Foundations of NLP

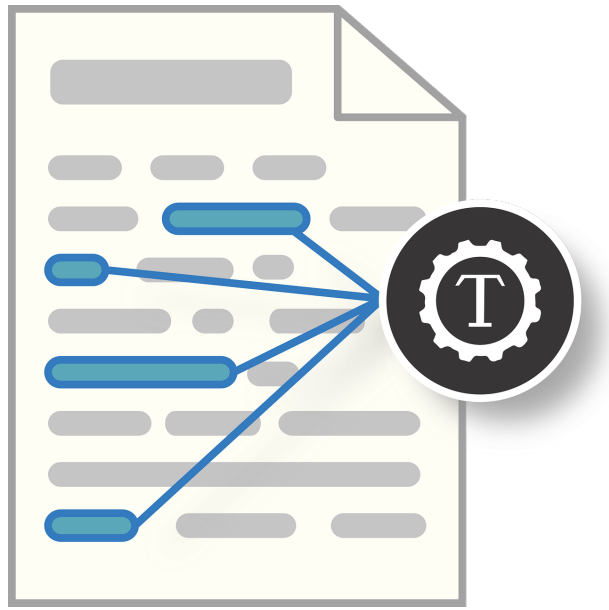
Jelke Bloem

Text Mining
Amsterdam University College

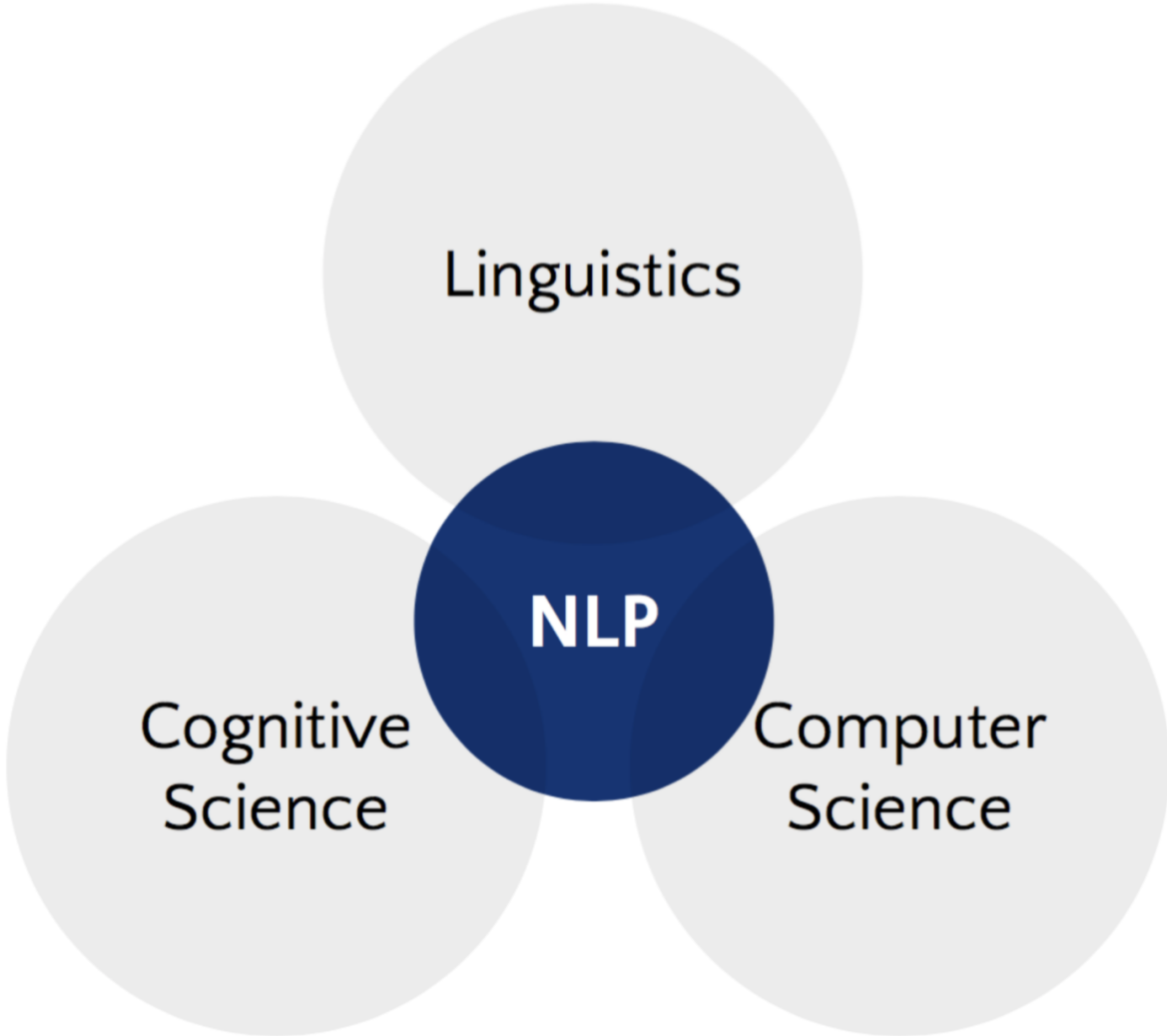
February 18, 2025

What this course is about

Natural Language Processing (NLP) is the set of methods for making human language accessible to computers. The goal of NLP is to perform tasks with language. NLP is at the intersection of linguistics, computer science, engineering and machine learning.



