732A47 Text Mining / 2015

Introduction to Computational Linguistics

Marco Kuhlmann

LINKÖPING UNIVERSITY

Terms abound

- Computational Linguistics (CL)

  most general term; focus on adequate models

- Natural Language Processing (NLP)

  origin within Artificial Intelligence; focus on computing

- (Human) Language Technology (HLT)

  a synonym favoured by the EU; focus on applications

- Natural Language Engineering (NLE)

  focus on performance and practical use

Major tasks in natural language processing

- coreference resolution
- discourse analysis
- machine translation
- morphological segmentation
- named entity recognition
- natural language generation
- natural language understanding
- optical character recognition
- part-of-speech tagging
- question answering
- relationship extraction
- sentence segmentation
- sentiment analysis
- speech recognition
- speech segmentation
- syntactic parsing
- text summarisation
- topic segmentation
- word segmentation
- word sense disambiguation

Source: Wikipedia

Why should you care?

- When mining text written in a natural language, knowing about natural language processing actually helps!

- It is tempting to view text as a bag of words.

  Many NLP people do it, too.

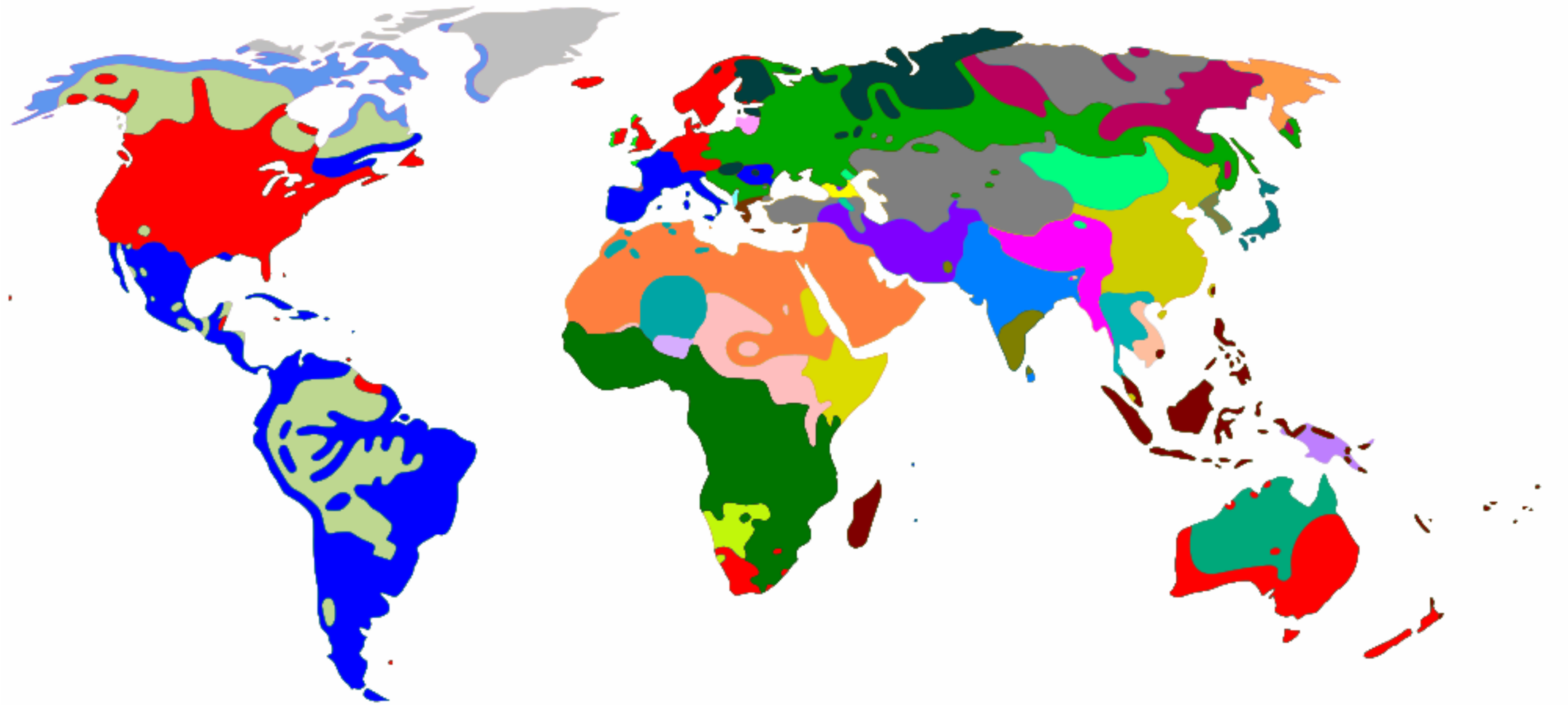
- However, doing so is to miss out on a lot of information.

  Google bought YouTube. YouTube bought Google.

- The language technology market is big and growing.

  $13.4 billion by 2020 (Source: marketsandmarkets.com)

7,000 languages



Source: Sorry, forgot

Five most populous language families

Language Family	Living Languages	Examples	Population
Indo-European	430	English, Welsh, Pashto, Bengali	44.78%
Sino-Tibetan	399	Mandarin Chinese, Sherpa, Burmese	22.28%
Niger-Congo	1,495	Swahili, Wolof, Bissa	6.26%
Afro-Asiatic	353	Arabic, Modern Standard Coptic, Somali	5.93%
Austronesian	1,246	Tagalog, Balinese, Hawaiian	5.45%
Total	**3,923**		**84.70%**

A bit of history

- 1950s: Fully automatic translation of Russian into English
- 1960s: Diaspora after funding cuts (ALPAC report)
- 1970s: Conceptual ontologies and chatterbots
- 1980s: Systems based on complex sets of hand-written rules
- 1990s: The surge of statistical techniques
- 2000s: Large corpora. Machine translation once more
- 2010s: Dominance of machine learning

What is a corpus?

lat. corpus, oris n ‘body’

- A body of texts, utterances or other specimens considered more or less representative of language and usually stored electronically.

  The Oxford Companion to the English Language

- In many cases, a standard linguistic data set used to train or evaluate computational models of language.

  traditionally, manually produced by experts

- In common parlance, the text or text file you are currently using.

Reading

- Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

  chapters 3, 5, 7

- Emily M. Bender. *Linguistic Fundamentals for Natural Language Processing. 100 Essentials from Morphology and Syntax*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, 2013.

  optional; covers much of the linguistic background

Structure

What is computational linguistics?

Prelude: From raw text to lists of words

From form to meaning

Presentation of the lab project

Prelude: From raw text to lists of words

Early stages in text mining

Stage	Description
Selection	Decide which texts to include.
Collection	Acquire the texts (e.g. by crawling web pages).
Unformatting	Strip off markup (such as XML) and metadata.
Segmentation	Divide the text into relevant units (such as tokens).
Normalisation	Make the text more uniform (e.g. by lowercasing).

Sentence segmentation

- **Sentence segmentation** refers to the task of dividing a text into individual sentences.
- Sentence segmentation is harder than splitting at periods.

  We visited the U.S. After that, we visited Canada.

Tokenisation

- **Tokenisation** refers to the task of segmenting a text into individual words or word-like units.

  numbers, punctuation marks

- Tokenisation is harder than splitting at whitespace.

  we’re; don’t; bl.a.; Hewlett-Packard; San Francisco

- Tokenisation is even harder for non-European languages.

