Gridspace

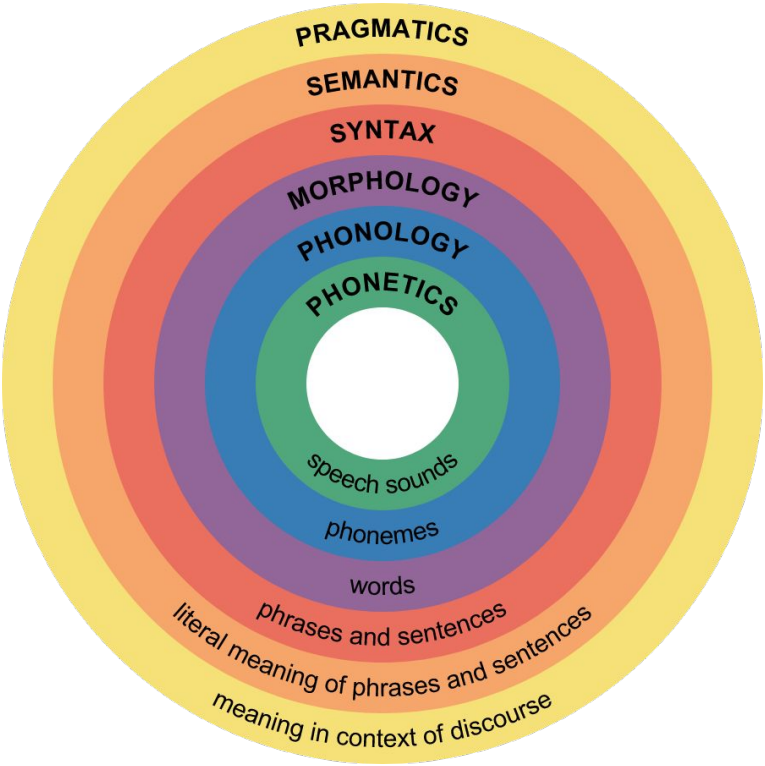
IAP Program 2023
Lecture 4: Morphology and Words

January 16th 2023

Course Logistics

- Project 1 due
- Project 2 released today - Fine Tune GPT-3
- Upcoming Lectures:
  - Syntax, Semantics and Word Embeddings
  - InstructGPT and Large Language Models
- Lecture videos and slides on website

