Introduction to Text Mining

Part II: Basics of Linguistics

Henning Wachsmuth

https://cs.upb.de/css

Basics of Linguistics: Learning Objectives

Concepts

- Get to know several fundamental phenomena in natural language.
- Learn about the different levels of language.
- Understand why natural language processing is complex.

Methods

- Get an overview of existing text analyses.
- Be ready 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
   - What Is Linguistics?
   - Morphology
   - Syntax
   - Semantics
   - Discourse and Pragmatics

III. Text Mining using Rules

IV. Basics of Empirical Research

V. Text Mining using Grammars

VI. Basics of Machine Learning

VII. Text Mining using Clustering

VIII. Text Mining using Classification and Regression

IX. Practical Issues

X. Text Mining using Sequence Labeling

What Is Linguistics?

Linguistics

What is 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

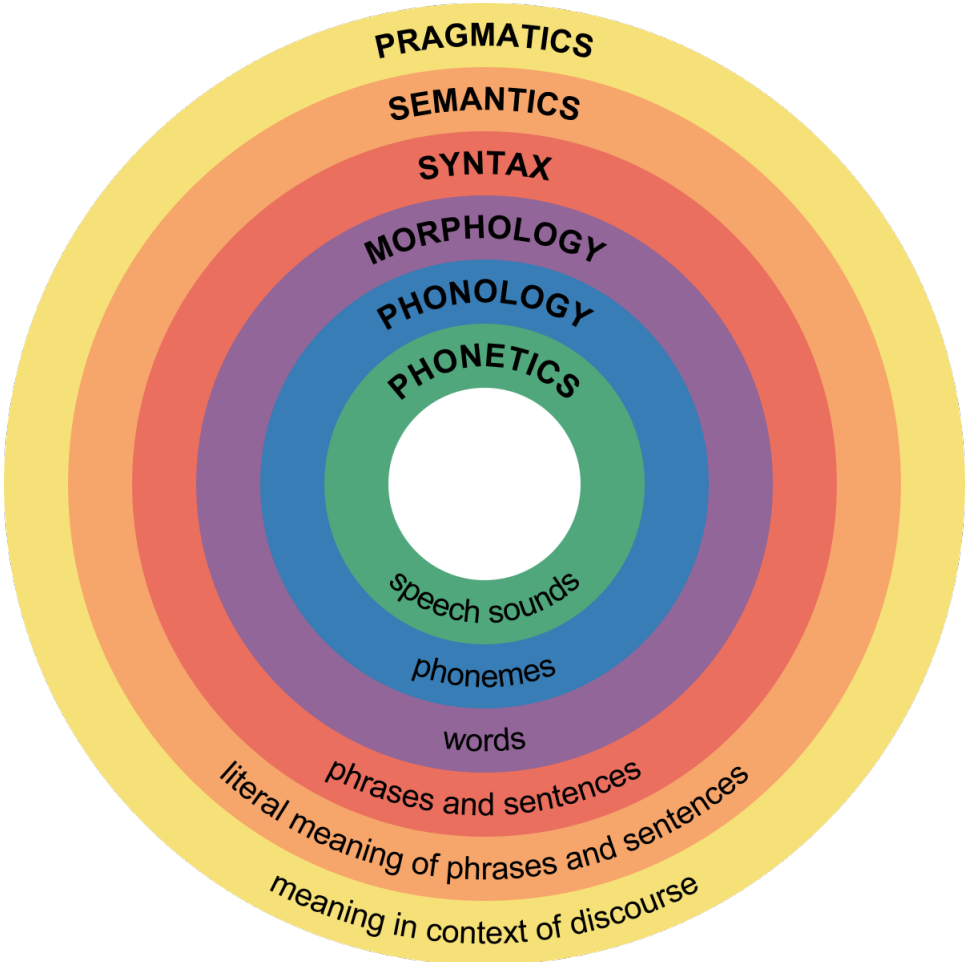
- Phonetics. The physical aspects of speech sounds.
- Phonology. The linguistic sounds of a particular language.

Levels of spoken and written language

- Morphology. The senseful components of words and wordforms.
- Syntax. The structural relationships between words, usually within a sentence (or a similar utterance).
- Semantics. The meaning of single words and compositions of words.
- Discourse. Linguistic units larger than a single sentence, such as paragraphs or complete documents.
- Pragmatics. How language is used to accomplish goals.

Linguistics

Levels of Language Analysis



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

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.

Written language in the focus

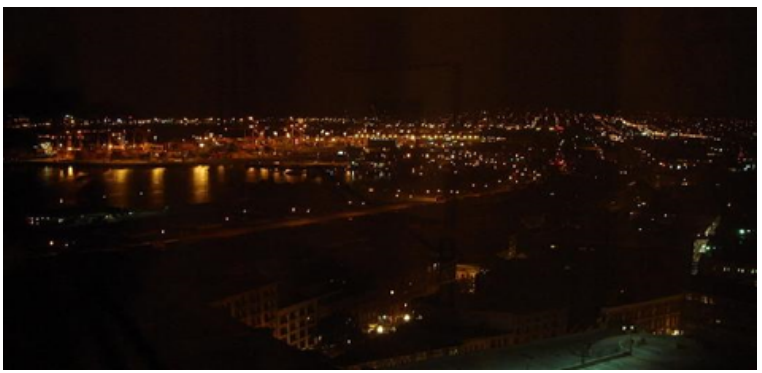
- Natural language is computationally analyzed mostly in text form.
  Where given, speech is transcribed to text before.
- Phonetics and phonology are largely disregarded in text mining, and they will play only a small role in this course.
- In the end, this just means: one problem less to deal with.

Selected Problems of Spoken Language Processing

Homophones

- A spoken word with two or more possible transcriptions.
- Example: /naɪt/

Knight



Night

Segmentation

- No punctuation in spoken language.
- Sentence (or other) segmentation must be based on breaks, pitch, etc.

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

Problems of *Non*-Spoken Language Processing

Drawbacks of focus on written language

- A restriction to text does not make the analysis easier only.
- Some important information is lost, especially *prosody*.

Prosody

- Prosody refers to language features of composed speech units.
- Features. Pitch, tone, stress, rhythm, loudness, tempo, ...
- Although there are ways to encode prosody in text, it is rarely done.

Consequences

- Text mining cannot analyze prosodic differences.

  "*I* never said she stole my money." vs. "I *never* said she stole my money."

- However, much language is given in text form only, anyway.

Linguistic Text Units

Overview

Language levels of units

- Morpohological level. Characters, syllables, morphemes, words.
- Syntactic level. Phrases, clauses, sentences.
- Discourse level. Paragraphs and 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

Morphemes

What is a 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.

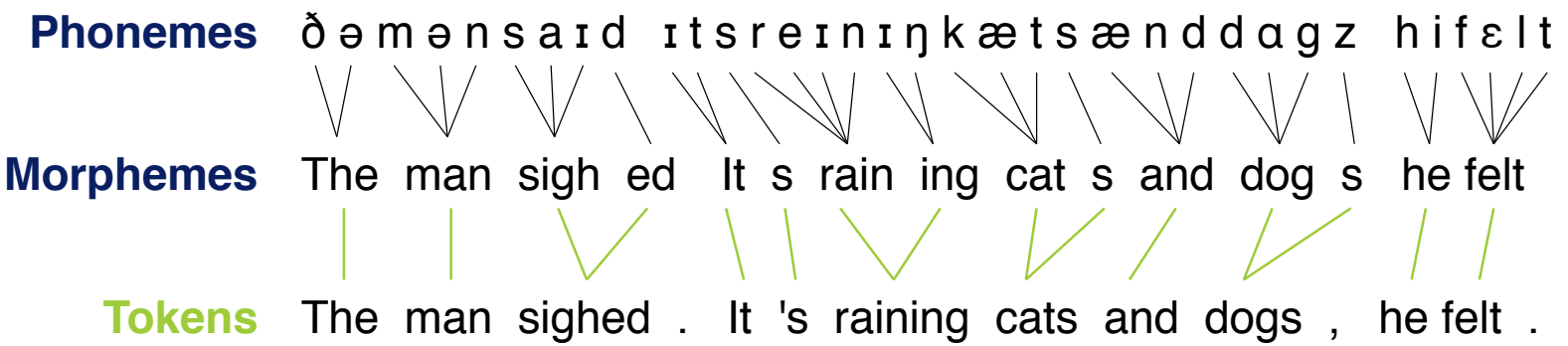
"speaker" → "speak" + "er" (morphemes) vs. "spea" + "ker" (syllables)

Free and bound morphemes

- Both "cat" and "cats" can be uttered in isolation, but "s" cannot.
- "cats" and "cat" are free morphemes, "s" is bound.
- Free morphemes can be used as words, bound morphemes cannot.

Linguistic Text Units

Words and Tokens



Words

What is a 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 letters, digits, spaces, punctuations, special symbols, or similar.
- In some languages, characters represent complete words (or syllables).

猫

Chinese "cat"

Words

Lemmas and Wordforms

What is a lemma?

- A lemma is the dictionary form of a word.

  A related term is *lexem*, denoting the unit of meaning of a word irrespective its form.

"be", "am", "was", ... → "be" "deriving", "derives", ... → "derive"

What is a wordform?

- The full 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

What is a stem?

- A stem is the part of a wordform that never changes.

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

- 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".

- Usually carries the main meaning of a word.

What is an affix?

- Any bound morpheme, such as "-s".
- Affixes add meanings of various kinds to the stem.
- Four types of affixes: Suffix, prefix, infix, and circumfix.

Words

Affix Types

Suffixes appear after the stem.

"cat" + "s" "nice" + "ly"

Prefixes appear before the stem.

"un" + "true" "pre" + "conceptions"

Infixes appear inside the stem.