Interdisciplinary topic



Connections

Linguistics $\longrightarrow$ **NLP**: Theory, research methods, data
NLP $\longrightarrow$ **Linguistics**: Testing theory (models), source of empirical evidence, data

Cognitive science $\longrightarrow$ **NLP**: Theory, research methods
NLP $\longrightarrow$ **Cognitive science**: Testing theory (models), metrics

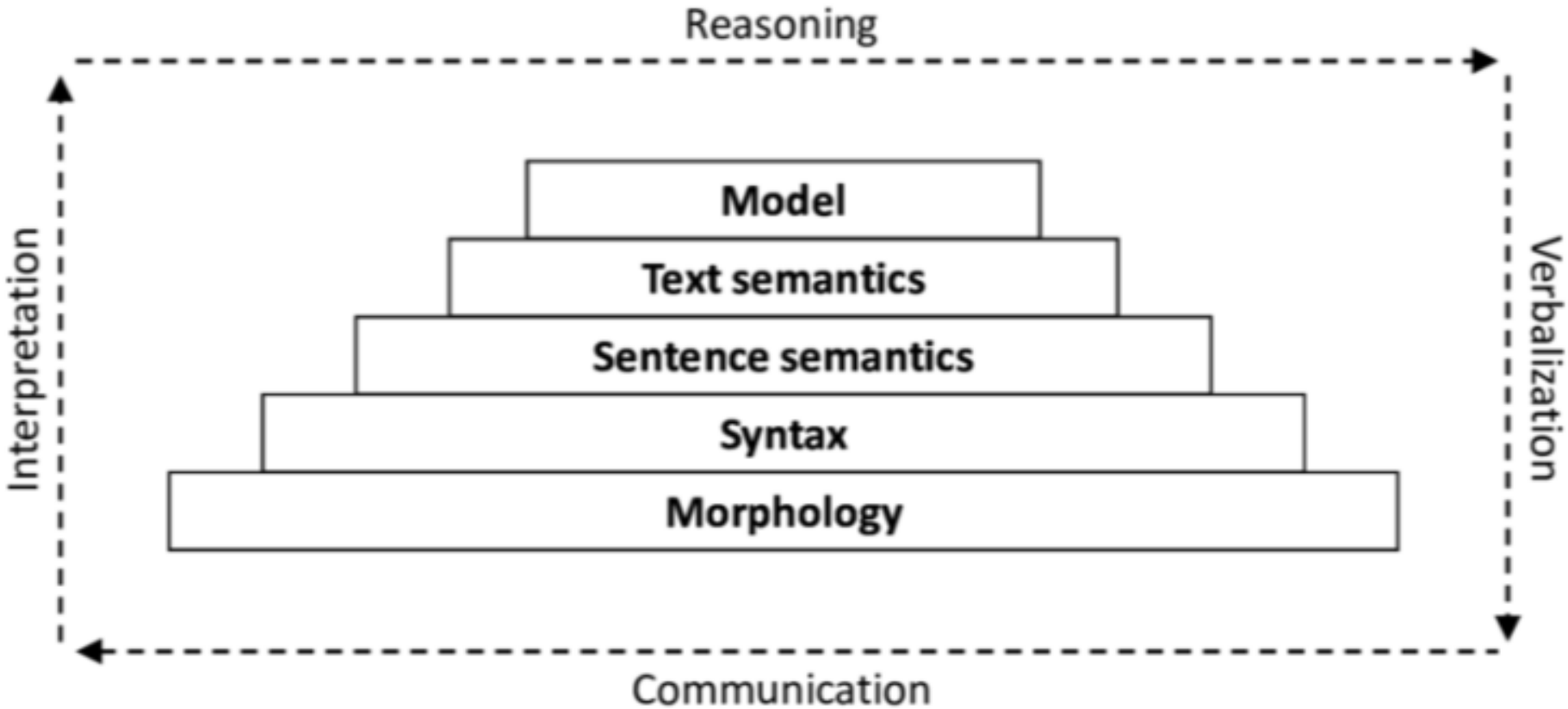
Computer science $\longrightarrow$ **NLP**: methods, systems (big data)
NLP $\longrightarrow$ **Computer science**: use case/application (language)
Machine learning $\Longleftrightarrow$ **NLP**: methods, interesting data type

On machine learning

When compared to other ML data (audio, video, images):

1. Language is discrete
   - Meaning is created by combinatorial arrangements of symbolic units. Unlike images or audio, this is fundamentally discrete. Much language-related data is categorial.
2. Language is compositional
3. Language is skewed
   - Few units are highly frequent and probable, many items have low frequency and low probability, stretching to infinity

Linguistics 101 (M&S, Ch. 3)



Linguistics 101

- **Language is a superset of English**
- There are many possible typological features of languages that English doesn't have
- Modality - spoken or signed?
- Word order - fixed or free?
- Morphology: Isolating/analytic, fusional, agglutinative, polysynthetic
- Many morphosyntactic features, such as null-subject languages
- Obligatory features to express - e.g. English always has tense/time, other languages have evidentiality
- Basic word order - does the subject come first or the object? Noun before adjective?
- Colour terms

