

FoLT Lecture 1 Summary

PART 1:

- Linguistic Analysis:

I/*Morphology*

II/*Syntax*

III/*Semantics*

IV/*Pragmatics*

V/*Discourse*

I/ Morphology:

Definition

- Morphology is the study of words, how they are built up from smaller meaning-bearing units, and how they are related to other words of the same language.
 - Examples:
 - cat(s)
 - read(ing)
 - (un)break(able)
 - (im)poss(ible)

Two types of morphemes:

1. Free morphemes: can stand alone as words
 - Examples: cat, read, break... (basically stems)
2. Bound morphemes: cannot stand alone as words
 - Examples: -s, -ing, un-, im- (basically affixes)

Types of affixes:

- Examples:
 1. Prefixes: un-, im-
 2. Suffixes: -s, -ing
 3. Infixes: -freakin' -, -freakin' -in-
 4. Circumfixes: en-...-en

Stemming:

- The process of removing affixes from a word to get its stem
 - Examples:
 - cats -> cat
 - reading -> read
 - impossible -> possible

Lemmatization:

- determines that two words have the same root, despite their surface differences
 - Examples:
 - is, are, am -> be
 - car, cars, car's, cars' -> car

II/ Syntax:

Definition

- Syntax refers to the way words are arranged together.
- Syntax == Ordering of sequences of words
 - Example:
cats chase mice **VS** mice chase cats
- Probabilities of sequences of words:
 - Example
cats eat fish **VS** fish eat cats **VS** eat cats fish
- Part of speech of words:
 - Example:
noun, verb, adjective, adverb, preposition, conjunction, interjection

The use of "Part of speech" tagging:

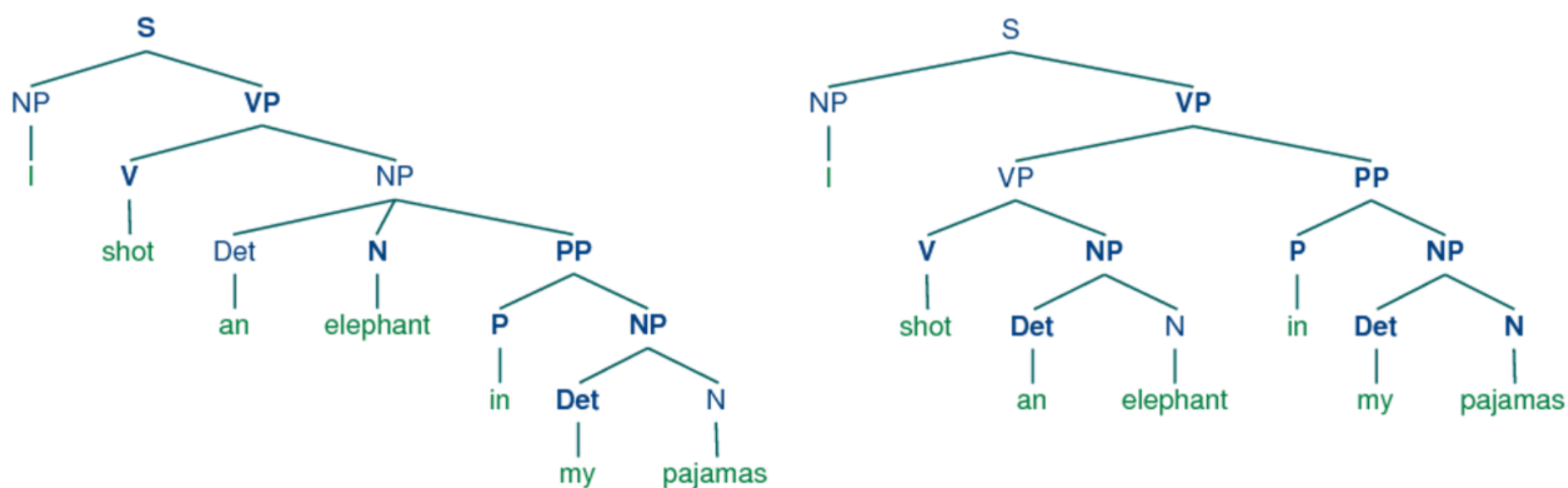
- For NLP tasks:
 - Examples:
sentiment analysis
question answering
text summarization
machine translation etc.
- For linguistic or language-analytic tasks:
 - study linguistic change, eg. new words, new meanings, etc.

