Einführung in die Computerlinguistik

Wiebke Petersen

WiSe 04/05

Folien teilweise entnommen aus

Cornelia Endriss: Einführung in die Computerlinguistik (SoSe 2001)

Hans Uszkoreit: Einführung in die Computerlinguistik (WiSe 01/02)

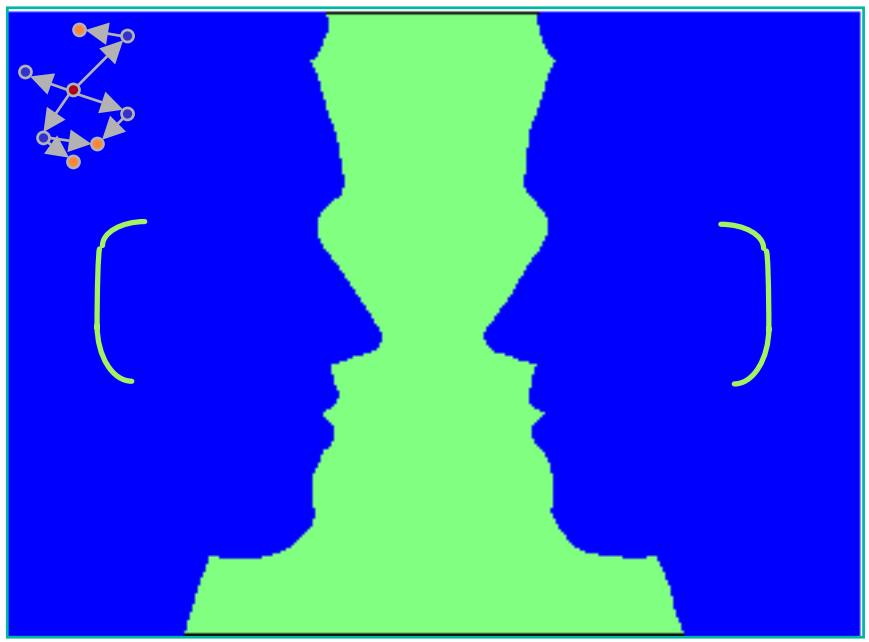
Faszination Sprache

Mehr noch als Denken ist die Sprache eine Fähigkeit, die nur der Mensch besitzt.

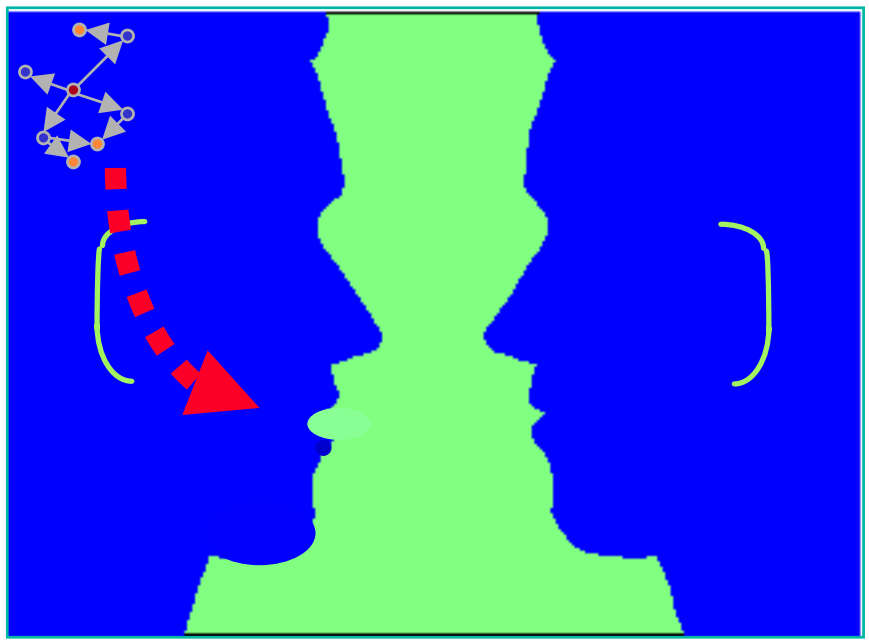
Es ist ein Wunder, wie wir in Sekundenschnelle komplexe Gedanken in einem Satz ausdrücken können.

Es ist nicht weniger erstaunlich, wie das Kind in nur wenigen Jahren zehntausende von Wörtern und eine komplexe Gramnmatik lernt.

