Computational Linguistics

Wolfgang Menzel

Department of Informatics
Hamburg University

Computational Linguistics

1. Natural Language and the Computer
2. Words and Wordforms
3. Phrases and Sentences
4. Discourse: Texts and Dialogs

Computational Linguistics

Words and Wordforms

1. Lexical items
2. Dictionary lookup
3. Word segmentation
4. Morphological analysis
5. Phonology
6. Lexical semantics
7. Distributed lexical representations
8. Part-of-speech tagging
9. Word-sense disambiguation

Computational Linguistics

Phrases and sentences

1. Language models
2. Chunking
3. Structural descriptions
4. Parsing with phrase structure grammars
5. Probabilistic parsers
6. Parsing with dependency grammars
7. Unification-based grammars
8. Semantics construction

Computational Linguistics

Discourse: Texts and Dialogs

1. Cross-sentential phenomena
2. Coreference resolution
3. Discourse representation theory
4. Rhetorical structure theory
5. Dialog modelling
6. Applications
    - Sentiment detection
    - Summarization

Computational Linguistics

Readings:

- Jurafsky, Daniel S., and James H. Martin (2009) Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.
- Chris Manning and Hinrich Schütze (1999) Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA.
- Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde und Hagen Langer (Hrsg.) (2009) Computerlinguistik und Sprachtechnologie – Eine Einführung, Heidelberg: Spektrum Akademischer Verlag, 3. Auflage.

Complementary courses

- 64-463(a) Projekt Sprachtechnologie (Timo Baumann, Arne Köhn, Wolfgang Menzel)
  Subtitling of lecture videos
  Mi 16-18, Do 12-18, F-429
- 64-418/419 Sprachverarbeitung (Timo Baumann)
  Mi 12-16, F-334
- Einführung in die Linguistik des Deutschen (Heike Zinsmeister)
  Do 14-16, Phil 1350 + Übung
- Textlinguistik (Heike Zinsmeister)
  Do 12-14, Phil 256

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

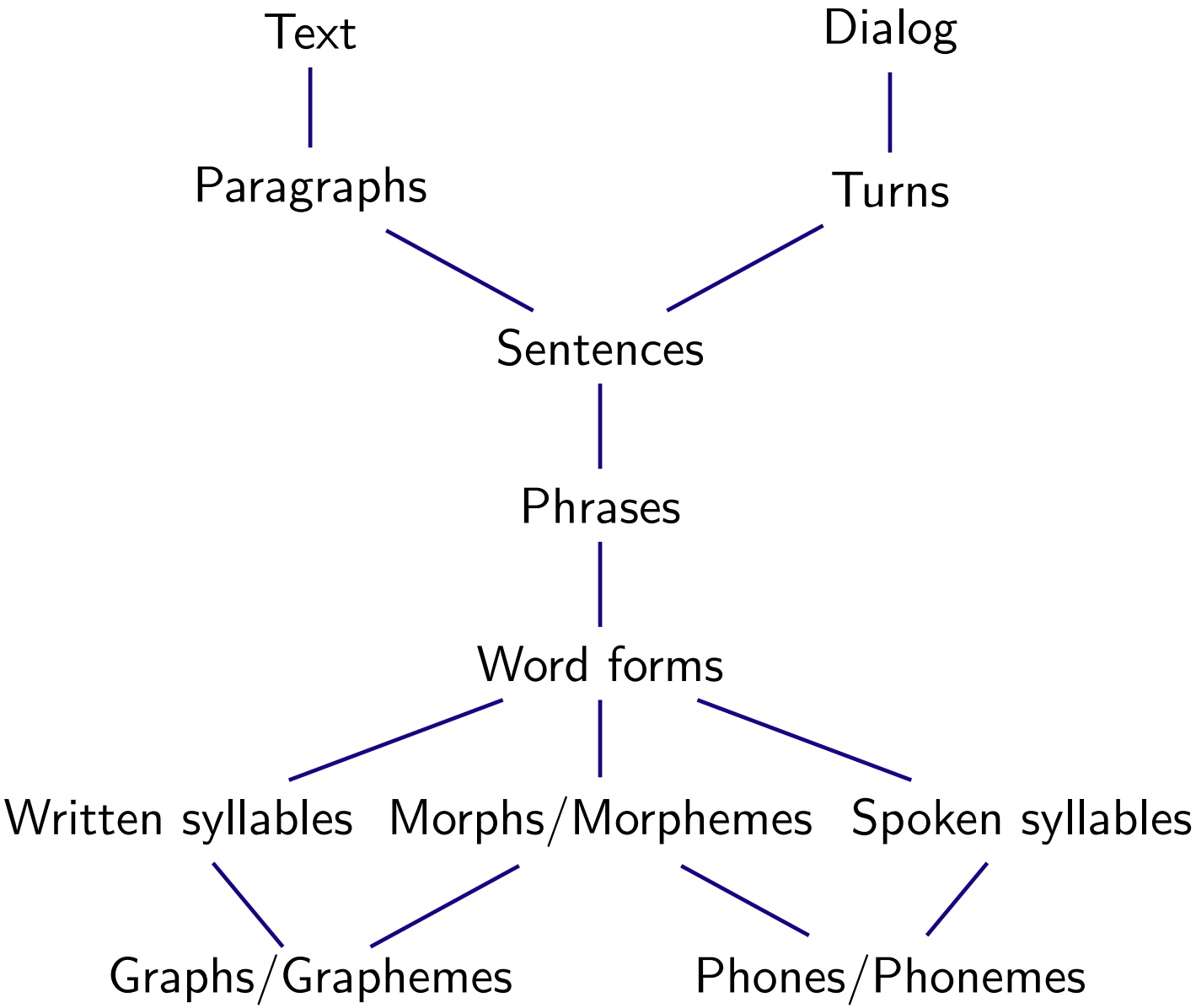
Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

