LING 334 - Introduction to Computational Linguistics

Week 5

Linguistic Structure, NLP “Tasks”, and Annotation

The Basic Approaches of Linguistics

It's all over the place! Low consensus field.

This makes some sense -
language has many parts and purposes.

Descriptivism

Maybe the one thing we can all agree on:
the object of study is how and what language is,
rather than what it "should be" (prescriptivism)

Descriptivism

Origins with Pāṇini, Sanskrit linguist ~400BC

Contrast with “experts,” Strunk and White etc.
(these are cultural norms and conventions)

Key (modern) ideas:

- Language change is normal and expected
- Everyone has a “dialect”
- There are very few cross-linguistic universals

Traditional Levels of Structure

Small to big units:

Phonetics	sounds
Phonology	ordering of sounds
Morphology	words and word parts
Syntax	ordering of words
Semantics	propositional meaning
Pragmatics	non-propositional meaning

But there are many more...

(very roughly) Small to big units:

Reference	pointing out things with words
Prosody	suprasegmental sounds like pitch
Discourse	sequences between large units
Social Meaning	social implicature of variation

The Concept of a “Task” in NLP

Research in NLP is often framed as solving a particular “task”, e.g. improving performance at some problem

Very frequent sort of task in traditional NLP:

> Given free text or speech audio,
> automatically generate a representation
> of some part of its linguistic structure

Phonetics

The physical production and perception of speech sounds

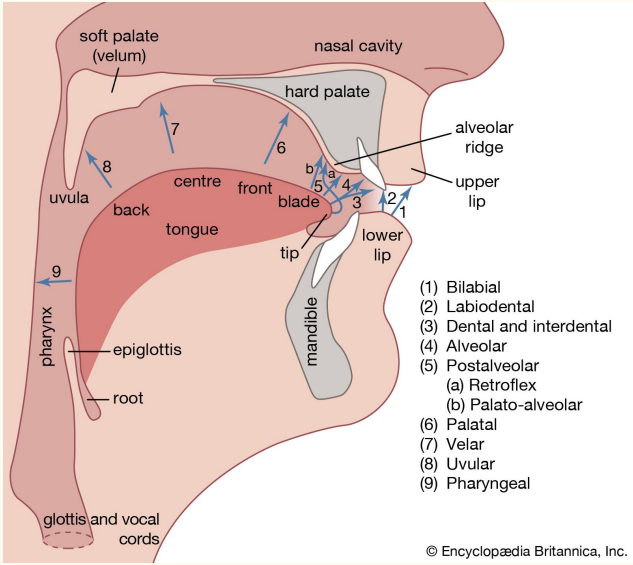
Unit of analysis: speech sound

NLP Tasks:

Speech synthesis

Automated transcription

https://dood.al/pinktrombone/



International Phonetic Alphabet (IPA) ˌɪntəˈnæʃn·l fəˈnetɪk ˈælfəˌbet

Consonants (pulmonic)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	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			

Phonology

The systematic organization of speech sounds

Unit of analysis: phoneme

Questions include:

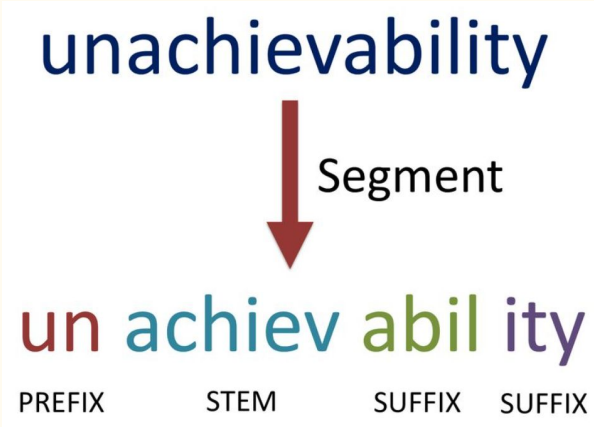
- Which set of sounds does a language use?
- What rules constrain their orderings?