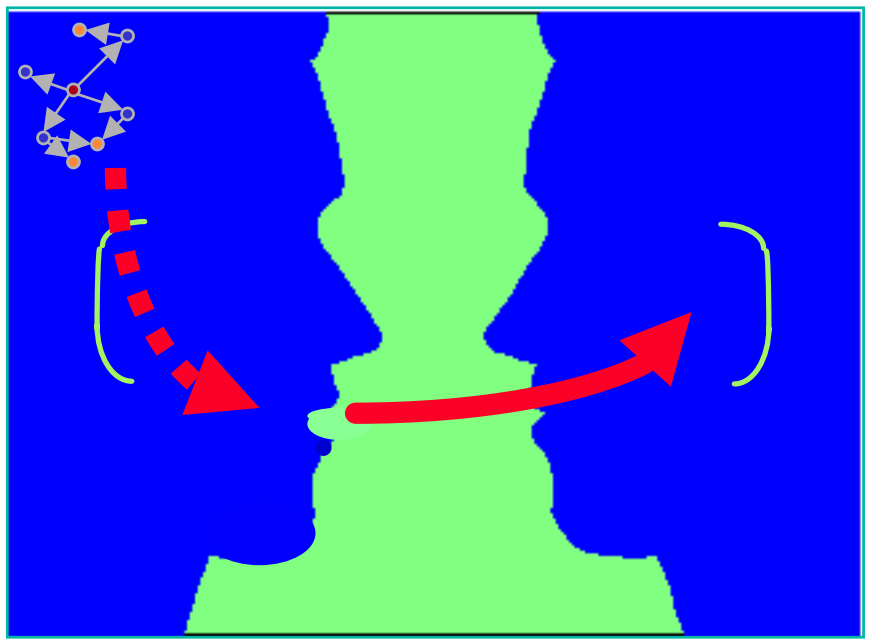
Sprachliche Kommunikation



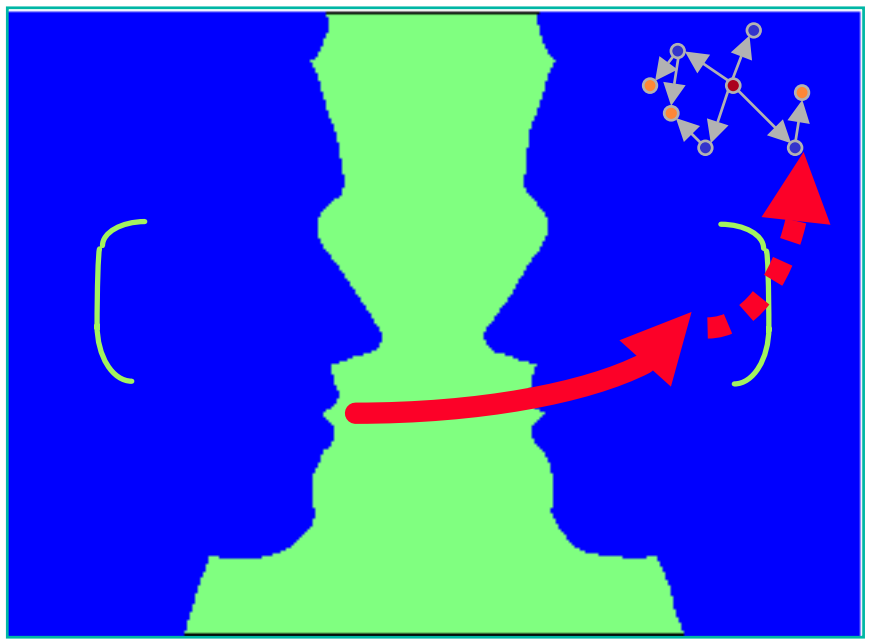
Sprachliche Kommunikation



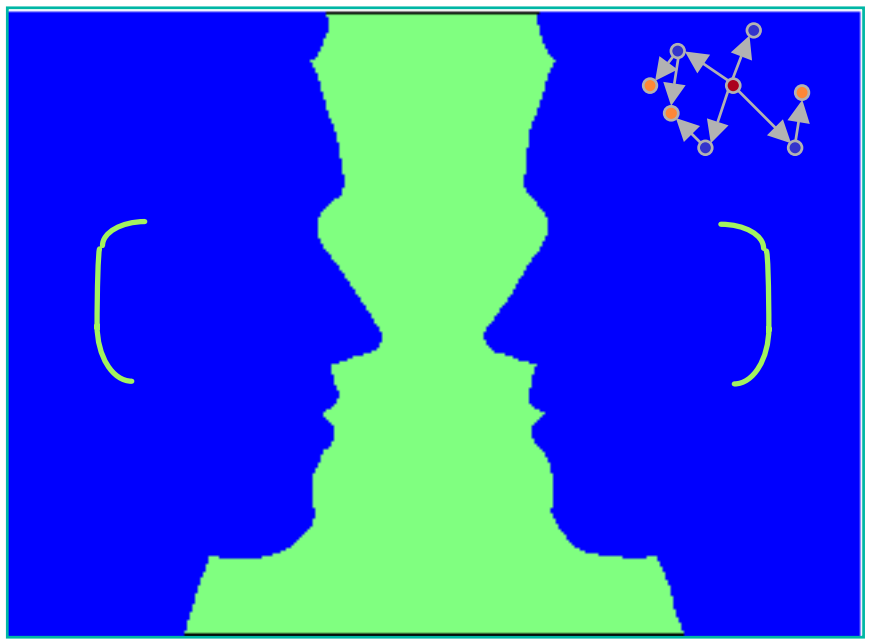
Sprachliche Kommunikation



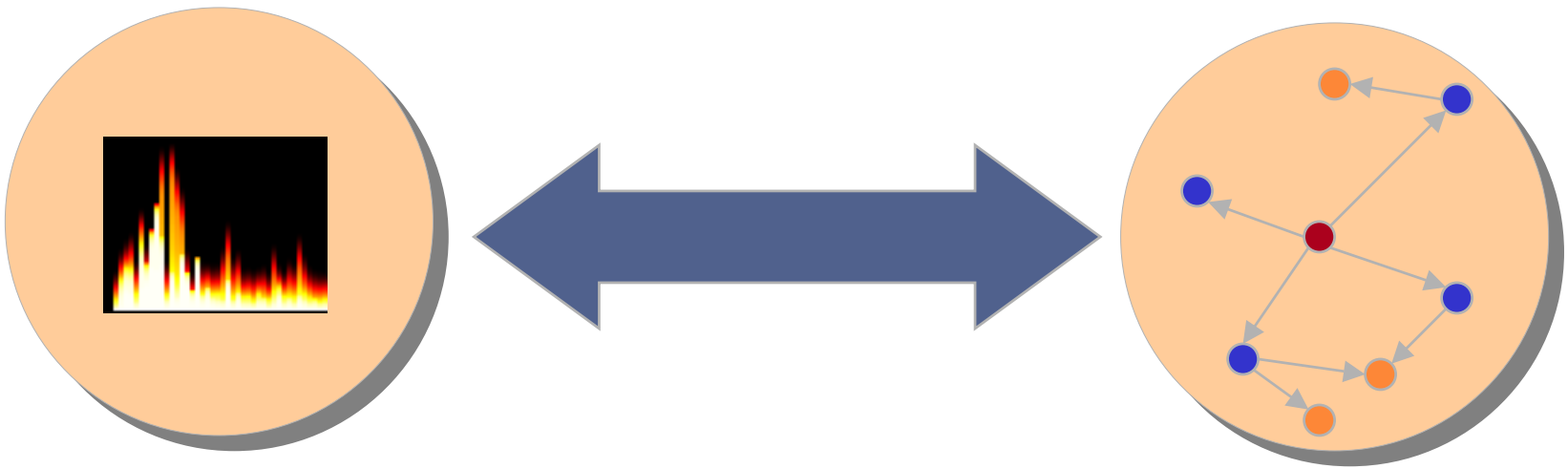
Sprachliche Kommunikation



Sprachliche Kommunikation



