Introduction to Natural Language Processing

Part II: Basics of Linguistics

Henning Wachsmuth

https://ai.uni-hannover.de

Learning Objectives

Concepts

- Several fundamental phenomena in natural language
- The different levels of language
- The complexity of natural language processing

Methods

- A first overview of common NLP tasks
- Preparation for processing natural language text

Notice

- While several of the introduced concepts exist in many or all languages, the focus is largely on English here.

Outline of the course

I. Overview

II. Basics of Linguistics

- Introduction
- Morphology
- Syntax
- Semantics
- Discourse
- Pragmatics

III. NLP using Rules

IV. NLP using Lexicons

V. Basics of Empirical Methods

VI. NLP using Regular Expressions

VII. NLP using Context-Free Grammars

VIII. NLP using Language Models

IX. Practical Issues

Introduction

Linguistics

Linguistics

- The study of spoken and written natural language(s) in terms of the analysis of form, meaning, and context

Levels of spoken language only

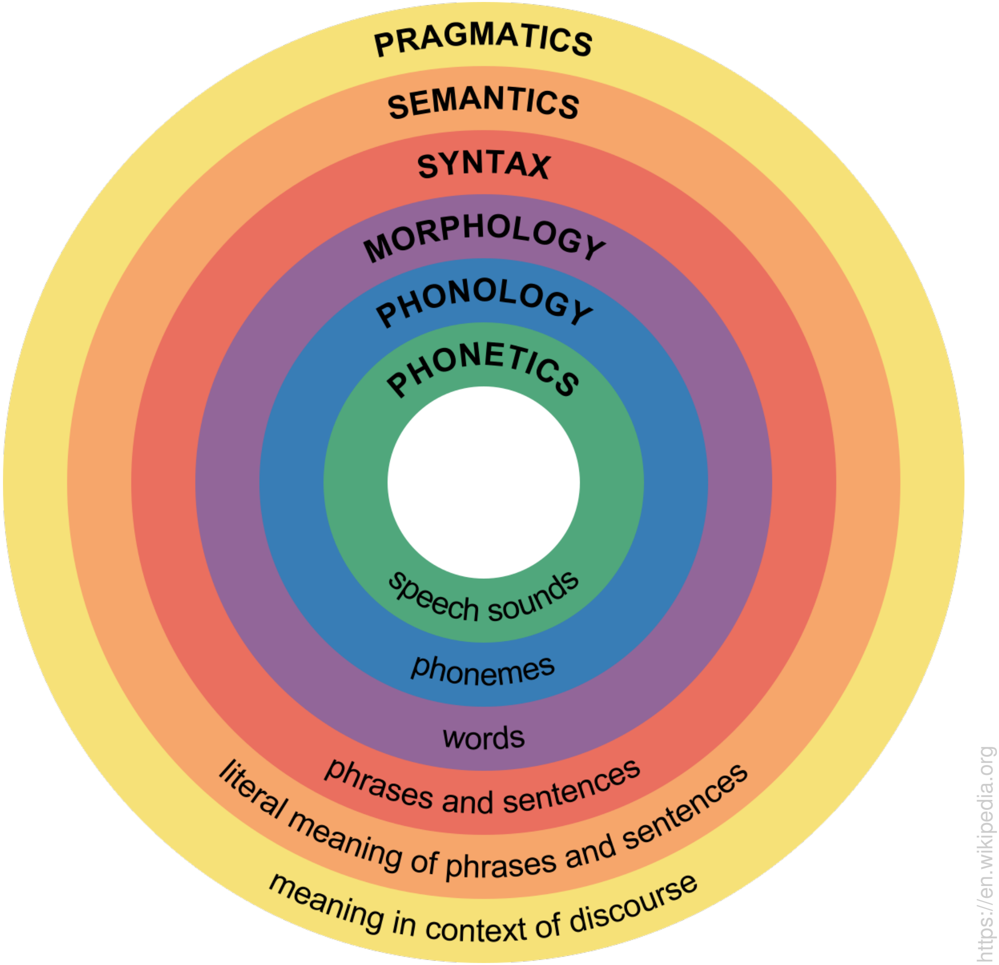
- Phonetics. Physical aspects of speech sounds
- Phonology. Linguistic sounds of a particular language

Levels of spoken and written language

- Morphology. Senseful components of words and wordforms
- Syntax. Structural relationships between words, usually in a sentence
- Semantics. Meaning of single words and compositions of words
- Discourse. Linguistic units larger than a single sentence
- Pragmatics. How language is used to accomplish goals

Linguistics

Levels of Language



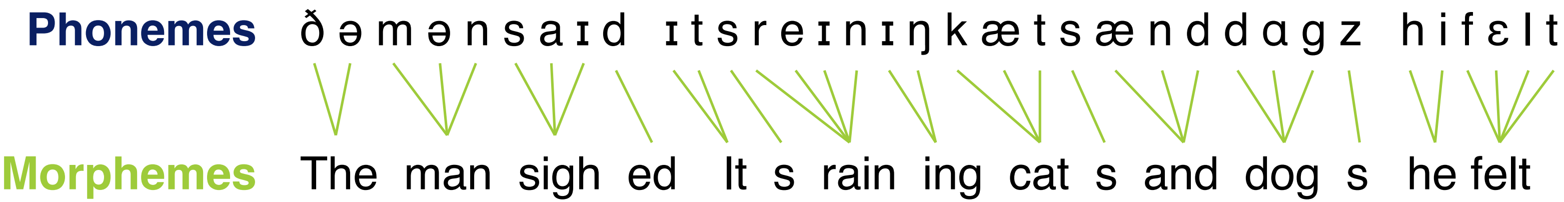
(*discourse* is on the boundary between semantics and pragmatics)

Linguistics

Spoken vs. Written Language

Basic linguistic units

- Phoneme. Smallest unit of spoken language (≈ one linguistic sound)
- Morpheme. Smallest unit with a meaning or grammatical function in both spoken and written language



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Focus on written language

- Natural language is processed computationally mostly in text form.

  Where given, speech is transcribed to text before, or vice versa.
- Phonetics and phonology are largely disregarded in NLP, and they will play only a small role in this course.

  This simplifies some aspects, but also has drawbacks. Which one?

Linguistic Text Units

Language levels of units

- Morpohological level. Characters, syllables, morphemes, words
- Syntactic level. Phrases, clauses, sentences
- Discourse level. Paragraphs, larger discourse units

Ordered by Size

- All paragraphs contain

  ≥ 1 sentences which contain

  ≥ 1 clauses which contain

  ≥ 1 phrases which contain

  ≥ 1 words which contain

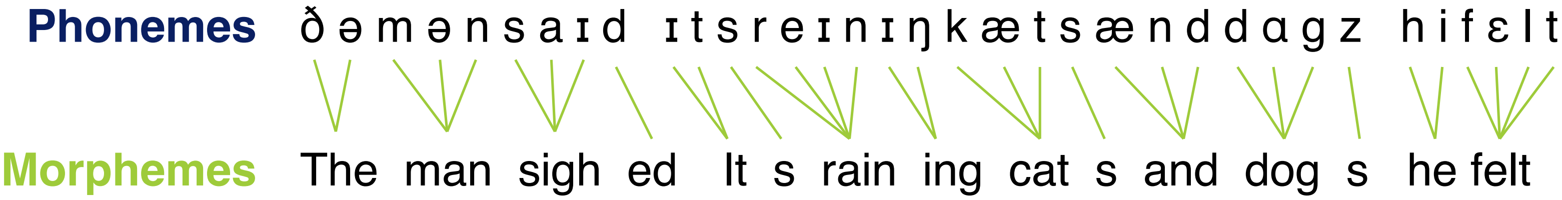
  ≥ 1 $\{$morphemes | syllables$\}$ which contain

  ≥ 1 characters

Morphology

Linguistic Text Units

Morphemes



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Morphemes

Morpheme

- The smallest lingustic unit with a meaning or grammatical function
- Corresponds to a character, syllable, word, or something in between

  Differs both within and across languages

“cats” → “cat” + “s”

“felt” → “felt”

Morphemes vs. syllables

- Syllables can be seen as the phonological building blocks of words.
- Similar concepts, but often lead to different word decompositions

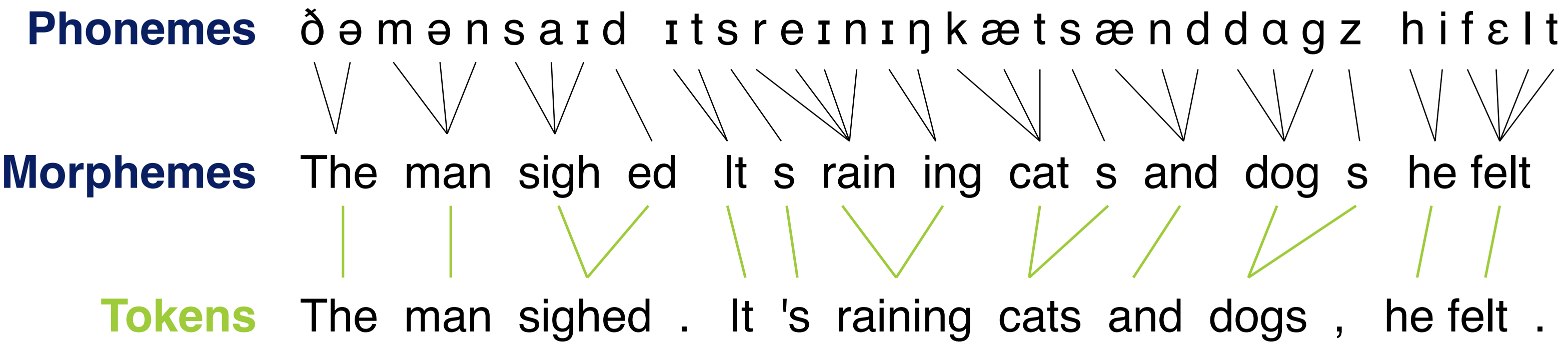
“speaker” → “speak” + “er” (morphemes) vs. “spea” + “ker” (syllables)

Free and bound morphemes

- “cat” can be uttered in isolation, but “s” cannot.
- “cat” is a free morpheme, “s” is bound.
- Free morphemes can be used as words, bound morphemes cannot.