In English used only in informal language, usually to show emotions.

"fan" + "bloody" + "tastic"

Circumfixes appear on both sides of the stem.

The change from *y* to *i* in "bodi" in the example is an inflection rather than an affix part.

"em" + "bodi" + "ed"

Inflection and Derivation

What is inflection?

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

"derive" → "derived"

What is derivation?

- The modification of a word to obtain a new word.

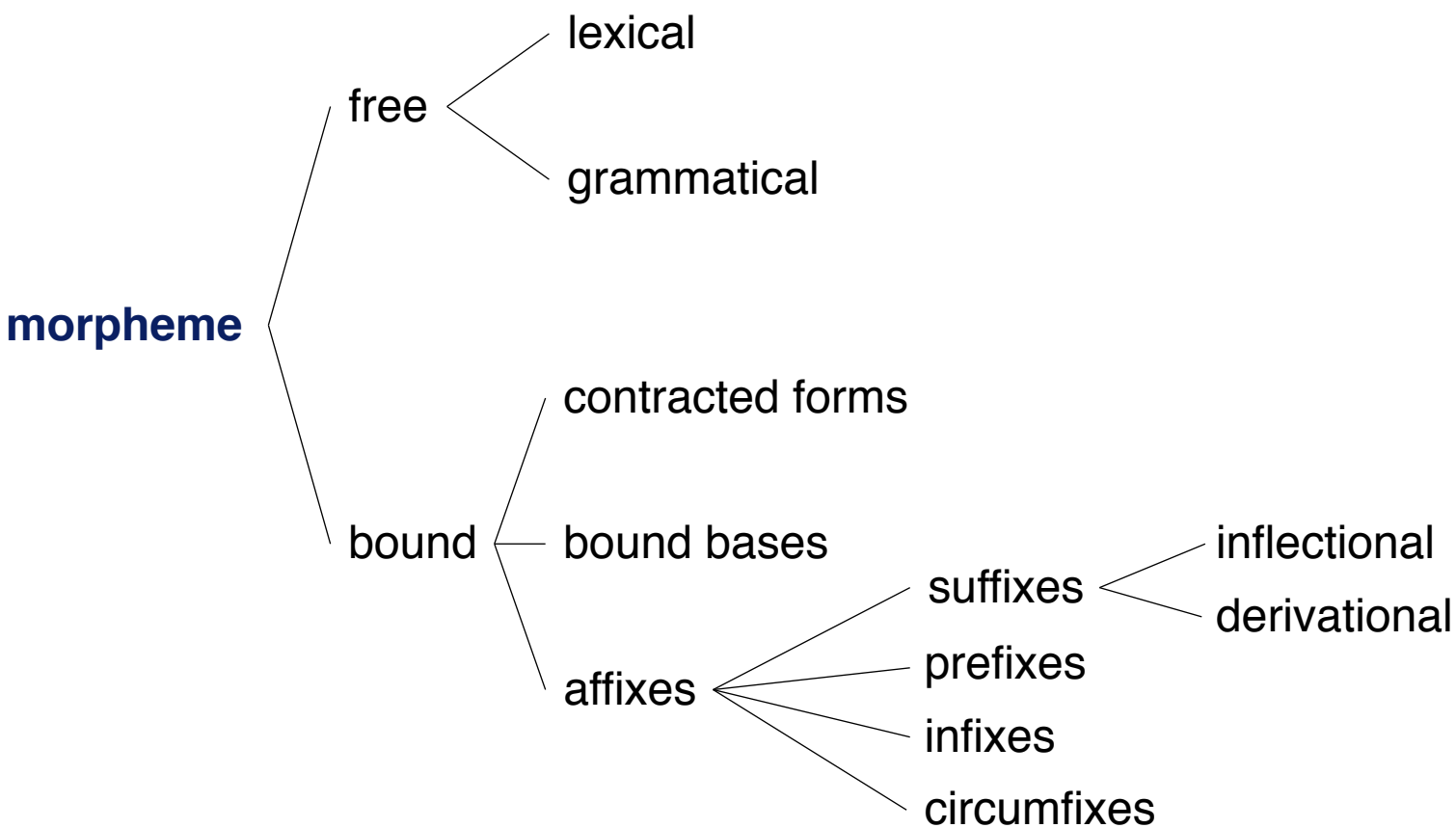
"derive" → "derivation"

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"

Classification of Morphemes



Tokens

What is a token?

- Smallest text unit usually considered in natural language processing.
- A token is a wordform, a number, a symbol, or similar.
- Whitespaces are usually *not* considered as tokens themselves.

Example tokens

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

Tokenization

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

Tokens

Lemmas vs. Tokens

Vocabulary

- The set of all different lemmas in a collection of text.

Some example collection sizes

Collection	# Lemmas	# Tokens
Switchboard phone conversation	20,000	$2.4 \cdot 10^6$
Shakepeare's works	31,000	884,000
Google n-gram corpus	$13 \cdot 10^6$	10^{12}

Ratio estimation (Church and Gray, 1990)

- $\#\text{vocabulary} > \mathcal{O}(\sqrt{\#\text{tokens}})$
- Implicitly based on *Zipf's law*.

Tokens

Zipf's Law

Empirical law according to George Kingsley Zipf

- Let all words be ordered by their frequency f in a large collection of text.
- Let r_i be the rank of the word w_i in the ordered list of words.
- Then the frequency of w_i is inverse proportional to r_i, i.e., $f_i \sim \frac{1}{r_i}$.

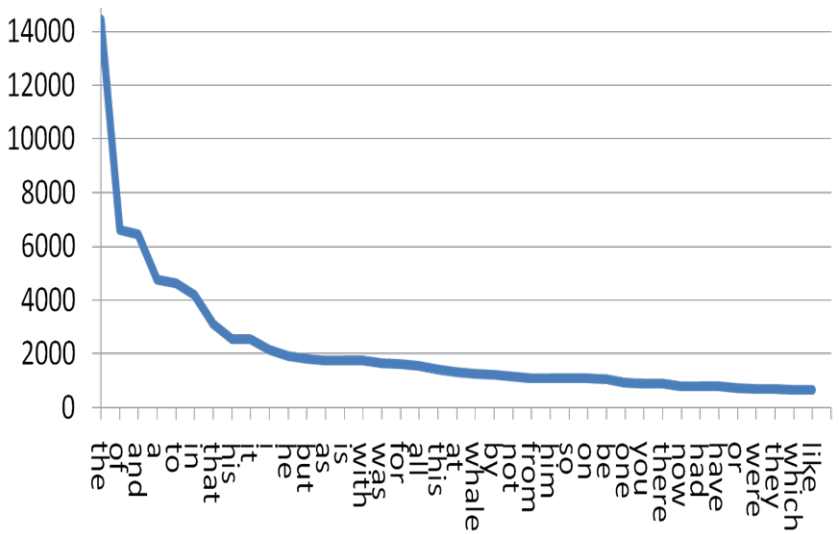
Example

- Words in "Moby Dick" follow Zipf's law roughly.

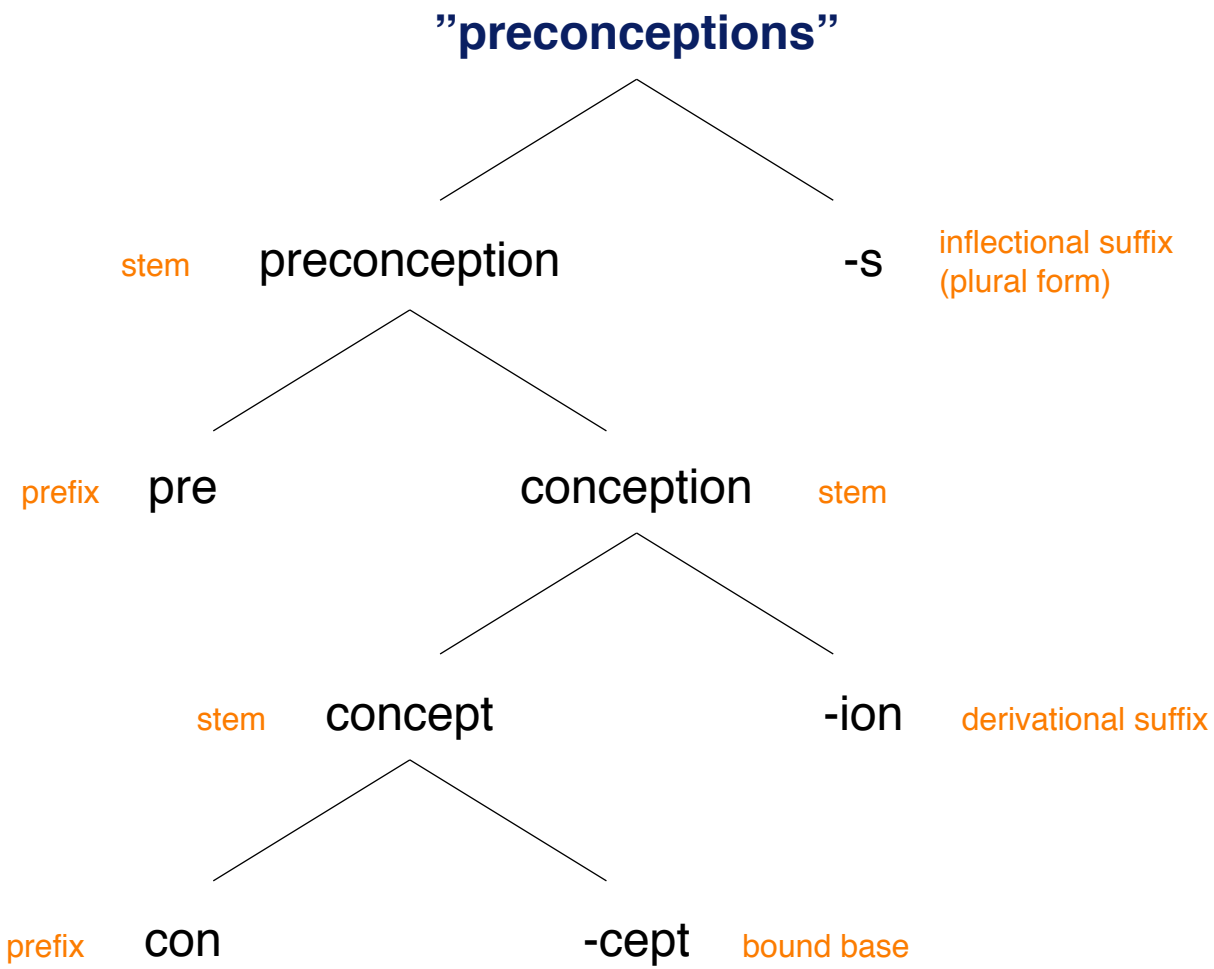
General observations

- Funtion words on the left
- Content words in the middle
- Rare words on the right

  Including misspelled words, very specific proper names, ...