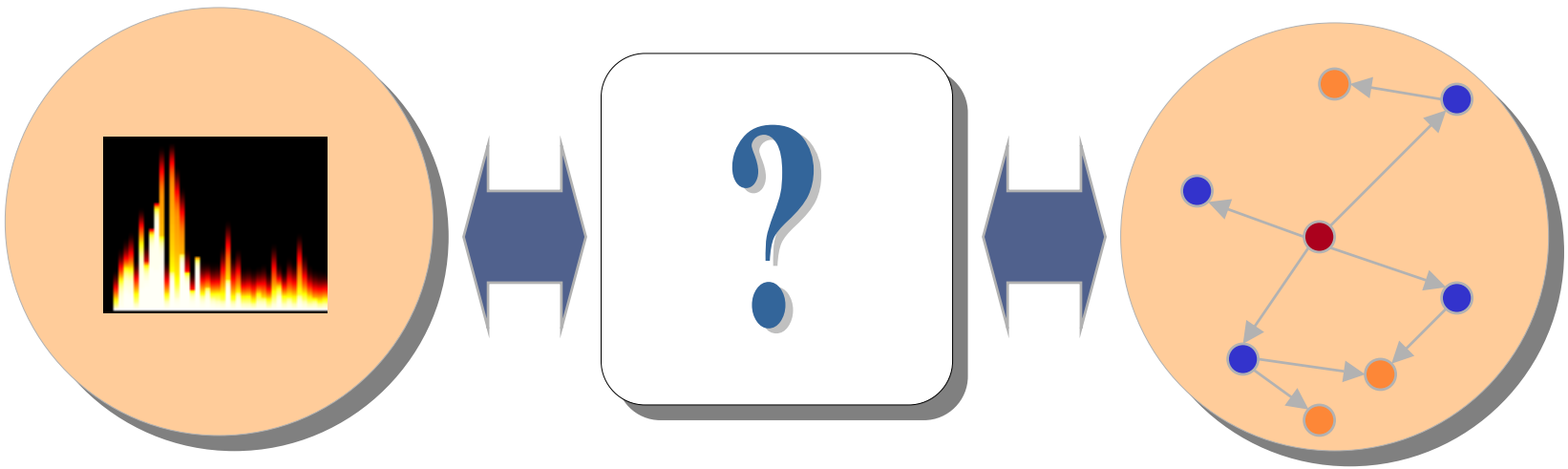
Grammatik



Schallwellen

Aktivation von Konzepten

Grammatik

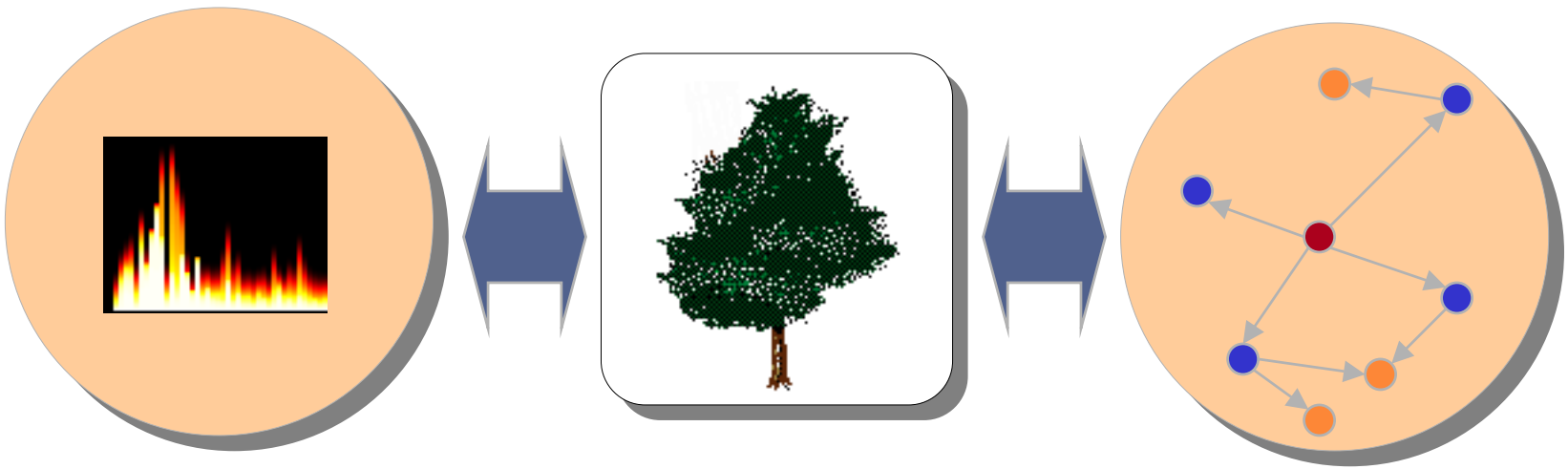


Schallwellen

Grammatik

Aktivation von Konzepten

Grammatik

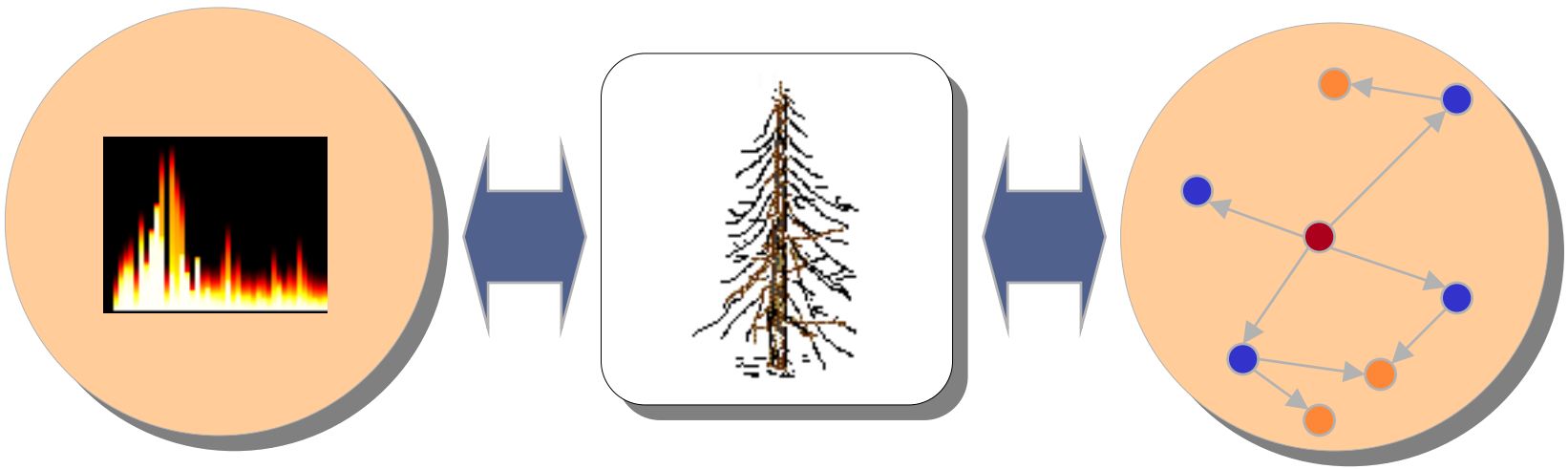


Schallwellen

Grammatik

Aktivation von Konzepten

Grammatik

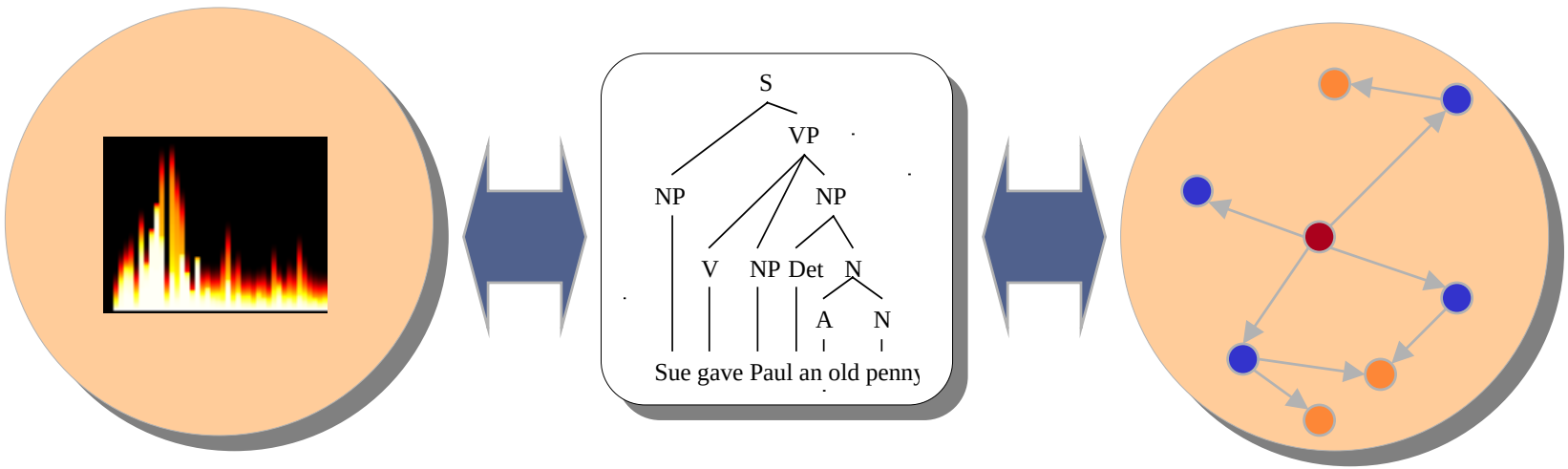