Syntactic Parsing:

- The process of analyzing a sentence into its component parts and describing their syntactic roles.
 - Examples:
cats chase mice -> (S (NP (NNS cats)) (VP (VBP chase) (NP (NNS mice))))

Constituent grammar:

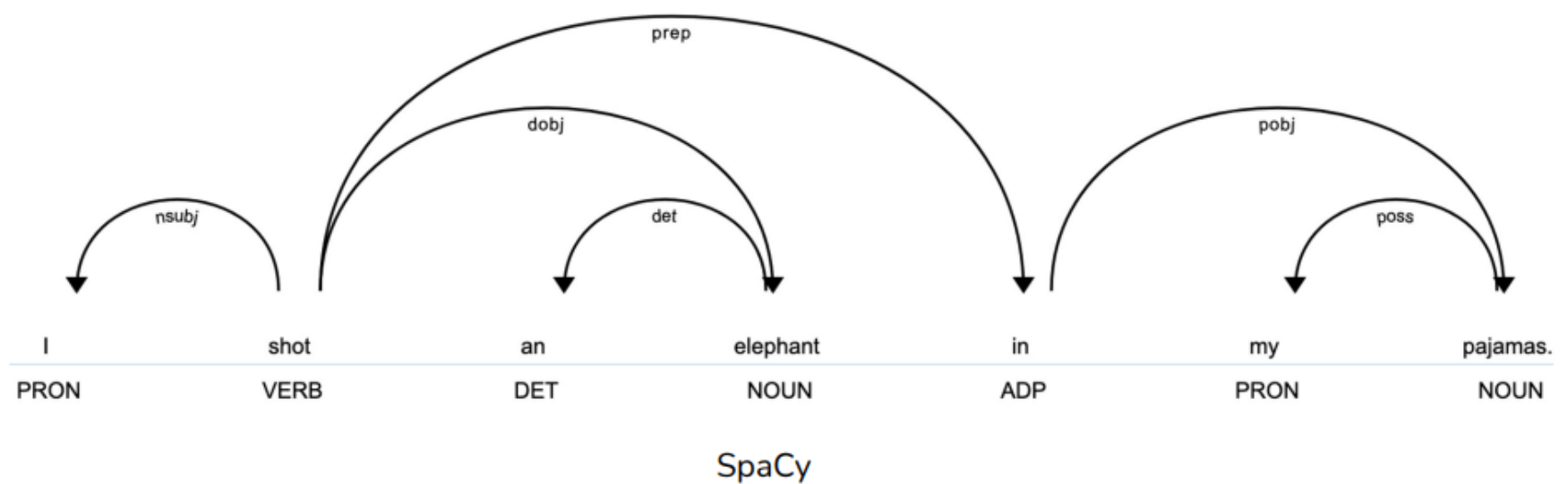
- A grammar that describes the syntactic structure of well-formed sentences.
 - Example:



Dependency grammar:

- Basically the relationship between words in a sentence
- SpaCy uses dependency grammar: <https://spacy.io/usage/linguistic-features#dependency-parse>
 - Spacy Example:

Directed binary grammatical relations between words



Why is parsing important?

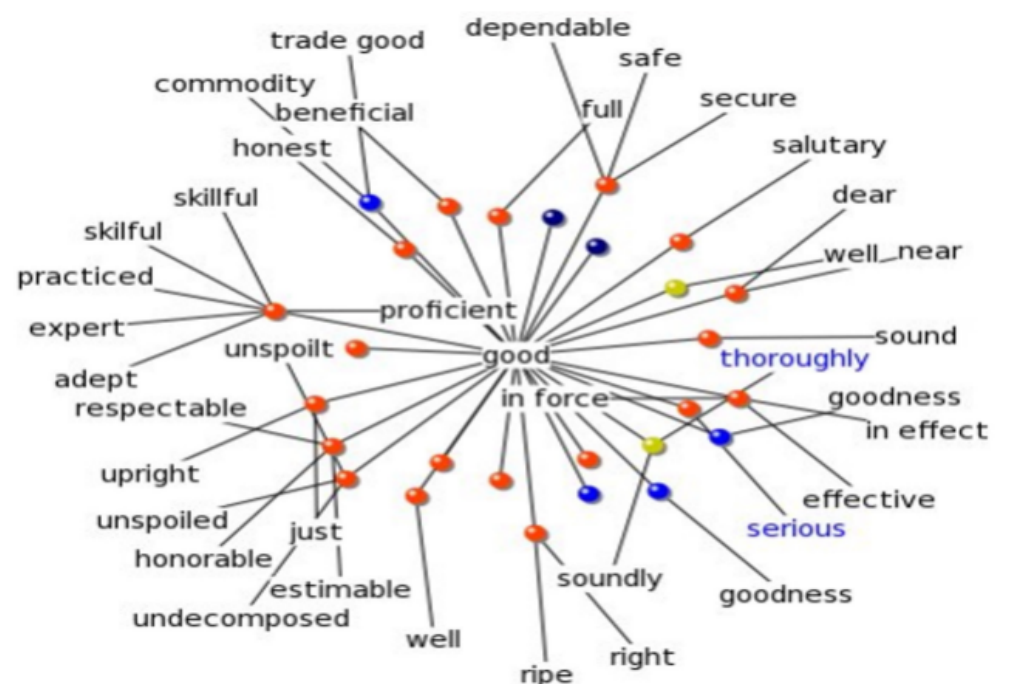
- Grammar checker
- Information extraction
- Question answering
- Machine translation, etc.