Morphological Analysis



Morphological Normalization

What is morphological normalization?

- Identification of a single canonical representative for morphologically related wordforms.
- Reduces inflections (and partly also derivations) to a common base.
- Used in text mining 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” $\rightarrow$ “deriv” (stem) vs. “derive” (lemma)

- For some tokens, the stem and lemma look entirely different.

  “am” $\rightarrow$ “am” (stem) vs. “be” (lemma)

Morphology Goes 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

And 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 + Aslf

Syntax

Syntax

What is 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.

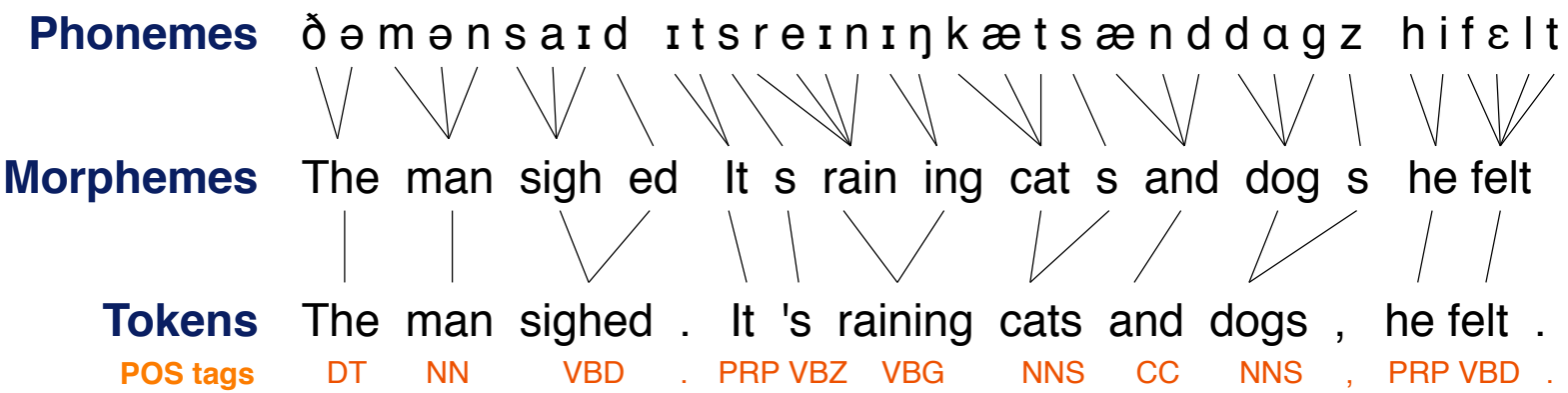
Structural relationships

- 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. Grammatical independent linguistics units.

Linguistic Text Units

Parts of Speech



Parts of Speech

What is a part of speech?

- A part of speech is a 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

- Natural language processing considers more fine-grained word classes, partly language-specific, and represents them 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" (???), ...

Part-of-speech tagging

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

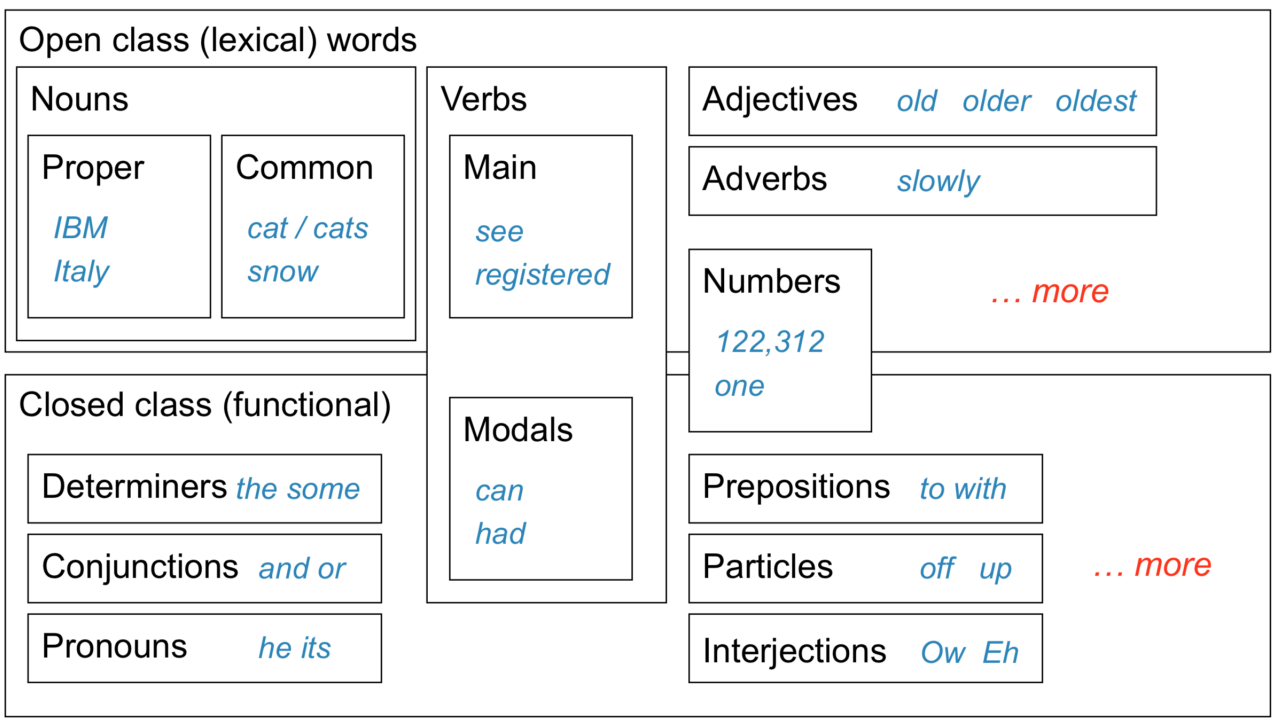
Parts of Speech

Open vs. Closed Word Classes

Two types of word classes

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

  As language evolves, changes rarely happen in closed classes, too.



Parts of Speech

Ambiguity

Observation

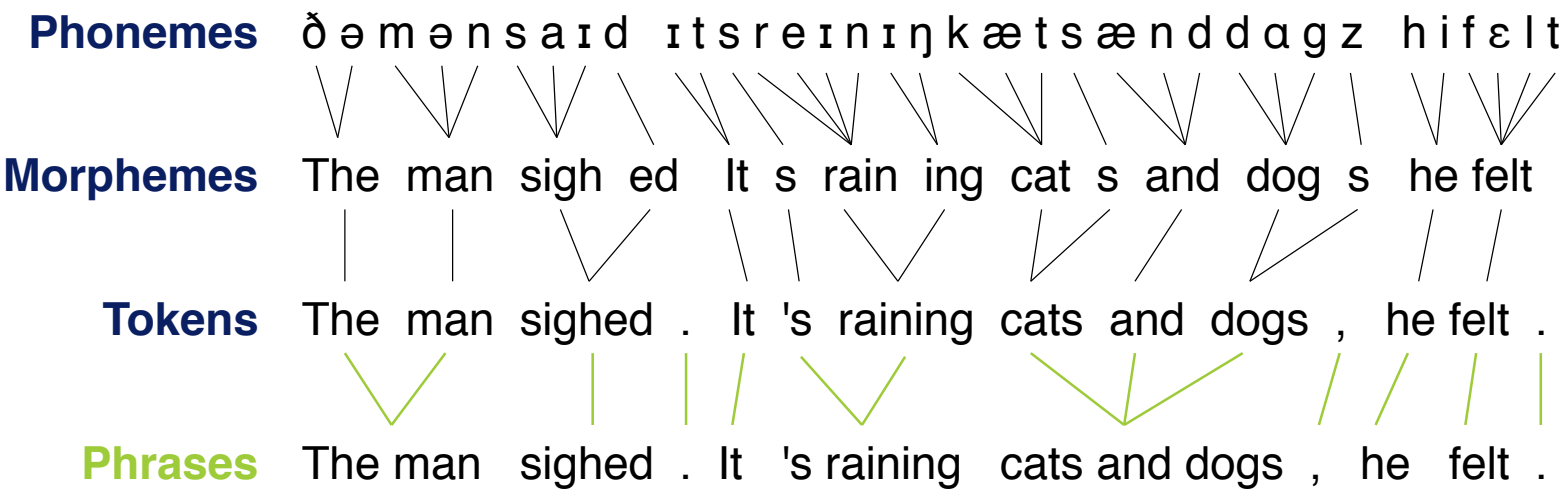
- ∼90% of all known wordforms have only one part-of-speech.
- The remaining wordforms and unknown words make tagging part-of-speech hard.