Schallwellen

Grammatik

Aktivation von Konzepten

Grammatik

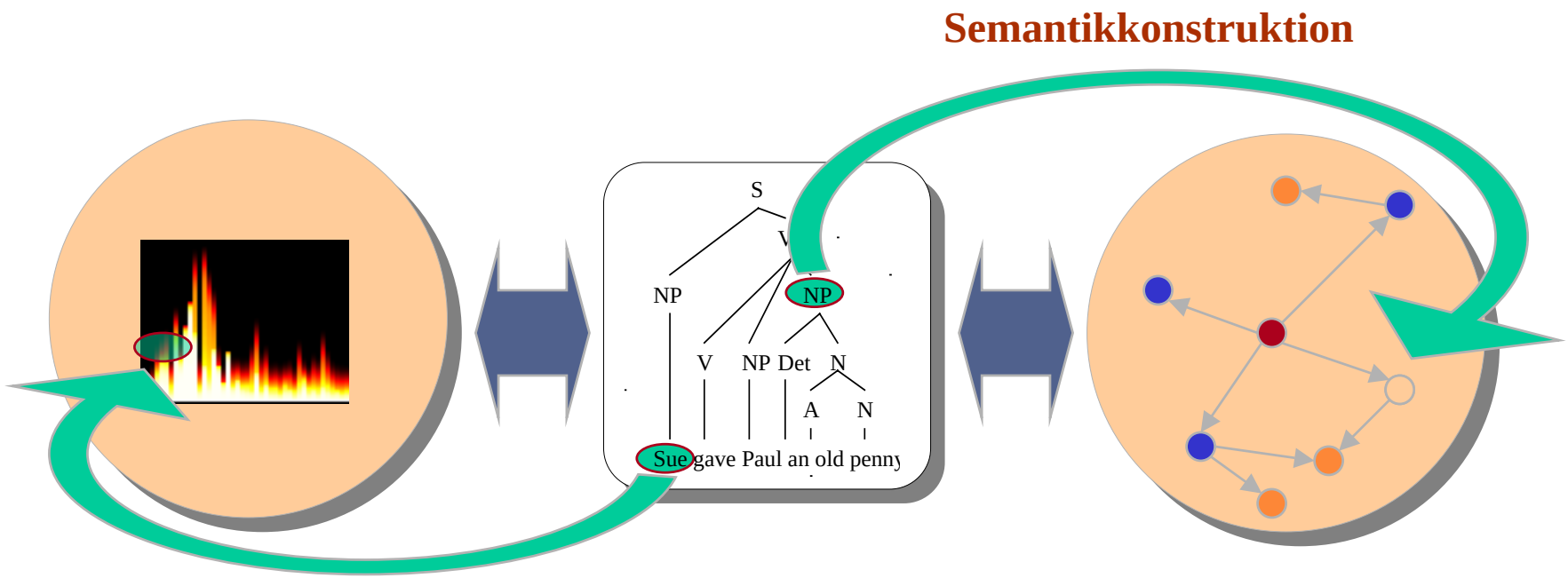


Schallwellen

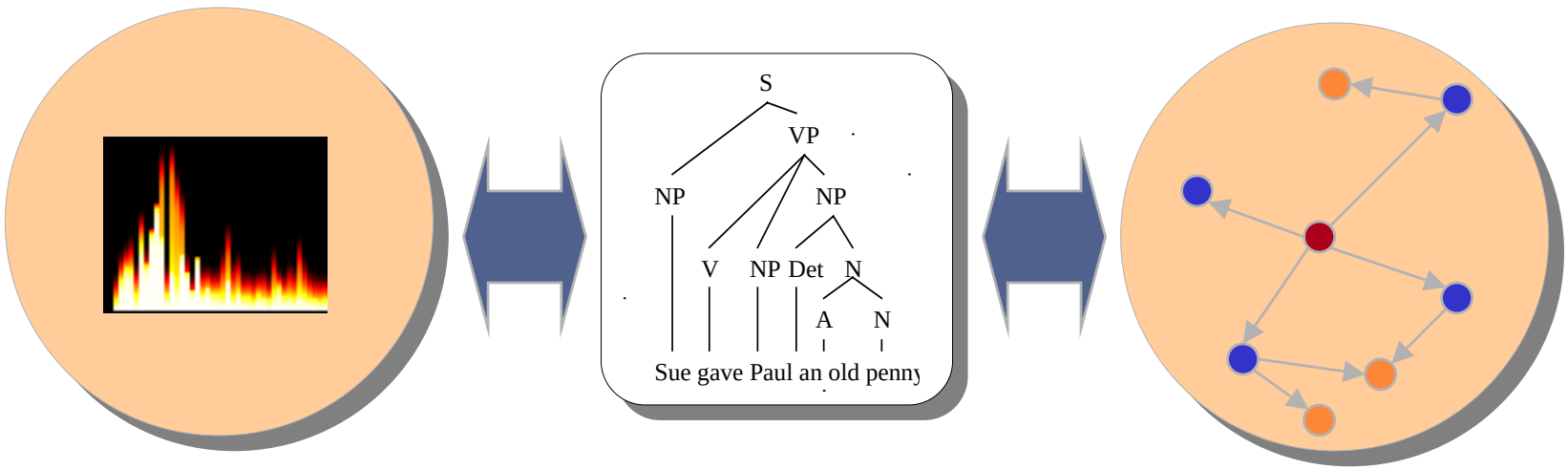
Grammatik

Aktivation von Konzepten

Grammatik



Grammatik



Schallwellen

Grammatik

Aktivation von Konzepten

Motivation

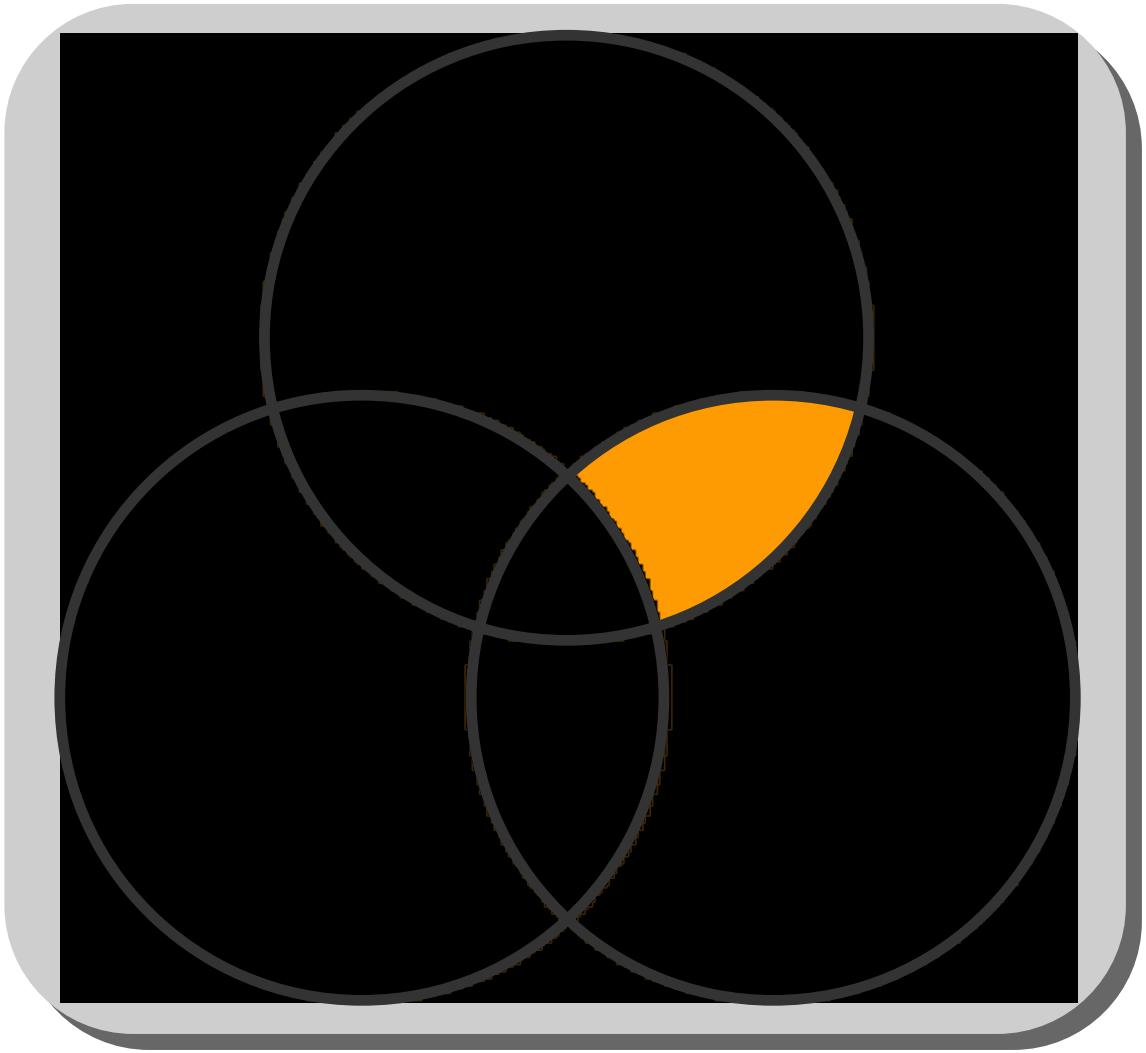