Linguistics 101 - Writing systems

- **English is not representative of all writing systems**
- Alphabetic writing systems: symbols represent phonemes
  - Latin alphabet, Cyrillic, Greek, Georgian, Armenian, Hangul (Korea)
- Alphasyllabic writing systems: symbols represent consonants, vowels are diacritics/modifications
  - Devanagari (Hindi), Tamil, Thai, Ge'ez (Ethiopia), Cree (Canada)
- Abjads: symbols represent consonants (vowels optional)
  - Arabic, Hebrew
- Logographic writing systems: glyphs represent words (or morphemes), with some phonetic elements
  - Chinese Hanzi, Kanji (Japan), Cuneiform, Mayan script
- Ideographic writing systems: glyphs represent concepts
  - Aztec, emoji (?)

Even when Latin alphabet is used, many diacritics not used in English may be used (ü, ê, ø)
Transparency of the spelling/orthography may differ (English is weird here)
→ *Notebook 2: section Tokenization*

Morphological level

- Word structures and word formation
- The minimal unit of meaning in a language is called the morpheme

Unfriendliness

1. *un-*: prefix denoting "not being"
2. *-ness*: suffix denoting "a state of being"
3. *-li* (modified from *-ly*): Turns word stem into an adverb
4. *friend*: the stem of this word, also a *free morpheme* (it can be a word on its own)

Bound morphemes (prefixes and suffixes like *-ness*) require a free morpheme to which it can be attached to, and cannot appear as a word on their own

Morphological level

Taken from: Bauer, Laurie (1983:20-21): *English word-formation*. Cambridge: Cambridge University Press.

'Root' and 'stem' are all terms used in the literature to designate that part of a word that remains when all affixes have been removed. Stems still have a clear lexical meaning, roots may not.

A root is a form which is not further analysable, either in terms of derivational or inflectional morphology. It is that part of word-form that remains when all inflectional and derivational affixes have been removed. A root is the basic part always present in a lexeme. In the form 'untouchables' the root is 'touch', to which first the suffix '-able', then the prefix 'un-' and finally the suffix '-s' have been added. In a compound word like 'wheelchair' there are two roots, 'wheel' and 'chair'. In Arabic/Hebrew, a root may be just consonants, like 'ktb' (relating to writing), and many derivatives exist mainly by using different vowels, e.g. in Arabic 'kitab' = book, 'katabtu' = I wrote.

A stem is of concern only when dealing with inflectional morphology. In the form 'untouchables' the stem is 'untouchable', although in the form 'touched' the stem is 'touch'; in the form 'wheelchairs' the stem is 'wheelchair', even though the stem contains two roots.

Morphological level

Wheelchair-s:

- Wheel, chair: roots
- Wheelchair: stem for -s

Untouchable-s:

- Touch: root
- Touchable: base for un-
- Untouchable: stem for -s

→ *Notebook 2: section Stemming and Lemmatization*

Linguistics 101

Word: basic linguistic unit which makes sense (in principle)
"is", "house", "jkqfbi", "ness" (is a morpheme but not a word)

Parts of speech: sets of words with similar syntactic behaviour

- Noun, Verb, Adjective, ...
- Substitution test

We often refer to the **Brown corpus** part-of-speech tags:
`https://en.wikipedia.org/wiki/Brown_Corpus`
→ *Notebook 2: section PoS Tagging*

Lexical level

- Words and their *dictionary meanings* (in common use)
- A language's *lexicon* is a collection of individual *lexemes*
- A lexeme represents the set of *senses*, or meanings (*synset*) taken by an individual stem word
- *Lexical words* are those words with independent meaning (nouns, verbs, adjectives, adverbs and prepositions)
- The lemma is the word form that is chosen to represent the lexeme (e.g, "to be")
- Ambiguity
  - *duck*: noun or verb. Its part-of-speech and lexical meaning can only be derived in context with other words used in the sentence

Lexical level

Word (lexical) semantics

A great resource we will encounter again is WordNet: `https://wordnet.princeton.edu` (a lexical database of synsets/lexemes and their relations)

Example of lexical relations:

- Polysemy (overlap in the synsets of different lemmas)
- Hyper-/hyponymy (more general, more specific sense)
- Collocations: frequently co-occurring words (e.g., "white whine", "New York", "couch potato")