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

- Analysis of syntactic structure helps disambiguating.

Linguistic Text Units

Phrases



Phrases

What is a phrase?

- A phrase is a contiguous sequence of related words, functioning as a single meaning unit.

  "I was inside the building" → "I", "was", "inside the building",

- Phrases can have nested phrases.

  "inside the building" (top-level phrase), "the building" (nested phrase),

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 text mining as preprocessing, e.g., for named entity recognition.
- *All* phrases are also a by-product of constituency parsing (see below).

Phrases

Standard tests

Phrases can be identified with standard tests

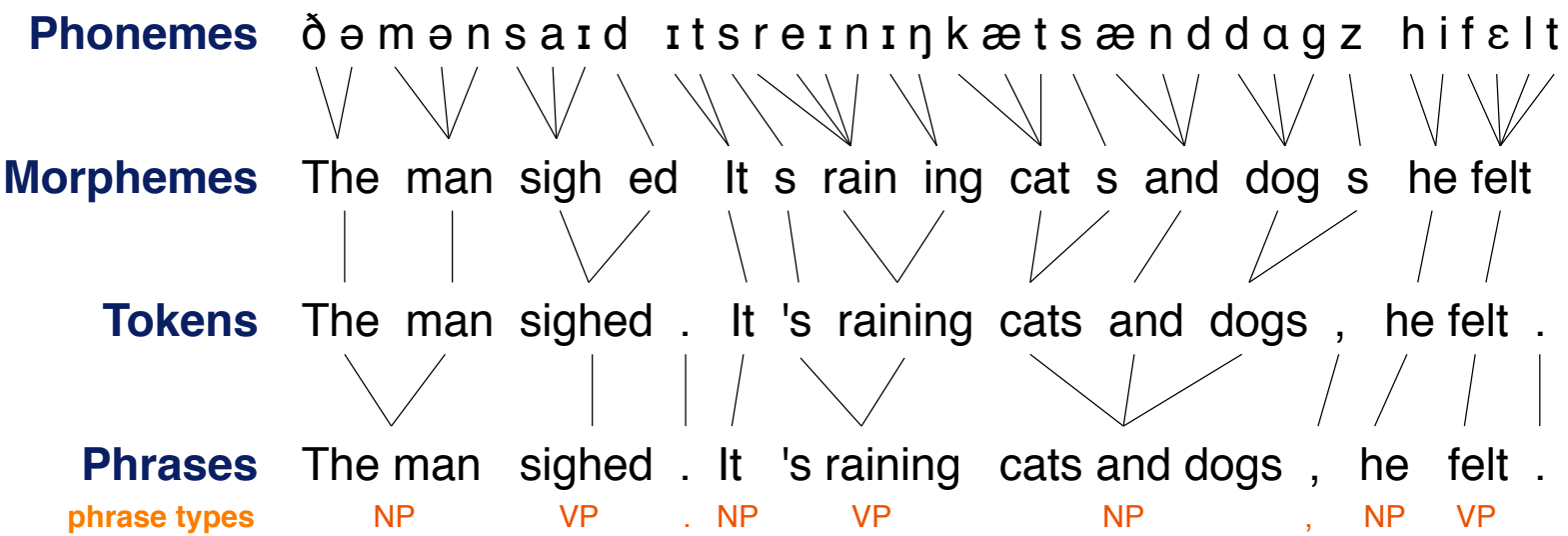
- Any phrase can be substituted, moved, coordinated, and asked for.

Example "The dog ate a cookie."

- Substitution. "The dog ate it."
- Movement. "A cookie was eaten by the dog."
  Also cases without other reformulations exist.
- Coordination. "The dog ate a cookie and a piece of cake."
- Question. "What did the dog eat?" $\rightarrow$ "A cookie."

Linguistic Text Units

Phrase Types



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

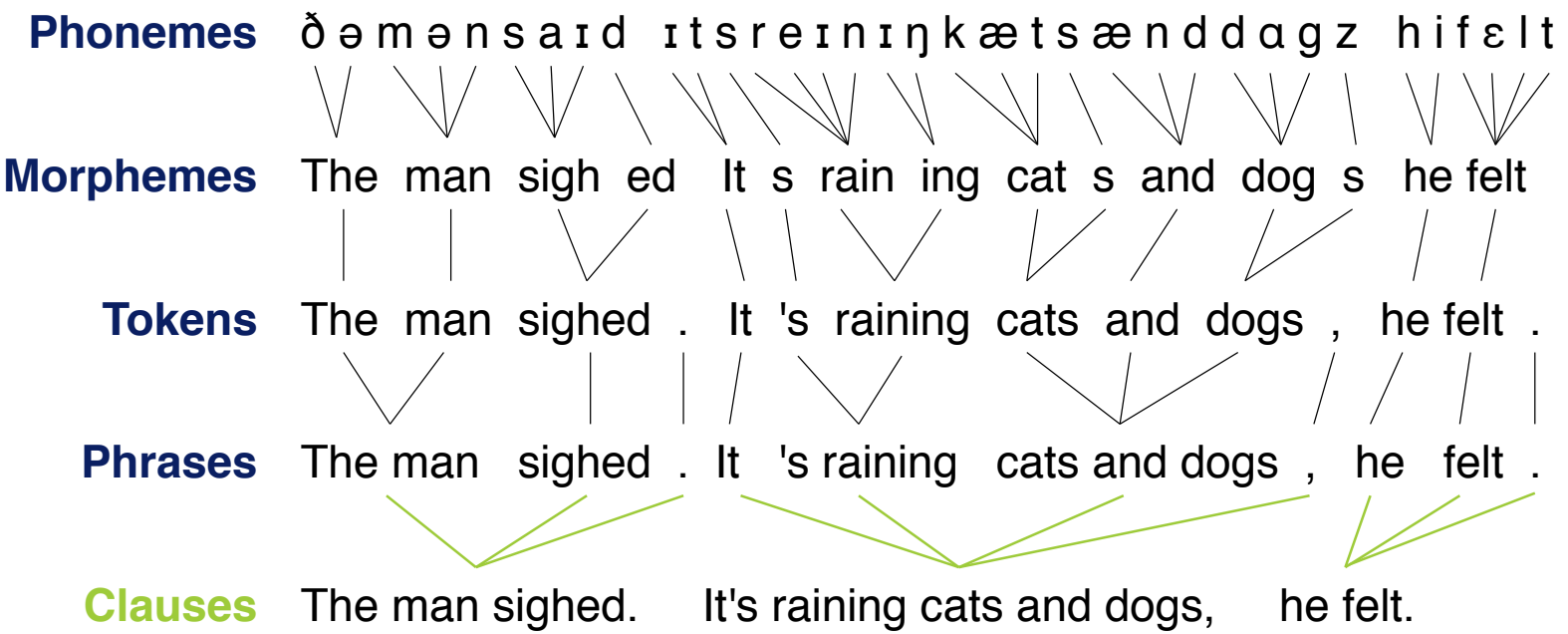
- Noun phrase (NP). “cats and dogs”, “he”, “cat on the mat”
- Verb phrase (VP). “felt”, “jump up and down”
- Prepositional phrase (PP). “in love”, “over the rainbow”
- Adjectival phrase (AP). “full of toys”, “fraught with guilt”
- Adverbial phrase (AdvP). “very carefully”

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



Clauses

What is a 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.

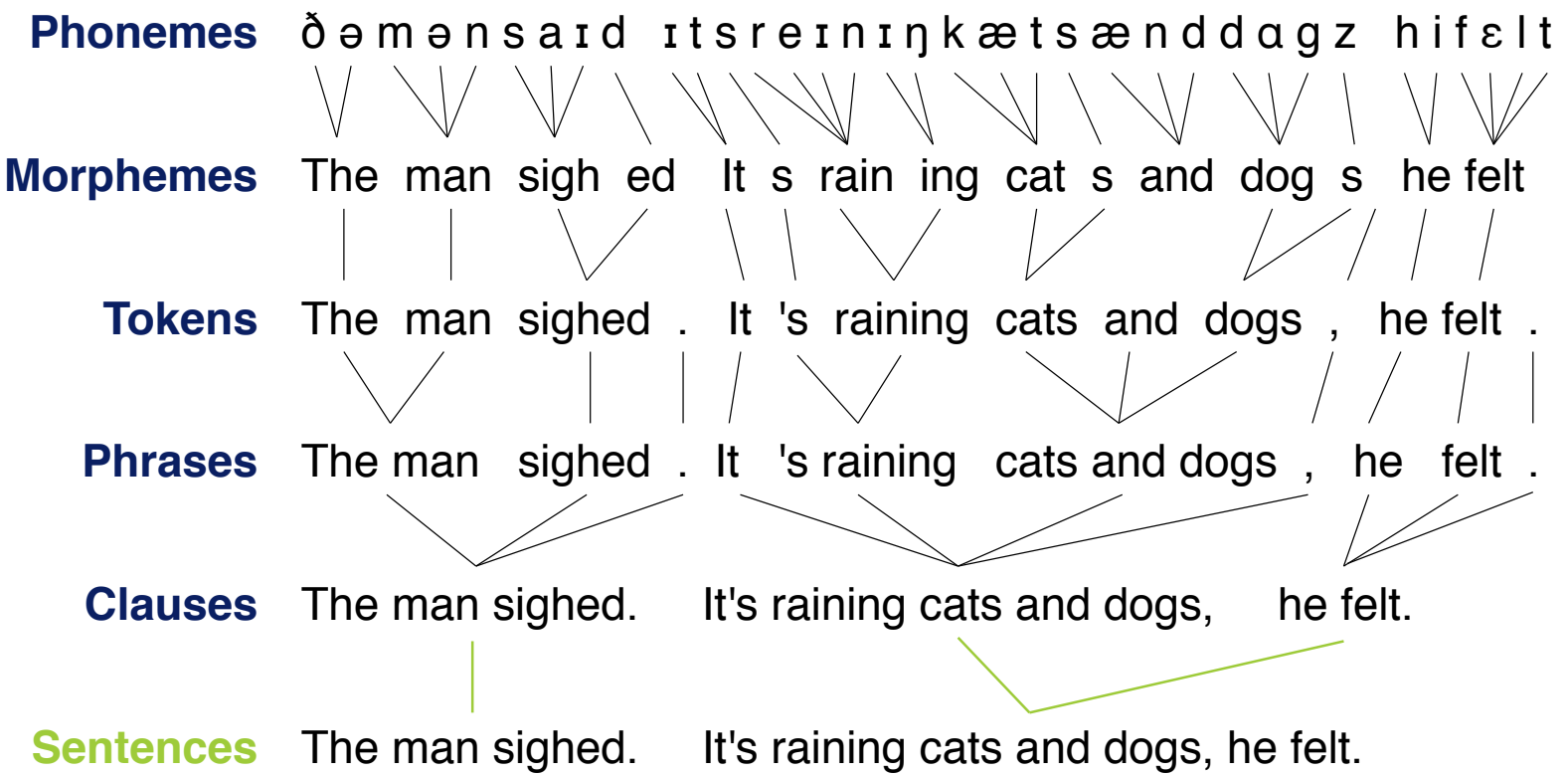
  “Although it rained” “because I was inside the building.”