Linguistic Text Units

Words and Tokens



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Tokens The man sighed . It 's raining cats and dogs , he felt .

Words

Word

- The smallest unit of language that is to be uttered in isolation
- Words have either a lexical function (open-class words) or a grammatical function (closed-class words).
- Every word is composed of one or more morphemes.

"cat" → "cat"

"cats" → "cat"+"s"   "unknowingly" → "un"+"know"+"ing"+"ly"

- The term *word* is used to refer to both *lemmas* and *wordforms*.

Words vs. characters

- A character is the smallest graphical unit of written language.
- May be a letter, digit, space, punctuation, special symbol, or similar
- In some languages, characters represent complete words (or syllables).

猫

Chinese "cat"

Words

Lemmas and Wordforms

Lemma

- The dictionary form of a word

  A related term is *lexeme*, i.e., the unit of meaning of a word irrespective of its form.

“be”, “am”, “was”, ... → “be”

“deriving”, “derives”, ... → “derive”

Wordform

- The fully inflected surface form of a lemma as it appears in a text.
- Mostly consists of one *stem* and zero or more *affixes*.

“am” → “am”

“derives” → “deriv” + “es”

- Bound base. Alternative to a stem, requiring an affix, such as “-ceive”
- Contracted form. Wordforms shortened by an apostrophe, such as “’s”

Words

Stems and Affixes

Stem

- The part of a wordform that never changes

"cat" and "catwalk", but not "cats"

- Usually carries the main meaning of a word
- Often composed of free morphemes, but not always, such as in "derive"

A related term is *root*, i.e., a minimal free morpheme, such as "cat" and "walk".

Affix

- Any bound morpheme
- Affixes add meanings of various kinds to the stem
- Affix types. Suffix, prefix, infix, and circumfix

cat + "s"

"pre" + "conceptions"

"fan" + "bloody" + "tastic"

"em" + "bodi" + "ed"

Words

Inflection and Derivation

Inflection

- The modification of a word to express different grammatical functions, such as tenses, cases, numbers, persons, ...

“derive” → “derived”

Derivation

- The modification of a word to obtain a new word

“derive” → “derivation”

Inflection vs. derivation

- Both inflection and derivation usually add affixes.

  Partly dropping letters that do not belong to a stem

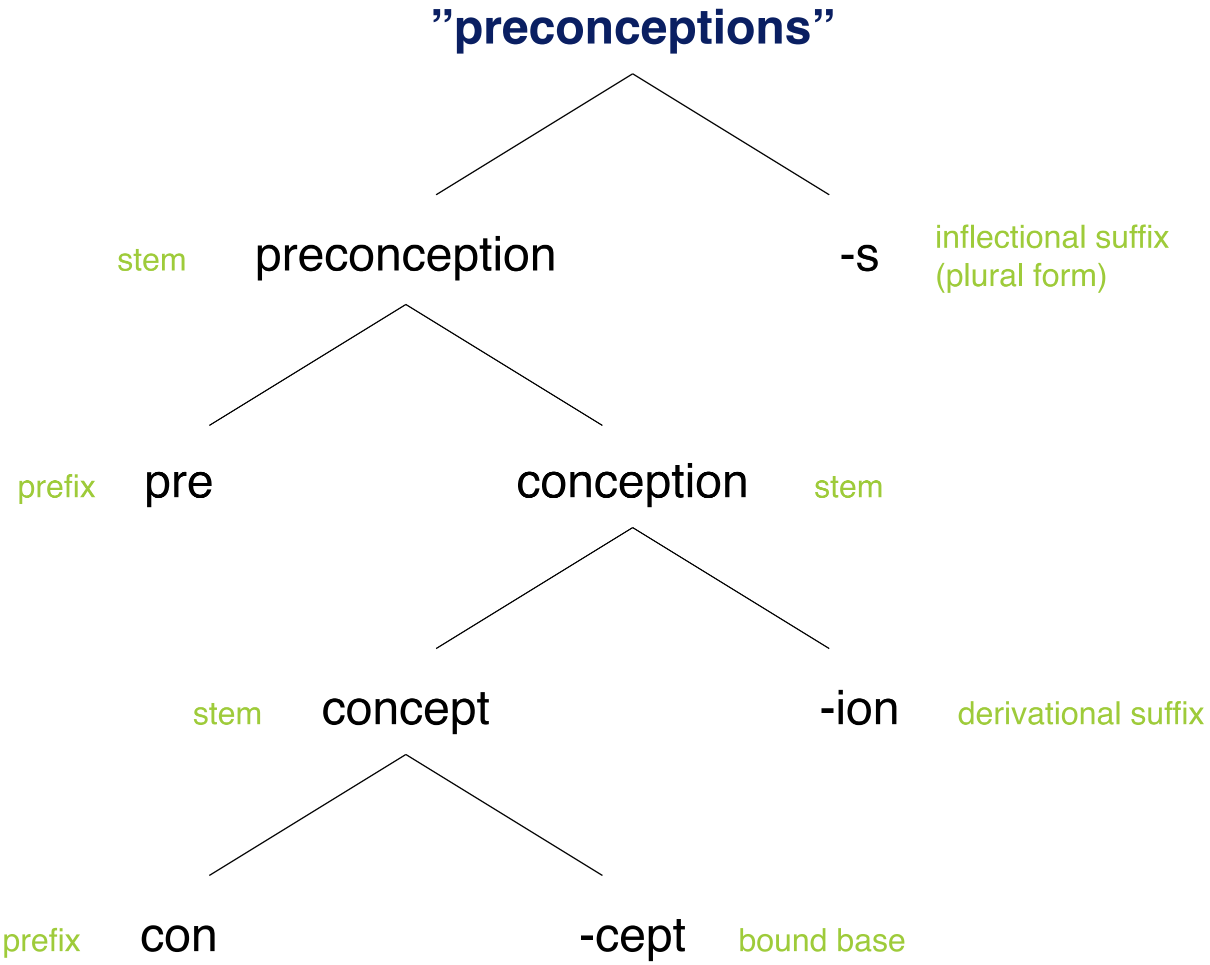
- Only inflection includes cases without affixes.

“be” → “am”

“mouse” → “mice”

Words

Morphological Analysis



”preconceptions”

stem preconception

-s inflectional suffix (plural form)

prefix pre

conception stem

stem concept

-ion derivational suffix

prefix con

-cept bound base

Tokens

Token

- A wordform, a number, a symbol, or similar
- The smallest text unit usually analyzed in NLP
- Whitespaces are *not* considered as tokens themselves.

Example tokens

- Simple cases. “The”, “the”, “sighed”, “sigh”, “42”, “-”, “–”, “‡”
- Complex cases. “i.e.”, “42.42”, “4 242”, “https://ai.uni-hannover.de’
- Other-language cases. “aujourd’hui” is one token, “本を読む” are four
- Controversial cases. “is”+“n’t”, “42%”, “argument-based”

  Usually favored over: “isn’t”, “42”+“%”, “argument”+“-”+“based”

Tokenization

- The text analysis that segments a span of text into its single tokens
- Used in NLP as one of the most basic preprocessing steps

Morphology

Morphological Normalization

Morphological normalization

- Identification of a single canonical representative for morphologically related wordforms
- Reduces inflections (and partly also derivations) to a common base
- Used in NLP to identify different forms of the same word

Normalization methods

- Stemming. The text analysis that identifies the stem of a token
- Lemmatization. The text analysis that identifies the lemma of a token

Stemming vs. lemmatization

- Many tokens will be reduced to the same form, but not all.

“derive” → “deriv” (stem) vs. “derive” (lemma)

“am” → “am” (stem) vs. “be” (lemma)

Morphology

Words Go Wild

German is notorious for its compounds

- “Lebensversicherungsgesellschaftsangestellter”
  “life assurance company’s employee”
- Side comment. The real specialty of German is the *ad-hoc* compound.