Linguistics/NLP Week



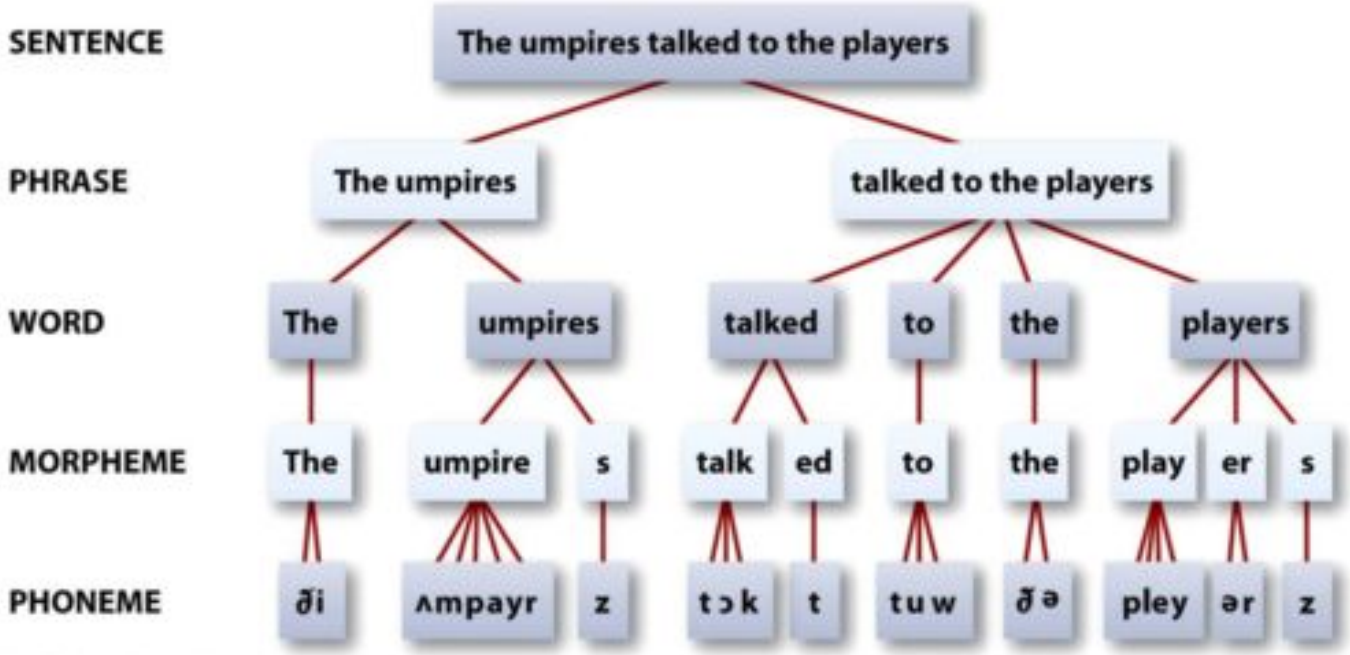
Phonetics

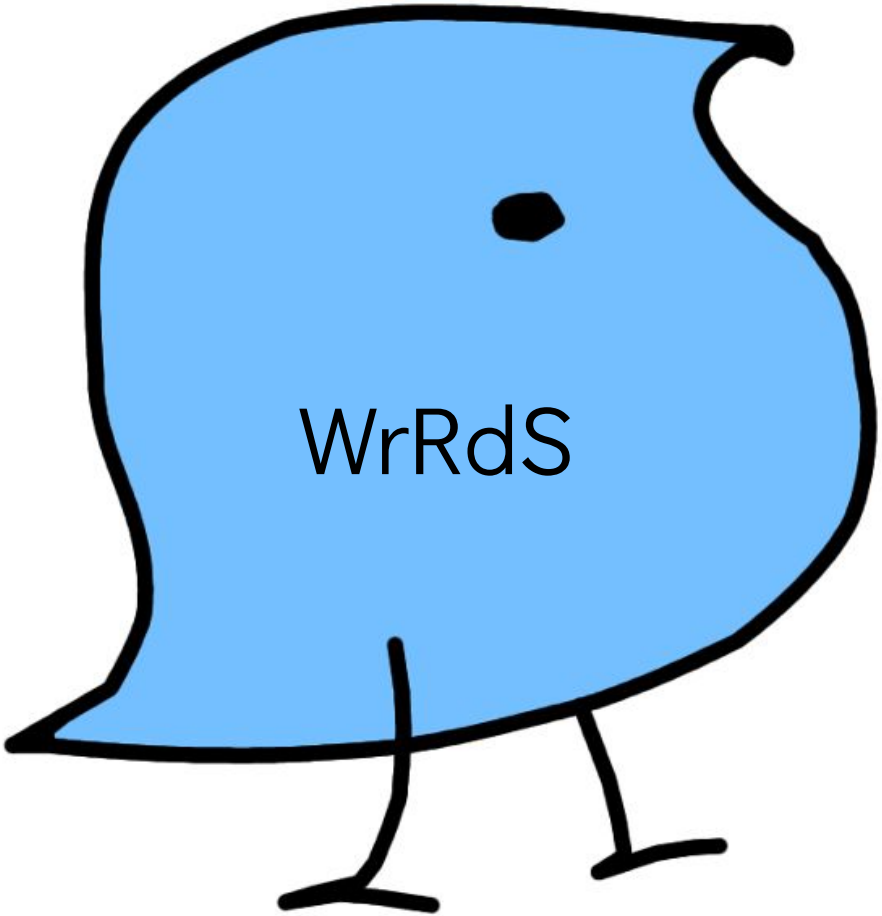
THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

CONSONANTS (PULMONIC)

	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.

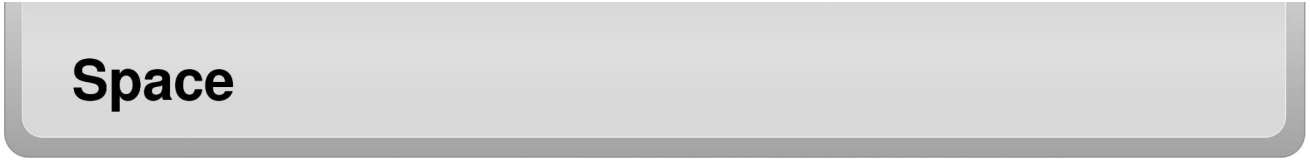




What is a word

A word is a unit of language that carries meaning and can be spoken or written. It is typically made up of one or more morphemes, which are the smallest units of meaning in a language. Words can be used to convey ideas, express emotions, or describe things and actions. They can be combined to form phrases, sentences, and paragraphs, which convey more complex meaning.

What defines a word?



Hypothesis: A word is defined by whitespace

Learning is so boring unless it's gossip. Teachers should just start every lecture like "omg did you hear about parabolas?"

@the.language.nerds

What defines a word?

“Oh my gosh, tell me everything, I’m all ears!”

9 words by whitespace

What defines a word?

"**Omg**, tell me everything, I'm **listening**!"

6 words by whitespace, same meaning

What defines a word?