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



Sentences

What is a sentence?

- A sentence is a grammatically independent linguistic unit consisting of one or more words.
- Contains at least one main clause.
- Most 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 text mining as one of the most basic preprocessing steps.

Grammars

What is a grammar?

- A grammar is 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.

  $A \rightarrow bC$ A structure A is composed of a word b followed by a structure C.
  $C \rightarrow de$ A structure C is composed of a word d followed by a word e.

- Rules consist of terminal and non-terminal symbols.
- Terminal symbols ($\approx$ 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.
- Two types: Constituency parsing and dependency parsing.
- Used in text mining as preprocessing for tasks like relation extraction.

Grammars

Toy Grammar of English

Rules

(convention: S is starting symbol)

S	$\rightarrow$	NP VP	PP	$\rightarrow$	P NP
S	$\rightarrow$	VP	N	$\rightarrow$	cats
VP	$\rightarrow$	V NP	V	$\rightarrow$	scratch
VP	$\rightarrow$	V NP PP	N	$\rightarrow$	claws
NP	$\rightarrow$	NP PP	N	$\rightarrow$	people
NP	$\rightarrow$	N	N	$\rightarrow$	scratch
NP	$\rightarrow$	N N	P	$\rightarrow$	with

Example generation

S $\rightarrow$ NP VP $\rightarrow$ NP V NP $\rightarrow$ N V NP $\rightarrow$ N V NP PP

$\rightarrow$ N V N P NP $\rightarrow$ N V N P N $\rightarrow$ cats V N P N $\rightarrow$ cats scratch N P N

$\rightarrow$ cats scratch people P N $\rightarrow$ cats scratch people with N

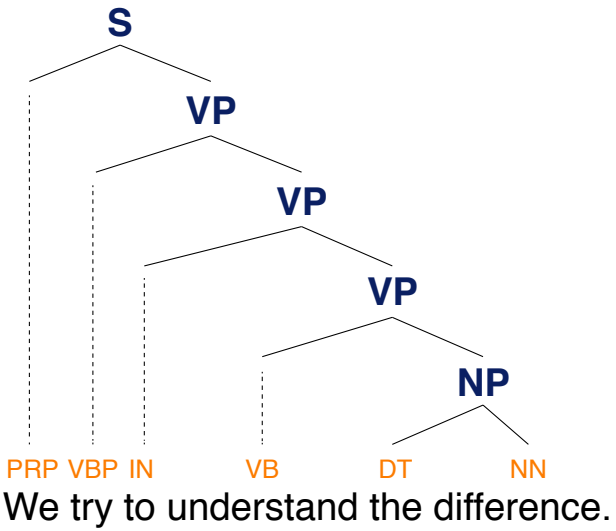
$\rightarrow$ cats scratch people with claws

Grammars

Phrase vs. Dependency Structure

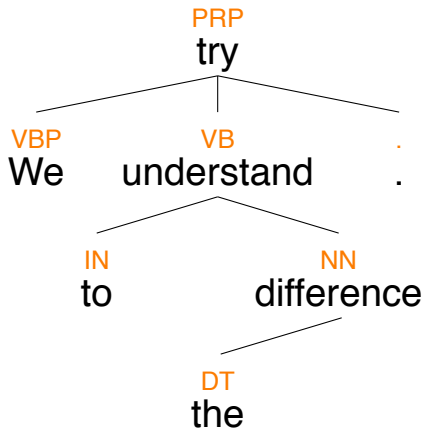
Phrase structure grammar

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

Multiple Valid Syntactic Structures

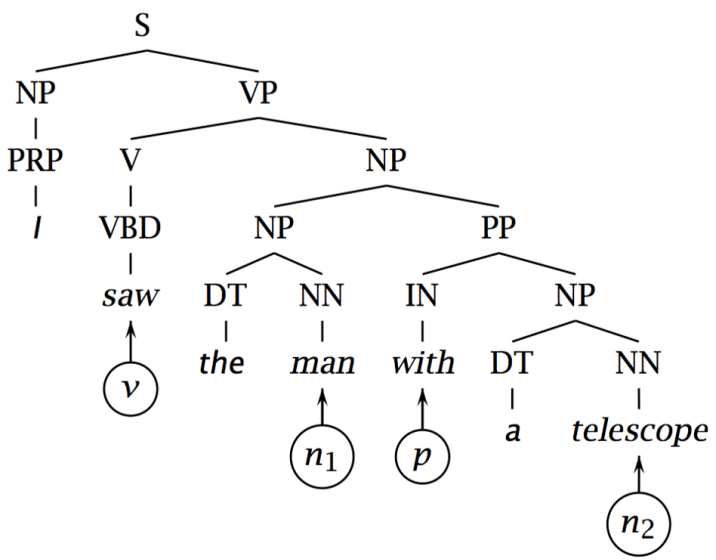
Syntactic ambiguity

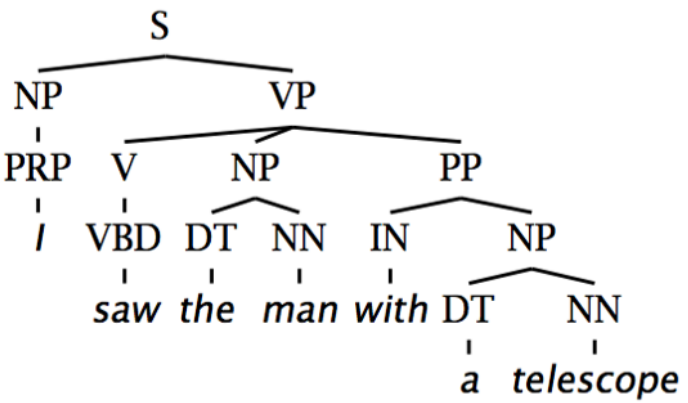
- Arises when one sentence has more than one syntactic structure.

Syntactic Ambiguity

Attachment

Example "I saw the man with a telescope."





Syntactic Ambiguity

Coordination

Coordination ambiguity

- Scope of the conjunction unclear.

Example: “If you love money problems show up.”

- “If you love, money problems show up.”
- “If you love money, problems show up.”
- “If you love money problems, show up.”

Observation

- Commas often help solve the problem.

Syntactic Ambiguity

Garden Paths

Garden Paths

- Sentences that lead you along a path that suddenly turns out to fail.

Examples

- “The man whistling tunes pianos.”
- “The cotton clothing is made of grows in Mississippi.”
- “The complex houses married and single soldiers and their families.”
- “The author wrote the novel was likely to be a best-seller.”
- “The tomcat curled up on the cushion seemed friendly.”
- “The man returned to his house was happy.”
- “The government plans to raise taxes were defeated.”
- “The sour drink from the ocean.”
- “The horse raced past the barn fell.”
- “The old man the boat.”

Semantics

Semantics

What is semantics?

- The meaning of single words and compositions of words.

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

Meaning

What is meaning?

- Propositional content in terms of validity or truth conditions.

"All cats are mortal."	$\forall x : cat(x) \rightarrow mortal(x)$
"Sunny is a cat."	$cat(Sunny)$
Sunny is mortal.	$mortal(Sunny)$

- Often requires common-sense reasoning based on world knowledge.

"Max can open Tim's safe. He knows the combination."

"Max can open Tim's safe. He should change the combination."

- Includes expressed emotional content.

"That poor cat!"

"Fortunately, Max can open Tim's safe."

Construction of meaning

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

Meaning

Linguistic Form

Meaning that can often be derived from linguistic form

- Constant meaning of language across different occasions of use.

  Linda: "Is it raining?"

  Max: "Yes."

- Meaning a speaker publicly commits to by using a certain form in a certain context.

  Linda: "But it's perfectly dry outside. You're mocking me?"

- Inferences about a speaker's private cognitive states.

  Linda: "So you want me to take an umbrella?"

  Max: "I didn't say that."

- Social meaning, such as politeness, formality, peer-group style, ...

  Linda: "Could you be serious, please?"

  Max: "Leave your umbrella. It's clear blue sky."

Meaning

Context of Use

Meaning that can often only be derived from context of use

- Scope of quantifiers
- Word sense ambiguities
- Semantic relations between nouns in compounds

... and many others...

Interpretation interacts with 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

What is lexical semantics?

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

Word senses

- Fine-grained 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.