English is *not* free of compounds

- “catwalk”, “girlfriend”, ...
- “pneumonoultramicroscopicsilicovolcanoconiosis”
  lung disease caused by the inhalation of very fine silica dust found in volcanoes

Turkish is an agglutinative language

- “uygarlaştıramadıklarımızdanmışsınızcasına”
  “(behaving) as if you are among those whom we could not civilize”
- uygar laş tır ama dık lar ımız dan mış sınız casına
  civilized + BEC + CAUS + NABL + PART + PL + P1PL + ABL + PAST + 2PL + AsIf

Syntax

Syntax

Syntax

- The structural relationships between words, usually within a sentence (or a similar utterance)
- Regularities and constraints of word order and phrase structure
- The syntax of a language is defined by a grammar.

  In principle. Actually, we fail to write down complete grammars for natural languages.

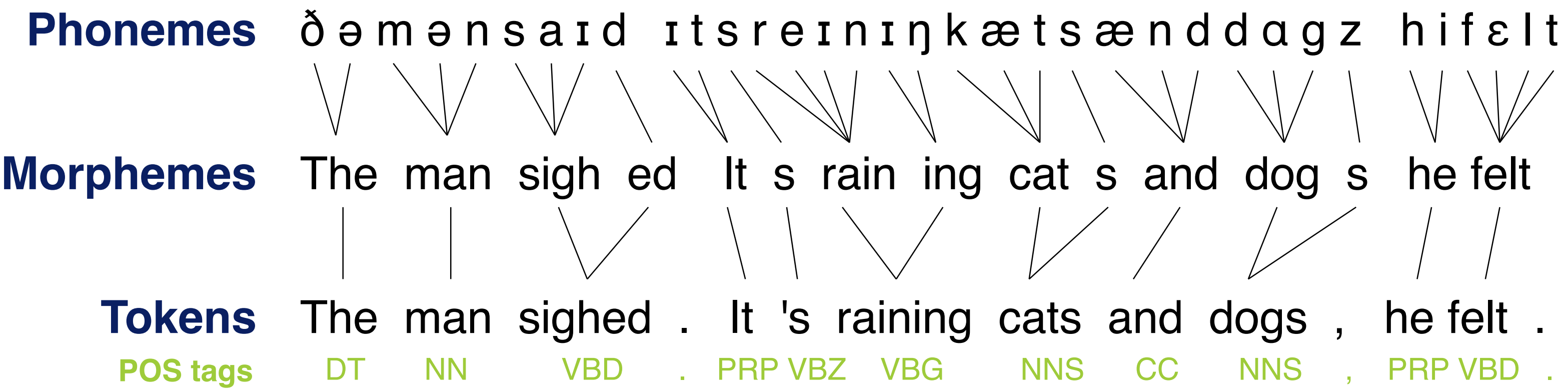
Structural relationships

- Part-of-speech. The class of a word is decided by its syntactic context

  Part-of-speech is on the boundary between morphology and syntax.
- Phrases. Sequences of words build meaning units
- Clauses. Grammatical units that express complete propositions
- Sentences. Grammatically-independent linguistic units

Linguistic Text Units

Parts of Speech



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Tokens The man sighed . It 's raining cats and dogs , he felt .

POS tags DT NN VBD . PRP VBZ VBG NNS CC NNS , PRP VBD .

Parts of Speech

Part of speech

- The lexical category of a word (also called *word class*)
- Abstract classes. Noun, verb, adjective, adverb, preposition, pronoun, conjunction, interjection, determiner.

Part-of-speech (POS) tags

- For analysis, more fine-grained (partly language-specific) word classes are considered, represented as token-level tags.

  Different tagsets exist, usually with 30–60 tags. Here, we use the PENN tagset.

  ”apple” (single noun, NN), ”apples” (plural noun, NNS), ”Apple” (proper noun, NNP), ”sigh” (verb base form, VB), ”sighed” (verb past tense *or* past participle, VBD or VBN), ”the” (determiner, DT), ”it” (personal pronoun, PRP), ”WHATZ” (???), ...

POS tagging

- The text analysis that assigns a part-of-speech tag to each token
- Used in NLP as a preprocessing step for several other analyses

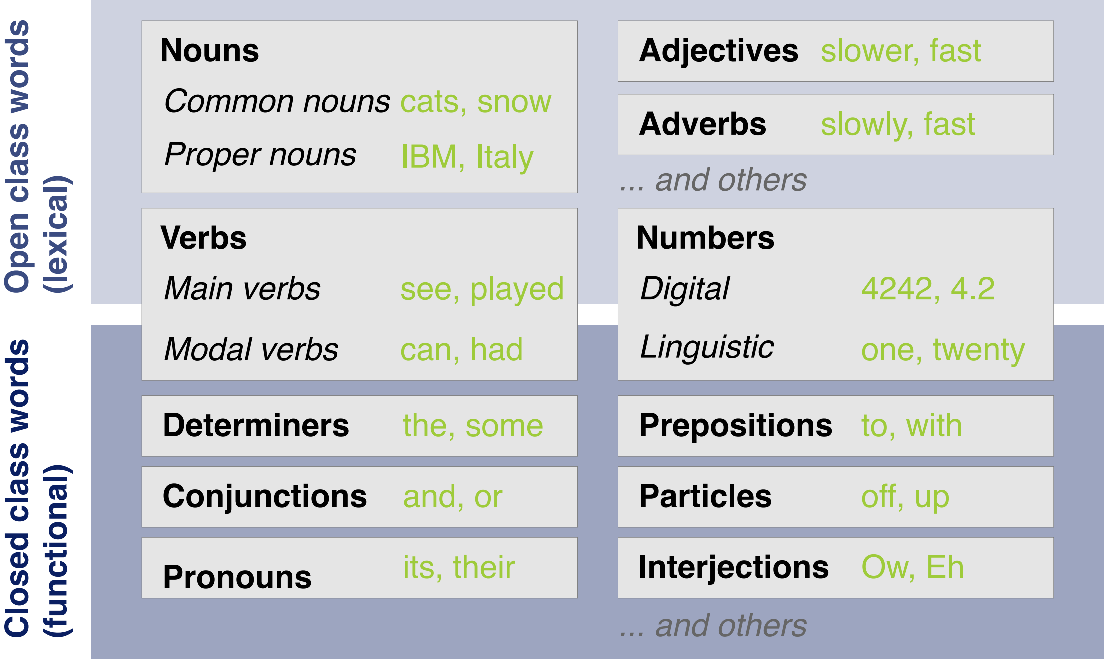
Parts of Speech

Open vs. Closed Word Classes

Two types of word classes

- Open (lexical words). Theoretically, infinitely many members per class
- Closed (functional words). Number of members is fixed in principle

  As language evolves, changes may happen occasionally in closed classes, too.



Open class words (lexical)

Nouns

Common nouns	cats, snow
Proper nouns	IBM, Italy

Adjectives	slower, fast
Adverbs	slowly, fast

... and others

Verbs

Main verbs	see, played
Modal verbs	can, had

Numbers

Digital	4242, 4.2
Linguistic	one, twenty

Closed class words (functional)

Determiners	the, some	**Prepositions**	to, with
Conjunctions	and, or	**Particles**	off, up
Pronouns	its, their	**Interjections**	Ow, Eh

... and others

Syntax

Part of Speech Goes Wild

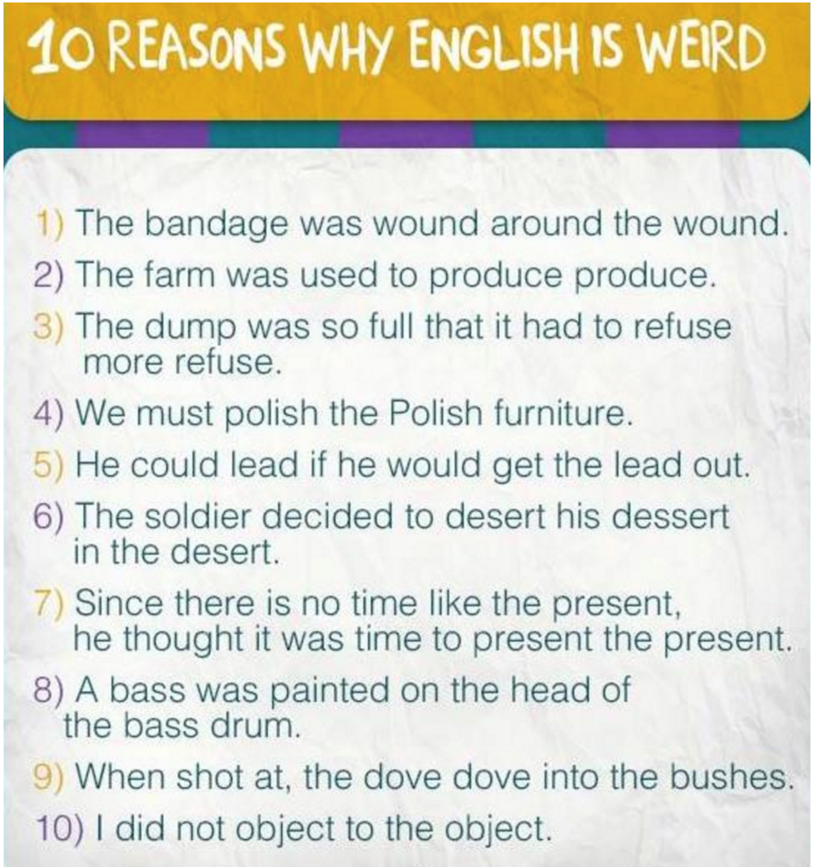
Observation

- ∼90% of all known wordforms have only one part-of-speech.
- The remaining wordforms and unknown words make tagging part-of-speech hard.