“Omg, tell me **every thing**, I’m listening!”

7 words by whitespace, same meaning

Hangry
Yeet

Catfish
Sick

Murder
Swagger
Skim-Milk

Eke
Pale

An archaic sense of the noun *pale* was fossilized in the English language in the 18th century in the expression *beyond the pale*. The noun is unrelated to the familiar adjective meaning "deficient in color"; it is ultimately derived, by way of Anglo-French, from the Latin word *palus*, meaning "stake." In its literal uses, *pale* referred to both stakes and fences and to boundaries made up of stakes.

Feckless —-> Wow, she is full of **feck**
Unkempt —-> He looks pretty **kempt** today
Disheveled —-> I woke up **heveled**
Antiquated —--> The wonders of **quated** technology

Morphology

The study of words, how they are formed, and their relationship to other words in the same language.

Morpheme

trimmings

trim / ing / s

Morpheme

trimmings

trim / ing / s

free

Morpheme

trimmings

trim / ing / s

bound

Morpheme

trimmings

trim / ing / s

root

Morpheme

trimmings

trim / ing / s

suffix

Morpheme

trimmings

trim / ing / s

Suffix

Inflectional vs Derivational

teach

teach-es

Inflectional Suffix

teach-er

Derivational Suffix

Morphemes Summary

Free morphemes

Can stand alone as own word

e.g. dog, gentle, picture gem

Bound morphemes

- **Derivational**
  - **Prefixes**
    - e.g. de- pre- in- un-
  - **Suffixes**
    - e.g. -ion -ly -able -er
- **Inflectional**
  - **Suffixes**
    - e.g. plural -s -ing -ed

What defines a word?

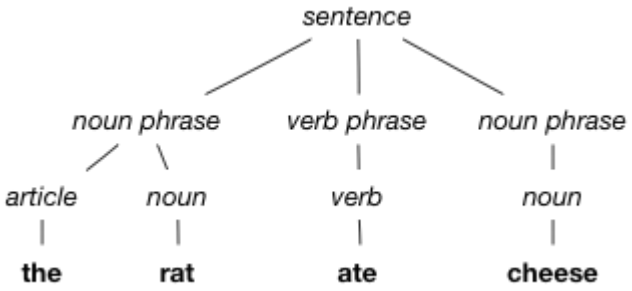
"**Omg**, tell me **every thing**, I'm **all ears**!"

~~A word is defined by whitespace~~

A word is something that can stand on it's own, and is composed of one or more morphemes.

Sequences

Sentence



Utterance

(**N**LP)

"the rat uh ate ate cheese"

Morphology of Languages

Morphological Typology

Classification of languages by the morphemes and how they interact.

Morphological Typology

The metric: average morphemes per word.

This forms a spectrum.

Morphological Typology

morphemes per word

Morphological Typology

Minimum=1

Very high

morphemes per word

Morphological Typology

Analytic

Synthetic

morphemes per word

Morphological Typology

Isolating

Gbogbo ènìyàn ni a bí ní òmìnira; iyì àti ẹ̀tọ́ kọ̀ọ̀kan sì dọ́gba. Wọ́n ní ẹ̀bùn ti làákàyè àti ti ẹ̀rí-ọkàn, ó sì yẹ kí wọn ó máa hùwà sí ara wọn gẹ́gẹ́ bí ọmọ ìyá.
All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Morphological Typology

Isolating

sub + adverb + verb + complement + obj

我以前爱过他

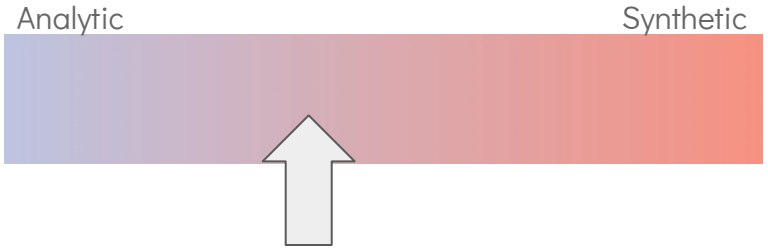
I loved him before

Morphological Typology

Analytic

Ek het nie geweet dat hy sou kom nie.	I did not know that he would come.