Lexical Semantics

Word senses

What is a word sense?

- The meaning of a word.
- Words can have multiple senses, due to *polysemy* and *homonymy*.

Example: “ride” has 16 senses, here is a selection

- ride over, along, or through
- sit and travel on the back of animal, usually while controlling its motions
- be carried or travel on or in a vehicle
- be contingent on
- harass with persistent criticism or carping
- keep partially engaged by slightly depressing a pedal with the foot
- continue undisturbed and without interference
- move like a floating object

Lexical Semantics

Polysemy vs. Homonymy

Constructional polysemy

- Related senses that have the same lexical entry.

"newspaper" (physical object vs. abstract content)

Sense extension polysemy

- Regular ways of deriving new word senses given a member of a class.

"chicken" (animal vs. meat of the animal)

Homonoymy

- Unrelated word senses that have the same lexical entry.

"bank" (river bank vs. money bank)

Lexical Semantics

Dropped Predicate Arguments

Missing predicate arguments can often be inferred

“Have you eaten?”	A meal.
“I drank all night.”	Alcohol.
“Max will bake tomorrow afternoon.”	A flour-based product.

Defeasible vs. lexically specific predicate arguments

- Defeasible. Arguments that can be inferred from the semantics.

  “He has symptoms of diabetes. For instance, he drinks all the time.” Not alcohol.

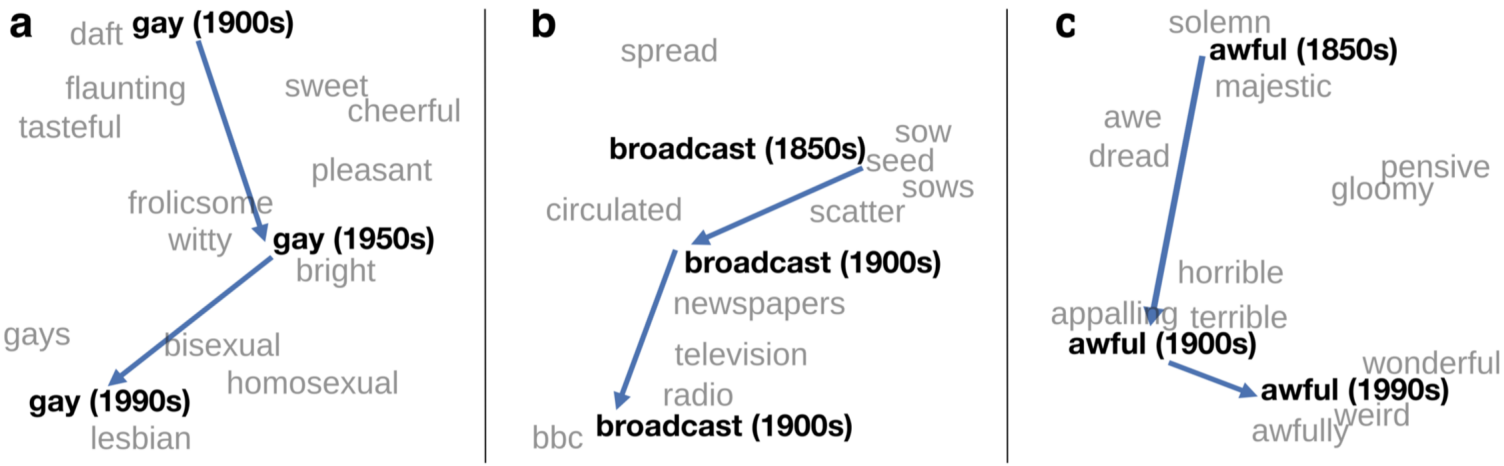
- Lexically specific. Arguments that are implicitly decided by a predicate.

  “Max sipped.” At his drink (or glass).

Lexical Semantics

Word Sense Goes Wild

Word senses may change over time



(Hamilton et al., ACL 2016)

Metaphoric word senses

- Metaphors add senses to words in (theoretically) unbounded ways.

"I have always despised politics.
But I have climbed to the top of that greasy pole."

Lexical Semantics

Semantic roles

What are 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 text mining when deeper language understanding is required.

Between Lexical and Compositional Semantics

Multi-Word Expressions

What is a multi-word expression?

- Lexical units larger than a word that can bear both compositional and idiomatic meanings.

"driving instructor"

"argumentation quality assessment"

"vice versa"

- On the boundary between lexical and compositional semantics.

"Kick the bucket."

"Long time no see."

Word n-grams

- Text mining often simply uses word bigrams, trigrams, or similar to capture multi-word expressions.
- Approaches to mine multi-word expressions exist, too.

Between Lexical and Compositional Semantics

Word n-grams

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."

Numbers of n-grams

- For a sequence of $m \geq n$ tokens, the number of n-grams is $(m - n) + 1$.
- Google's freely available 5-gram corpus in version 1:

1-grams	2-grams	3-grams	4-grams	5-grams	Tokens	Sentences
13.6 million	314.8 million	977.1 million	1.3 billion	1.2 billion	1.0 trilion	95.1 billion

- The most frequent 3-gram on the English web: "all rights reserved".
- n-grams with less than 40 occurrences are not included.

Between Lexical and Compositional Semantics

Entities

What is an entity?

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

Entity types

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

  Persons, locations, organizations, products, ...

  "Jun.-Prof. Dr. Henning Wachsmuth" "in Paderborn" "at Paderborn University"

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

  Temporal and monetary expressions, phone numbers, ...

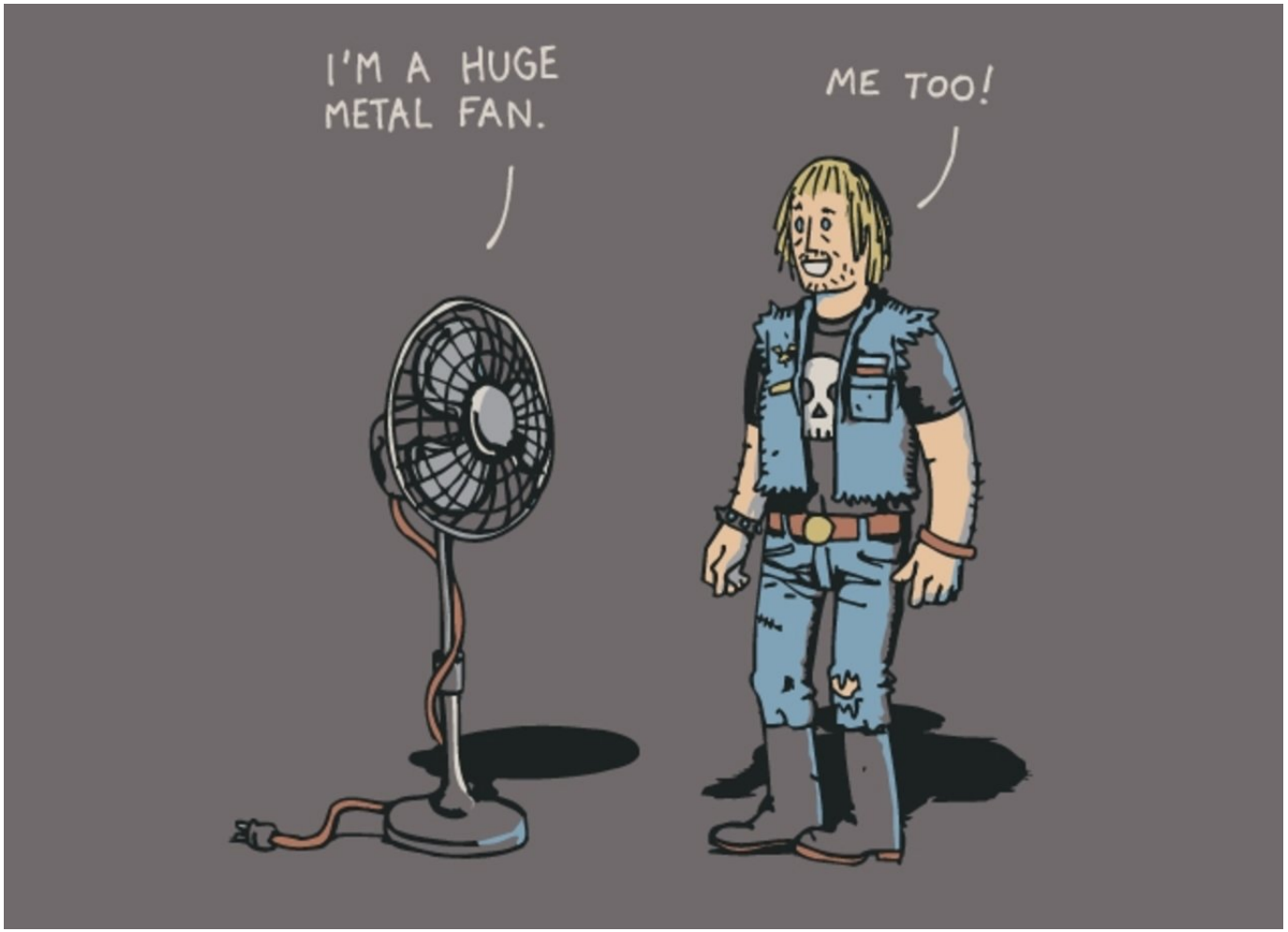
  "in this year" "2018-10-18" "$ 100 000" "60-68 44"

Named and numerical entity recognition

- The analyses that mine respective entities from text.
- Used in text mining as a key analysis for information extraction tasks.

Semantic Ambiguity

Multi-Word Expressions



Compositional Semantics

What is compositional semantics?

- The meaning of the composition of words in phrases, sentences, ...
- The translation of natural language strings to logical forms.
- Covers relations, scopes of operators, collocations, and much more.