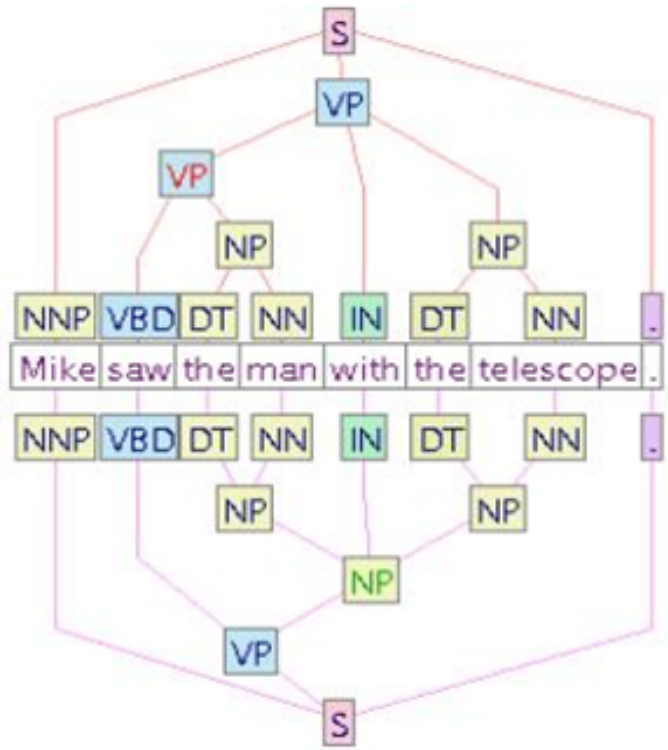
WordNet example
→ *Notebook 2: section Stemming and Lemmatization*

Syntactic level

Considers the structure of groups of words into phrases and sentences. *Syntactic analysis* or *parsing* allows to determine if and how a sequence of words is well-formed.

I saw the man with the telescope. // With man the the I telescope saw.

Ambiguity: there are usually many legit ways to parse the same sentence.

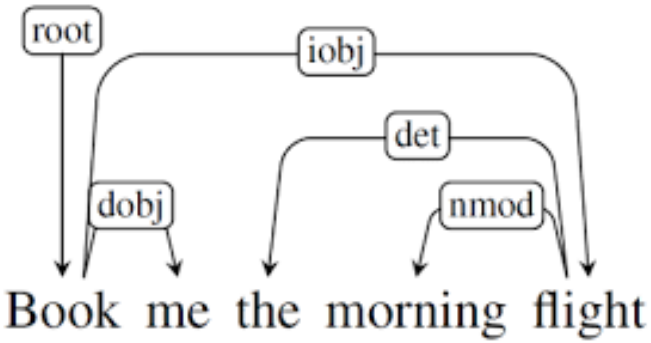


Syntactic structure

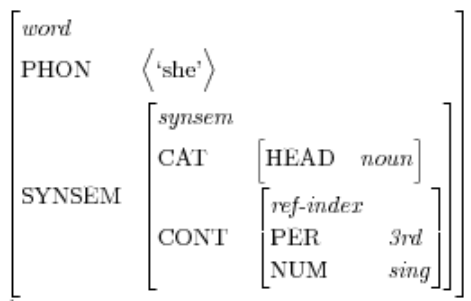
- **Phrases**: groups of words clumped together into units.
  - A sentence is a self-contained set of phrases, most often ending with a full stop.
- **Syntax** studies phrase structure using parse/dependency trees (also called "derivations" or "parses").
- **Noun phrase**: all about a noun, e.g., *The old friend of mine that I called yesterday.*
- **Verb phrase**: same, but about a verb.
- Note **compositionality**: the example above contains 2 NP and 1 VP.

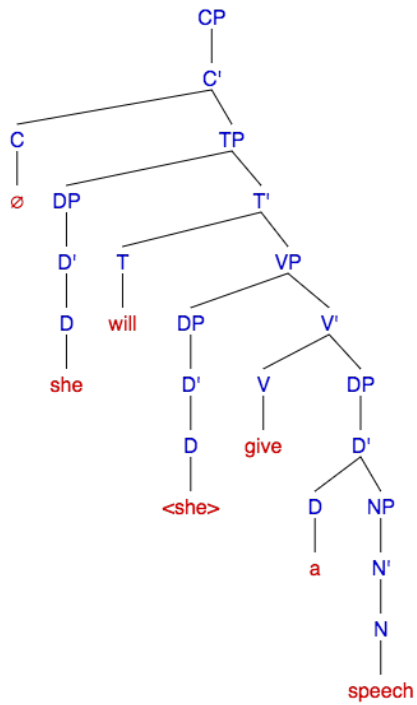
Syntactic level

There are many different theories of syntax and corresponding formalizations.



- Dependency parsing
- Head-driven phrase structure grammar (HPSG)
- Lexical functional grammar (LFG)
- Minimalist syntax
- Construction grammar (CxG, e.g. FCG)