III/ Semantics:

Definition:

- Semantics are the study of words, phrases, sentences, or documents.
 - Lexical semantics: the meaning of words; Eg. how close are words in meaning?
 - Semantic role labeling: the meaning of the predicate with respect to its arguments; Eg. who did what to whom?
- Example (WordNet):

WordNet is a semantically-oriented dictionary of English
Similar to a thesaurus, but richer in structure
155,287 words
117,659 synonym sets



IV/ Pragmatics:

Definition:

- Pragmatics are the study of language use in context, and the context-dependence of various aspects of linguistic interpretation.
- Example:
 - Utterance: "It's hot in here."
 - Implicature: "Please open the window."
 - Exchange: Have you seen spider man? I don't like Marvel.
 - Premise: Spider man is from Marvel.
 - Conclusion: I has not seen it and, maybe, does not intend to see it.

V/ Discourse:

Definition:

- A discourse is a coherent structured group of sentences.
- Coherence is the relationship between sentences that makes real discourses different than just random assemblages of sentences.

Example :

- The movie is interesting. → Tim wants to watch it.
- The movie is interesting. × Tim likes pizza.

Types of Coherence:

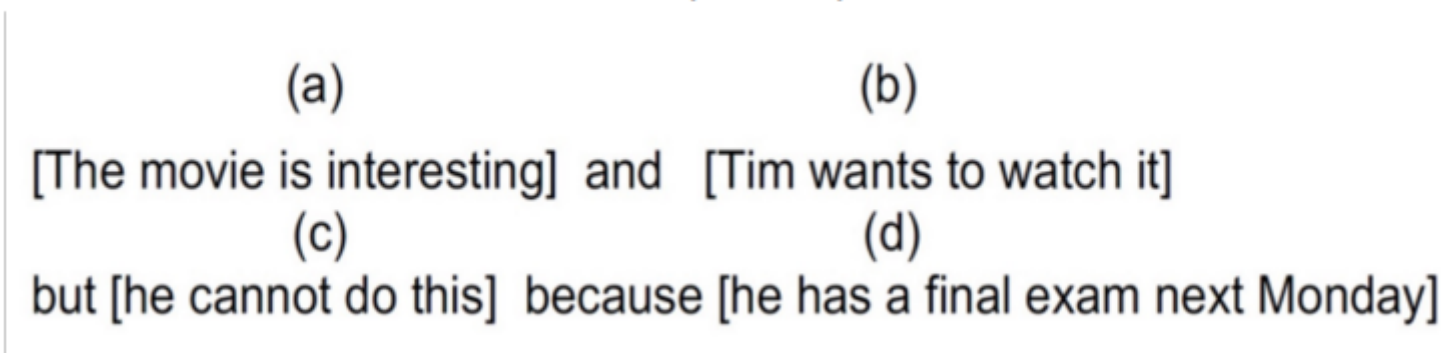
- Local coherence: sentence/clause coherence, entity-based coherence; topical coherence
 - Example:
The movie(1) is interesting. Tim wants to watch it(1).
- Global coherence: conventional discourse structures.
 - Examples:
Academic articles: abstract->introduction->method->result->discussion->conclusion

Discourse parsing:

- The process of analyzing the discourse structure of a text.
- Examples:

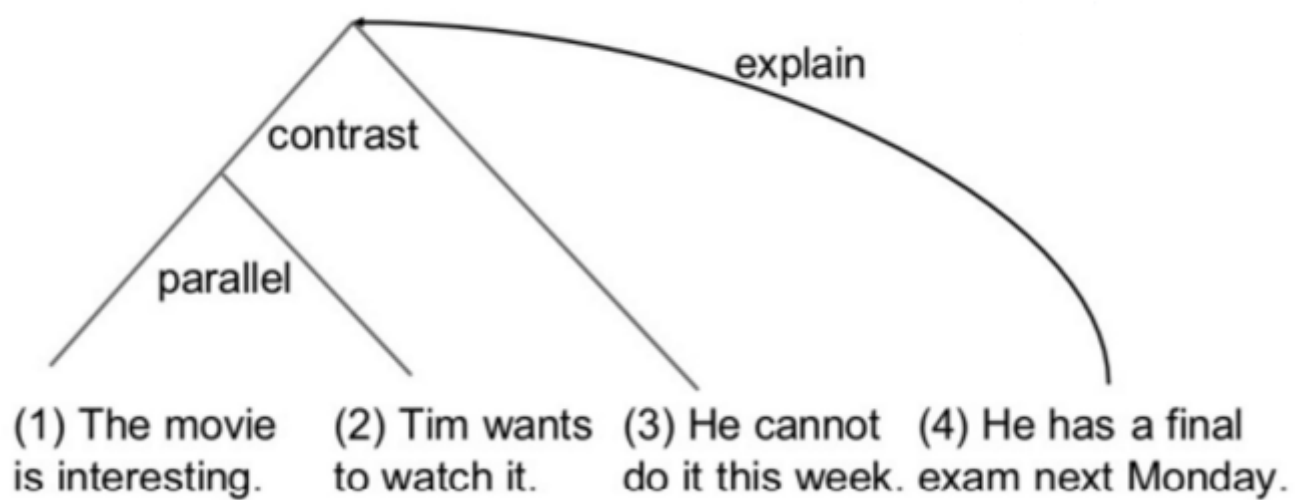
Penn Discourse Treebank (PDTB) with examples: <https://www.seas.upenn.edu/~pdtb/>

Penn Discourse TreeBank (PDTB)



Rhetoric structure theory (RST): <https://www.aclweb.org/anthology/J93-2002.pdf>

Rhetoric structure theory



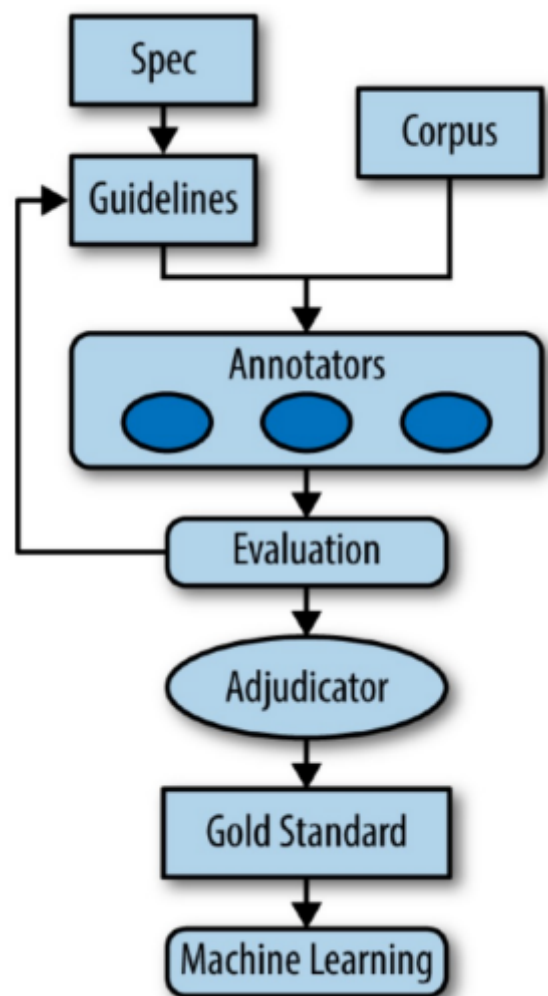
PART 2:

Data Annotation:

- Meta Data about the text: author, title, date, etc.
- User contributed data: comments, ratings, etc.
- Factors in human annotation:
 - source data(genre,size?)

- annotation scheme(guidelines? ambiguities?)
 - annotators(with training?)
 - annotation tool
 - quality control(multiple?), etc.
- Annotation is not easy: any annotation scheme for language will have some difficult cases, grey areas, and ambiguities, because human language needs to be flexible, it cuts corners and is reshaped over time.