Die Computerlinguistik beschäftigt sich mit den formalen Aspekten natürlicher Sprache. Sie wird aus unterschiedlichen Richtungen motiviert:

Sprachwissenschaftliches Interesse Modelle der Grammatik

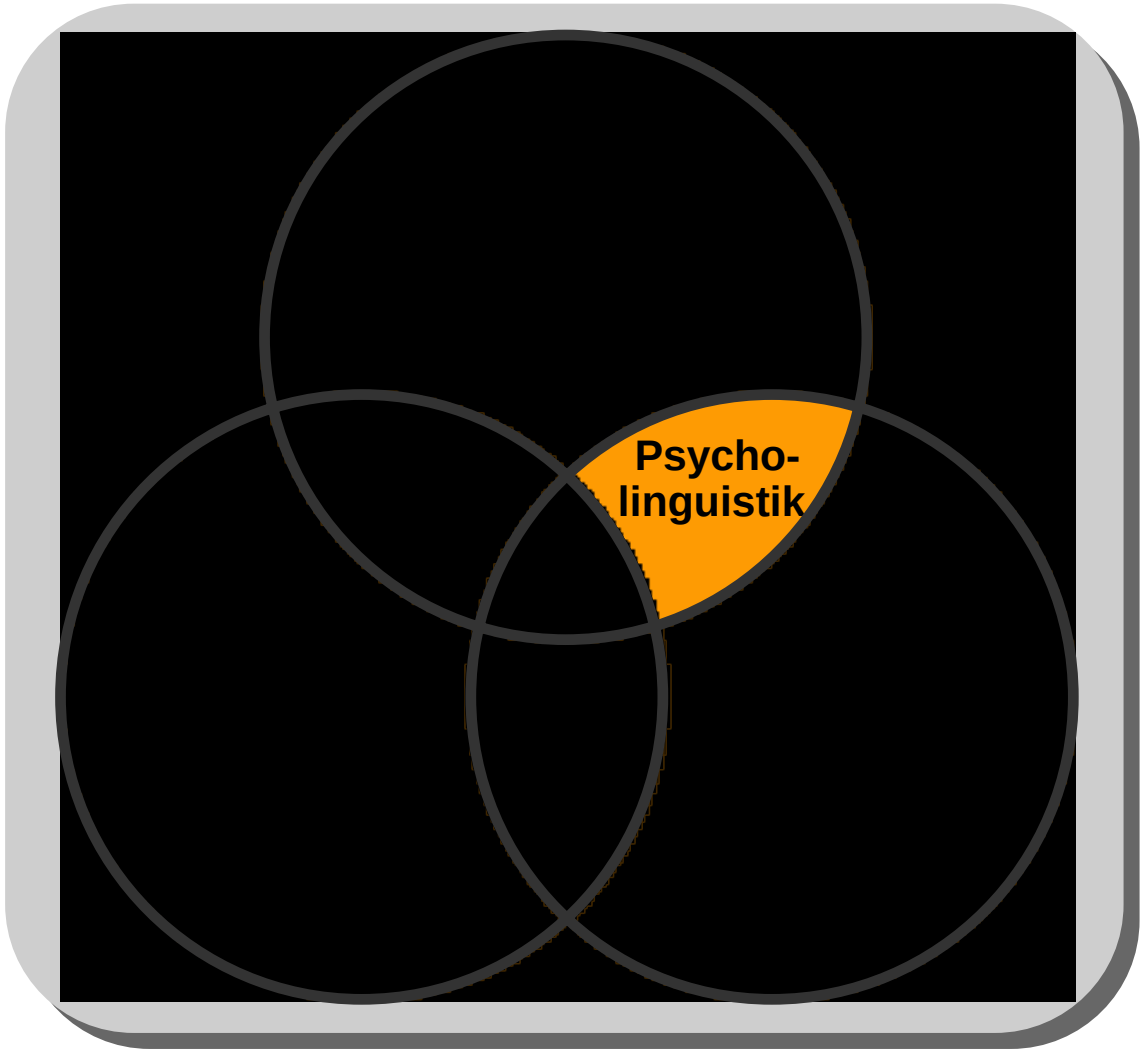
Kognitionswissenschaftliches Interesse
Modelle der menschlichen Sprachfähigkeit