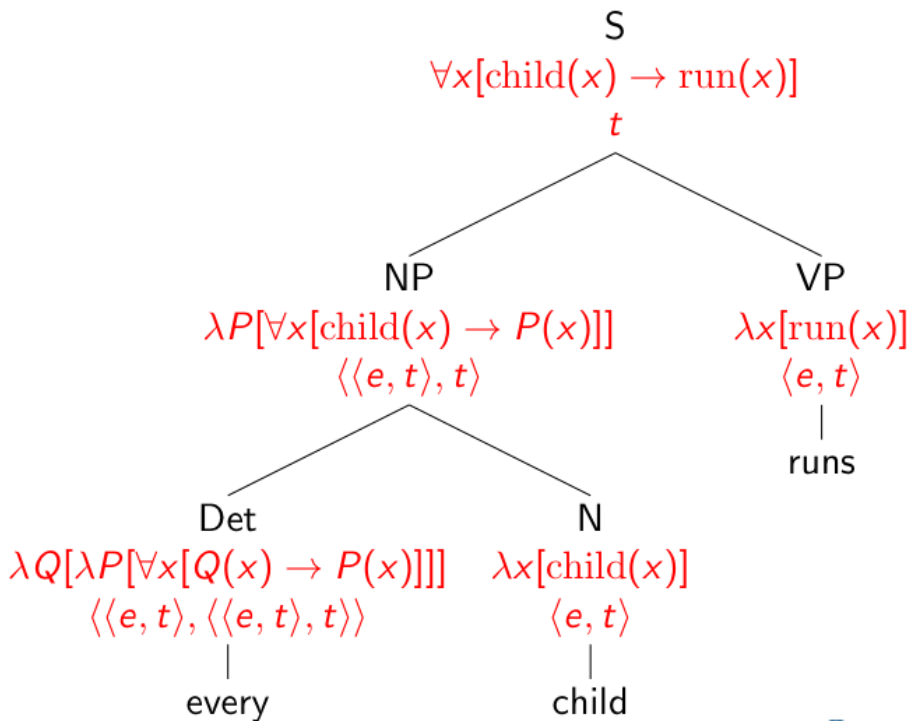


Semantic level

Deals with the meaning of words in context (also called lexical semantics) and of whole sentences. E.g., *word-sense disambiguation*.

The soldier ducked in order to avoid friendly fire.

Can be modeled through formal logic, or composing word representations learned from data



Discourse level

Deals with the analysis of structure and meaning of a text beyond a single sentence.

- *Anaphora resolution* (most commonly in the form of, but not limited to, a pronoun).
  - Take your onion and remove the outer skin. Take your knife and chop **it** in half.
- *Metadiscourse* (the verbal expressions that highlight the discourse structure or the writer's/speaker's interaction with the addressees)
  - *and, because, firstly, in addition, perhaps*

Pragmatics level

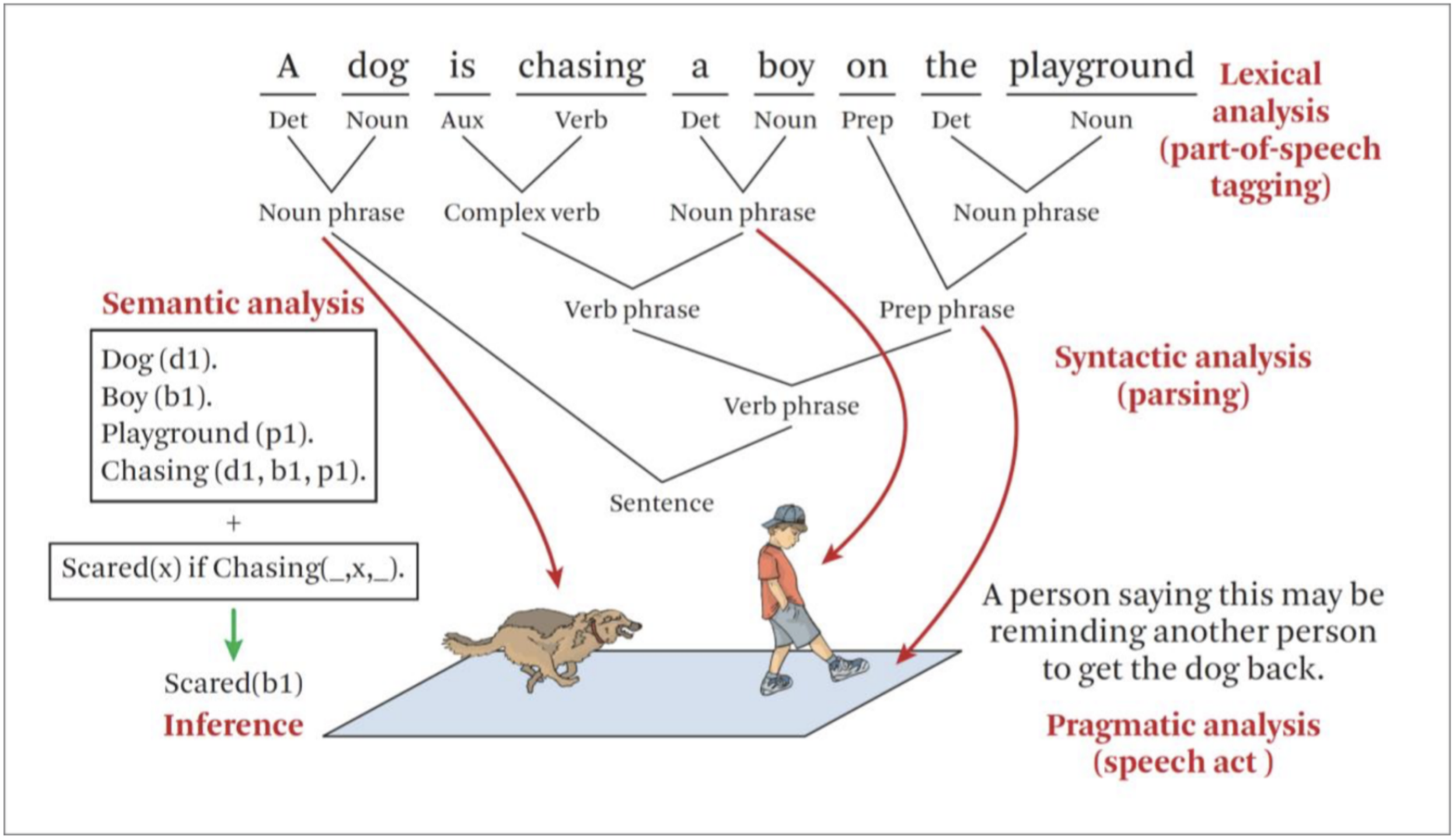
Deals with the real-world use of language in a given context or situation.