  Chinese word segmentation

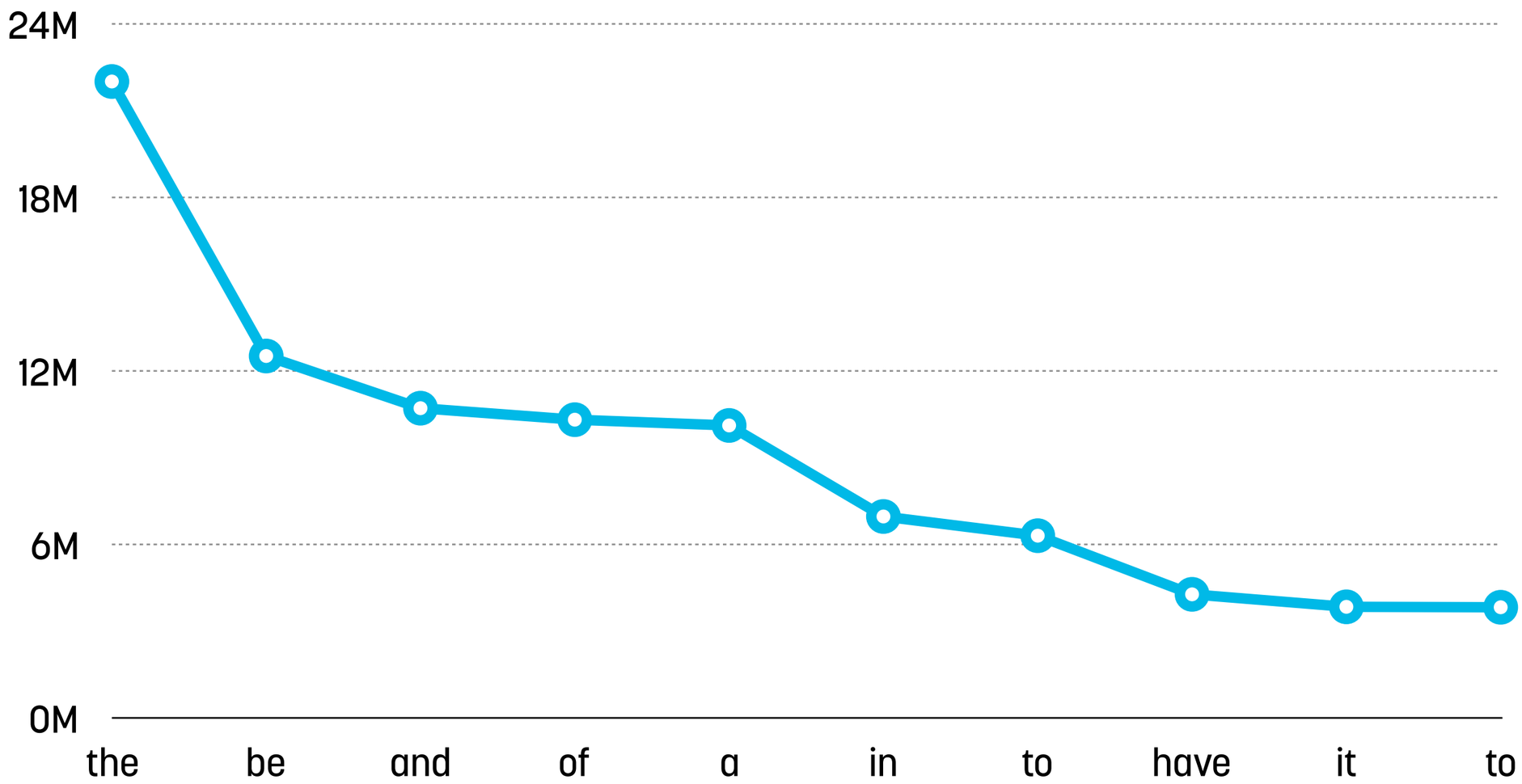
‘Rose is a rose is a rose is a rose.’

Gertrude Stein (1874–1946)

Word tokens and word types

Corpus	Number of tokens	Number of types
Shakespeare	ca. 884,000	ca. 31,000
Riksmöte 2012/2013	4,645,560	96,114
Google Ngrams	1,176,470,663	13,588,391

Zipf’s law



Most frequent words in COCA Corpus | Source: wordfrequency.info

Normalisation

- Lowercasing

  windows vs. Windows

- Harmonisation of spelling variants

  colour, color; gaol, jail; metre, meter

- Stemming (suffix removal)

  wanted → want, wanting → want, happily → happily

Stop words = words with high frequency but little relevance

och det att i en jag hon som han på den med var sig för så till är men ett om hade de av icke mig du henne då sin nu har inte hans honom skulle hennes där min man ej vid kunde något från ut när efter upp vi dem vara vad över än dig kan sina här ha mot alla under någon eller allt mycket sedan ju denna själv detta åt utan varit hur ingen mitt ni bli blev oss din dessa några deras blir mina samma vilken er sådan vår blivit dess inom mellan sådant varför varje vilka ditt vem vilket sitta sådana vart dina vars vårt våra ert era vilkas

Regular expressions

- **Regular expressions** (‘regexps’) can be used to find and manipulate text in files and streams.
- Many Unix-tools have regular expressions built in. Most programming languages provide support for regular expressions.

Structure

What is computational linguistics?

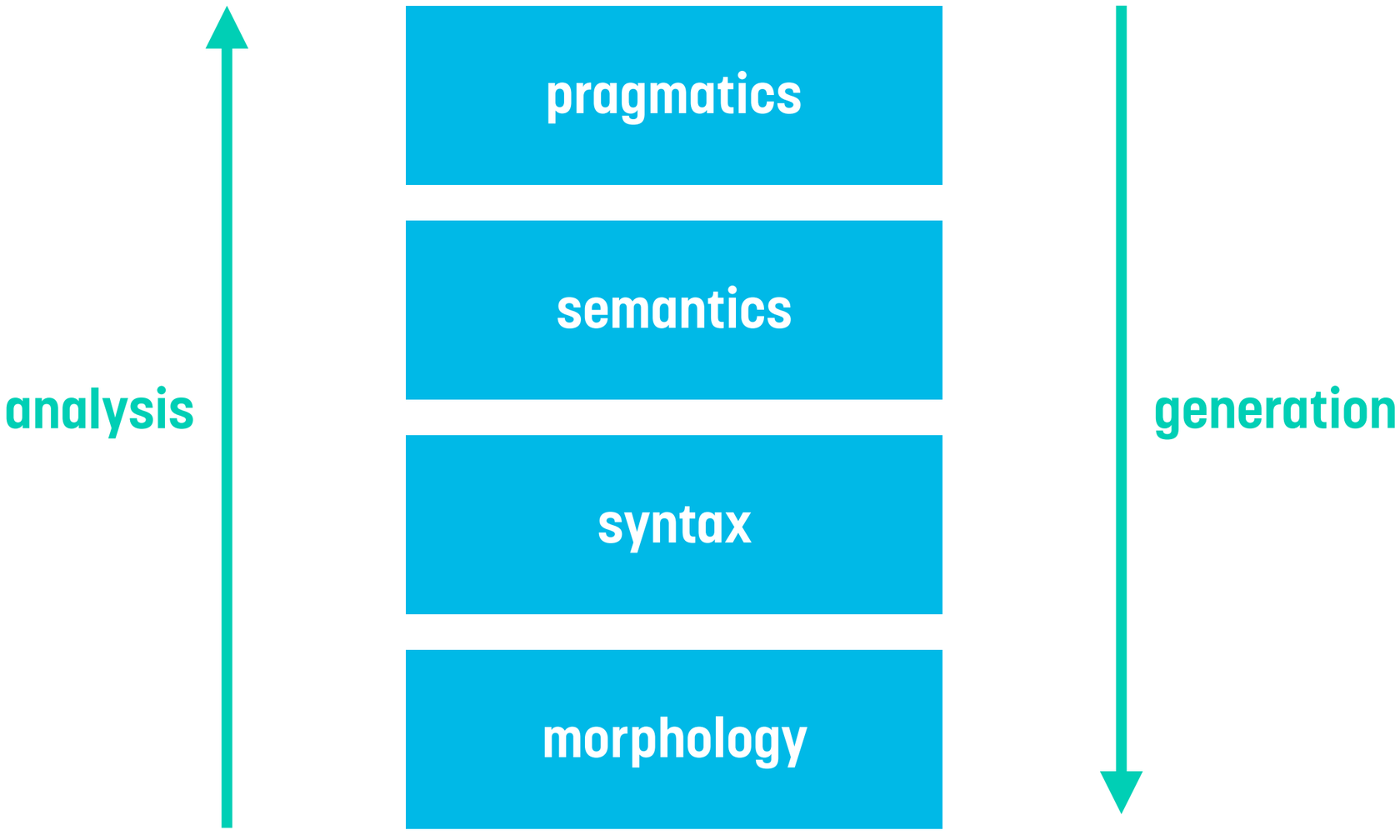
Prelude: From raw text to lists of words