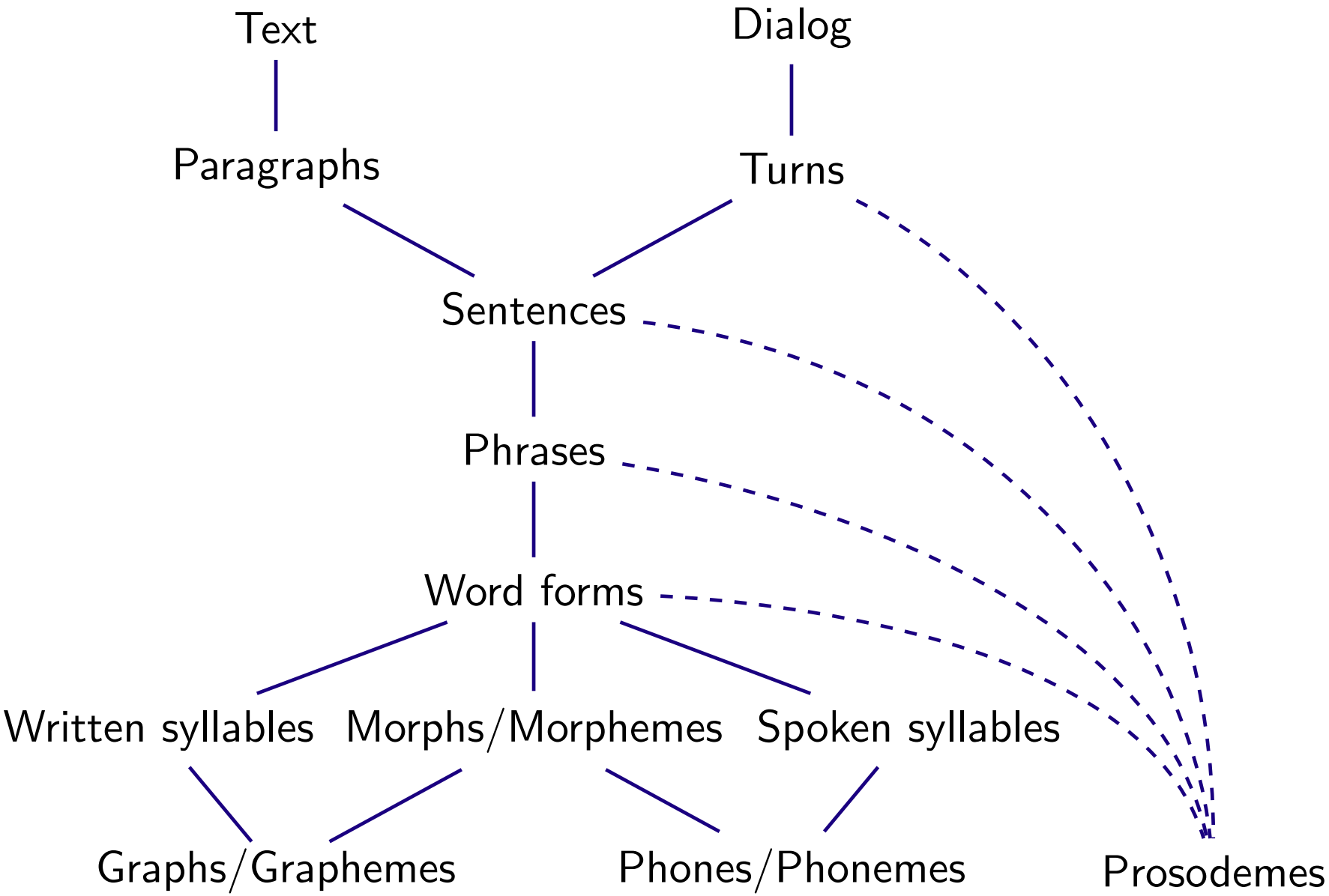
Compositionality

- language is compositional
  - smaller units combine into larger ones
  - the meaning of a complex expression is determined by its structure and the meanings of its constituents (Gottlob Frege, 1879)
- alternative kinds of decomposition are possible
  - no strict hierarchical organisation