- *Speech acts*: Language that performs an action
  - Please pass me the chili sauce
  - May you live an interesting life
- *Implicature*: Information that is not literally expressed
  - *"Do you know the time?"* *"Yes"*
  - Violates cooperative principle
- *Conversation analysis*: Structures of social interaction
  - Turn-taking structure
  - Interruptions
  - Repair

In summary: Meaning and compositionality

- Lexical semantics: meaning of words in isolation
- Compositional semantics: meaning of sentences
- Language in context: meaning of dialogues and discourses

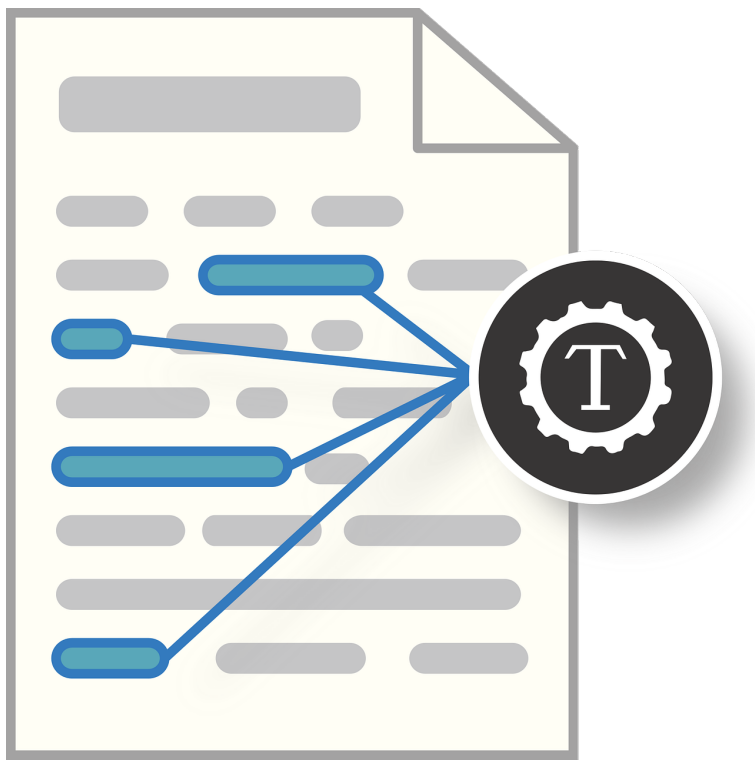
Examples of NLP applications



(Zhai & Massung 2016)

Examples of NLP applications

`http://nlpprogress.com`



Brief history of NLP

- Seminal ideas: *Translation memorandum* (Weaver, 1949); *Computing machinery and intelligence* (Turing, 1950, introducing the Turing test)
- Dartmouth workshop (1956, Minsky, Shannon, McCarthy, Simon, Solomonoff)

  "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. **An attempt will be made to find how to make machines use language, form abstractions and concepts**, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer."

- AI winters: *Perceptrons* (Minsky and Papert, 1969); Lighthill report (1973); challenges to fit neural networks; expert systems
- Classic machine learning: from the late 1990s
- "Deep learning": 2010s
- "Generative AI": 2020s

Two paradigms

Symbolic

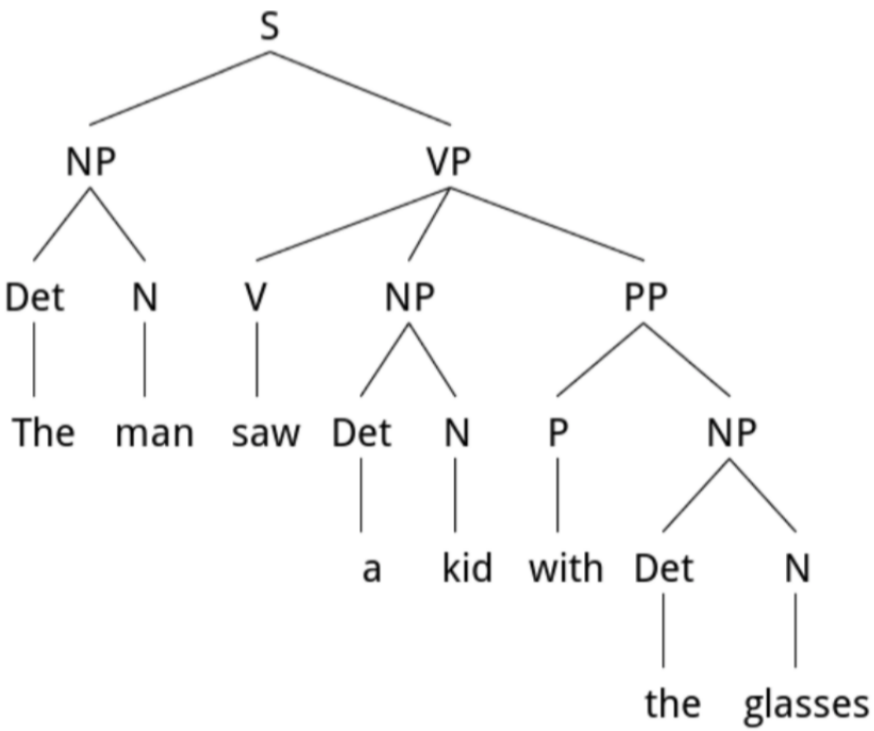
- Driven by theory
- Denotational view, *WordNet*
- Compositional, focused on rules/grammars
  - Lambda calculus (Church)
  - Formal grammars (Chomsky)