From form to meaning

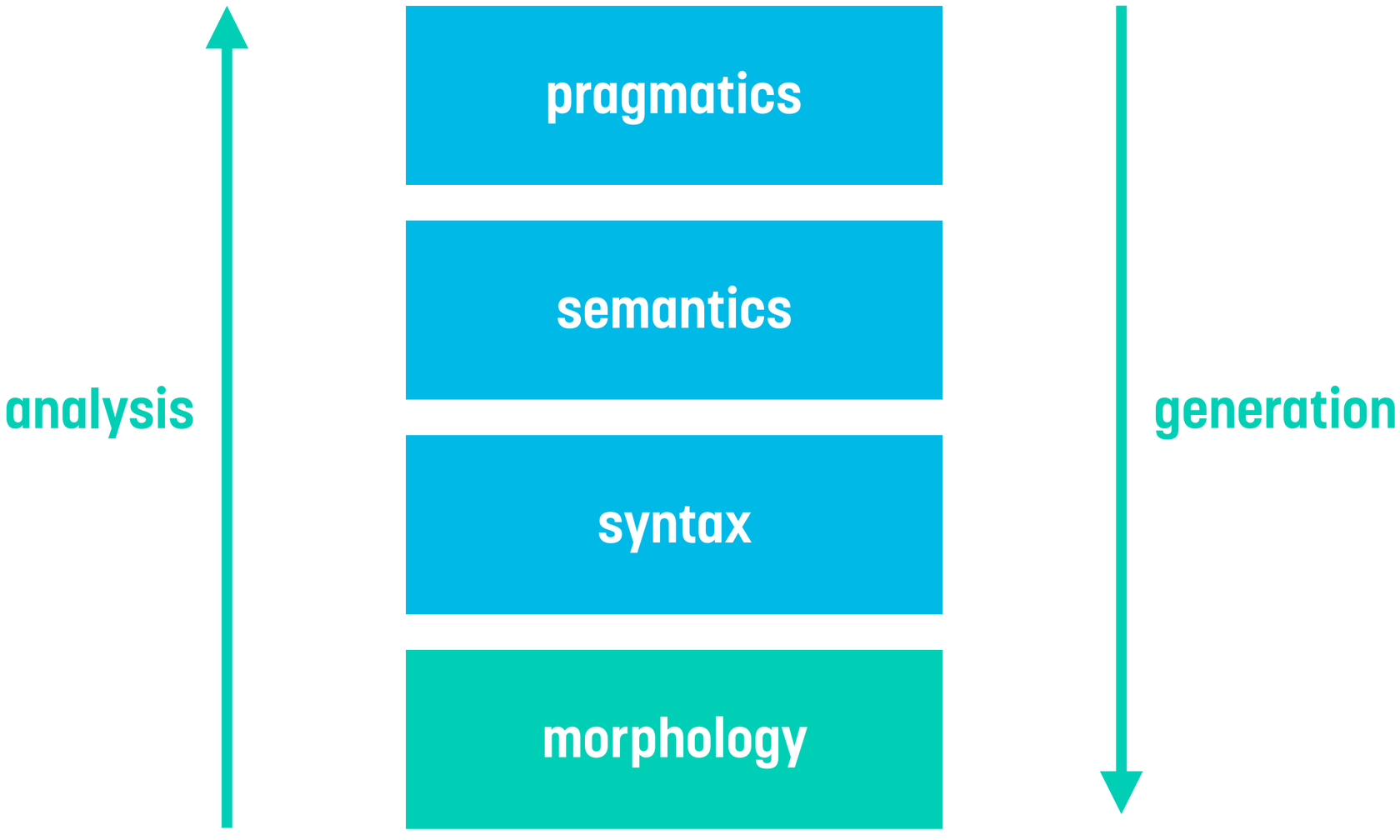
Presentation of the lab project

From form to meaning

From form to meaning



Morphology


pragmatics
semantics
syntax
morphology
analysis
generation

Morphemes

- **Morphemes** are the smallest meaningful units of language.

  Morpheme+s are the small+est mean+ing+ful unit+s of language.

- A word consists of one **root** morpheme and zero or more **affixes**.

  draw, draw+s, draw+ing+s, un+draw+able

- The sounds making up a morpheme do not have to be contiguous.

  Hebrew root к-т-в ‘write’: כתב (katav) ‘wrote’, מכתב (mixtav) ‘a letter’

- The form of a morpheme may even be null.

  French: je mange+∅ ‘I eat’, tu mange+s ‘you eat’, il/elle mange+∅ ‘he/she eats’

A case for morphology: Agreement

- Verbs commonly agree with one or more of their arguments in person, number, and gender.

  He/she/it jumps. I/we/you/they jump.

- Determiners and adjectives commonly agree with nouns in number, gender, and case.

  ett stort hus. de stora husen.

- By detecting agreement, we can detect which words go together.

Three problems in morphological analysis

- Regularity

  How to represent regularities in morphological variation?

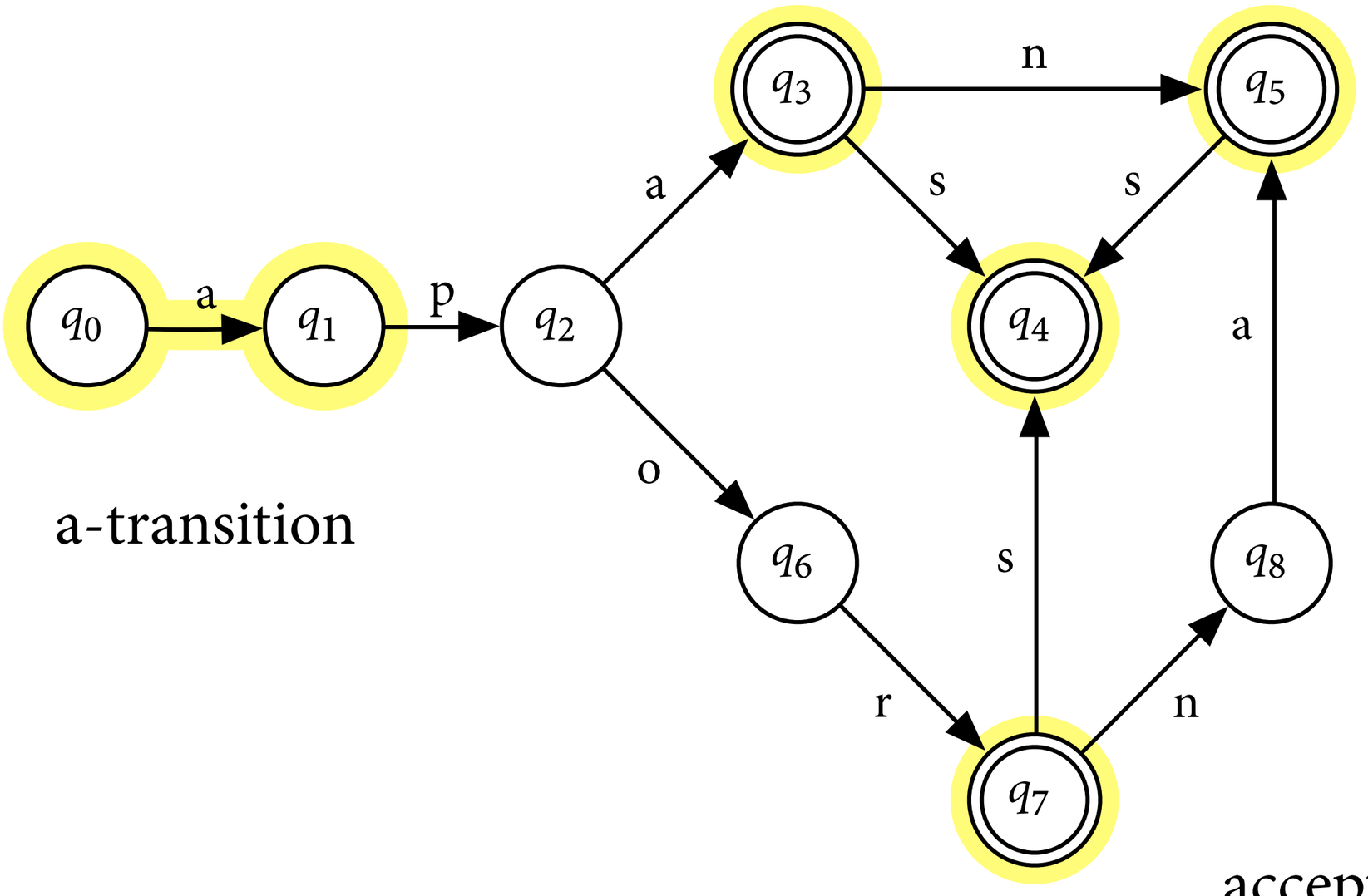
- Ambiguity

  Time flies like an arrow; fruit flies like a banana.

- Dynamicity

  manspreading, weak sauce, cupcakery

Finite automata as morphological lexicons



Related: Lexemes and lemmas

- The term **lexeme** refers to the set of all word forms that have the same fundamental meaning.