Ingenieurwissenschaftliches Interesse
Softwareentwicklung

NACHBARWISSENSCHAFTEN



NACHBARWISSENSCHAFTEN



Die Disziplin

Computerlinguistik im weiteren Sinne

ist ein zwischen Linguistik und Informatik liegendes interdisziplinäres Forschungsgebiet, das sich mit der maschinellen Verarbeitung natürlicher Sprachen beschäftigt.

Computerlinguistik im engeren Sinne

ist ein Teilgebiet der modernen Linguistik, das berechenbare Modelle menschlicher Sprache entwirft, implementiert und untersucht.

Forschungsgebiete der CL

- Theoretische Computerlinguistik
  - Teilgebiet der Linguistik
  - Entwurf, Implementierung und Untersuchung berechenbarer Modelle natürlicher Sprache
  - Ziel: Beitrag zur Verbesserung der zugrundeliegenden linguistischen und psychologischen Theorien

  $\Rightarrow$ Computational Linguistics (CL)

- Angewandte Computerlinguistik
  - Interdisziplinäres Forschungsgebiet zwischen Informatik und Linguistik
  - Maschinelle Verarbeitung natürlicher Sprache
  - Ziel: Softwareanwendungen, die einen Teil der natürlichen Sprache simulieren

  $\Rightarrow$ Natural Language Processing (NLP)

Fragestellungen der theoretischen CL

- Wie können Phänomene der natürlichen Sprache adäquat repräsentiert werden? Welche Mechanismen stehen zur Beschreibung der Phänomene zur Verfügung?

- Welche Eigenschaften muss ein Formalismus aufweisen, um relevante Aspekte natürlicher Sprache angemessen repräsentieren zu können? Wie komplex ist der entsprechende Formalismus?

- Welche Komplexität weist die natürliche Sprache bzw. ein bestimmter Phänomenbereich derselben auf, und inwieweit kann die Komplexität effektiv bewältigt werden?

$\Rightarrow$ Wie sieht eine adäquate Modellierung natürlicher Sprache aus? Welche Komplexität besitzt sie?

Fragestellungen der angewandten CL

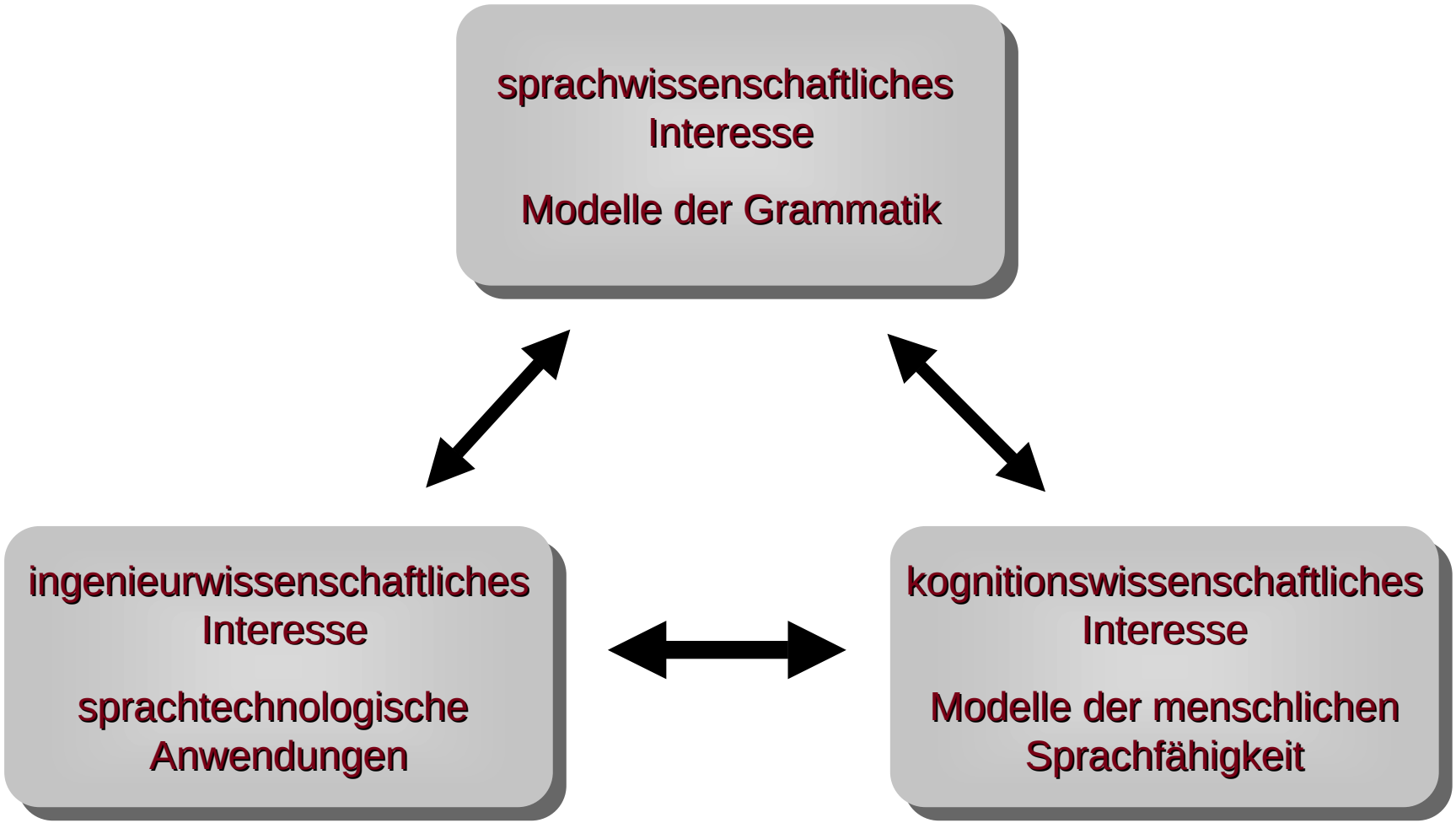
- Wie kann sprachliches Wissen erfolgreich auf einer Maschine modelliert werden?