Example: /P/ aspiration

- ‘pin’ - the ‘p’ sound has a puff of air [p^h]
- ‘spin’ - it doesn’t [p]

NLP Tasks: Similar to Phonetics

Morphology



credit Ryan Cotterell

The structure and constituent parts of words

Unit of analysis: morpheme

(smallest meaning-bearing unit)

Morphemes can be:

Free can stand alone, words like ‘cat’ and ‘banana’

Bound can’t stand alone, word-parts like ‘un-’ and ‘-est’

NLP Tasks:

- Morphological Segmentation (very important in synthetic langs!)
- Lemmatization and Inflection

Syntax

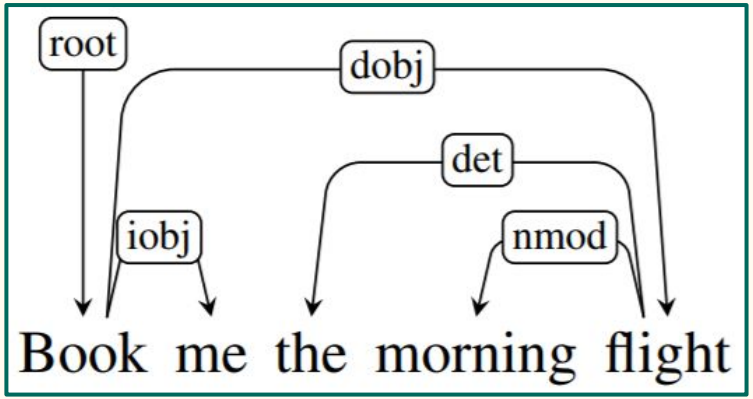
The systematicity of word orderings

"The sloth ate the cupcake." != "The cupcake ate the sloth."

* "Cupcake sloth ate the the."

NLP Tasks:

- Syntactic Parsing
- Downstream applications, e.g.:
  - Machine Translation
  - Semantic Similarity



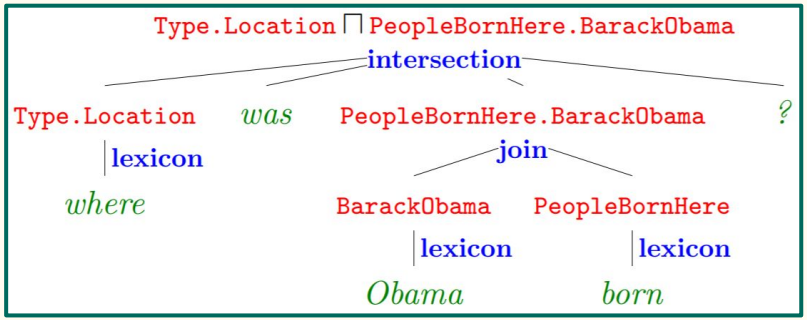
Semantics

The propositional (e.g., literal) meanings of words and larger units (frequently sentences)

Table = 

NLP Tasks:

- Recognizing Textual Entailment
- Semantic Parsing



credit Berant et al (2013)
Semantic Parsing on Freebase from Question-Answer Pairs

We're digging into semantics next week!

Pragmatics

The beyond-propositional meanings of words and larger units

Among the many possibilities:

Implicature "I'm sad." "Here's a popsicle."

Performatives "I now pronounce you X and Y."

Deference "Please follow me, your majesty."

Information Structure

a. A large book was sitting on the desk.
b. On the desk a large book was sitting.
c. On the desk was sitting a large book.
d. It was a large book that was sitting on the desk.
e. What was sitting on the desk was a large book.
f. There was a large book sitting on the desk.
g. Sitting on the desk was a large book.

credit Gregory Ward

NLP Tasks:

Many social/applied!

Traditional Levels of Structure

Small to big units:	Phonetics	sounds
	Phonology	ordering of sounds
	Morphology	words and word parts
	Syntax	ordering of words
	Semantics	propositional meaning
	Pragmatics	non-propositional meaning

Reference

What entity in the world does a linguistic expression point out?

