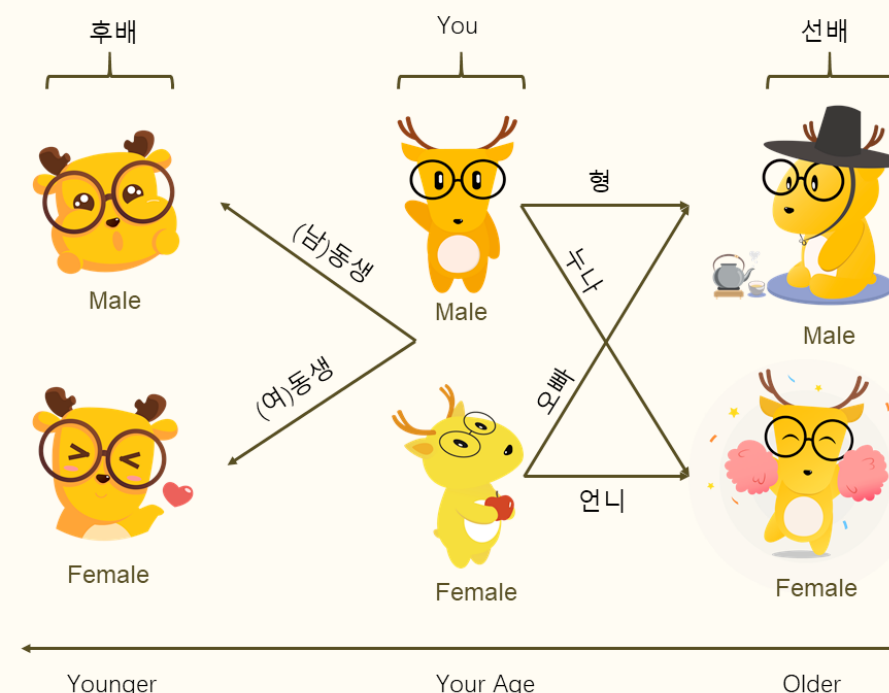
Irish

Malagasy deixis

		proximal		medial		distal		
Adverbs (here, there)	NVIS	atỳ	àto	ào	àtsy	àny	aròà*	arỳ
	VIS	etỳ	èto	èo	ètsy	èny	eròà	erỳ
Pronouns (this, that) (these, those)	NVIS	izatỳ*	izàto*	izào	izàtsy*	izàny	izaròà*	izarỳ*
	VIS	itỳ	ìto	ìo	ìtsy	ìny	iròà*	irỳ
	VIS.PL	irèto		irèo	irètsy	irèny	ireròà*	irerỳ*

English gloss	Standard Thai	Clerical vocabulary	Royal vocabulary
'hand'	/mān/ (มือ)	/mān/ (มือ)	/pʰráʔ hət/ (พระหัตถ์)
'house'	/bān/ (บ้าน)	/kūʔ.tìʔ/ (กุฏิ)	/wān/ (วัง)
'mother'	/māy/ (แม่)	/jō:m māy/ (โยมแม่)	/pʰráʔ tɕʰōn.náʔ.nīʔ/ (พระชนนี)
'to give'	/hāj/ (ให้)	/tʰàʔ.wǎj/ (ถวาย)	/tʰàʔ.wǎj/ (ถวาย)
'to speak'	/pʰūt/ (พูด)	/pʰūt/ (พูด)	/trət/ (ตรัส)
'to sleep'	/nō:n/ (นอน)	/tɕām wát/ (จำวัด)	/bān.tʰōm/ (บรรทม)

How to Address Other People



LINGODEER

Korean honorifics

Challenge: Language Change



Basilect ("Singlish")
*Wah lau! This guy Singlish
si beh
hiong sia.*

Mesolect
*This guy Singlish
damn good leh.*

Acrolect ("Standard")
*This person's Singlish
is very good.*

Singlish

走

1. to go; to walk; to go on foot

走吧！ — **Zǒu ba!** — Let's go!

他走在我的面前。 [MSC, trad. and simp.]

Tā zǒu zài wǒ de miànqián. [Pinyin]

He **walked** before me.

不要走得那麼快！ [MSC, trad.]

不要走得那么快！ [MSC, simp.]

Bù yào zǒu de nàme kuài! [Pinyin]

Don't **walk** so fast!

走著去火車站要多久？ [MSC, trad.]

走着去火车站要多久？ [MSC, simp.]

Zǒu zhe qù huǒchēzhàn yào duōjiǔ? [Pinyin]

How long does it take to **walk** to the station?

2. (literary or dialectal Mandarin, Cantonese, Hakka, Min, Wu) to run; to jog

未學行，先學走 [Cantonese, trad.]

未学行，先学走 [Cantonese, simp.]

mei⁶ hok⁶ haang⁴, sin¹ hok⁶ zau² [Jyutping]

(figurative) to learn to **run** before one can walk

Word	Vowel pronunciation	
	Late Middle English before the GVS	Modern English after the GVS
bīte	[i:]	[aɪ]
meet	[e:]	[i:]
meat	[ɛ:]	
serene		
mate	[a:]	[eɪ]
out	[u:]	[aʊ]
boot	[o:]	[u:]
boat	[ɔ:]	[oʊ]
stone		