Relations

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

Linguistic operators

- Quantifiers. Indicating the quantity of objects.
- Hedges. Lessening the impact of a proposition.
- Negation. Inverting an adjective, predicate, or similar.

Collocation

- Words appearing together overproportionally often.

Compositional Semantics

Semantic Relations

What are 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.

  reads(agent, theme) $\rightarrow$ "Max reads a book."

  founded(organization, time) $\rightarrow$ "Google was established in 1998."

- Events. Relations with multiple arguments, possibly nested relations.

  reads(agent, theme, date, time, location, origin) $\wedge$ 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 text mining as a key analysis for information extraction tasks.

Compositional Semantics

Scope

What is scope?

- The range of text affected by a linguistic operator.

  Important in many tasks, such as information extraction and sentiment analysis.

Selection of operator types

- Quantifiers. Scope depends on syntax, but is not decided by it.

  "Every student reads some book."

  $\forall(x, student(x), \exists(y, book(y), read(x, y)))$ vs. $\exists(y, book(y), \forall(x, student(x), read(x, y)))$

- Hedging. Scope decided by syntax, if used correctly.

  "I worked only tonight." vs. "I only worked tonight."

  "Probably every student reads a book." vs. "Every student probably reads a book."

- Pronouns. Syntax helps, but resolution of scopes can be very complex.

  "A person got run over on market square. He really got angry about it."

- Negation. Ditto.

  "It's not good manners I don't care about."

Compositional Semantics

Collocation

What is collocation?

- Sequences of two or more words that appear together with greater frequency than their individual frequencies would predict.

“do homework”	Similarly frequent as “homework” alone.
“in my opinion”	Most typical phrase including “opinion”.
“vice versa”	Sometimes, multiple words may be more frequent, e.g., in idioms.

Observations

- Collocations are often less ambiguous than the words taken in isolation.

  Due to knowledge about preferred linguistic forms, contexts, and meanings.

  “heavy smoker” (meaning clear, although “heavy” has almost 30 word senses)

- Multi-word expressions are a particular type of collocations.

Discourse and Pragmatics

Discourse

What is (linguistic) discourse?

- Discourse describes linguistic units that are larger than a sentence.
- Usually referring to the entirety of a given text.
- Discourse needs to be *coherent*, in order to be understandable.

Discourse vs. dialogue

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

Linguistic vs. societal discourse

- The notion of discourse also plays an important role in the humanities.
- Such societal discourse is a related but not the same concept.

Discourse

Coherence

What is coherence?

- Coherence is the continuity of meaning in discourse.

"Max hid Bill's car keys. He drank too much." Coherent.
"Max hid Bill's car keys. He likes spinach." Coherent?

Global vs. local coherence

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

Local coherence does not guarantee global coherence

"Max hid Bill's car keys. He drank too much. " Locally coherent.
"He drank too much. No water was left." Locally coherent.
"Max hid Bill's car keys. He drank too much. No water was left." Globally coherent?

Coherence vs. Cohesion

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

Discourse

Coreference

What is coreference?

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

Common types of coreference

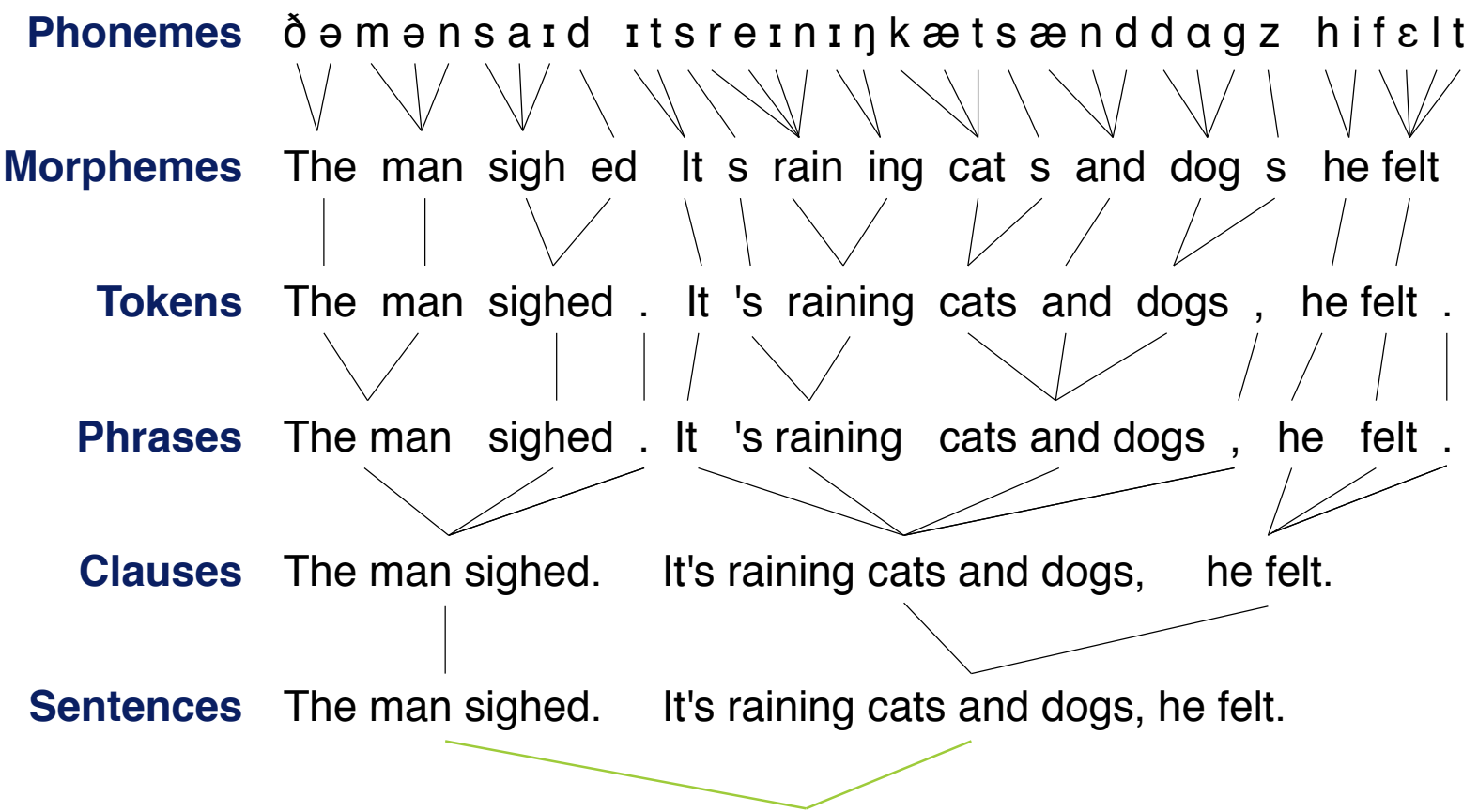
- Anaphora. “Max walked in. He sat down.”
- Cataphora. “After he walked in, Max sat down.”
- Split antecendents. “Max asked Linda to join. They arrived together.”
- Coreferring noun phrases. “Apple is based in Cupertino. The company is actually called Apple Inc.

Coreference resolution

- The text analysis that maps all references to umambiguous identifiers.
- Used in text mining as preprocessing for tasks like entity recognition.
- Coreference resolution is a very hard task.

Linguistic Text Units

Paragraphs



Discourse

Paragraphs and Other Discourse-Level Units

What is a paragraph?

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

Discourse-level linguistic units

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

Discourse

Segments and Relations

Discourse segment

- Building block of a discourse in terms of linguistic units.
- May consist of multiple smaller adjacent discourse segments.
- Elementary discourse unit (EDU). Smallest discourse segments, usually clauses or sentences.

Coherence relations (aka rhetorical or discourse relations)

- Describes how two segments are related to each other.
- Can be semantic (subject matter) or pragmatic (presentational).
- Can be coordinating (paratactic) or subordinating (hypotactic).

Discourse parsing

- The text analysis that infers the discourse structure of a text.
- Used in text mining for tasks where structure is important.

Discourse Structure

Discourse structure

- Represents the organization of an entire text.
- Coherence relations between the contents of discourse segments.
- The most common model is the *rhetorical structure theory*.

Rhetorical Structure Theory (RST)

- Models discourse structure hierarchically via coherence relations between adjacent segments.
- A coherent text is supposed to have a fully connected RST tree.
- The original RST considers 22 relation types.

Circumstance	Volitional cause	Anthithesis	Evidence
Solutionhood	Non-volitional cause	Concession	Justify
Elaboration	Volitional result	Condition	Restatement
Background	Non-volitional result	Otherwise	Summary
Enablement	Purpose	Interpretation	Sequence
Motivation		Evaluation	Contrast

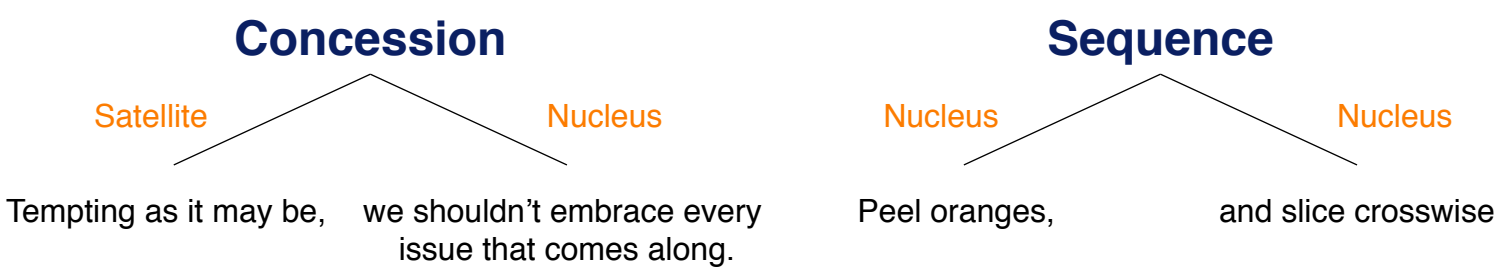
Discourse Structure

Relation Types

Subject matter vs. presentational relations

- Subject matter. Relations between the content of text spans.
  Cause, purpose, condition, summary, ...
- Presentational. Relations describing the rhetorical effect on the reader.
  Motivation, antithesis, background, evidence, ...