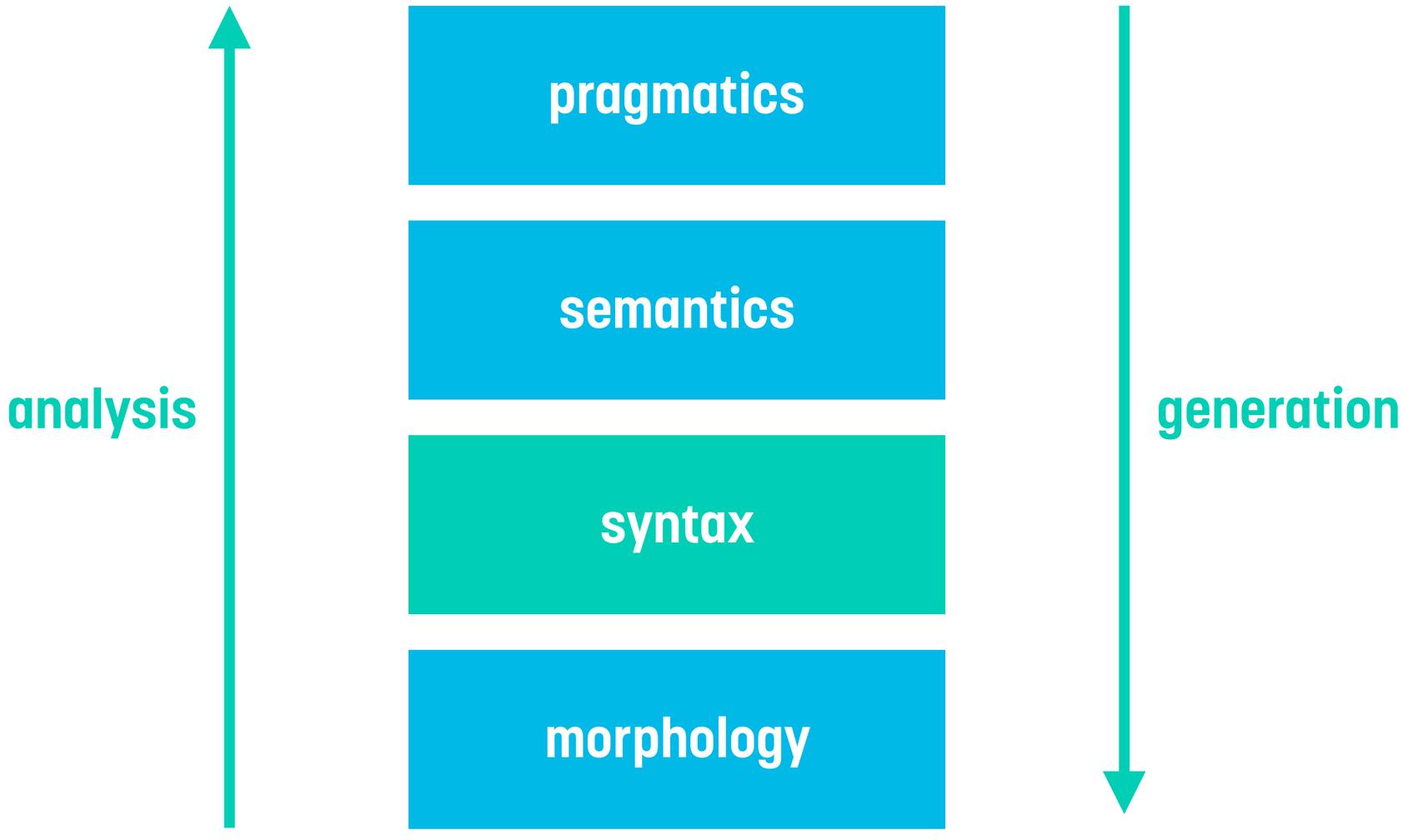
  run, runs, ran, running

- The term **lemma** refers to the particular word form that is chosen by convention to represent a given lexeme.

  what you would put into a lexicon (run)

- The process of determining the lemma for a given word is called **lemmatisation**.

Syntax


pragmatics
semantics
syntax
morphology
analysis
generation

Syntax

- Syntax is what puts constraints on possible sentences.
- Theoretical syntacticians study what formal mechanisms are required in order to describe natural languages in this way.

  connection to formal language theory

- These questions are not directly relevant to natural language processing, which needs to be robust to ill-formed input.

  obvious exceptions: grammar checkers, text generation systems

Parts of speech

- A **part of speech** is a category of words that play similar roles within the syntactic structure of a sentence.

- Parts of speech can be defined distributionally or functionally.

  Kim saw the {elephant, movie, mountain, error} before we did.

  verbs = predicates; nouns = arguments; adverbs = modify verbs, …

- There is no universal set of parts of speech.

Google's universal part-of-speech tags

Tag	Category	Example
VERB	verb	*throw*
NOUN	noun	*pudding*
PRON	pronoun	*she*
ADJ	adjective	*happy*
ADV	adverb	*not*
ADP	adposition (preposition, postposition)	*over*
CONJ	conjunction	*and*
DET	determiner	*the*
NUM	cardinal numbers	*three*
PRT	particle or other function word	*to*
X	other (foreign word, typo, abbreviation)	*fika*
.	punctuation	*?*

Ambiguity causes combinatorial explosion

jag	bad	om	en	kort	bit
PN	VB	PP	DT	JJ	NN
NN	NN	SN	PN	AB	VB
		PL	RG	NN	
		AB	NN		

384 potential analyses

Part-of-speech tagging as classification

- Part-of-speech tagging can be cast as a classification problem.

  Requires corpora manually tagged with parts of speech.

- In this form, it can be attacked with a large number of techniques.

  Hidden Markov Models, linear classifiers (perceptron), logistic regression

- Can be evaluated using accuracy, precision, and recall.

  baseline (most frequent class): 91% accuracy; inter-annotator agreement: 96%

Stockholm Umeå Corpus (SUC)

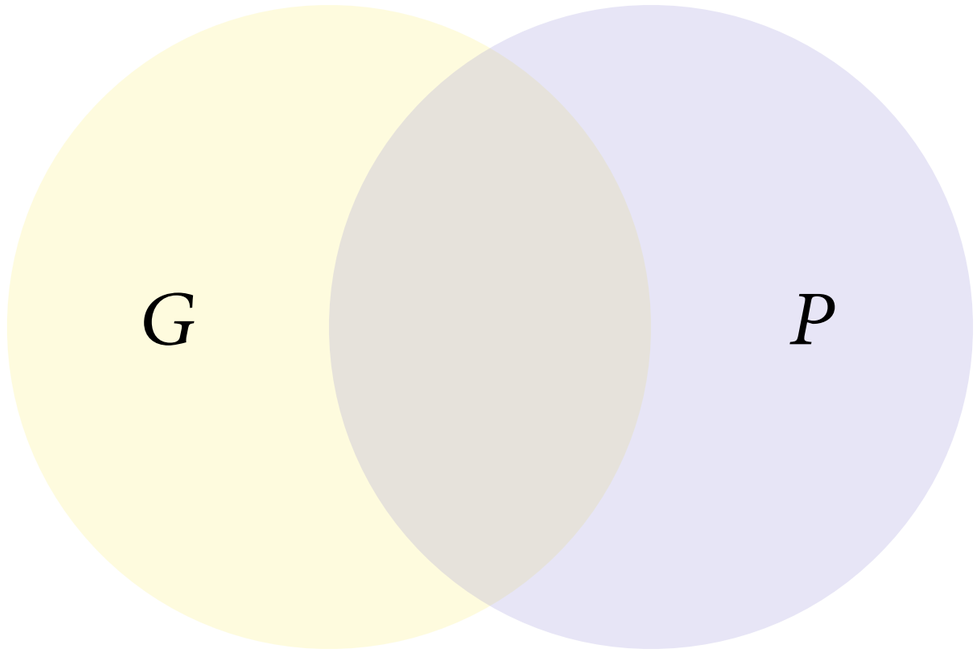
- The largest manually annotated corpus for written Swedish; a collaboration between Stockholm and Umeå University.

  but a bit dated now (1990s)

- Contains 1.2 million words (tokens) annotated with part-of-speech, morphological information and lemma.

- Balanced corpus with texts from various genres.

Precision and recall



$$\text{precision} = \frac{|G \cap S|}{|S|} \qquad \text{recall} = \frac{|G \cap S|}{|G|}$$

Constituents

- Words within sentences form groupings called **constituents**.

  Kim read [a very interesting book about grammar]. Kim read [it].

- Each constituent is projected by a **syntactic head**, which determines its internal structure and external distribution.

  [The war on drugs] is controversial. / *[The battle on drugs] is controversial.

  [The war on drugs] is controversial. / *[The war on drugs] are controversial.

Chunking

- Chunk patterns can be learned from annotated corpora.

  Requires corpora manually annotated for chunks.

- A common strategy is to reduce chunking to tagging.

  IOB format: B – beginning, I – inner word, O – word outside of a chunk

- Not easy to evaluate multi-word chunks in this setting.

Evaluating chunking

		guldstandard	system
1	First	B-NP	O
2	Bank	I-NP	B-NP
3	of	I-NP	I-NP
4	Chiacgo	I-NP	I-NP
5	announced	B-VP	B-VP
	…		

NP 1-4

NP 2-4

The system makes 2 errors: 1 with respect to precision, 1 with respect to recall.

Syntactic parsers

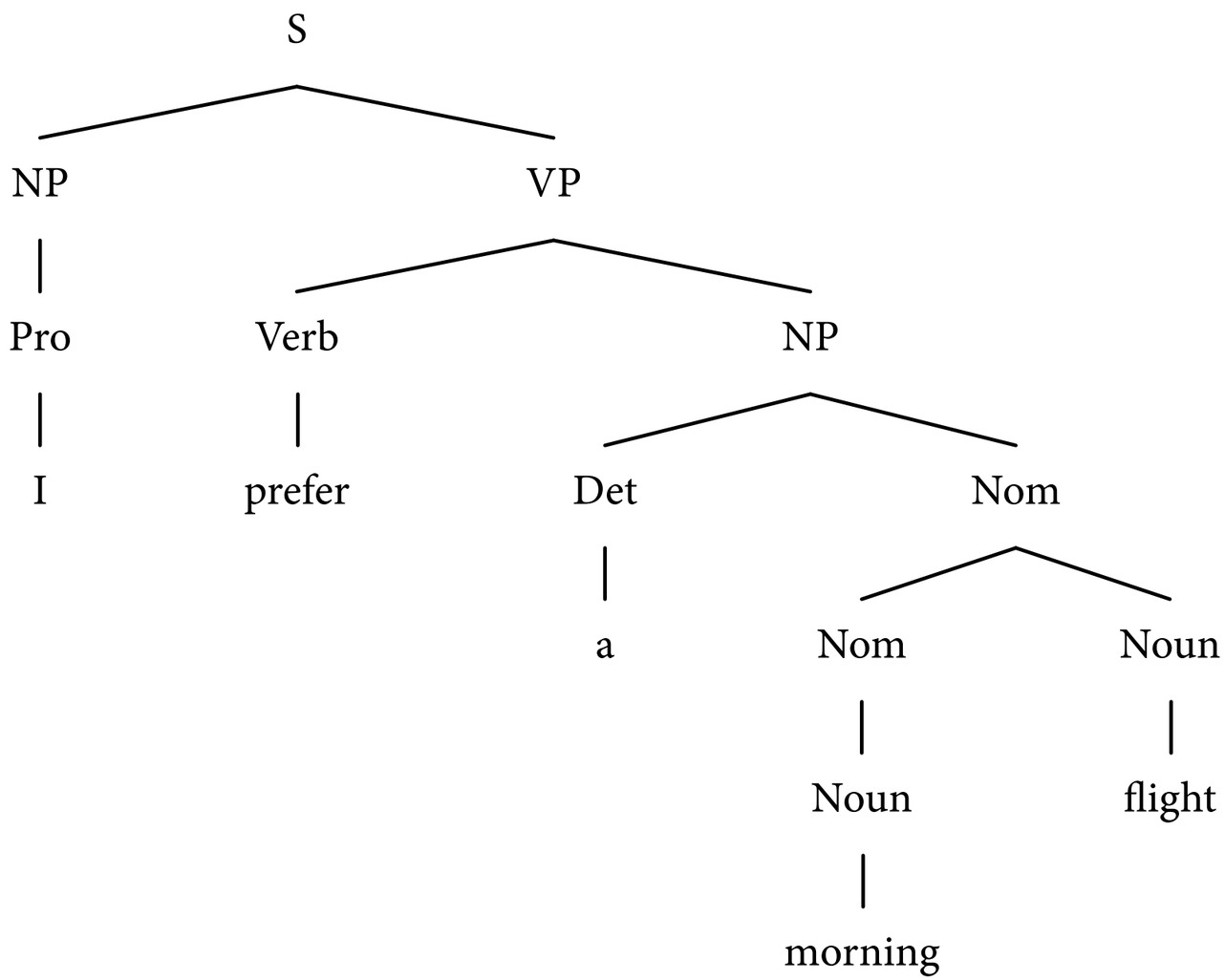
- A **syntactic parser** is a computer program that maps a natural language sentence to a syntactic representation.