- Was sind die Probleme, die sich bei der konkreten Implementierung stellen, und wie können sie bewältigt werden?

- Welche Formalismen eignen sich für die Modellierung welcher (evtl. einzelsprachlichen) Phänomene bzw. welcher Aspekte einer Sprache?

⇒ Wie sieht ein konkreter Algorithmus zur Verarbeitung natürlichsprachlicher Äußerungen aus?

Motivationen



Methoden der CL

Symbolische Verfahren stützen sich auf formale theoretische Grundlagen wie Formale Sprachtheorie, Logik etc.

Stochastische Verfahren stützen sich auf wahrscheinlichkeitstheoretische Grundlagen unter Benutzung stochastischer Methoden wie das Training von Modellen anhand eines Korpus

Es wurde lange Zeit (fälschlicherweise) angenommen, dass ein stochastisches Verfahren den Unterschied zwischen folgenden beiden Sätzen nicht erfassen könnte:

1. Colorless green ideas sleep furiously.

2. Furiously sleep ideas green colorless.

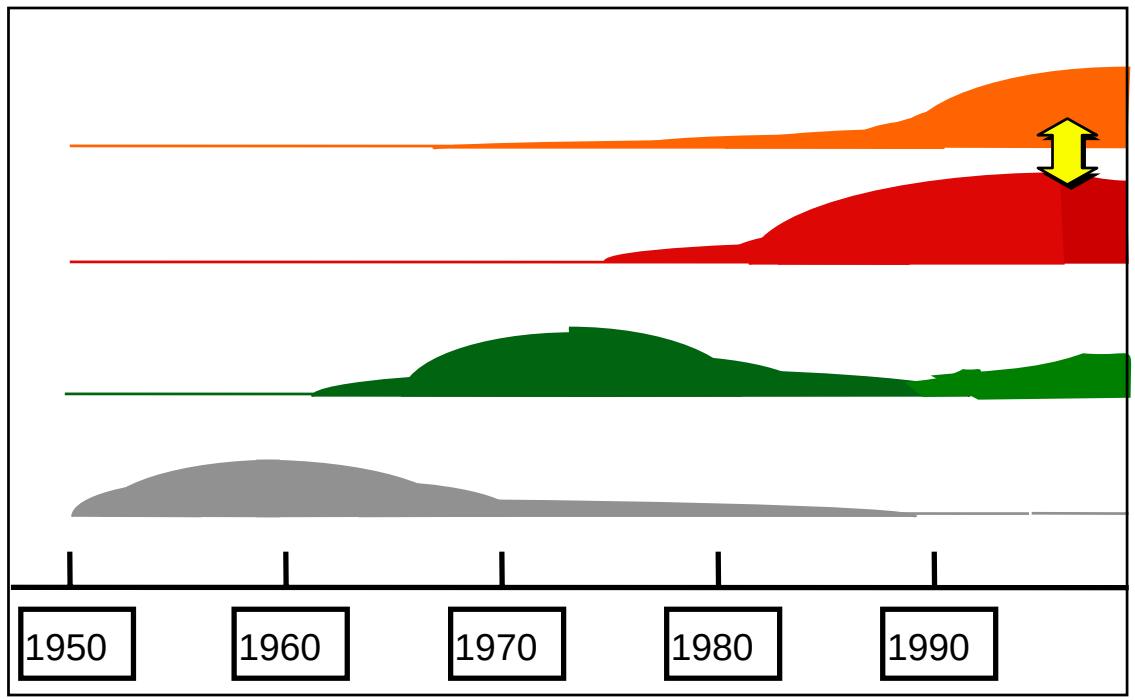
Hauptansätze der CL

statistische und konnektionistische Methoden in der CL

deklarative linguistische Formalismen in der CL