Compositionality



Compositionality



Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Linguistic Descriptions

Semiotics

- linguistic expressions can be considered as (complex) signs
- dyadic notion of a sign (Ferdinand de Saussure)
  - relationship between the form of the sign (the signifier) and its meaning (the signified)
  - essentially arbitrary, motivated only by social convention

Linguistic Descriptions

- triadic notion of a sign (Charles S. Peirce)
  - "something that stands for something, to someone in some capacity"
  - form: (semiotic) syntax
    - how the sign is composed of (less complex) signs
  - meaning: semantics
    - what the sign tells about the world
  - communicative function: pragmatics
    - what the sign means to a recipient
    - how the sign is interpreted

Pragmatics

- the context shapes the meaning
  - the speaker
    - communicative intent/speaker meaning vs. informative intent/sentence meaning
  - the hearer
    - prior knowledge, prejudices
  - the relationship between the speaker and the hearer
    - social status and social distance, the cultural habits (politeness)
  - the situation
    - place, time, privacy, ...

Pragmatics

Friedemann Schulz von Thun

- four aspects of a message
  - rational content (Sachinhalt)
  - self-presentation (Selbstoffenbarung)
  - appeal (Appell)
  - relationship (Beziehung)

Linguistic Descriptions

		Semiotics		
		form syntax	meaning semantics	function pragmatics
Linguistics	discourse text/dialog			
	syntax sentence/phrase			
	morphology word/morpheme			
	phonology phoneme/prosodeme			

Linguistic descriptions

- distinction between type$_{informatics}$ and instance
  - abstract class vs. individual instance
  - Phoneme vs. Phone
  - Grapheme vs. Graph
  - Morpheme vs. Morph
  - X-eme: linguistic unit establishing a meaning difference
    vs. X(-e): observable unit

Linguistic descriptions

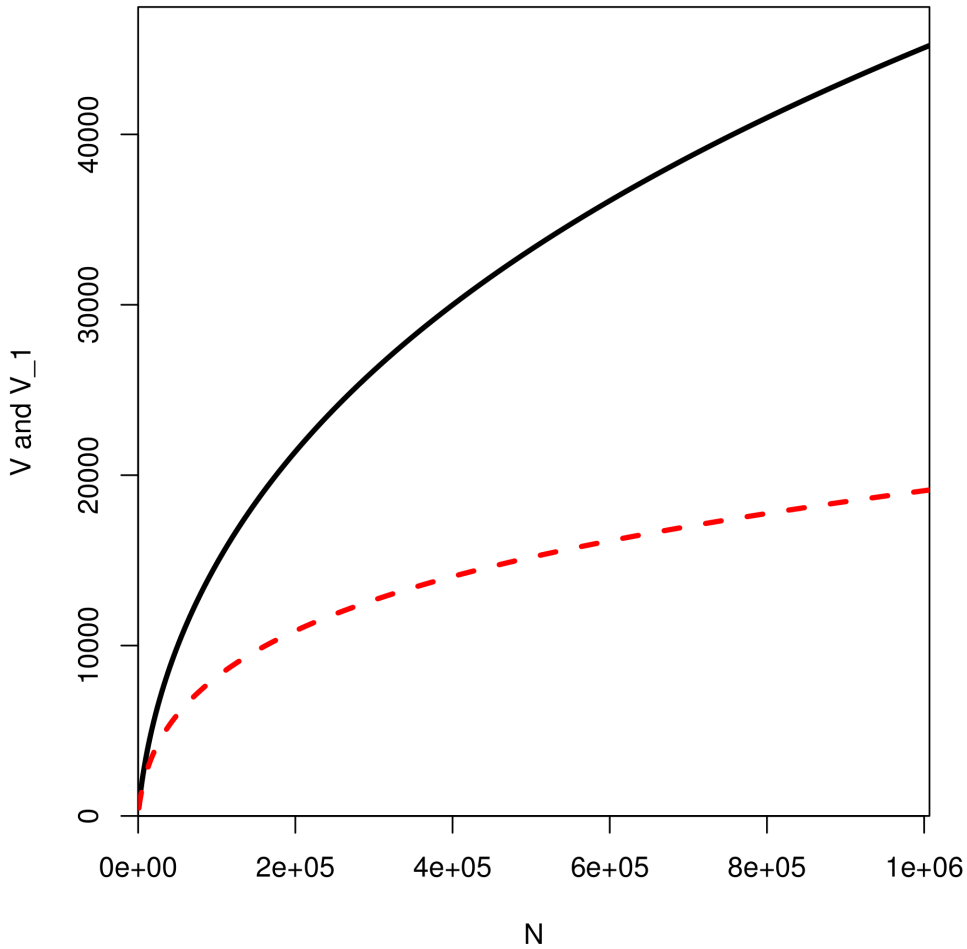
- corpus linguistics:
  - type$_{linguistics}$ refers to a wordform which might occur several times in the data set
  - a token is a specific occurrence of a wordform.