  phrase structure parsing, dependency parsing

- Modern parsers are learned from corpora of manually crafted syntactic analyses called **treebanks**.

  Penn Treebank, Swedish Treebank, …

Phrase structure trees

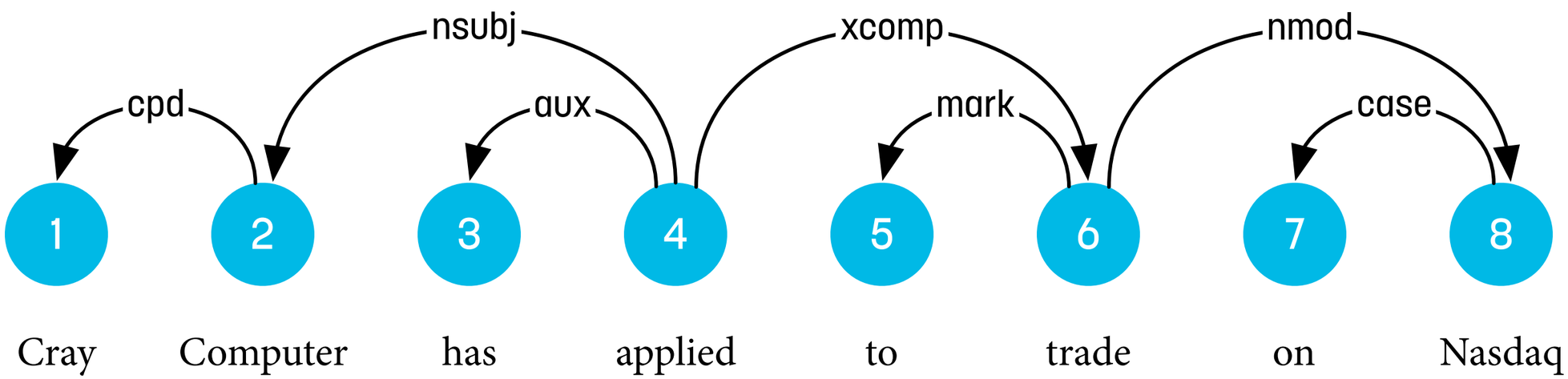


Penn Treebank

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
  (VP (MD will)
    (VP (VB join)
      (NP (DT the) (NN board) )
      (PP-CLR (IN as)
        (NP (DT a) (JJ nonexecutive) (NN director) ))
      (NP-TMP (NNP Nov.) (CD 29) )))
  (. .) ))
```

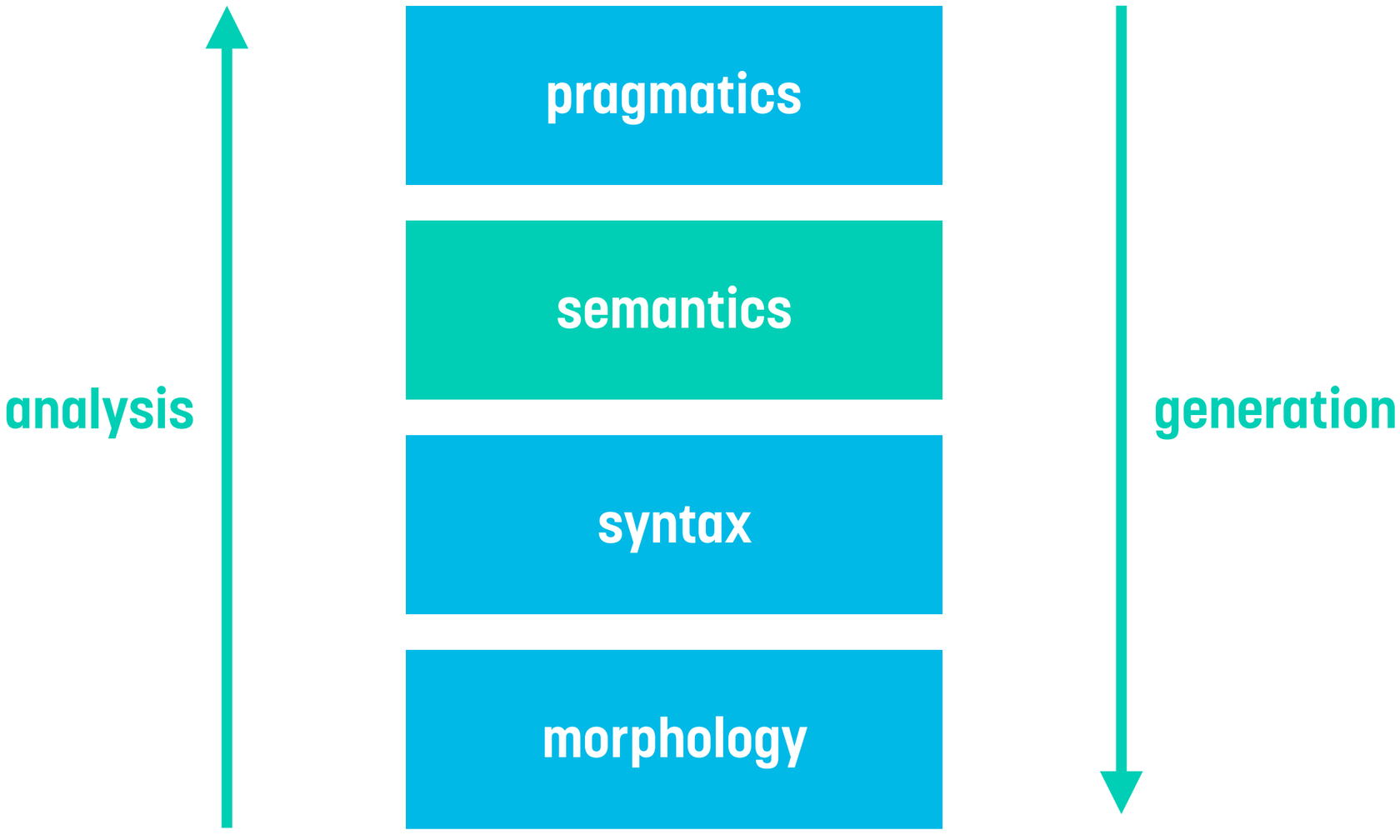
Grammar rule	Constituent
S → NP-SBJ VP .	Pierre Vinken … Nov. 29.
NP-SBJ → NP , ADJP ,	Pierre Vinken, 61 years old,
VP → MD VP	will join the board …
NP → DT NN	the board

Syntactic dependency trees



PTB section 00, document 18, item 026; Stanford dependencies (basic)

Semantics


pragmatics
semantics
syntax
morphology
analysis
generation

Syntax and semantics

- **The Principle of Compositionality**

  The meaning of a complex expression is determined by its structure and the meanings of its parts.

  challenges: idiomatic expressions, contextuality

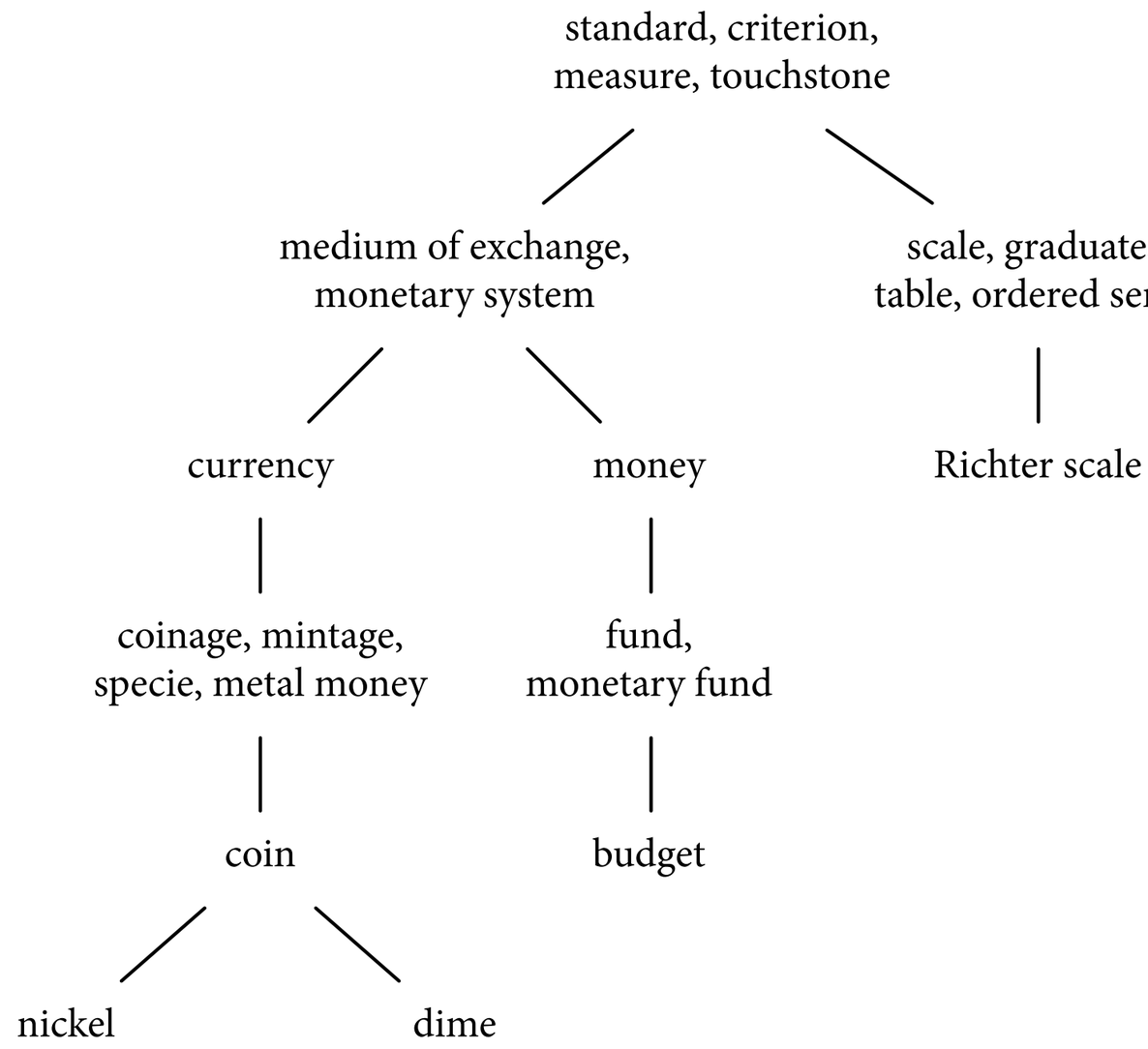
- Syntax provides the scaffolding for semantic composition.

- The way that information is combined into semantic structures is highly dependent on syntax.

  The brown dog on the mat saw the striped cat through the window.