Includes pronouns, honorifics, naming and nicknaming

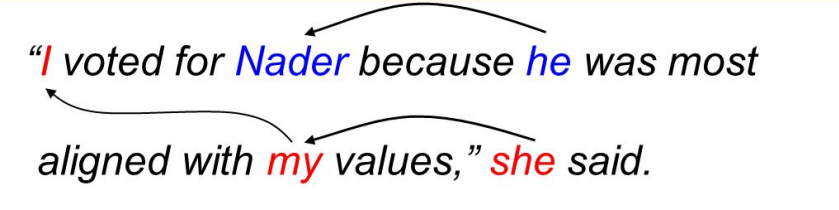
Winograd Schema Challenge:

"The goose wouldn't fit in the boat because **it** was too big."

"The goose wouldn't fit in the boat because **it** was too small."

NLP Tasks:

- Coreference Resolution
- Named Entity Recognition

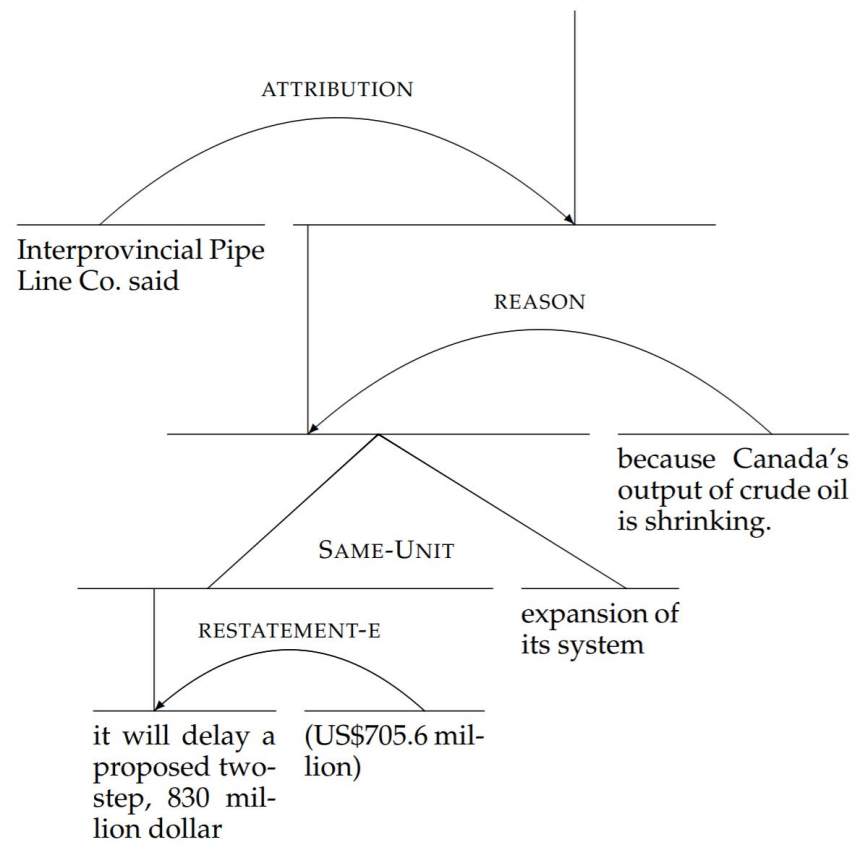


Discourse

The relations between clauses and propositions

NLP Tasks:

- Discourse Parsing
- Argumentation Mining



Social Meaning

Many sorts of complex socially enmeshed meaning-making:

- Sentiment and stance
- Regional variation
- Identity performance
- Memes and spread of ideas

Each can be an NLP Task!

Data in Linguistics

Introspection, and/or “native speaker intuitions”

Collected observations of language in use (e.g. corpora)

Laboratory data (experimentally collected or manipulated)

All of the above potentially augmented with *annotations*

Linguistic Annotations

To train a relevant model, we need training data

So, we hand-label some!

Traditionally, most commonly done by experts

Today, frequently done with crowdsourcing as well

Which is more appropriate depends on the task!

See relevant readings re: wisdom of the crowd - Naive annotators can do a great job!