Morphological Typology

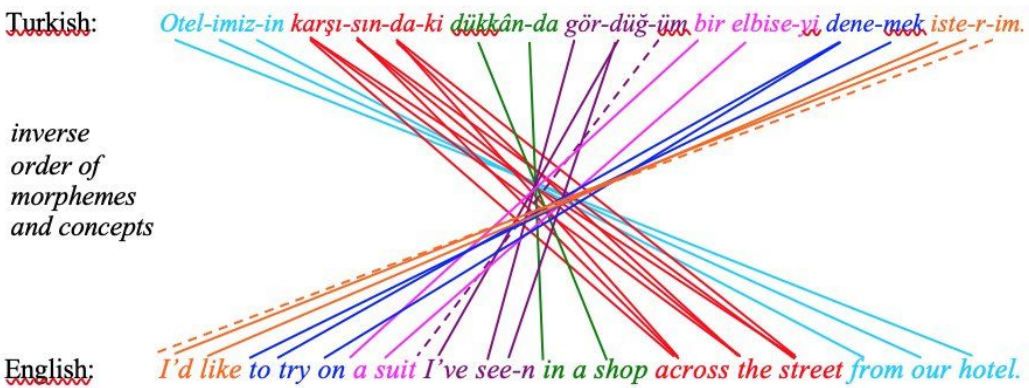


Fusional

hablo	'I am speaking'
habla	'S/he is speaking'
hablé	'I spoke'
hablamos	'We are speaking'
hablan	'They are speaking'

Morphological Typology

Agglutinative



Turkish	English
Muvaffak	Successful
Muvaffakiyet	Success
Muvaffakiyet**siz**	**Un**successful ('**without** success')
Muvaffakiyetsiz**leş***(-mek)*	*(To)* **become** unsuccessful
Muvaffakiyetsizleş**tir***(-mek)*	*(To)* **make one** unsuccessful
Muvaffakiyetsizleştiri**ci**	**Maker of** unsuccessful ones
Muvaffakiyetsizleştirici**leş***(-mek)*	*(To)* **become** a maker of unsuccessful ones
Muvaffakiyetsizleştiricileş**tir***(-mek)*	*(To)* **make one** a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiri**ver***(-mek)*	*(To)* **easily/quickly** make one a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştirivere**bil***(-mek)*	*(To)* **be able to make** one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştirivere**meye**bil*(-mek)*	**Not** *(to)* be able to make one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebil**ecek**	**One who is** not able to make one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebilecek**ler**	**Those** who are not able to make one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebilecekleri**miz**	Those who **we** cannot make easily/quickly a maker unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimiz**den**	**From** those we can not easily/quickly make a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizden**miş**	**(Would be)** from those we can not easily/quickly make a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmiş**siniz**	**You** would be from those we can not easily/quickly make a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsiniz**cesine**	**Like** you would be from those we can not easily/quickly make a maker of unsuccessful ones

Morphological Typology

Polysynthetic

Aliikkusersuillammassuaanerartassagaluarpaalli	However, they will say that he is a great entertainer, but

Morphological Challenges in NLP

Stemming

"the boy's cars are different colors"

"the boy car be differ color"

Manning, Raghavan & Schütze, 2008

Morphological Challenges in NLP

Lammas
and
Lemmatization

better is the enemy of the good
good is the LEMMA of the better

Morphological Challenges in NLP

Evaluating the Impact of Sub-word Information and Cross-lingual Word Embeddings on Mi'kmaq Language Modelling

Jeremie Boudreau,[1] Akankshya Patra,[2] Ashima Suvarna[1] and Paul Cook[1]
1. Faculty of Computer Science, University of New Brunswick
2. University of Southern California
jeremie@boudreau.me, patra@usc.edu, asuvarna31@gmail.com, paul.cook@unb.ca

Abstract

Mi'kmaq is an Indigenous language spoken primarily in Eastern Canada. It is polysynthetic and low-resource. In this paper we consider a range of n-gram and RNN language models for Mi'kmaq. We find that an RNN language model, initialized with pre-trained fastText

8. Conclusions

In this paper we explored a variety of approaches to language modelling for Mi'kmaq, which is particularly challenging due to its rich morphology, and because it is a low-resource language.

We considered n-gram and RNN language models, with a variety of parameter settings in an effort to establish a strong baseline. We then considered the use of pre-trained fastText embeddings to initialize the input layer of the RNN language models. This gave substantial improvements over the baseline, highlighting the importance of sub-word information, and approaches that can repre-