  The brown cat saw the striped dog through the window on the mat.

WordNet: Hand-crafted ontologies



‘You shall know a word
by the company it keeps.’

John Rupert Firth (1890–1960)

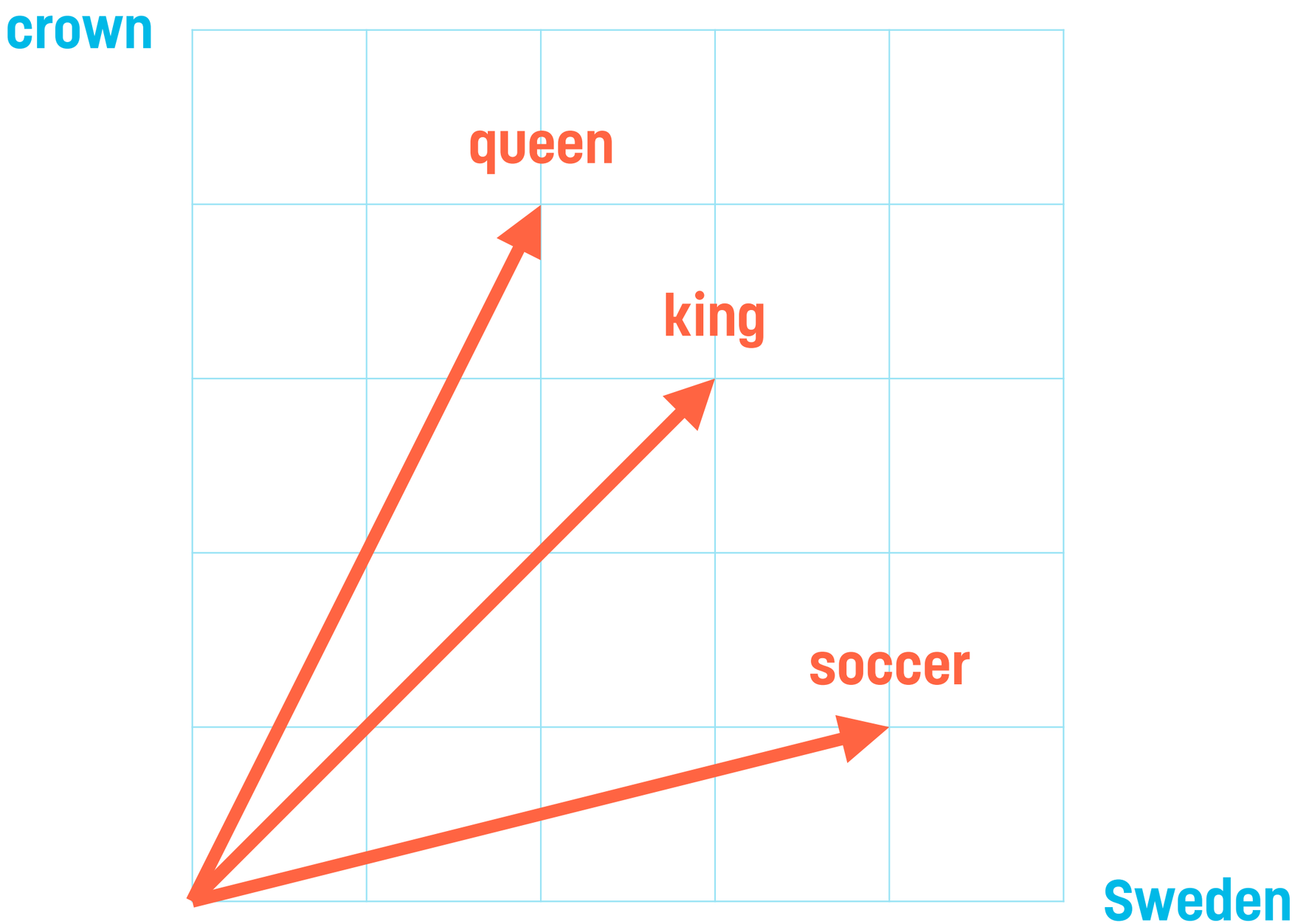
Co-occurrence matrix

context / target	king	throne	reign	Sweden	match	goal	play
queen	4	1	1	2	0	0	0
king	3	2	1	3	1	0	0
soccer	1	0	0	4	3	4	2
hockey	0	1	0	1	2	1	1

Word vectors

context / target	king	throne	reign	Sweden	match	goal	play
queen	4	1	1	2	0	0	0
king	3	2	1	3	1	0	0
soccer	1	0	0	4	3	4	2
hockey	0	1	0	1	2	1	1

Words as vectors



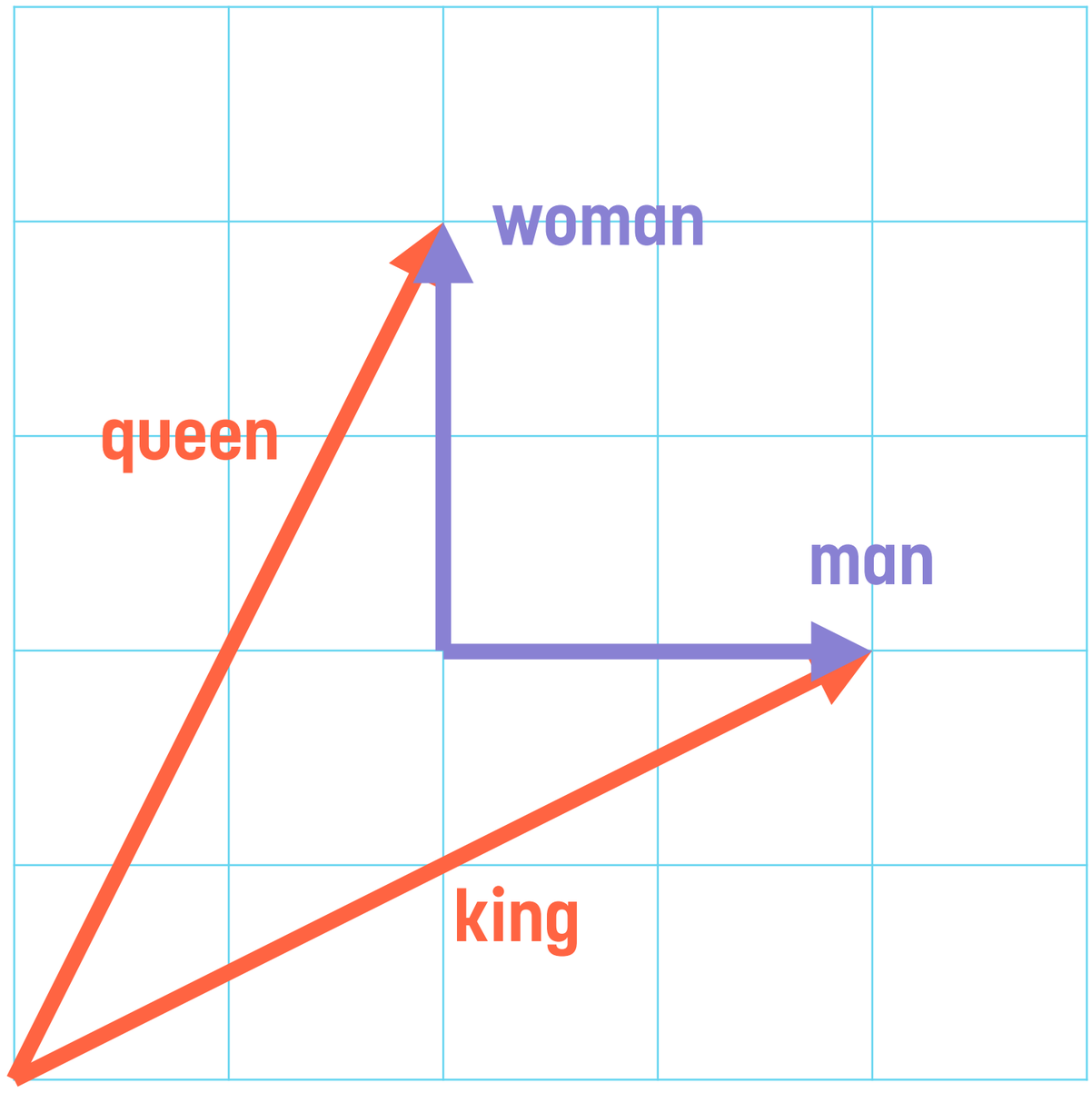
Word vectors

- Raw co-occurrence counts yield word vectors that are high-dimensional but sparse (few non-zero entries).
- **Approach 1:** Dimensionality reduction

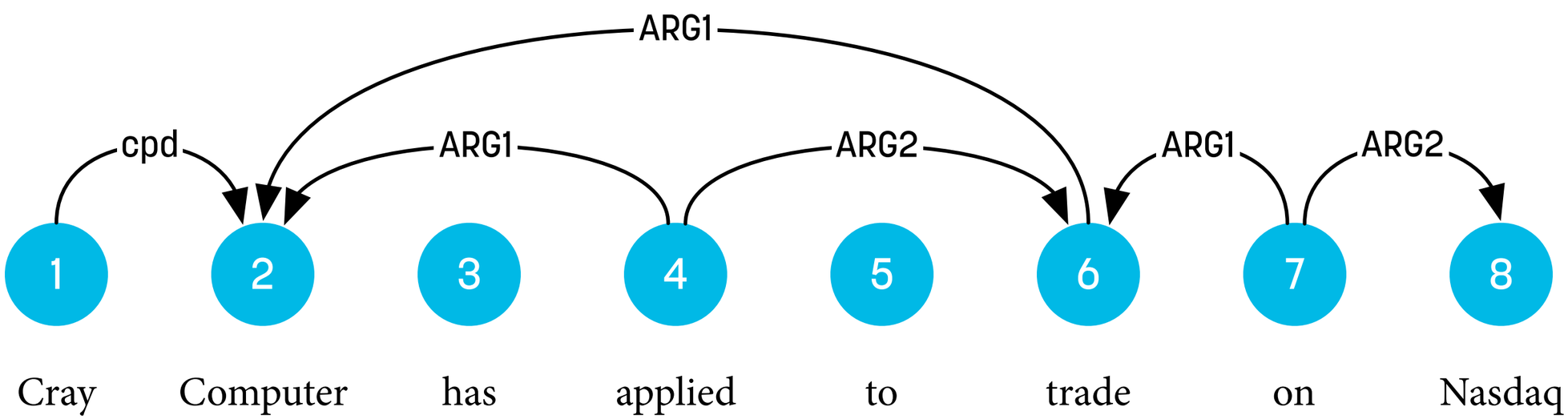
  singular value decomposition
- **Approach 2:** Direct learning of low-dimensional, dense vectors

  word2vec

Computing with word vectors

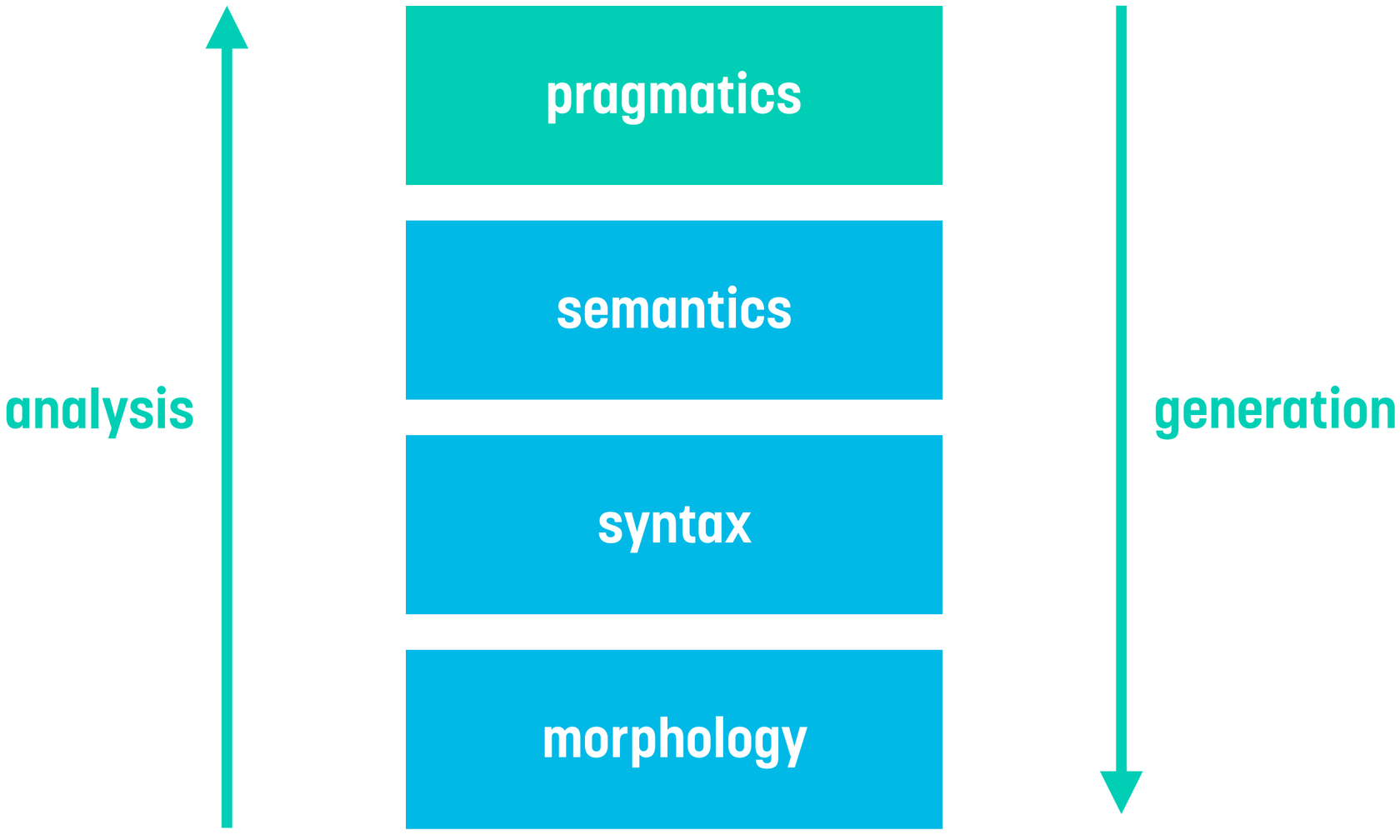


Semantic dependency graphs



PTB section 00, document 18, item 026; SDP/DM

Semantics


pragmatics
semantics
syntax
morphology
analysis
generation

Contextuality

- How an utterance is interpreted depends on information that exceeds what is mediated by morphosyntax.

  A: Are you coming tonight? B: I need to study.

- Even grammaticality depends on context.

  A: What time is it? B: Five, I think.