Disambiguation

- Analysis of syntactic structure helps:

"The back door"	→ adjective, JJ
"On my back"	→ noun, NN
"Win the voters back"	→ adverb, RB
"Said to back the bill"	→ verb, VB

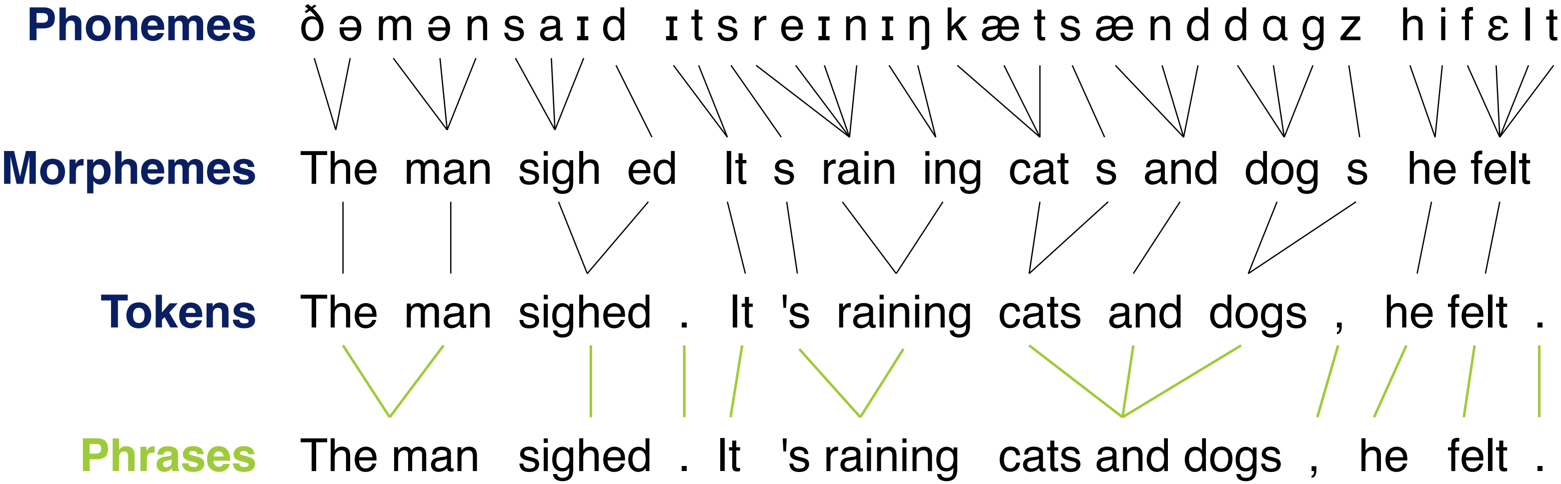


10 REASONS WHY ENGLISH IS WEIRD

1) The bandage was wound around the wound.
2) The farm was used to produce produce.
3) The dump was so full that it had to refuse more refuse.
4) We must polish the Polish furniture.
5) He could lead if he would get the lead out.
6) The soldier decided to desert his dessert in the desert.
7) Since there is no time like the present, he thought it was time to present the present.
8) A bass was painted on the head of the bass drum.
9) When shot at, the dove dove into the bushes.
10) I did not object to the object.

Linguistic Text Units

Phrases



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

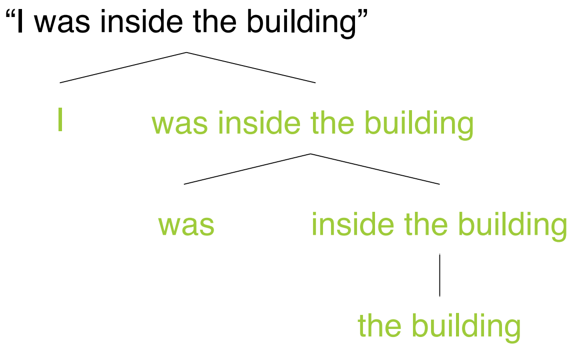
Tokens The man sighed . It 's raining cats and dogs , he felt .

Phrases The man sighed . It 's raining cats and dogs , he felt .

Phrases

Phrase

- A contiguous sequence of related words, functioning as a single meaning unit
- Phrases can have nested phrases.
- Any phrase can be substituted, moved, and asked for, such as ”Where was I?”



“I was inside the building”

I

was inside the building

was

inside the building

the building

Phrases vs. constituents

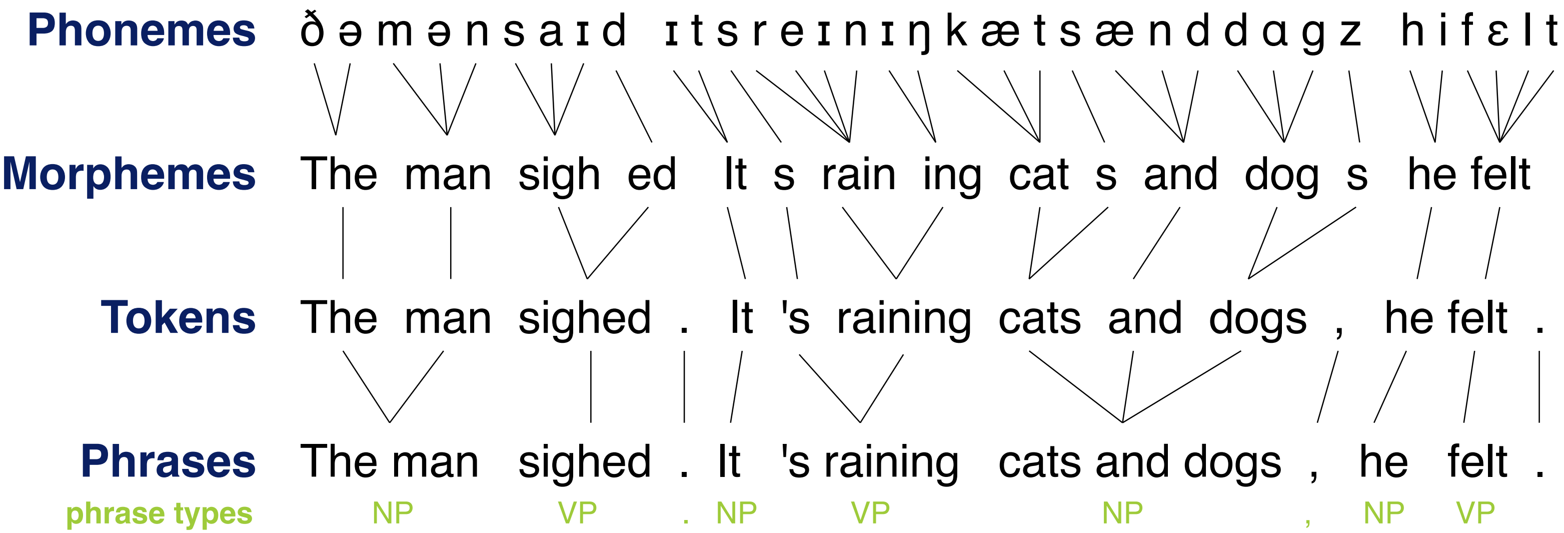
- Phrases represent the constituents in the syntax of a sentence.
- More or less, the two terms are used synonymously.

Phrase chunking (aka shallow parsing)

- The text analysis that segments a sentence into its top-level phrases
- Used in NLP as preprocessing, e.g., for named entity recognition
- *All* phrases are also a by-product of constituency parsing (see below).

Linguistic Text Units

Phrase Types



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Tokens The man sighed . It 's raining cats and dogs , he felt .

Phrases The man sighed . It 's raining cats and dogs , he felt .

phrase types NP VP . NP VP NP , NP VP

Phrase Types

Head-driven phrases

- The head of a phrase is the word which determines the syntactic type.
- Phrases are classified by the part-of-speech of their head.