Morphological Challenges in NLP

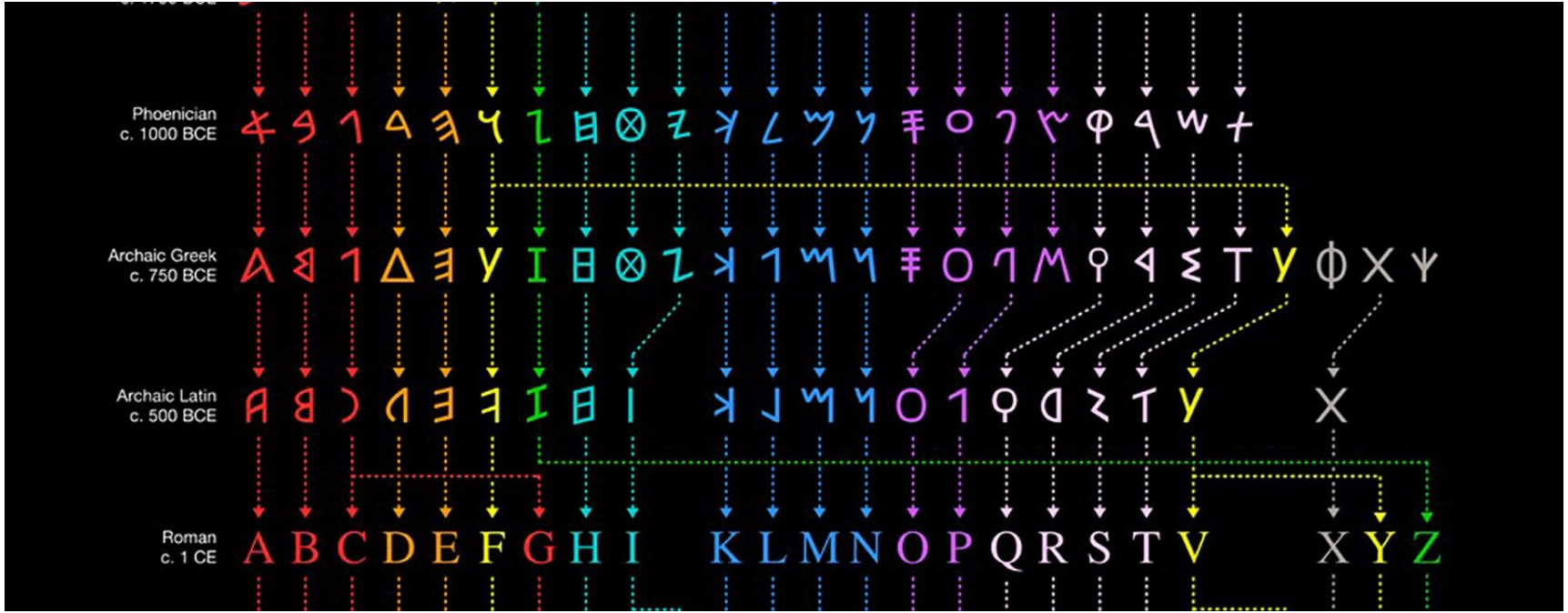
Glove vs Handschuhe

Morphological Challenges in NLP

GloVe = Global Vectors for Word Representation

...that's just pragmatics ;)

Writing Systems



Writing Systems

Type	Each symbol represents	Example
Logographic	word	Chinese characters
Syllabary	syllable	Japanese *kana*
Abjad	consonant	Arabic alphabet
Alphabet	consonant or vowel	Latin alphabet
Abugida	consonant accompanied by specific vowel, modifying symbols represent other vowels	Indian *Devanagari*
Featural system	distinctive feature of segment	Korean *Hangul*

Daniels and Bright, 1996

Logographic

漢語

汉语

中文

Syllabary

あア	かカ	さサ	たタ	なナ	はハ	まマ	やヤ	らラ	わワ	んン
a	ka	sa	ta	na	ha	ma	ya	ra	wa	n
いイ	きキ	しシ	ちチ	にニ	ひヒ	みミ		りリ	ゐヰ	
i	ki	shi	ti	ni	hi	mi		ri	(wi)	
うウ	くク	すス	つツ	ぬヌ	ふフ	むム	ゆユ	るル		
u	ku	su	tsu	nu	fu	mu	yu	ru		
えエ	けケ	せセ	てテ	ねネ	へヘ	めメ		れレ	ゑヱ	
e	ke	se	te	ne	he	me		re	(we)	
おオ	こコ	そソ	とト	のノ	ほホ	もモ	よヨ	ろロ	をヲ	
o	ko	so	to	no	ho	mo	yo	ro	(w)o	