    Harris and I had been hard at work on our German during several weeks at that time, and although we had made good progress, it had been accomplished under great difficulty and annoyance, for three of our teachers had died in the mean time. A person who has not studied German can form no idea of what a perplexing language it is.

  - the text sample contains 62 tokens and 49 types
  - # of tokens: corpus size
  - # of types: (corpus specific) dictionary/vocabulary size

Linguistic descriptions

- Brown corpus: 53.076 types vs. 1.015.945 tokens

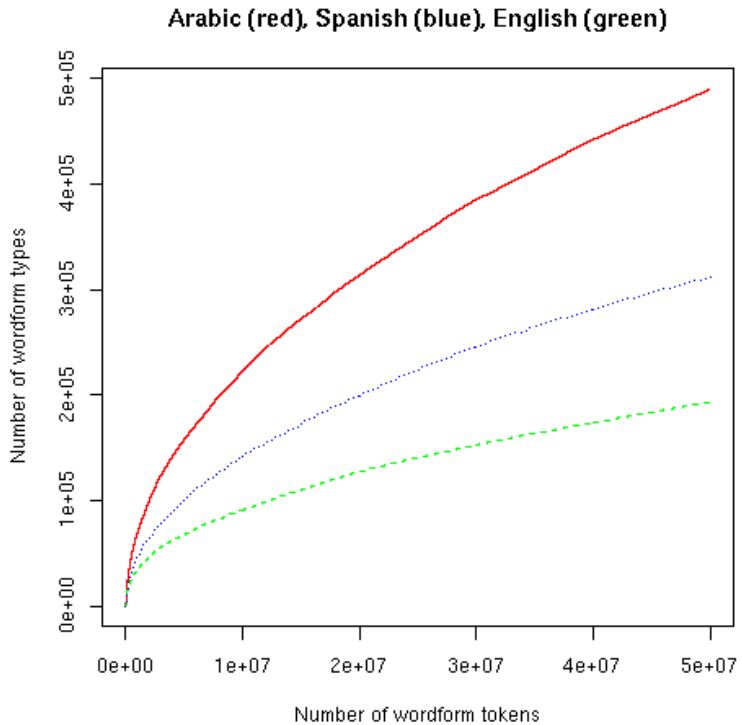


black: vocabulary size
red: # of hapax legomena

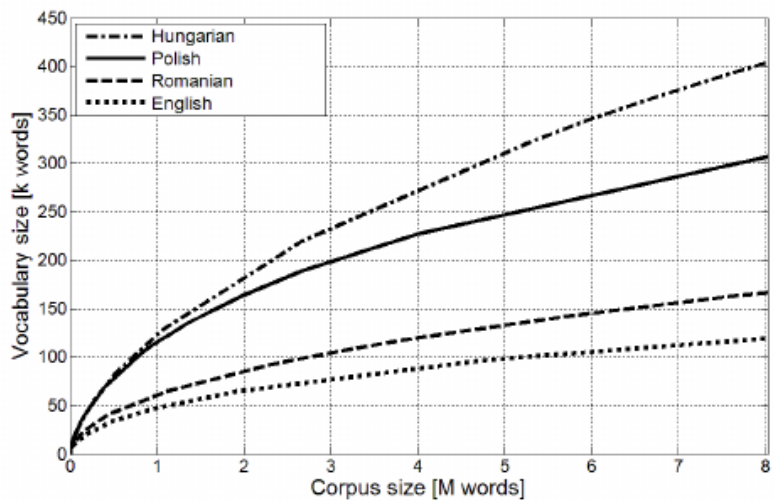
Baroni/Evert

Linguistic descriptions

- languages differ in the vocabulary growth rate, but not in the general characteristics of growth



Liberman (2006)



Tarjan et al. (2012)

Linguistic descriptions

Allo-X:

- one of several possible X-s that can be used to produce a X-eme
  - language dependent
  - positional variants: choice depends on the context

    ich / ach
    Kasten / Kessel / Kisten / Kosten / Kunst
    in-dispensible / im-possible / il-literate / ir-respective

  - free variants: arbitrary choice

    Zungen-r vs. Zäpfchen-r

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Ambiguity

- is pervasive in natural language
- occurs on all levels of linguistic description
- often can be resolved by contextual knowledge
- mostly goes unnoticed by speakers of the language

- most processing tasks can be understood as a task of disambiguation
- disambiguating information is often not easily accessible for computational approaches

Ambiguity

- Phonology
  - final devoicing: /d/ vs. /t/, /b/ vs. /p/, /g/ vs. /k/

    Hand, Händ-e, hand-lich
    Leib, Leib-er, leib-lich
    Sieg, sieg-en, sieg-ten

  - vowel quality:

    will read / has read
    to write, has been written
    modern, rasten
    weg / Weg, die Sucht / er sucht

  - word stress

    to increase, the increase
    umfahren, übersetzen, ...