Five different phrase types

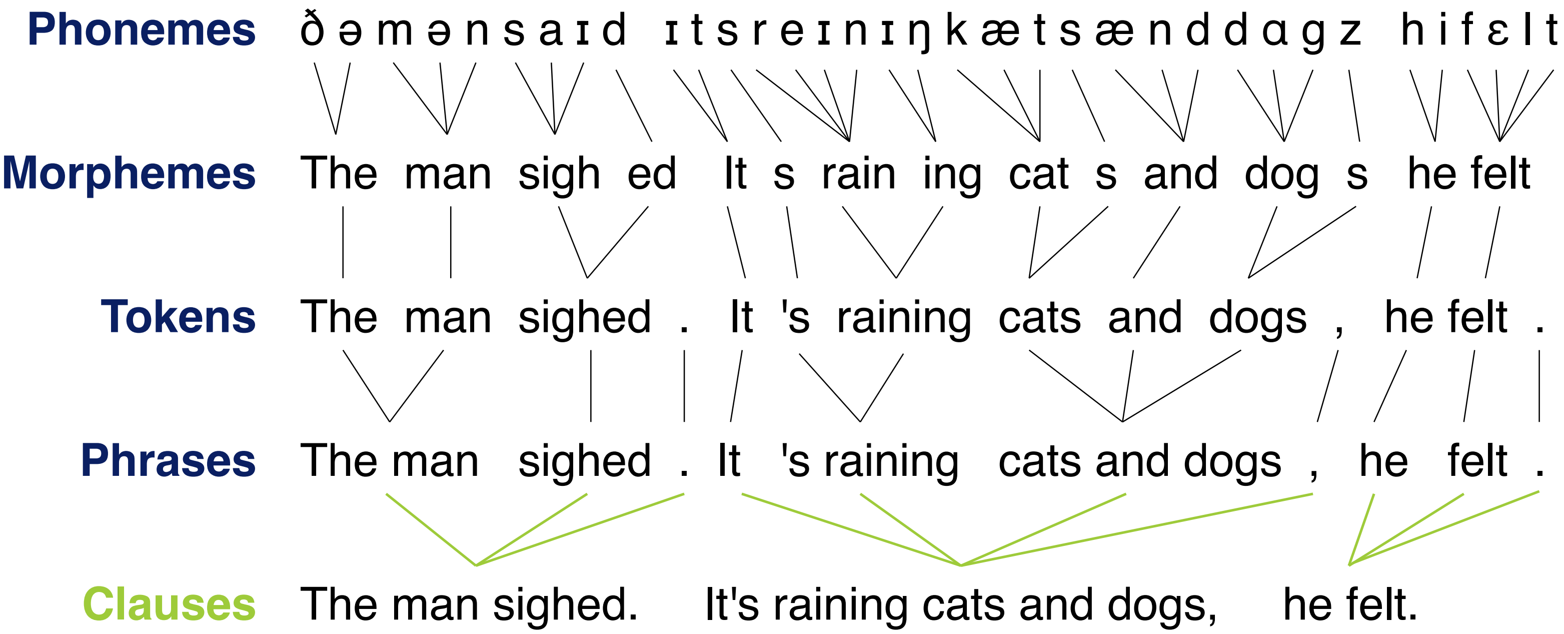
- Noun phrase (NP). ”he”, ”cat on the mat”, ”cats and dogs”
- Verb phrase (VP). ”felt”, ”jump up and down”
- Prepositional phrase (PP). ”in love”, ”over the rainbow”
- Adjectival phrase (AP). ”full of toys”, ”fraught with guilt”
- Adverbial phrase (AdvP). ”very carefully”

Three top-level phrase types

- Only NP, VP, and PP considered as top-level phrases.
- AvdP goes with VP.
- AP usually goes with NP or PP.

Linguistic Text Units

Clauses



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Tokens The man sighed . It 's raining cats and dogs , he felt .

Phrases The man sighed . It 's raining cats and dogs , he felt .

Clauses The man sighed. It's raining cats and dogs, he felt.

Clauses

Clause

- The smallest grammatical unit that can express a complete proposition

Two basic types of clauses

- Main clause. Independent, can stand alone as a sentence

  Usually, one proposition with subject and verb

  ”I remained dry”

- Subordinate clause. Is reliant on a main clause and thus depends on it

  Usually starts with a subordinating conjunction

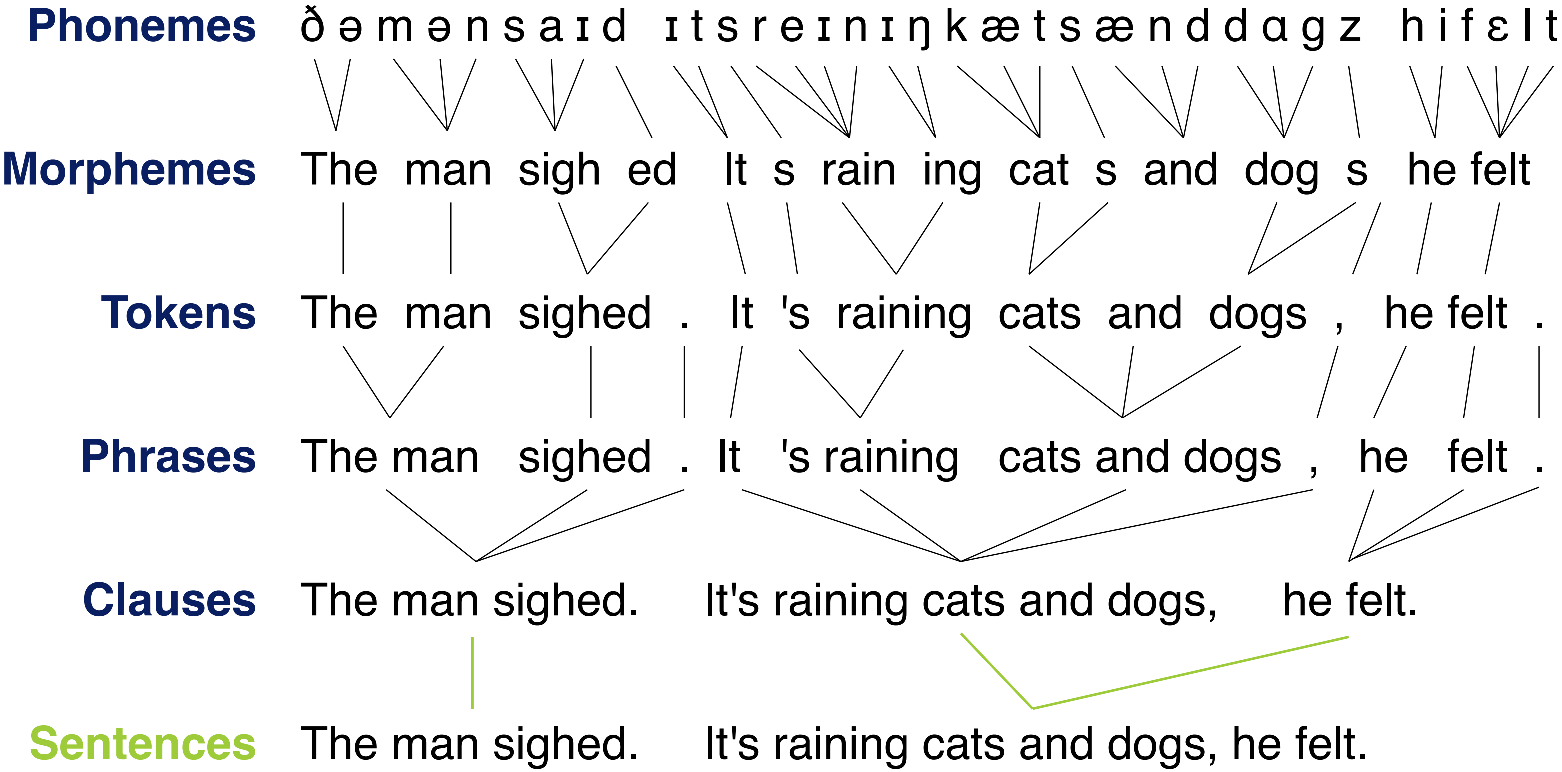
  ”Although it rained” ”because I was inside the building.”

Clause recognition

- The text analysis that identifies the clauses of a sentence
- Not a common analysis; rather, clauses are identified as a by-product of constituency parsing (see below).

Linguistic Text Units

Sentences



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Tokens The man sighed . It 's raining cats and dogs , he felt .

Phrases The man sighed . It 's raining cats and dogs , he felt .

Clauses The man sighed. It's raining cats and dogs, he felt.

Sentences The man sighed. It's raining cats and dogs, he felt.

Sentences

Sentence

- A grammatically independent linguistic unit consisting of multiple tokens
- Contains at least one main clause
- Many text analyses process a text sentence by sentence.

  The concept of sentences basically exists across all languages.

Observation

- There are infinitely many ways to compose words in sentences.
- Yet, we can understand sentences we have never heard or read before.

Sentence splitting (aka sentence segmentation)

- The text analysis that segments a text into its single sentences
- Used in NLP as one of the most basic preprocessing steps

Grammars

Grammar

- A description of the valid structures of a language

  Not always this means natural language structures.

- A grammar is defined by a set of rules.

A → bC	A structure *A* is composed of a word “b” followed by a structure *C*.
C → de	A structure *C* is composed of a word “d” followed by a word “e”.

- Rules consist of terminal and non-terminal symbols.
- Terminal symbols (≈ words) cannot be rewritten any further.
- Non-terminals express clusters or generalizations of terminals.