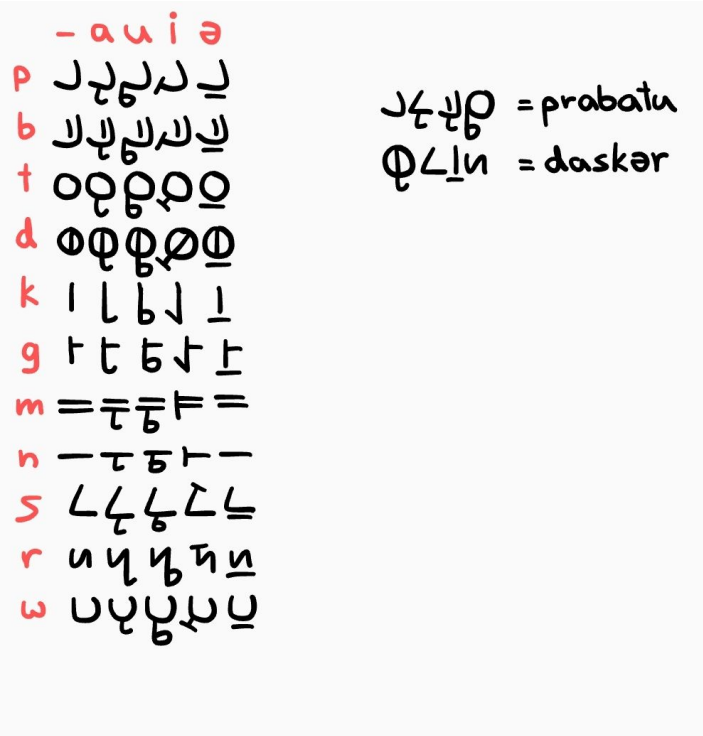
Abjad

خ	ح	ج	ث	ت	ب	ا
ḫ	ḥ	ǧ	ṯ	t	b	ʾā
ص	ش	س	ز	ر	ذ	د
ṣ	š	s	z	r	ḏ	d
ق	ف	غ	ع	ظ	ط	ض
q	f	ġ	ʿ	ẓ	ṭ	ḍ
ي	و	ه	ن	م	ل	ك
y	w	h	n	m	l	k

Alphabet

	EGYPTIAN			SEMITIC		LATER EQUIVALENTS		
Values	Hieroglyphic		Hieratic	Phœnician		Greek	Roman	Hebrew
a	eagle					A	A	א
b	crane					B	B	ב
ḳ (*g*)	throne					Γ	C	ג
ṭ (*d*)	hand					Δ	D	ד
h	mæander					E	E	ה
f	cerastes					Y	F	ו
z	duck					I	Z	ז
χ (*kh*)	sieve					H	H	ח
θ (*th*)	tongs					Θ	...	ט
i	parallels					I	I	י
k	bowl					K	K	כ
l	lioness					Λ	L	ל
m	owl					M	M	מ
n	water					N	N	נ
s	chairback					Ξ	X	ס