Ambiguity

- Morphology
  - pseudo-affixes:

    Genom / *ge-nommen*
    Gerste / *ge-stehen*
    End-ung, *Kuh-dung*

  - segmentation:

    be-in-halten / *Bein-halten*
    Wacht-raum / *Wach-traum*

Ambiguity

- Morpho-syntax
  - syncretism

    der, die, das, ...
    die / wegen der / mit der / die Frau
    der / mit dem / den / die / wegen der / die Teller
    das / mit dem / das / die / wegen der / die Fenster ...
    wir / sie schauen, er / ihr schaut,

  - part-of-speech ambiguity

    green peas / the green / to green
    der Hund / der da / der Schulden hat
    wir laufen / beim Laufen

Ambiguity

- Syntax
  - segmentation

    ..., weil dem Sohn des Meisters Geld fehlt.

  - attachment

    ... das Bild mit dem Buch auf dem Stapel.

  - role assignment

    Die Mutter pflegt die Tochter.

Ambiguity

- Semantics/Discourse
  - homonymy (homographs/homophones)

    der / die See, der / die Otter, der Tau / das Tau,
    rasten / rasten
    die Lerche / die Lärche, statt / die Stadt
    malen / mahlen

  - polysemy

    school (building, abstract institution, organizational unit)

  - referential ambiguity

    Herr X begrüßt den Chef. Er hat ...

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Variability

- natural language provides many different ways to convey the same (or a similar) communicative message
- strict vs. partial synonymy
  - strict synonymy is almost non-existent:

    Streichholz / Zündholz
    Klempner / Installateur

  - usually connotative differences

Intension	Extension	Konnotation	Example
same	same	same	Streichholz/Zündholz
same	same	different	Gymnasium/Penne
different	same	—	Morgenstern/Abendstern

Variability

- lexical vs. phrasal synonymy
  - paraphrase

    der humpelnde Mann / der Mann, der humpelt / ...

  - metapher

    angeben / den dicken Maxen markieren

- often regional/social variants, cross-lingual influence

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Rules and exceptions

- natural language is obviously governed by regularities
  - strong consensus about well-formedness within a language community
    - codified in (prescriptive) grammar handbooks
    - can be captured by rules
  - all language are productive: new words and sentences can be created and understood
  - rule violations can be noted and corrected
- there is no rule without an exception
  - even exceptions have their exceptions
  - often the more rare phenomena are considered as exceptions

Rules and exceptions

- a vowel is short if it is followed by at least two consonants, none of them separated by a morpheme boundary

  Hast vs. *Glas*, *hast-en* vs. *glas-t-en*
  but: *(du) ha-st*

- the word final substring '-en' is an inflectional ending, unless it is part of the root

  (zu) geb-en, (wir) lauf-en, (mit) Pferd-en ...
  but: *Gen, Kien, Hafen,* , ...

- the past tense of a verb is built by attaching the suffix '-t' to the root, unless its root ends on 'd' or 't', or it is a strong one

  bau(-en), bau-t(-en), sag(-en), sag-t(-en)
  but: *trag(-en), trug(-en), sing(-en), sang(-en)*
  but: *rast(-en), rast-et(-en), rett(-en), rett-et(-en)*

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Natural Language and the Computer

- Compositionality
- Linguistic descriptions
- Ambiguity
- Variability
- Rules and exceptions
- Research methodology

Research methodology

Linguistic research questions

- What's constitutive for a natural language? What have all the human languages in common?
- What distinguishes different natural languages?
- How linguistic knowledge is used for production and understanding?
- What's the time course of language processing?
- How language is acquired?

Research methodology

Computational research questions

- What kind of linguistic knowledge is relevant for a particular task?
- Can a task also be solved by means of partial linguistic knowledge?
- What kind of models/algorithms lend themselves to a particular task?
- How should linguistic knowledge be represented to facilitate efficient processing?
- How can linguistic knowledge be acquired?
- What kind of system architecture is required for a particular task?

Research methodology

Abstraction

- Any scientific discipline needs abstraction to gain fundamental insights. e.g.
  - Newtonian mechanics: frictionless motion
  - Economics: rational markets
  - Computer Science: uniform operation costs
  - Linguistics: grammatical competence CHOMSKY (1965)

Research methodology

- grammatical competence abstracts away
  - individual linguistic deficits
  - limited processing capacity
  - contextual influences
  - differences between linguistic communities (regional, social, ...)
- grammatical competence is
  - (largely unconcious) knowledge of an idealized speaker about the regularities and components of her language
  - the capability to produce correct sentences and to use them to convey content (meanings)