Syntactic parsing (aka full parsing)

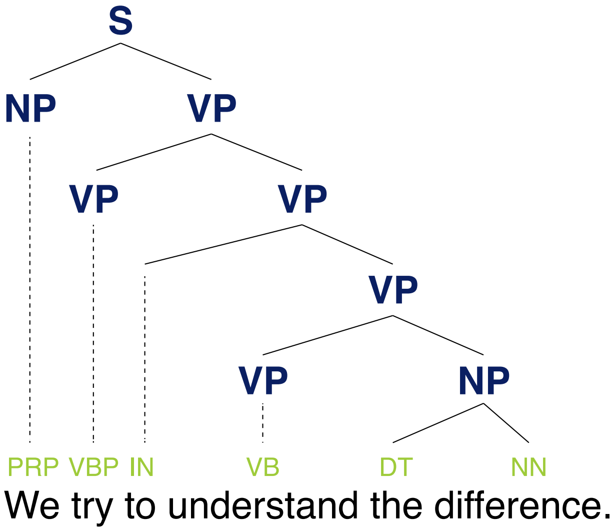
- The text analysis that determines the grammatical structure of a sentence with respect to a given grammar
- Used in NLP as preprocessing for tasks like relationship extraction
- Types. Constituency parsing and dependency parsing

Grammars

Phrase vs. Dependency Structure

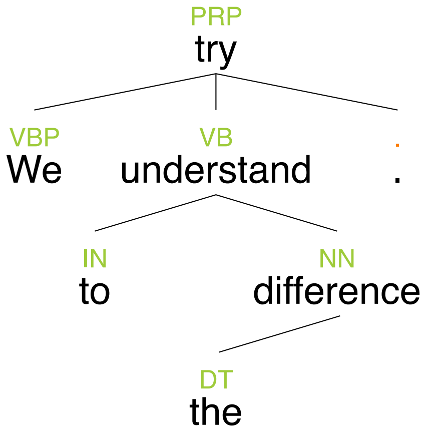
Phrase structure grammar

- Models the constituents of a sentence and how they are composed of other constituents and words
- Constituency (parse) tree. Inner nodes are non-terminals, leafs are terminals



Dependency grammar

- Models the dependencies between the words in a sentence
- Dependency (parse) tree. All nodes are terminals, the root is nearly always the main verb (of the first main clause).

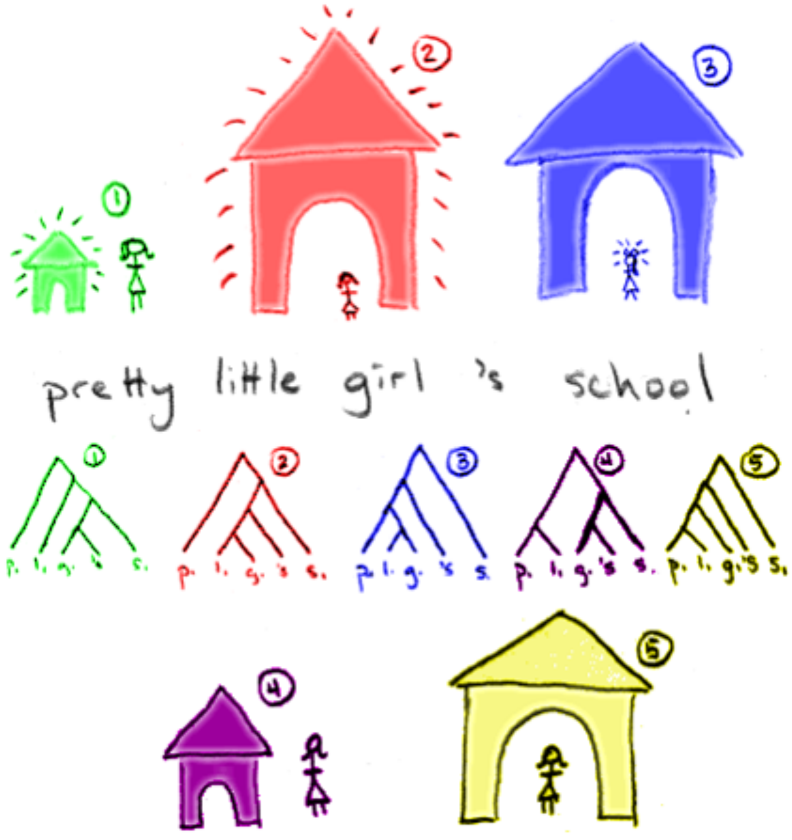


Syntactic Ambiguity

Phrase Structure Goes Wild

Syntactic ambiguity

- Arises when a piece of text has more than one valid syntactic structure



Semantics

Semantics

Semantics

- The meaning of single words and compositions of words

"The man sighed.
It's raining cats and dogs, he felt."

Semantics

Meaning

Meaning

- Propositional content in terms of validity or truth conditions

“All cats are mortal.”
“Sunny is a cat.”

Sunny is mortal.

$$\begin{array}{l} \forall x : cat(x) \rightarrow mortal(x) \\ cat(Sunny) \\ \hline mortal(Sunny) \end{array}$$

- Often requires common-sense reasoning based on world knowledge

“Max can open Timon’s safe.

He knows the combination.”

“Max can open Timon’s safe.

He should change the combination.”

- Includes expressed emotional content

“That poor cat!”

“Fortunately, Max can open Timon’s safe.”

Construction of meaning

- Linguistic form vs. context of use
- Lexical semantics vs. compositional semantics

Semantics

Linguistic Form vs. Context of Use

Meaning implied by linguistic form

- Context-independent meaning, such as “It’s raining.”
- What a speaker publicly commits to, such as “It is wet outside.”
- A speaker’s private state, such as “You should take an umbrella.”

Meaning dependent on context of use

- Scope of quantifiers, such as “Every student reads some book”
- Word sense ambiguities, such as “I’m making it”
- Semantic relations between nouns in compounds, such as “play book”

Meaning dependent on non-linguistic perception

- Time, such as “now”, “tomorrow”, ...
- Location, such as “here”, ”there”, “That’s a beatiful city.”
- Speaker and hearer, such as “I”, “you”, ...

Lexical Semantics

Lexical semantics

- The meaning of words and multi-word expressions
- Covers word senses, semantic roles, and connotations

Word senses

- Distinctions in meaning between different uses of the same form
- Shared meanings between different forms

Semantic roles

- Number of arguments of a predicate
- Specific relationship the arguments bear to the predicate

Connotation

- What word choice conveys beyond truth-conditional semantics, such as "insightful results" vs. "interesting results"

Lexical Semantics

Word Senses

Word sense

- A meaning of a word
- Words can have multiple senses, due to *polysemy* and *homonymy*.
- Moreover, metaphors add senses to words in unbounded ways.

Example: Selected senses of "ride"

- ride over, along, or through
- sit and travel on the back of animal
- be carried or travel on or in a vehicle
- be contingent on
- harass with persistent criticism or carping
- continue undisturbed and without interference
- move like a floating object

16 words senses in total

https://commons.wikimedia.org

https://de.wikipedia.org

Lexical Semantics

Complexities of Word Senses

Polysemy vs. Homonymy

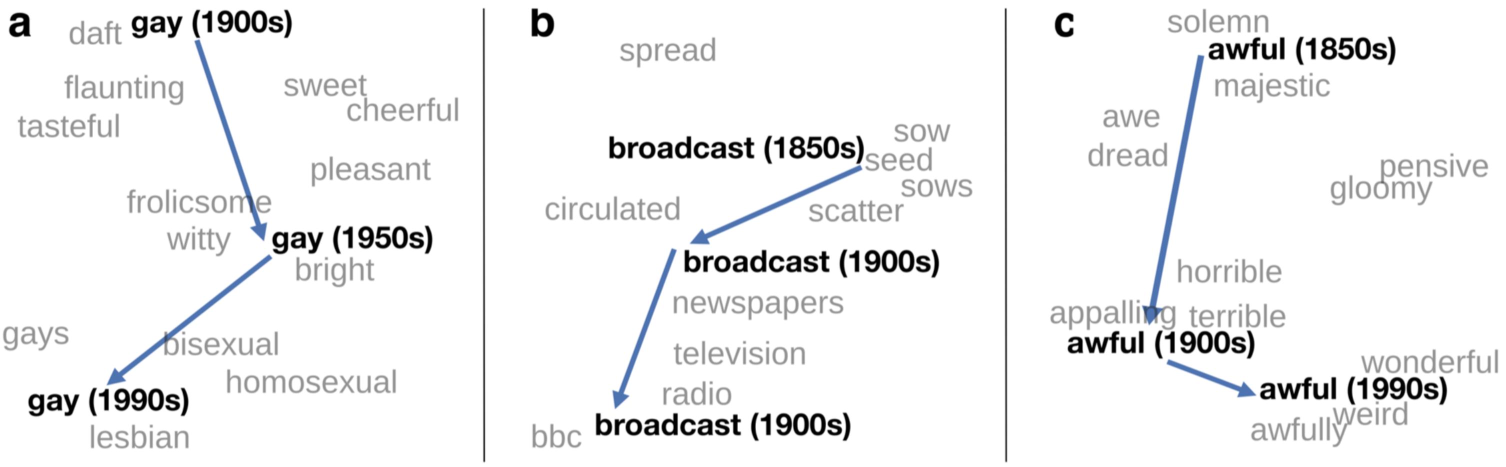
- Polysemy. Related word senses with the same lexical entry

  "newspaper" (physical object vs. abstract content)

- Homonymy. Unrelated word senses that have the same lexical entry

  "bank" (river bank vs. money bank)

Word senses may change over time



(Hamilton et al., ACL 2016)

Lexical Semantics

Semantic Roles

Semantic roles

- The roles the arguments of a predicate have in the state or activity captured by the predicate
- Not to be confused with syntactic roles, such as subject or object
- Different predicates have different semantic roles.

“She saw Max.” vs. “She kissed Max.” vs. “She ressembled Max.”

Why is this *lexical* semantics?

- Syntax is important for identifying what roles an argument plays.
- But the predicate defines the semantic roles.