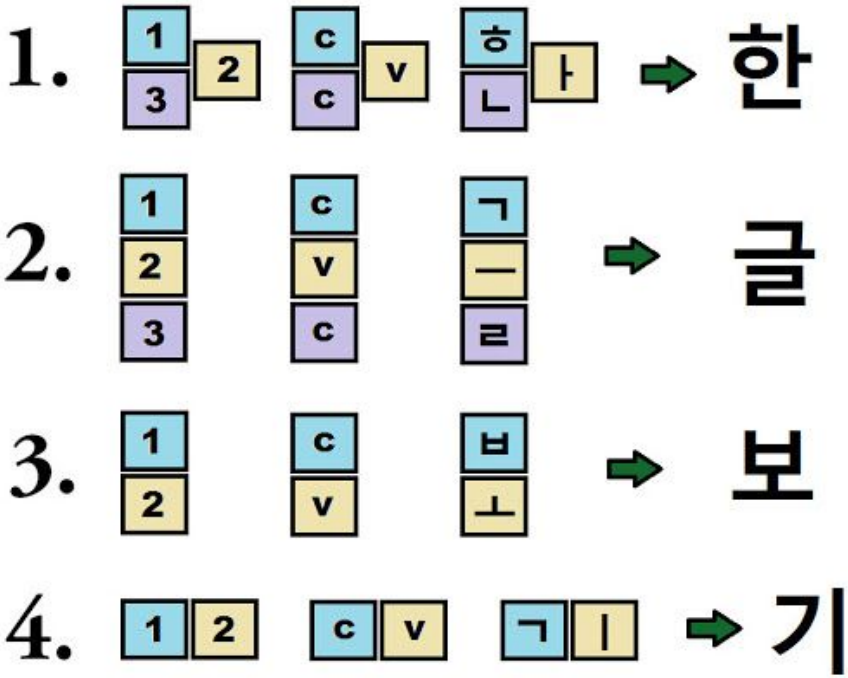
Abugida



Featural system

+	ㄱ g	ㄴ n	ㄷ d	ㄹ r	ㅁ m	ㅂ b	ㅅ s	ㅇ o	ㅈ j	ㅊ ch	ㅋ k	ㅌ t	ㅍ p	ㅎ h
ㅏ a	가 ga	나 na	다 da	라 ra	마 ma	바 ba	사 sa	아 a	자 ja	차 cha	카 ka	타 ta	파 pa	하 ha
ㅑ ya	갸 gya	냐 nya	댜 dya	랴 rya	먀 mya	뱌 bya	샤 sya	야 ya	쟈 jya	챠 chya	캬 kya	탸 tya	퍄 pya	햐 hya
ㅓ eo	거 geo	너 neo	더 deo	러 reo	머 meo	버 beo	서 seo	어 eo	저 jeo	처 cheo	커 keo	터 teo	퍼 peo	허 hya
ㅕ yeo	겨 gyeo	녀 nyeo	뎌 dyeo	려 ryeo	며 myeo	벼 byeo	셔 syeo	여 yeo	져 jyeo	쳐 chyeo	켜 kyeo	텨 tyeo	펴 pyeo	혀 hyeo
ㅗ o	고 go	노 no	도 do	로 ro	모 mo	보 bo	소 so	오 o	조 jo	초 cho	코 ko	토 to	포 po	호 ho
ㅛ yo	교 gyo	뇨 nyo	됴 dyo	료 ryo	묘 myo	뵤 byo	쇼 sya	요 yo	죠 jyo	쵸 chyo	쿄 kyo	툐 tyo	표 pyo	효 hyo
ㅜ u	구 gu	누 nu	두 du	루 ru	무 mu	부 bu	수 su	우 u	주 ju	추 chu	쿠 ku	투 tu	푸 pu	후 hu
ㅠ yu	규 gyu	뉴 nyu	듀 dyu	류 ryu	뮤 myu	뷰 byu	슈 syu	유 yu	쥬 jyu	츄 chyu	큐 kyu	튜 tyu	퓨 pyu	휴 hyu
ㅡ eu	그 geu	느 neu	드 deu	르 reu	므 meu	브 beu	스 seu	으 eu	즈 jeu	츠 cheu	크 keu	트 teu	프 peu	흐 heu
ㅣ i	기 gi	니 ni	디 di	리 ri	미 mi	비 bi	시 si	이 i	지 ji	치 chi	키 ki	티 ta	피 pi	히 hi
ㅐ ae	개 gae	내 nae	대 dae	래 rae	매 mae	배 bae	새 sae	애 ae	재 jae	채 chae	캐 kae	태 tae	패 pae	해 hae

Featural system



Writing Systems

Glyphs

Characters

Punctuation

Graphemes

N-grams

Unigram - "i"
Bigram - "i am"
Trigram - "i am waiting"
4-gram - "i am waiting for"

Writing Systems and NLP

Spelling and ASR dictionaries.

Spelling and token ambiguity.

Corpus quality and consistency.

Corpora

Corpus - a collection of written texts

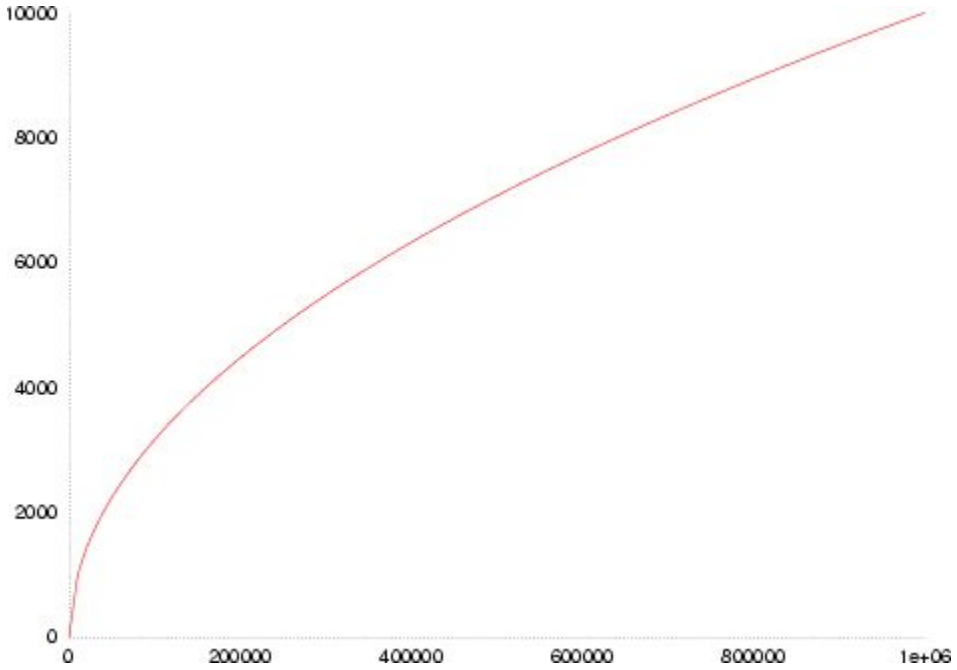
Lexicon - unique tokens present in the corpus.