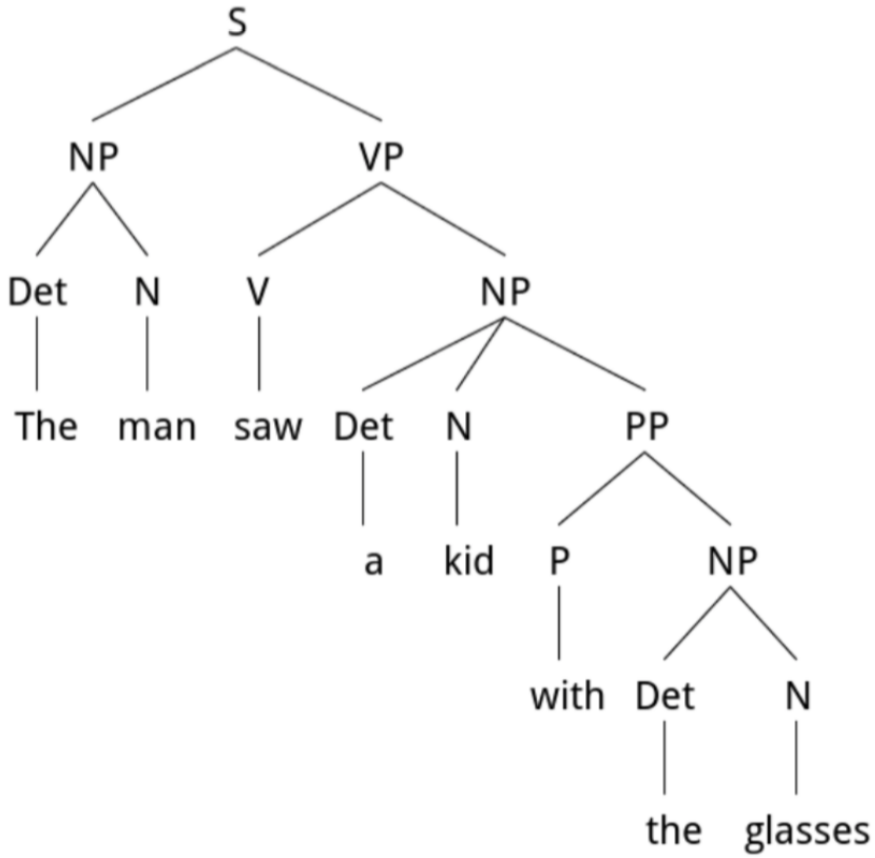
Probabilistic

- Driven by data
- Distributional view
  - *You shall know a word by the company it keeps* - Firth
  - *The meaning of a word is its use in the language* - Wittgenstein
- Statistical learning

Both agree there are **rules** in language that we need to understand and study, but differ in the approach they take. Grounded in the cognitive debate around **rationalist/innate** vs **empiricist/acquired** perspectives on language.