Structure

What is computational linguistics?

Prelude: From raw text to lists of words

From form to meaning

Presentation of the lab project

Presentation of the lab project

Lab project

- Find a web page, and extract some 300–500 words of text content.
- Segment and tokenize the text content, have it part-of-speech-tagged, and named entities recognised.
- Evaluate the performance and report your findings.

Named entities

Danskägda Foss stänger anläggningen i Höganäs och flyttar produktionen till Kina medan utvecklingsavdelningen koncentreras till Hillerød. 163 anställda i Höganäs berörs av beskedet.

I mitten av december i fjol tog koncernstyrelsen beslutet att stänga i Höganäs. Och på tisdagsmorgonen fick samtliga anställda veta att deras arbetsplats ska slå igen senast första januari 2015.

– Det är inget lätt beslut eftersom en stängning berör många medarbetare och deras familjer, förklarar koncernchef Torben Ladegaard.

Enligt Ladegaard är det inte bristande lönsamhet som ligger bakom nedläggningen men han ser ändå beslutet som nödvändigt för att bevara den danskägda koncernens konkurrenskraft och lönsamhet.

Named entities

- … can be indexed and linked from
- … take part in semantic relations
- … are common answers in question answering systems

JEOPARDY!®

This New York University alumnus has won several Academy Awards.

Query against DBPedia (SPARQL-format)