Heaps' law

The size of the lexicon as a function of the total tokens in the corpus.

$$L(n) = Kn^{\beta}$$

Heaps' law



Heaps' law

For English, K is 10-100, and β is 0.4-0.6.

$$L(n) = Kn^{\beta}$$

Zipf's law

The frequency of tokens in a corpus are power law distributed.

$$F(n) = 1/n$$

TF-IDF

Numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Term x within document y

Tokenization

Writing System for Machines

Sub-Word Tokenization

Robust to misspellings
Generalizable to multi-lingual systems

Requires careful pre-processing of corpus
High memory and Unused sub-words

BPE - Byte-Pair Encoding

1. Split into words
2. Split into characters
3. Iteratively count frequency of consecutive character pairs and merge

BPE - Byte-Pair Encoding

1. Split into words
2. Split into characters
3. Iteratively count frequency of consecutive character pairs and merge

BPE - Byte-Pair Encoding

("hug", 10), ("pug", 5), ("pun", 12), ("bun", 4), ("hugs", 5)

↓

Vocabulary: ["b", "g", "h", "n", "p", "s", "u", "g"]
Corpus: ("h" "u" "g", 10), ("p" "u" "g", 5), ("p" "u" "n", 12), ("b" "u" "n", 4), ("h" "u" "g" "s", 5)

BPE - Byte-Pair Encoding

“u” + “g” -> “ug”

Vocabulary: ["b", "g", "h", "n", "p", "s", "u", "ug"]

Corpus: ("h" "ug", 10), ("p" "ug", 5), ("p" "u" "n", 12), ("b" "u" "n", 4), ("h" "ug" "s", 5)

BPE - Byte-Pair Encoding

“u” + “n” -> “un”

↓

Vocabulary: ["b", "g", "h", "n", "p", "s", "u", "ug", "un"]

Corpus: ("h" "ug", 10), ("p" "ug", 5), ("p" "un", 12), ("b" "un", 4), ("h" "ug" "s", 5)

BPE - Byte-Pair Encoding

“u” + “n” -> “un”

Vocabulary: ["b", "g", "h", "n", "p", "s", "u", "ug", "un"]

Corpus: ("h" "ug", 10), ("p" "ug", 5), ("p" "un", 12), ("b" "un", 4), ("h" "ug" "s", 5)

BPE - Byte-Pair Encoding

“h” + “ug” -> “hug”

Vocabulary: ["b", "g", "h", "n", "p", "s", "u", "ug", "un", "hug"]

Corpus: ("hug", 10), ("p" "ug", 5), ("p" "un", 12), ("b" "un", 4), ("hug" "s", 5)

BPE - Byte-Pair Encoding

.....til desired size of corpus is reached

↓

Vocabulary: ["b", "g", "h", "n", "p", "s", "u", "ug", "un", "hug", “pug”, “pun”, “bun”]

Corpus: ("hug", 10), ("pug", 5), ("pun", 12), ("bun", 4), ("hug" "s", 5)

OOV Words

Vocabulary: ["b", "g", "h", "n", "p", "s", "u", "ug", "un", "hug", “pug”, “pun”, “bun”]

“bug” —> “b”, “ug”

“Mug” —> <UNK>, “ug”

BPE - Byte-Pair Encoding

Pro: Less space wasted for unused subwords - Only store the most useful 'byte-pairs' as a sub-word.

Con: Tokenization specific to corpus and number of iterations. Different tokenizations can result for same corpus

This affects our embeddings and models.... More on this next lecture!

Answers from Last Time

- How do you handle randomness in JAX?
  - By creating and manually passing around PRNGKeys and using jnp.split to progress the random state manually
- Why would JAX not support JIT compiling side-effects (printing and globals) *nor* dynamically-sized argument-based values (e.g. passing in a length as an argument to use for a tensor)? (Hint: look at the purpose of the library as a whole.)
  - JAX Autograd is optimized for pure functions (so no side-effects). The argument-based values are unsupported for JIT code for the reason of avoiding recompilation.
- What are Flax and Optax and where would we use them in the example application?
  - Flax is a network module layer for JAX, similar to Pytorch, and Optax is a library of NN optimizers. It would replace the manual gradient descent (Optax) and the prediction of a linear regression model (Flax) in the example application.

Exercises for Next Time

- Split the word “antidisestablishmentarianism” into its morphemes. What does the word mean?

- Build your own, brand new word in Turkish

- Run the Byte Pair Encoding algorithm on the string **aaabdaaabac.** What is the smallest number of characters needed to encode this in a compressed form?

Morpheme Match-Ups

Cut out the morphemes below and create as many words as you can. Use the Morpheme Match-Ups Guide to help you determine the meanings of these words.

phone	tele	vision
scope	micro	graph
auto	sub	mobile
way	scribe	re