Annotation Schemes

An annotation scheme or ontology instantiates a theory of language.

Example - Part of Speech Tagging:

36 Penn Treebank Tags

Implicit Proposal: these are what's important

I/PRP love/VBP eating/VBG noodles/NNS

Penn Treebank POS Tags

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Annotation Schemes (cont.)

Frequently developed over multiple rounds of piloting

Common tradeoff between specificity and speed/expense/scale

Do I want 40 categories and 400 annotations,

or 5 categories and 4,000 annotations?

Zipf's Law - vanishing returns as we get many categories

Annotation Evaluation

Linguistic categories are purely abstract human creations!

There is no ground truth. (rut roh)

So we usually evaluate with **Inter-Annotator Agreement**

Have some proportion of the data annotated by multiple people

Obtain a measurement of consistency -
how often do people make the same judgment?

Inter-Annotator Agreemeent

Common - Cohen's Kappa

Compare the expected agreement to the actual:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e}$$

p_o = probability of observed agreement

p_e = probability of expected agreement

Inter-Annotator Agreemeent

Say we have a task with two labels, POS and NEG,

and two annotators, A and B, with these labels:

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Annotator A	POS	POS	NEG	POS	POS	NEG	NEG	POS
Annotator B	POS	NEG	NEG	POS	NEG	NEG	POS	POS

... etc.

Inter-Annotator Agreemeent

Count up each category:

Annotator B

Annotator A	POS	NEG
POS	45	15
NEG	25	15

Inter-Annotator Agreemeent

Get totals:

Annotator B

Annotator A	POS	NEG	*total*
POS	45	15	60
NEG	25	15	40
total	70	30	N = 100

Inter-Annotator Agreemeent

They agreed 60% of the time

p_o = (45 POS + 15 NEG) / 100 total = 60%

Annotator B

Annotator A

	POS	NEG	*total*
POS	45	15	60
NEG	25	15	40
total	70	30	N = 100

Inter-Annotator Agreemeent

Probability of expected is trickier - calculate expected freq for each category: E_{freq} = (row_total * col_total) / N

Annotator B

Annotator A	POS	NEG	*total*
POS	45 (42)	15	60
NEG	25	15 (12)	40
total	70	30	N = 100

Inter-Annotator Agreemeent

Now we can get p_e:

$p_e = (42 \text{ POS exp} + 12 \text{ NEG exp}) / 100 = 0.54$

Annotator B

Annotator A

	POS	NEG	*total*
POS	45 (42)	15	60
NEG	25	15 (12)	40
total	70	30	N = 100

Inter-Annotator Agreemeent

And calculate Kappa: $\frac{p_o - p_e}{1 - p_e} = \frac{0.6 - 0.54}{1 - 0.54} = 0.13$

Annotator B

Annotator A		POS	NEG	*total*
	POS	45 (42)	15	60
	NEG	25	15 (12)	40
	total	70	30	N = 100