```
SELECT DISTINCT ?x WHERE {
 ?x dbpedia-owl:almaMater dbres:New_York_University.
 ?x dbpedia-owl:award dbres:Academy_Award.
}
```

Types of named entities in DBPedia

- Persons

  Actor, Curler, FictionalCharacter

- Organisations

  Band, Company, SportsTeam

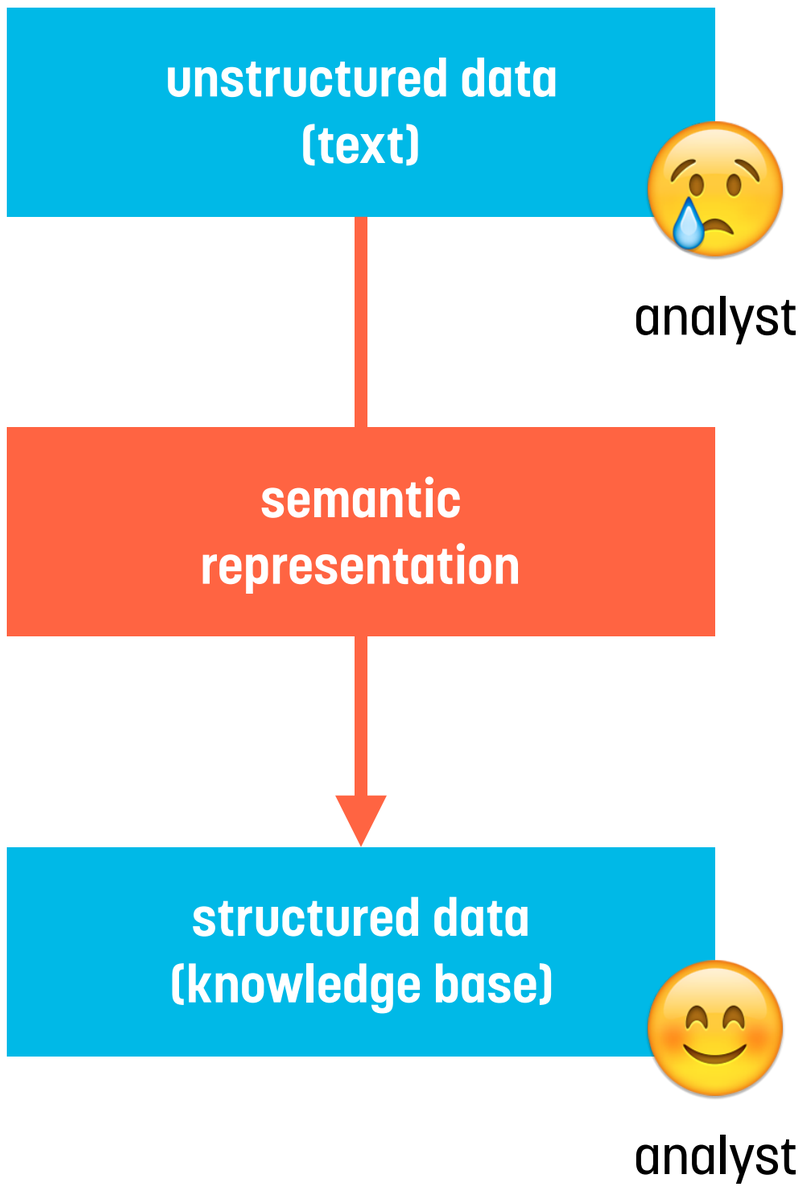
- Places

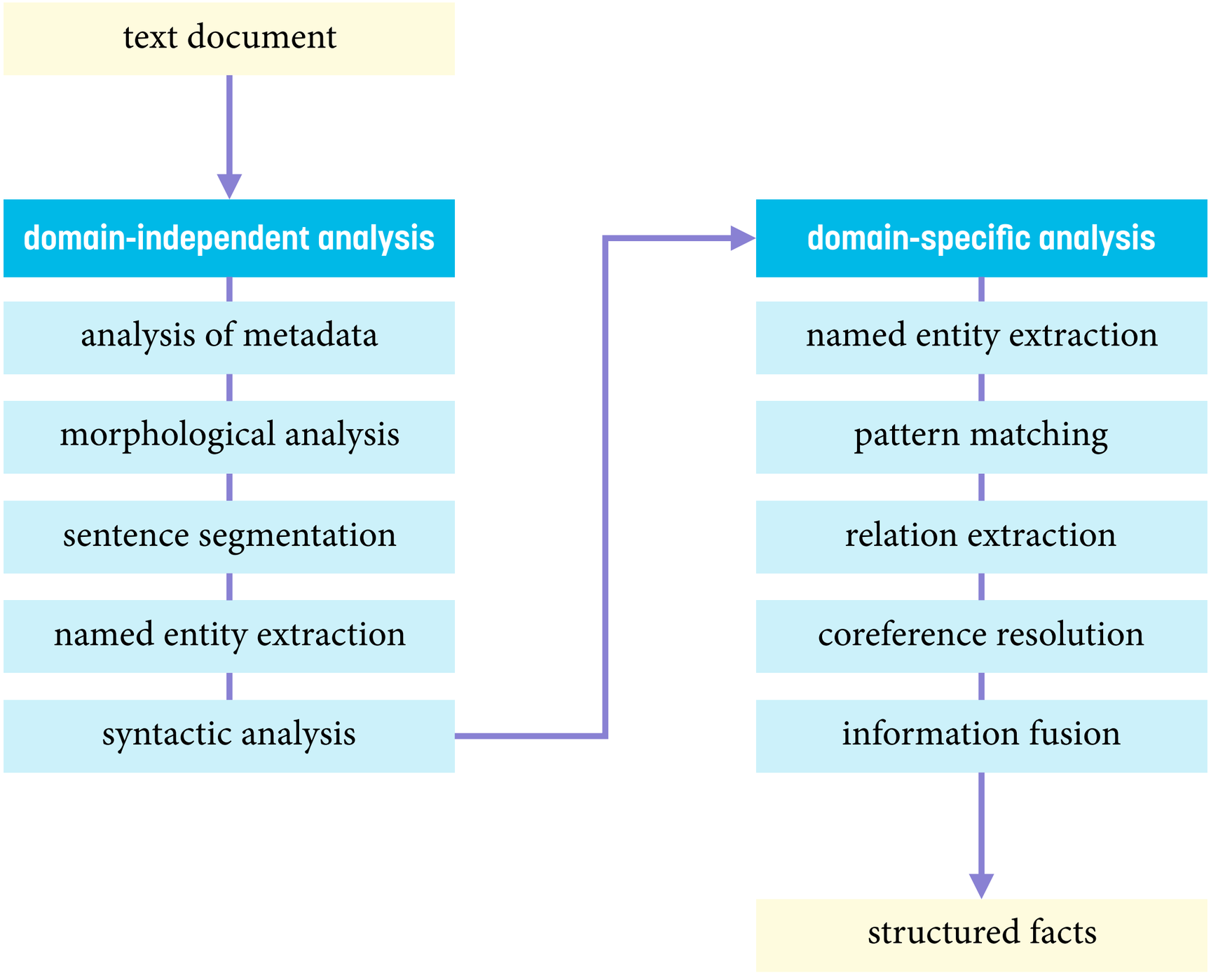
  Building, Mountain, Country

- Medical terms

  Muscle, Enzyme, Disease

Information extraction




text document
domain-independent analysis
analysis of metadata
morphological analysis
sentence segmentation
named entity extraction
syntactic analysis
domain-specific analysis
named entity extraction
pattern matching
relation extraction
coreference resolution
information fusion
structured facts

Why NER is not easy: Name forms in Polish

Kasus	Form
Nominative	Muammar Kaddafi
Genitive	Muammara Kaddafiego
Dative	Muammarowi Kaddafiemu
Accusative	Muammara Kaddafiego
Instrumental	Muammarem Kaddafim
Locative	Muammarze Kaddafim
Vocative	Muammarze Kaddafi