Research methodology

- competence-based linguistic capabilities (Grewendorf et al. 1987)
  - to decide whether two utterances are the same
  - to correctly segment an utterance (into phones, syllables, word forms, phrases)
  - to determine whether an utterance is well-formed
  - to decide whether two utterances have the same meaning
  - to notice an ambiguity
  - to notice gradual deviation from the norm
  - to distinguish different types of deviating utterances
  - to recognize different structural relationships within different sentences

Research methodology

- performance: the use of competence to produce speech acts
  - imperfect realizations under resource limitations
  - but: performance cannot be reduced to an error-prone competence
  - linguistic performance is based on pragmatic competence

Research methodology

- pragmatic competence (Grewendorf et al. 1987)
  - to identify the intention of a speaker
    (promise, proposal, consent, rejection, refusal, ...)
  - to recognize emotionally biased utterances

    Will you help me to tidy up the kitchen?
    Will you at least help me to tidy up the kitchen?
  - to recognize implicit speaker intentions

    Could you tell me the time, please?
  - to recognize hidden speaker intentions

    Der Steuermann ist heute betrunken.
    Der Kapitän ist heute nicht betrunken.
  - to recognize violated presuppositions
  - to recognize inconsistent speaker intentions

Research methodology

Linguistic data

- correct utterances
- incorrect utterances (performance errors, language learners)
- human judgements/annotations
- language acquisition
- speech disorders
- structure/constituency/dependency tests
- data from psycholinguistic experiments (reading, visual world paradigm)

Research methodology

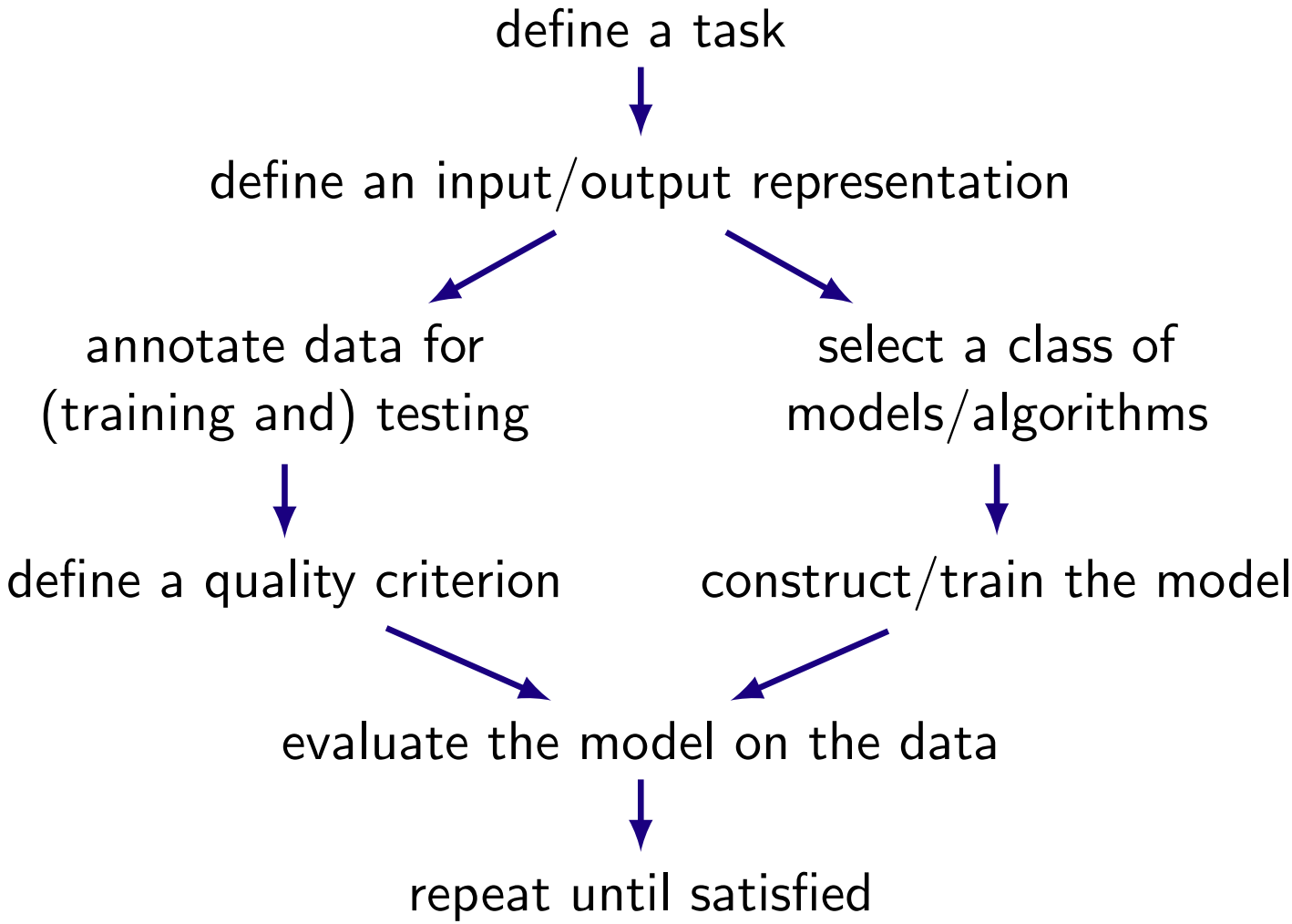
- linguistic data are systematically collected and published as linguistic resources
- lexical databases
  - definitions, semantic relations (synonymy, hyponymy, ...), thematic roles, ...
- corpora of spoken/written language, usually annotated
  - raw text collections
  - treebanks (syntactic/semantic)
  - translations (bi- or multilingual): unaligned, sentence aligned, word aligned
  - transcriptions: spoken words, non-verbal cues (hesitations, repairs, gestures, ...)
  - special purpose annotations: error corrections, reading times
- most languages are highly under-resourced