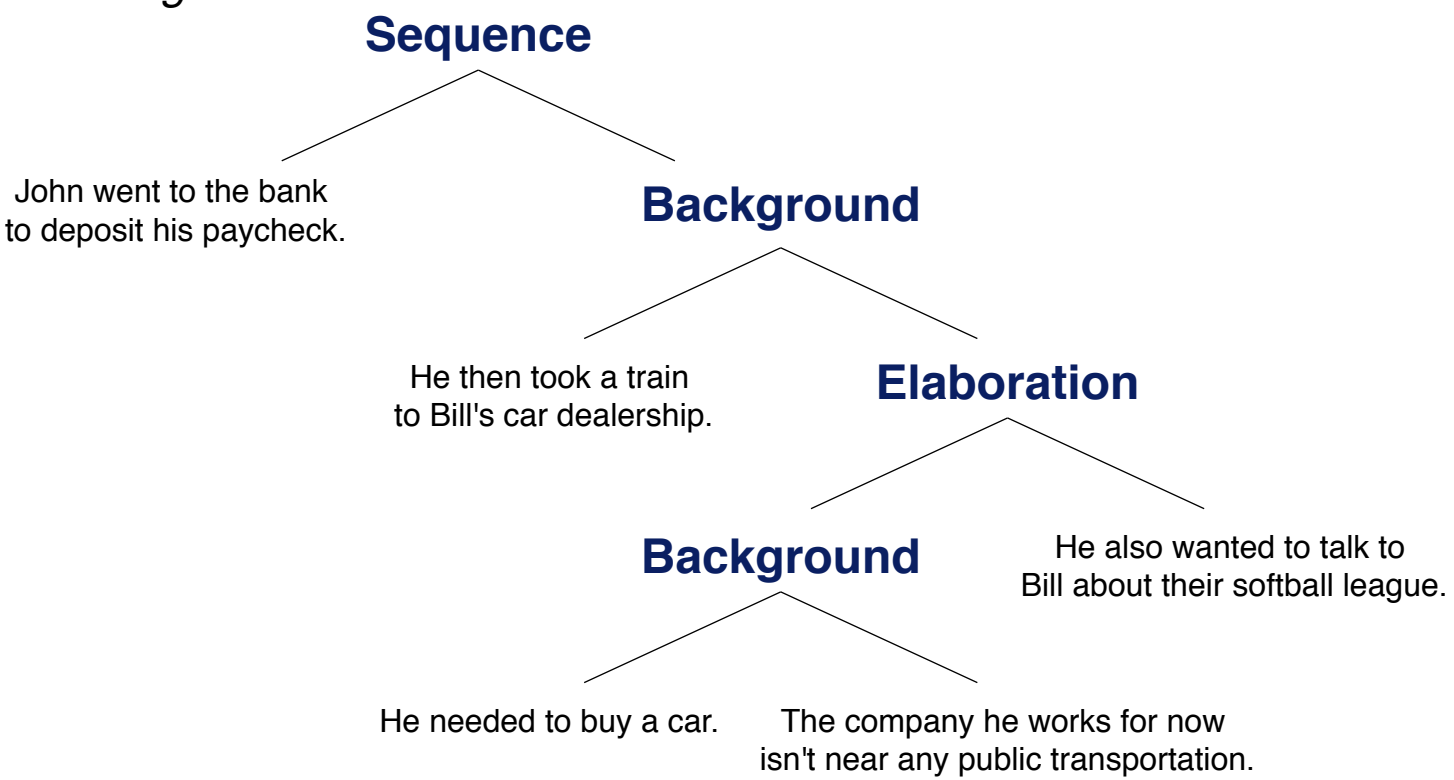
Paratactic vs. hypotactic



Discourse Structure

Example RST Tree

"John went to the bank to deposit his paycheck. He then took a train to Bill's car dealership. He needed to buy a car. The company he works for now isn't near any public transportation. He also wanted to talk to Bill about their softball league."



Discourse Structure

Implicitness and Explicitness

Discourse markers

- Connectives and particles used to indicated discourse relations.
- Connectives. “because”, “as a result”, “and”, “whereas”, “but”, ...
- Particles. “well”, “you know”, “I mean”, ...

Discourse relations are often implicit

Implicit. “I took my umbrella. It was raining outside.”

Explicit. “I took my umbrella, because it was raining outside.”

Also arguments of discourse relations may be implicit

Implicit. “Sunny is a cat. So, Sunny is mortal.”

Explicit. “Sunny is a cat, and all cats are mortal. So, Sunny is mortal.”

Pragmatics

What is 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.
- Covers speech acts, information status and structure, presupposition and implicature, ...

Speech Acts

- Linguistic utterances with a performative function.

Information status and structure

- Status. Relation of information to a common ground.
- Structure. Linguistic structure to clarify information status.

Presupposition and implicature

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

Pragmatics

Intention of the Author/Speaker

Intention of "I never said she stole my money."

I never said she stole my money.	Someone else said it, but I didn't.
I *never* said she stole my money.	I simply didn't ever say it.
I never *said* she stole my money.	I might have implied it in some way. But I never explicitly said it.
I never said *she* stole my money.	I said someone took it. But I didn't say it was her.
I never said she *stole* my money.	I just said she probably borrowed it.
I never said she stole *my* money.	I said she stole someone else's money.
I never said she stole my *money*.	I said she stole something of mine. But not my money.

Pragmatics

Goal of an Author/Speaker

Example: Argumentation

- Argument. A conclusion (claim) supported by premises (reasons) that conveys a stance on a controversial issue.

  Conclusion: "The death penalty should be abolished."
  Premise 1: "It legitimizes an irreversible act of violence."
  Premise 2: "As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."

- Argumentation. The usage of arguments and rhetorical means in the discussion a controversial issue.
- Goals. Persuasion, justification, agreement, deliberation, or similar.

Influencing factors (according to Aristotle)

- Good arguments, credibility, and emotions.
- The clarity and appropriateness of the used language.
- The arrangement of argumentation.

Speech Acts

What is a speech act?

- A speech act is 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.

Information Status and Structure

Information status

- Relationship of referents to common ground.
- Predominantly expressed by choice of determiners.

  In some languages also by presence/absence of case marking or specific morphology.

"a man" vs. "the man" vs. "that man" vs. "him" vs. ...

Information structure

- Distinguishes what is presented as given vs. new
- What's expressed as given might not be mutually known.

"It must have been Max who said that."

- What's expressed as new should also be new.

"Who voted for option A? Max. Linda voted for option B."

Presupposition

What is presupposition?

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

"Max' cousin took an aspirin." $\rightarrow$ Max has a cousin, someone's called Max

Selection of linguistic triggers

- Lexical items. "know", "regret", "manage (to)"
- Proper names. "Max"
- Definite descriptions. "the cat"
- Possessives. "Max' cousin"
- Iterative adverbs. "also", "again", "too"
- Ordinals. "second", "third"
- Domain of quantification. "all the kids are happy"

Implicature

What is implicature?

- What is suggested by a linguistic utterance, even though neither expressed nor entailed.
- In cooperative conversations, utterances can be interpreted based on the assumption that people try to follow the *Gricean Maxims*.

Gricean Maxims (after Paul Grice, 1975)

- Maxim of Quality. Do not say what you believe to be false. Do not say that for which you lack adequate evidence.
- Maxim of Quantity. Make your utterance as informative as is required. Do not make it more informative than is required.
- Maxim of Relation. Be relevant.
- Maxim of Manner. Avoid obscurity of expression. Avoid ambiguity. Be brief. Be orderly.

Implicature

Conversational Implicature

Conversational implicature is calculable from what is said

Linda: "Did the students pass the exam."
Max: "Some of them did." → Not all of them.

Linda: "I'm out of gas."
Max: "There's a gas station around the corner." → Linda can get gas from there.

Linda: "Are you coming out tonight?"
Max: "I have to work." → Max won't come.

Conversational implicature is cancellable

Linda: "Are you coming out tonight?"
Max: "I have to work. But I'll come out anyway."

Linda: "I'm out of gas."
Max: "There's a gas station around the corner. However, it's closed."

Implicated agreement and denial

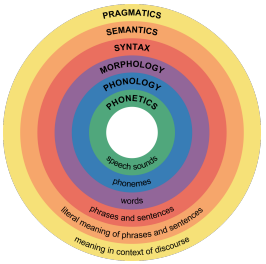
Linda: "He's brilliant and imaginative."
Max: "He's imaginative."

Conclusion

Summary

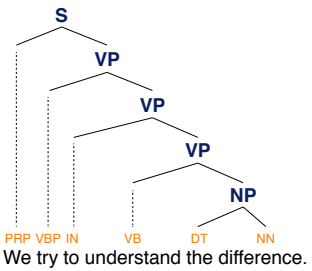
Linguistics in text mining

- Text mining analyzes natural language text.
- Linguistic concepts define the basis of all analyses.
- The analysis can take place at several different levels.



Morphology and syntax

- How words are formed and grammar is constructed.
- Central concepts are tokens, phrases, and sentences.
- Text mining analyzes these levels for preprocessing.



Semantics and pragmatics

- How meaning is conveyed and language is used.
- Central concepts are entities, relations, and discourse.
- Text mining targets the results of these analysis 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.