What symbolic approaches can handle

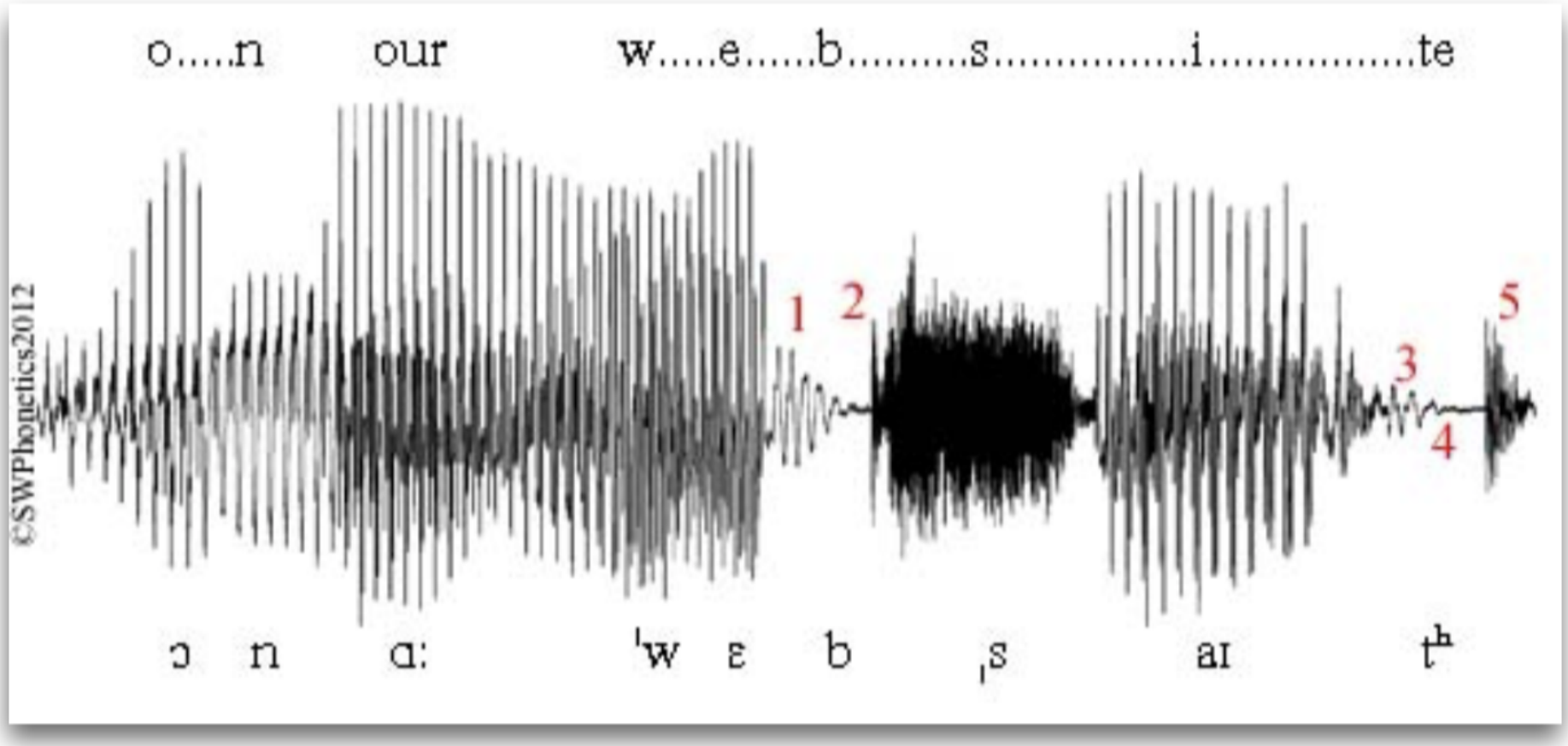




What symbolic approaches *can't* handle

Ms James:	when they went in there again
	it was all blocked up wasn't it
	they [couldn't get through
Sean:	[*((nods emphatically))*
Ms James:	why is that
	why- why did that happen (.)
	this group why do you think that happens
Sean:	becau:[s:e
Ben:	[erm
Ms James:	right Sean was just about to say something
	(3)
Sean:	I don't know
Ms James:	you're not sure
	o:ka:y
	maybe you can add something in a moment
	Ben
Ben:	I think

What symbolic approaches *can't* handle



What symbolic approaches *can't* handle

- Scale
- Incompleteness and uncertainty
  - Novel/unseen words
  - Generalization
- Continuity
- Performing well in the wild only on controlled tasks
- Not robust

Language and probability

- Probabilistic approach to grammar and other tasks
- Deterministic production rules are substituted by rules equipped with probabilities
- Probabilities can be estimated from (text) data by observing frequencies
- Problem of vocabulary skewness
- Grammaticality

Language and probability

```
S → NP VP
NP → Det N
VP → V NP PP
PP → P NP
V → saw
N → man | kid | glasses
Det → the | a
P → with
```

S → NP VP	1.0
NP → Det N	0.5
VP → V NP PP	0.3
PP → P NP	0.9
V → saw	0.01
N → man | kid | glasses	0.3
Det → the | a	0.6
P → with	0.02

Corpora

- For all these data-driven methods, we need data, provided by corpora.
- Corpora can be written text, or transcripts of speech, providing a sample of naturally occurring language.
- A corpus should be representative of the variations of a language phenomenon we want to model. Corpora are often used for *shared tasks (leader boards)*.
- In reality, corpora are incomplete and potentially flawed.
- Consequence: know your corpus (and its limitations)!

Corpora

The more data we can analyze, (usually) the better.

- Corpus size went from a few million tokens in the 60s and 70s to several billion of tokens today.
- The Brown corpus (1964) was the first machine-readable corpus, with about 1 million tokens, all English-language text.

Modern examples:

- Google Books ngrams `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`
- TenTen Web corpora `https://www.sketchengine.eu/documentation/tenten-corpora/`
- Many more: I will make a list in project datasets file

An uneven playing field: bigger datasets require larger compute resources.

For the next class

- Please make sure you have done all the previous labs or are comfortable with their contents
- Assignment 1 will be available soon (after lab)
- Reading assignment 1 (in-class test) on **Friday** - read before class

References

- **E, ch. 1**
- **MS, ch. 1 and 3**