Annotation pipeline:



- Annotation Quality: Gold data will have some tarnish, how can we measure it?
 - Inter-annotator agreement: the degree to which two or more annotators give the same annotation to the same data.
 - The agreement rate can be thought of as an upper bound on accuracy of a system evaluated on the same data.
- Validity and Reliability:
 - Validity is the degree to which an annotation scheme measures what it is supposed to measure.
 - Reliability is the degree to which an annotation scheme produces consistent results.
 - Higher reliability is a prerequisite for higher validity.

Measuring agreement:

- Formula for Observed Agreement:

$$\frac{agreement(item1) + agreement(item2) + \dots + agreement(itemN)}{item1 + item2 + \dots + itemN}$$

- Example :

Observed Agreement: proportion of items on which the two annotators agree

	Apple	Orange	Total
Apple	30	15	45
Orange	10	45	55
Total	40	60	100

$$\text{Agreement:} \\ (30 + 45) / 100 = 0.75$$

- Chance Agreement:
 - Some agreement is expected by chance: two annotators are asked to pick Apple and Orange randomly, they might agree with each other half of the time
- Chance-corrected agreement:
 - Agreement beyond chance
 - Observed agreement A_0 : proportion of actual agreement
 - Expected agreement A_e : expected value of A_0
 - Amount of agreement beyond chance: $A_0 - A_e$
 - Maximum agreement beyond chance: $1 - A_e$
 - Proportion of the possible agreement beyond chance: $\frac{A_0 - A_e}{1 - A_e}$

• **How to get expected Agreement A_e**

Expected agreement A_e is the probability of the two annotators c_1 and c_2 agreeing on any given category k

$$A_e = \sum_{k \in K} P(k | c_1) \cdot P(k | c_2)$$

• **How to get Cohen's Kappa**

Cohen's κ assumes the random assignment of categories to items is governed by prior distributions that are unique to each other (annotator bias).

$$P(k | c_i) = \hat{P}(k | c_i) = \frac{\mathbf{n}_{c_i k}}{\mathbf{i}}$$

The actual number of assignment to k by c_i
The number of items

$$A_e^\kappa = \sum_{k \in K} \hat{P}(k | c_1) \cdot \hat{P}(k | c_2) = \sum_{k \in K} \frac{\mathbf{n}_{c_1 k}}{\mathbf{i}} \cdot \frac{\mathbf{n}_{c_2 k}}{\mathbf{i}} = \frac{1}{\mathbf{i}^2} \sum_{k \in K} \mathbf{n}_{c_1 k} \mathbf{n}_{c_2 k}$$

- Cohen's Kappa quality table:

“Good” values are subject to interpretation, but rule of thumb:

0.80-1.00	Very good agreement
0.60-0.80	Good agreement
0.40-0.60	Moderate agreement
0.20-0.40	Fair agreement
< 0.20	Poor agreement

- Cohen's Kappa examples with steps and illustrations:

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

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$

annotator A

		puppy	fried chicken
annotator B	puppy	7	4
	fried chicken	8	81

$$p_e = P(A = \text{puppy}, B = \text{puppy}) + P(A = \text{chicken}, B = \text{chicken})$$

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

P(A=puppy)	15/100 = 0.15
P(B=puppy)	11/100 = 0.11
P(A=chicken)	85/100 = 0.85
P(B=chicken)	89/100 = 0.89

$$= 0.15 \times 0.11 + 0.85 \times 0.89$$

$$= 0.773$$

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

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - 0.773}{1 - 0.773}$$

$$= 0.471$$

annotator A

	puppy	fried chicken
annotator B	puppy	7
	fried chicken	8

annotator A

	puppy	fried chicken
annotator B	puppy	7
	fried chicken	8

- Adjudication:
 - The process of deciding on a single annotation for a piece of text, using information about the independent annotations.

Part 3 (Corpus Statistics) :

Definitions:

- Token: a sequence of characters that we want to treat as a group.
 - Example:
 - "I love you" -> "I", "love", "you"
 - "Learn from yesterday" -> "Learn", "from", "yesterday"
- Tokenization: Segmenting a text into an ordered sequence of tokens.
 - A system which splits texts into word tokens is called a tokenizer.
 - Very simple example:
 - Input text: Learn from yesterday
 - Tokens: {"Learn", "from", "yesterday"}
 - Issues in tokenization: periods don't always mean the end of a sentence, single quotes, Celtics (We're = We are), Multiword Expressions (New York, Rock 'n' Roll)
- How to deal with periods:
 - Common Solution:

Common algorithm:

Tokenize first: use rules or ML to classify a period as either

- (a) part of the word
- (b) a sentence-boundary

An abbreviation dictionary can help