Research methodology

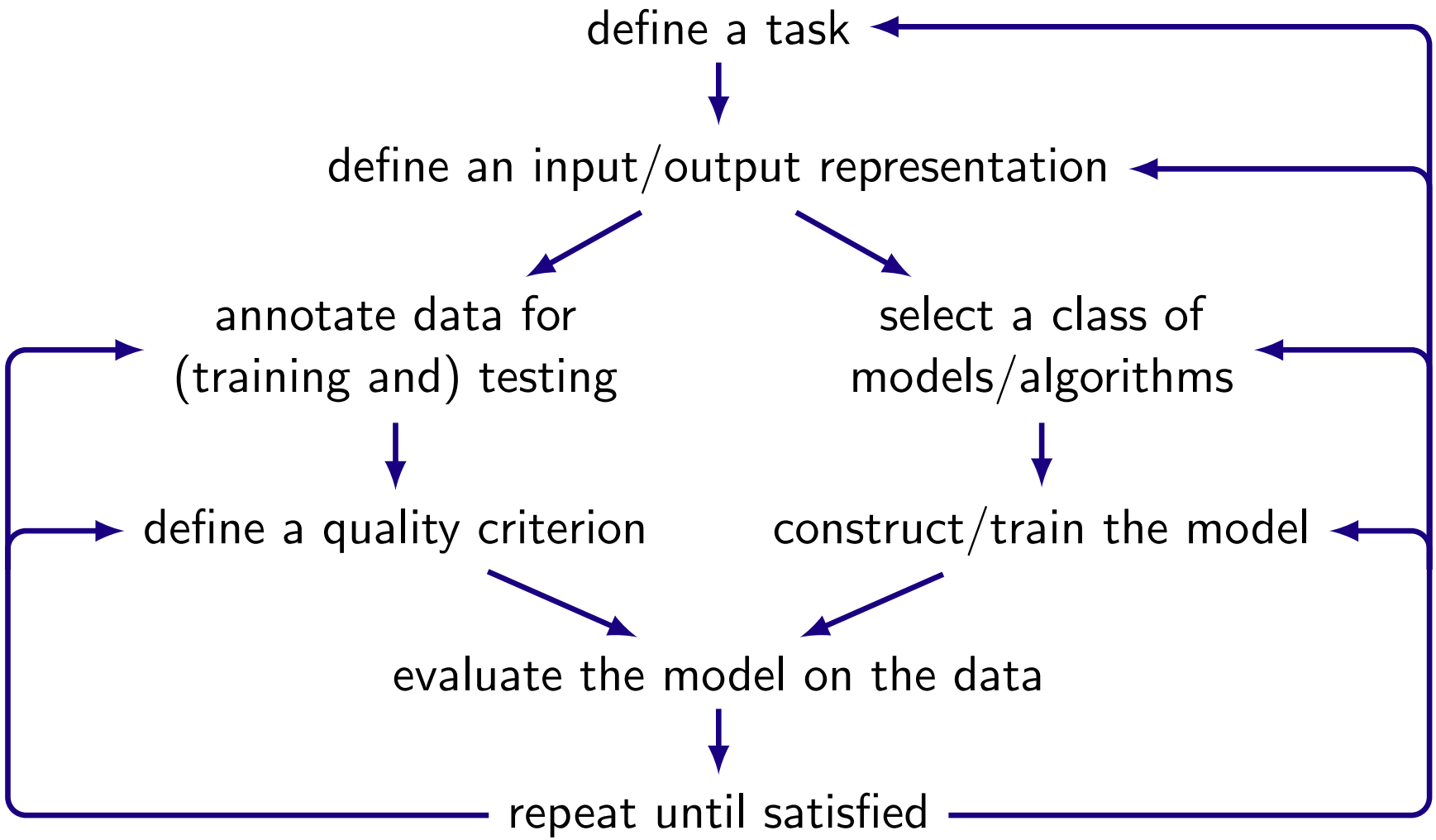
Natural Language Processing

- building and evaluating systems
- comparing different system versions
- cyclical refinement
- training models by means of machine learning techniques
- training models on different data sets
- comparing hand-crafted models with trained ones
- modifying processing tasks (towards higher degrees of difficulty)
- porting system solutions to similar tasks