Semantic role labeling

- The text analysis that finds the arguments taking on the semantic roles in a predicate
- Used in NLP when deeper language understanding is required.

Between Lexical and Compositional Semantics

Multi-Word Expressions

Multi-word expression

- A lexical unit larger than a word that can bear both compositional and idiomatic meanings

“driving instructor” “vice versa”

- On the boundary between lexical and compositional semantics

“Kick the bucket.” “Long time no see.”

Word n-grams

- An alternative to identifying multi-word expressions is to simply use word bigrams, trigrams, or similar instead.

Example “The quick brown fox jumps over the lazy dog.”

- 1-grams. “The”, “quick”, “brown”, “fox”, ..., “dog”, “.”
- 2-grams. “The quick”, “quick brown”, ..., “lazy dog”, “dog.”
- 3-grams. “The quick brown”, “quick brown fox”, ..., “lazy dog.”

Between Lexical and Compositional Semantics

Entities

Entity

- An entity represents an object from the real world.
- The basic semantic concept in NLP

Entity types

- Named entities. Objects that can be denoted with a proper name

  Persons, locations, organizations, products, ...

“Prof. Dr. Henning Wachsmuth” “in Hannover” “at Leibniz University Hannover”

- Numeric entities. Values, quantities, proportions, ranges, or similar

  Temporal and monetary expressions, phone numbers, ...

“in this year” “2024-04-11” “$ 100 000” “762-123 77”

Named and numeric entity recognition

- The analyses that mine respective entities from text
- Used in NLP as key steps in information extraction tasks

Compositional Semantics

Compositional semantics

- The meaning of word compositions in phrases, clauses, and sentences
- Covers relations, linguistic operators, collocations, and more

  These concepts can, in principle, be represented in logical forms.

Relations

- Semantic. Relations between entities from the world
- Temporal. Relations describing courses of events

Linguistic operators

- Quantifiers. Indicating quantities of objects, such as "*Every* man reads"
- Hedges. Limiting the impact of propositions, such as "*Probably* every"
- Negation. Inverting adjectives/predicates, such as "Probably *not* every"

Collocations

- Words that cooccur overproportionally often, such as "bluntly speaking"

Compositional Semantics

Semantic Relations

Semantic relations

- Word compositions that capture relational predicates with arguments
- Typically: Who did what to whom, where, when, how, and why?

Common relation types

- Binary relations. Relations with two arguments

  founded(organization, time) → “Google was established in 1998.”

- Events. Relations with multiple arguments, possibly nested relations

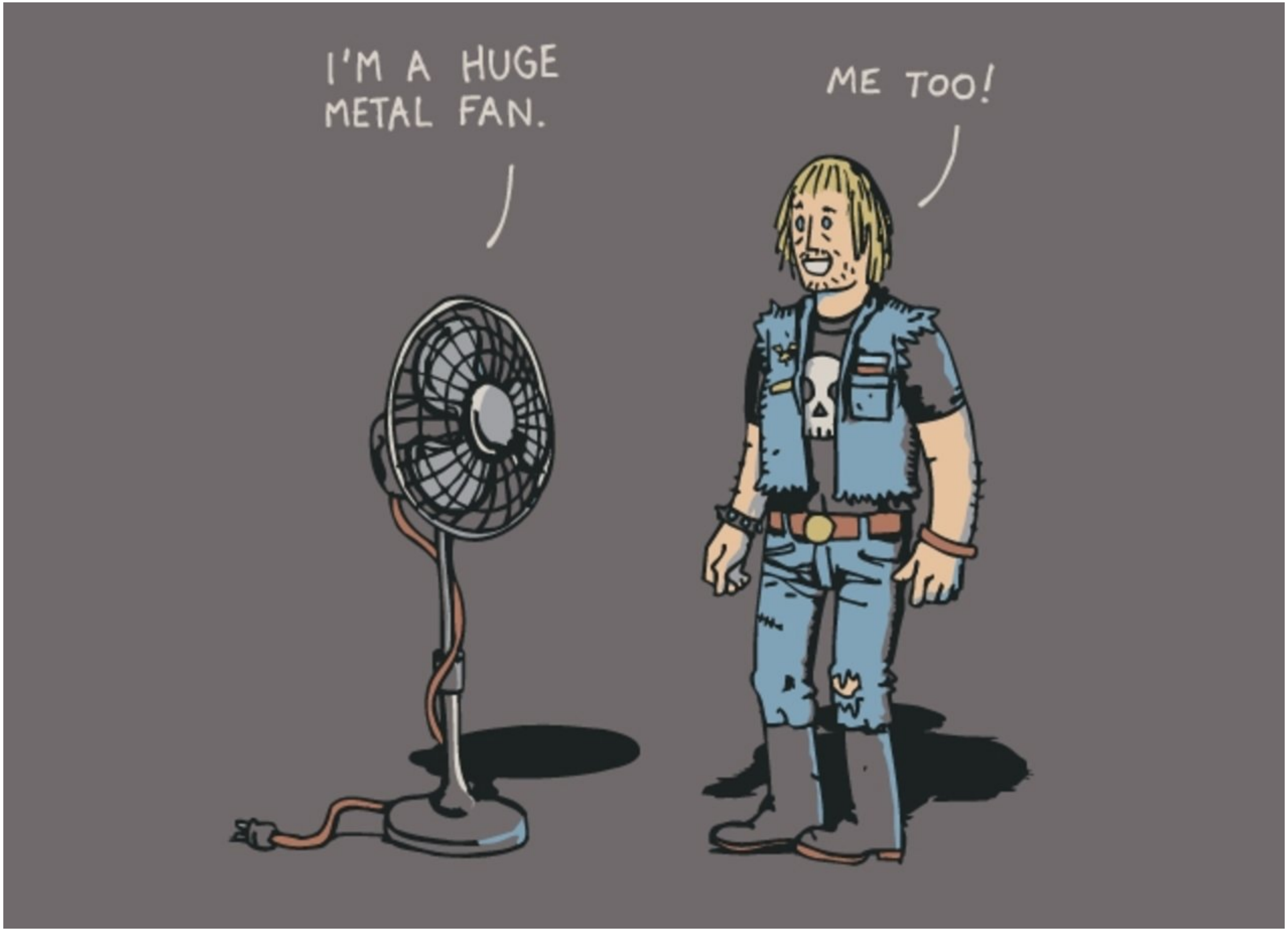
  reads(agent, theme, date, time, location, origin) ∧ origin(theme, author)
  → “Max reads a book in the garden on Monday at midnight. It is from Shakespeare.”

Relationship and event extraction

- The text analyses that mine relations and events from text
- Used in NLP as key steps in information extraction tasks

Semantics

Multi-Word Expressions Go Wild



Discourse

Discourse

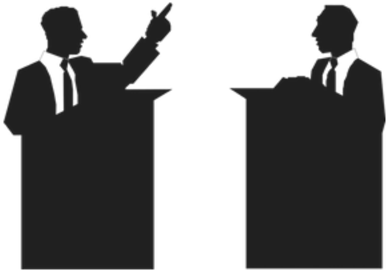
Discourse

- Discourse describes linguistic units that are larger than a sentence.
- Usually, it refers to the entirety of a given text.

Discourse vs. dialogue

- Discourse. The term *discourse* is usually used to refer to monologues.
- Dialogue. A conversational discourse with two or more parties