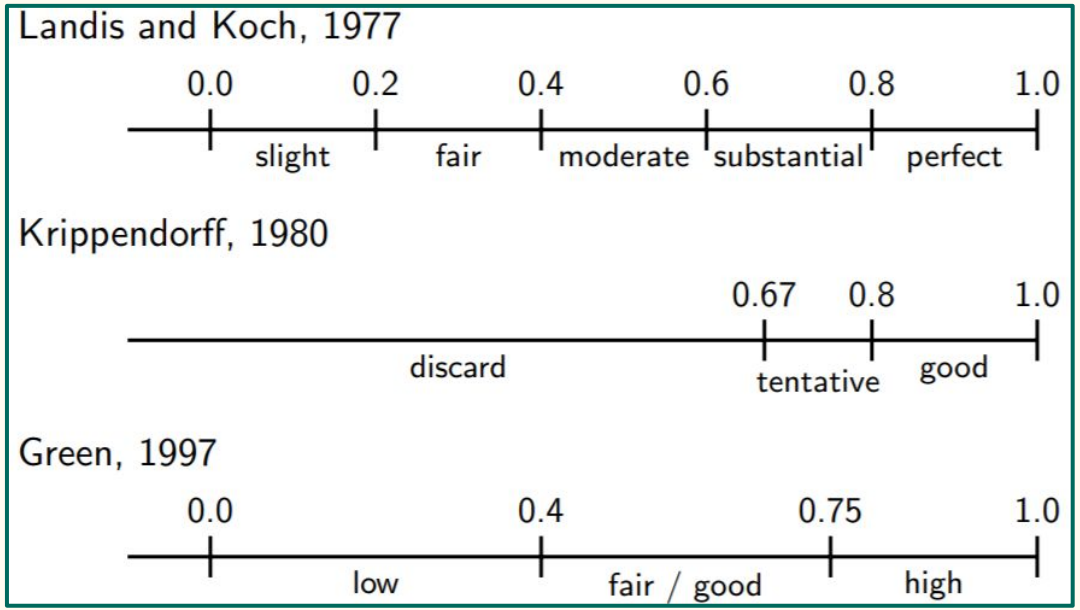
Interpretation of Agreement Metrics

Usually scaled 0.0 - 1.0:
What counts as good?

Differing opinions!

Ultimately, it's made up, so it depends on the task

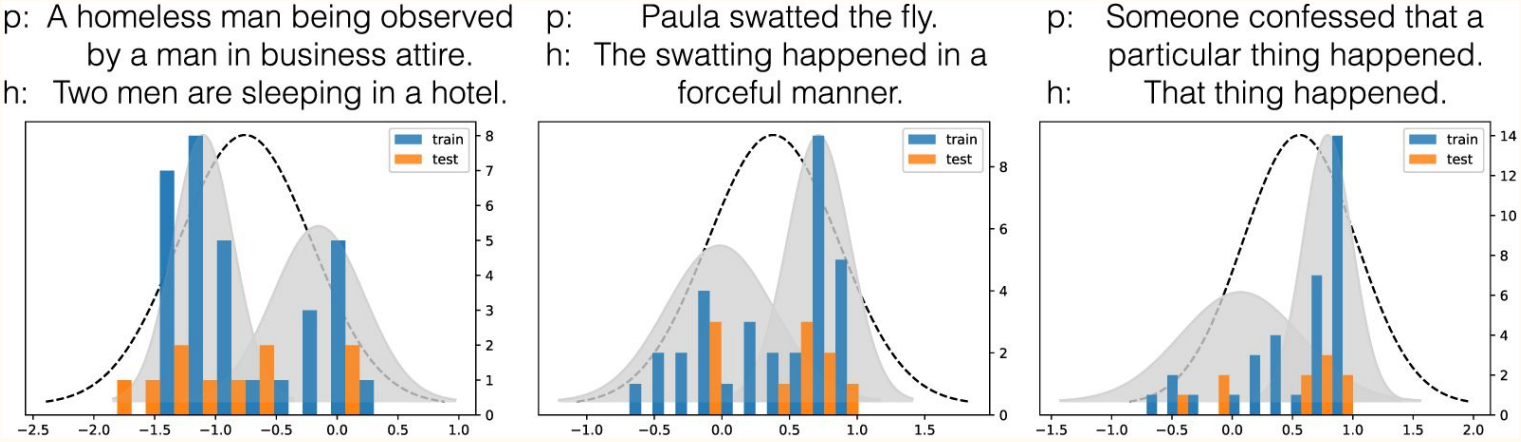
credit Marie Meteer



Disagreement is natural and real!

There are many real-world cases for which we would not necessarily expect or even want perfect agreement

Relevant reading: Inherent Disagreements in Human Textual Inferences



Who are the Annotators?

Especially for more subjective / social tasks
(e.g. hate speech / toxic language detection)

Annotators will assign labels differently based on:

- Demographics (Kuwalty et al. 2020)
- Attitudes or beliefs (Sap et al. 2022)
- Exactly how the question is framed (Jakobsen et al. 2022)

Who are the Annotators?

Therefore always important to ask / report who the annotators are in data collection as thoroughly as possible

... and be thoughtful about what annotators are right for a given task