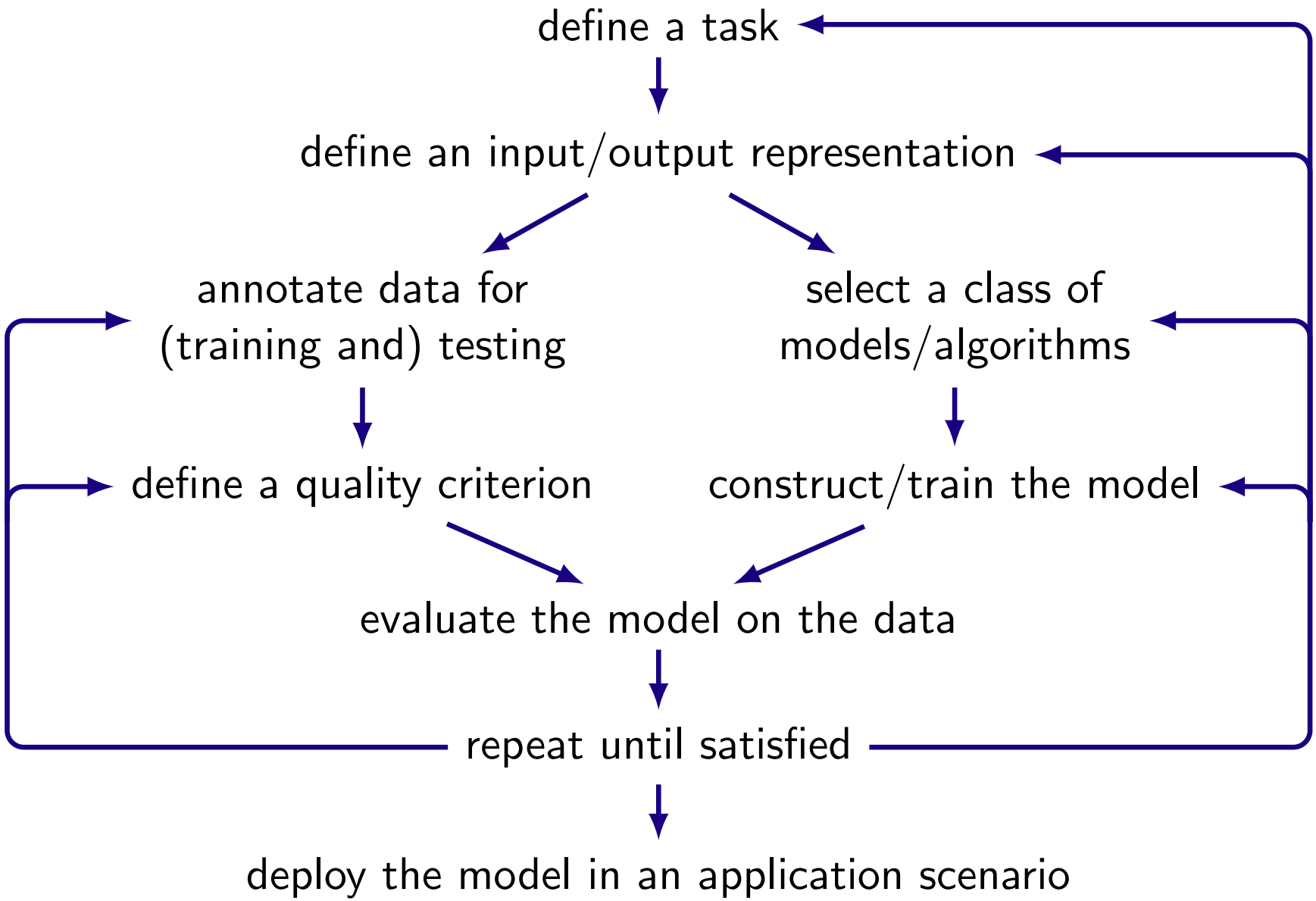
Research methodology



Research methodology



Research methodology



Research methodology

A model

- captures domain knowledge relevant for a particular research question or processing task
  - task specific
  - influenced by subjective opinion
- is implemented by means of a formalism
  - high level programming language
  - runs on a dedicated virtual machine

Research methodology

Sample formalisms for Natural Language Processing

- (weighted) finite state automata/transducer
- (probabilistic) context-free grammars
- unification-based grammars
- transformation rules for strings, trees, and graph structures
- Bayesian classifier, neural networks, support vector machines, ...
- (hidden) Markov models, Dynamic Bayesian networks

Research methodology

Models

- can be hand-crafted or trained on data
- training can be supervised or unsupervised
- can be combined to solve a processing task in a hybrid system architecture