https://commons.wikimedia.org



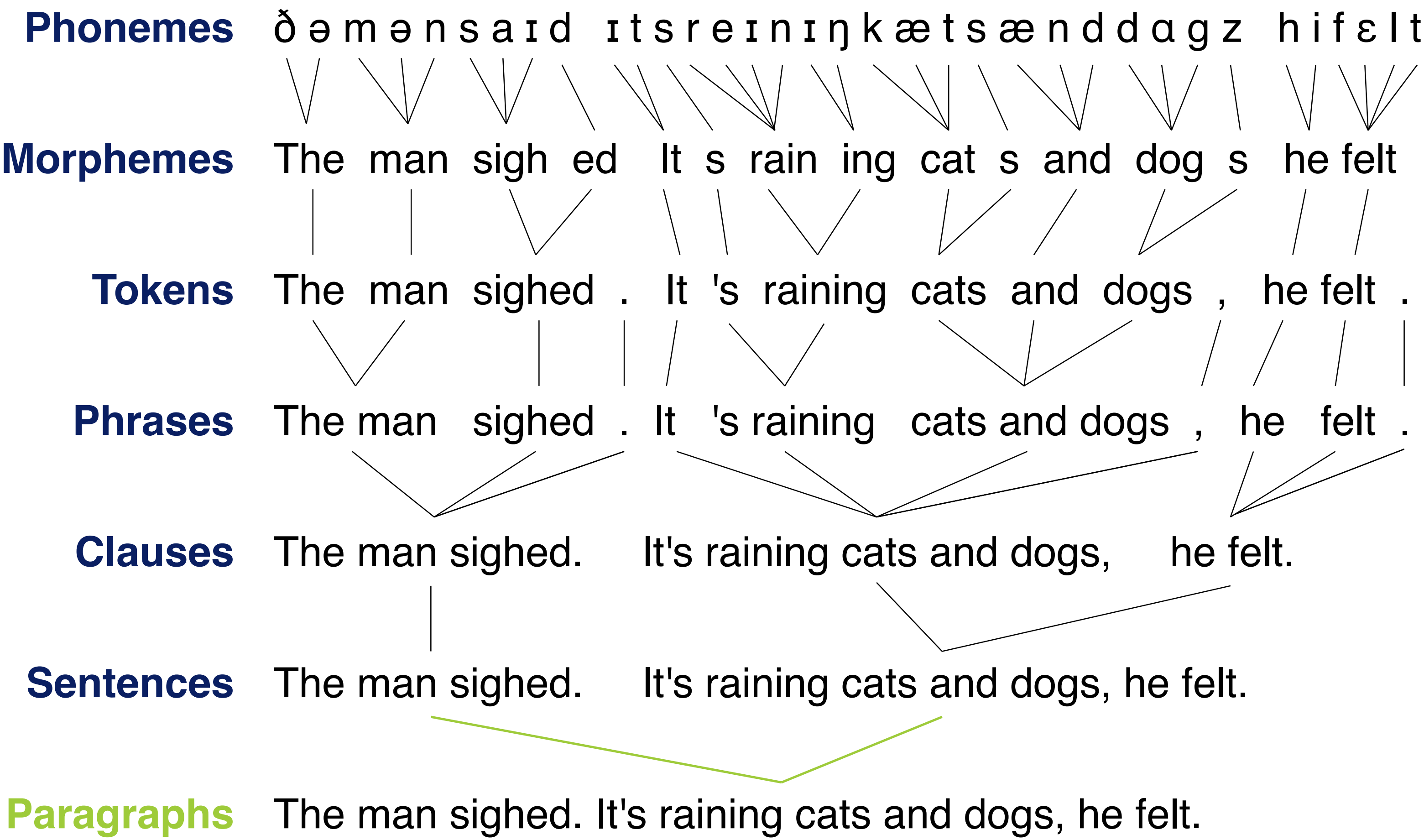
https://de.wikipedia.org

Discourse-level semantics

- Coreference. Different expressions may be used to refer to one thing.
- Coherence. Understandable discourse has continuity in meaning.

Linguistic Text Units

Paragraphs



Phonemes ð ə m ə n s a ɪ d ɪ t s r e ɪ n ɪ ŋ k æ t s æ n d d ɑ g z h i f ɛ l t

Morphemes The man sigh ed It s rain ing cat s and dog s he felt

Tokens The man sighed . It 's raining cats and dogs , he felt .

Phrases The man sighed . It 's raining cats and dogs , he felt .

Clauses The man sighed. It's raining cats and dogs, he felt.

Sentences The man sighed. It's raining cats and dogs, he felt.

Paragraphs The man sighed. It's raining cats and dogs, he felt.

Discourse

Discourse-Level Units

Linguistic discourse units

- General. Paragraphs, entire texts
- Genre-specific. Sections, chapters, parts, books, or similar

Paragraph

- Gramatically, a paragraph is a sequence of one or more sentences, whose boundaries are denoted by line breaks.
- Ideally, each paragraph represents one thought, argument, or similar.
- Practically, paragraphs are not consistently used.

Discourse structure

- The structure that represents the organization of an entire text
- Consists of *discourse segments* and *coherence relations* between them

  A common discourse model is the Rhetorical Structure Theory (RST).

Discourse

Discourse Structure

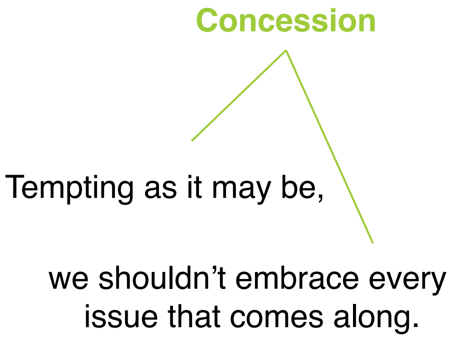
Discourse segment

- A linguistic unit serving as a single building block of a discourse
- May consist of multiple smaller adjacent segments
- Elementary discourse unit. Minimum segment, usually a clause

Tempting as it may be, | we shouldn't embrace every issue that comes along.

Coherence relation (aka rhetorical/discourse relation)

- Describes how two segments relate to each other
- May be semantic or pragmatic
- May be coordinating or subordinating



Discourse parsing

- The text analysis that infers the discourse structure of a text
- Used in NLP for tasks where structure is important
- Implicit segments and relations are what makes parsing hard.

Discourse

Coreference

Coreference

- Two or more expressions in a text that refer to the same thing

Common types of coreference

- Anaphora. “Max walked in. He sat down.”
- Cataphora. “After he walked in, Max sat down.”
- Split antecendents. “Max asked Leandra to join. They arrived together.”
- Coreferring noun phrases. “Apple is based in Cupertino. The company is actually called Apple Inc.”

Coreference resolution

- The text analysis that maps all references to umambiguous identifiers
- Used in NLP as preprocessing for tasks like entity recognition
- Coreference resolution may require deep text understanding.

Discourse

Coherence

Coherence

- The continuity of meaning in discourse

“Max hid Timon’s keys. He drank too much.”	Coherent.
“Max hid Timon’s keys. He likes spinach.”	Coherent?

Local vs. global coherence

- Local. Coherence in adjacent discourse segments
- Global. Coherence of the entire discourse of a given text

From local to global coherence?

“Max hid Timon’s keys. He drank too much. ”	Locally coherent.
“He drank too much. No water was left.”	Locally coherent.
“Max hid Timon’s keys. He drank too much. No water was left.”	Globally coherent?

Coherence vs. cohesion

- Cohesion is the continuity of grammatical structure, not meaning.

Pragmatics

Pragmatics

Pragmatics

- Pragmatics deals with how language is used to accomplish goals.
- Relates to the author's (or speaker's) intention and to the context of use

> "Have you emptied the dishwasher?"

> "I never said she stole my money."

- Covers speech acts, presupposition and implicature, and much more

Speech Acts

- Linguistic utterances with a performative function

Presupposition and implicature

- Presupposition. Linguistic utterances presuppose things.
- Implicature. Linguistic utterances entail things.

Pragmatics

Speech Acts

Speech act

- A linguistic utterance with a performative function
- The terms is mostly used to refer to *illocutionary* speech acts.

Three types of speech acts

- Locutionary act. The act of saying something meaningful

  "Smoking is bad for your health."

- Illocutionary act. A direct or indirect act performed by performing a locutionary act

  Assertion that smoking is bad for your health (direct)

  Warning not to smoke (indirect)

- Perlocutionary act. An act which changes the cognitive state of the interlocutor

  Causing you to adopt the intention to stop smoking.

Implicature *

Presupposition and Implicature

Presupposition

- Implicit assumption about the world related to an utterance whose truth is taken for granted

“Max’ cousin took an aspirin.” → Max has a cousin, someone’s called Max

Implicature

- What is suggested by a linguistic utterance, even though neither expressed nor entailed.

Maja: “Did the students pass the exam.”
Max: “Some of them did.” → Not all of them.

Maja: “Are you coming out tonight?”
Max: “I have to work.” → Max won’t come.

Max: “I have to work. But I’ll come out anyway.” (cancelling implicature)

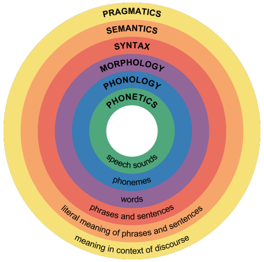
Maja: “He’s brilliant and imaginative.”
Max: “He’s imaginative.” (implicit agreement/denial)

Conclusion

Conclusion

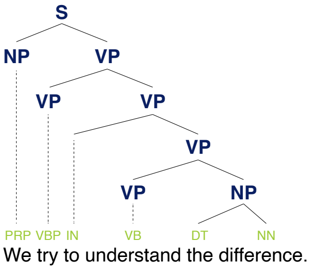
Linguistics in NLP

- NLP analyzes and synthesizes natural language text.
- Linguistic concepts define the basis of all processing.
- Analyses and synthesis can take place at several levels.



Morphology and syntax

- How words are formed and grammar is constructed
- Central concepts are tokens, phrases, and sentences.
- NLP deals with these levels for proper processing.



Semantics and pragmatics

- How meaning is conveyed and language is used
- Central concepts are entities, relations, and discourse.
- NLP applications target these levels.

References

Some content and examples taken from

- Emily M. Bender (2018). 100 Things You Always Wanted to Know about Semantics & Pragmatics But Were Afraid to Ask. Tutorial at the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). http://faculty.washington.edu/ebender/papers/Bender-ACL2018-tutorial.pdf.
- Daniel Jurafsky and Christopher D. Manning (2016). Natural Language Processing. Lecture slides from the Stanford Coursera course. https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html.
- Matthias Hagen (2018). Natural Language Processing. Slides from the lecture at Martin-Luther-Universität Halle-Wittenberg. https://studip.uni-halle.de/dispatch.php/course/details/index/8b17eba74d69784964cdefc154bb8b95.
- Daniel Jurafsky and James H. Martin (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall, 2nd edition.
- Christopher D. Manning and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- Henning Wachsmuth (2015): Text Analysis Pipelines — Towards Ad-hoc Large-scale Text Mining. LNCS 9383, Springer.