spezielle Verfahren für die CL

direkte Programmierung, keine Trennung von Beschreibung und Verarbeitung



Komponenten eines Sprachmodells

1. Spracherkennung und -ausgabe (**Phonetik** und **Phonologie**)

2. Struktur und Verarbeitung von Wortformen (**Morphologie**)

3. Satzstruktur (**Syntax**)

4. Bedeutung einzelner Wörter (**Lexikalische Semantik**) und ihrer Kombination (**Kompositionelle Semantik**)

5. Wissen über Äußerungszusammenhänge (**Diskurs**) und über konventionelle Verhaltensweisen (**Pragmatik**)

6. Hintergrundwissen/Weltwissen

Komponenten eines Sprachmodells

akustische Form

geschriebene Form

phonetische Verarbeitung

orthographische Verarbeitung

phonetische o. graphemische Repräsentation

morphonologische Verarbeitung

morphonologische Repräsentation

syntaktische Verarbeitung

syntaktische Repräsentation

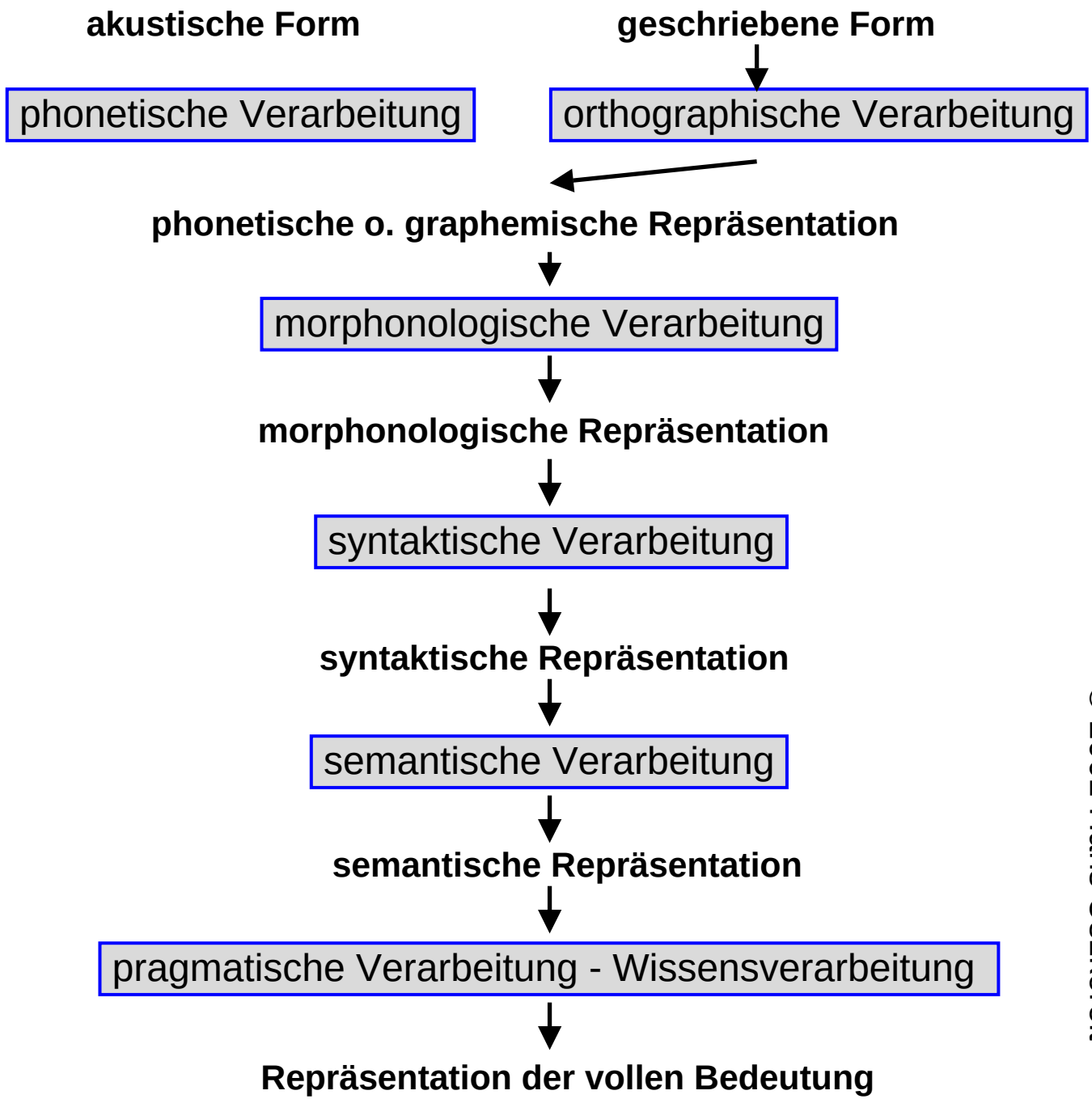
semantische Verarbeitung

semantische Repräsentation

pragmatische Verarbeitung - Wissensverarbeitung

Repräsentation der vollen Bedeutung

Textverstehen

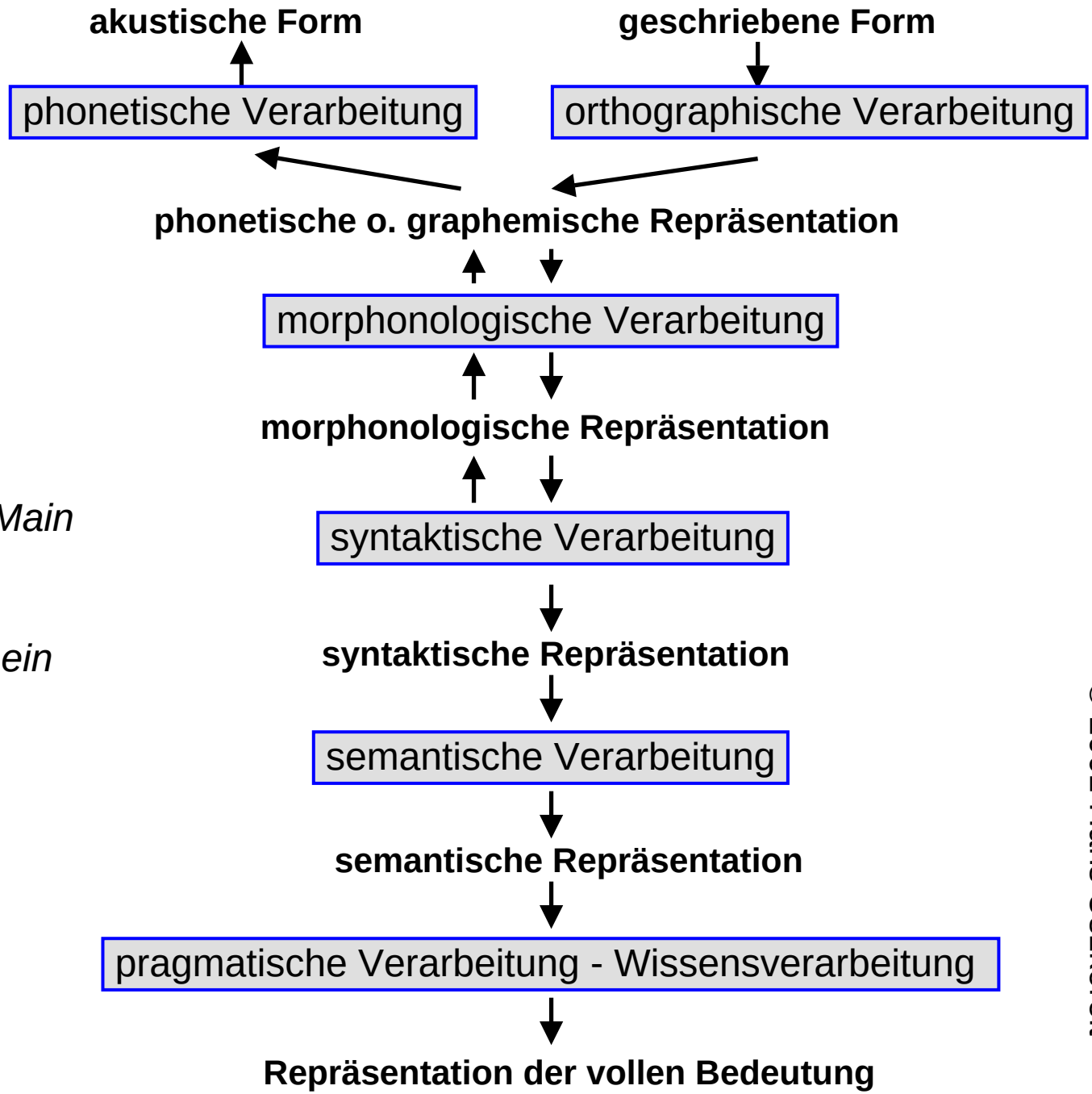


Diktat

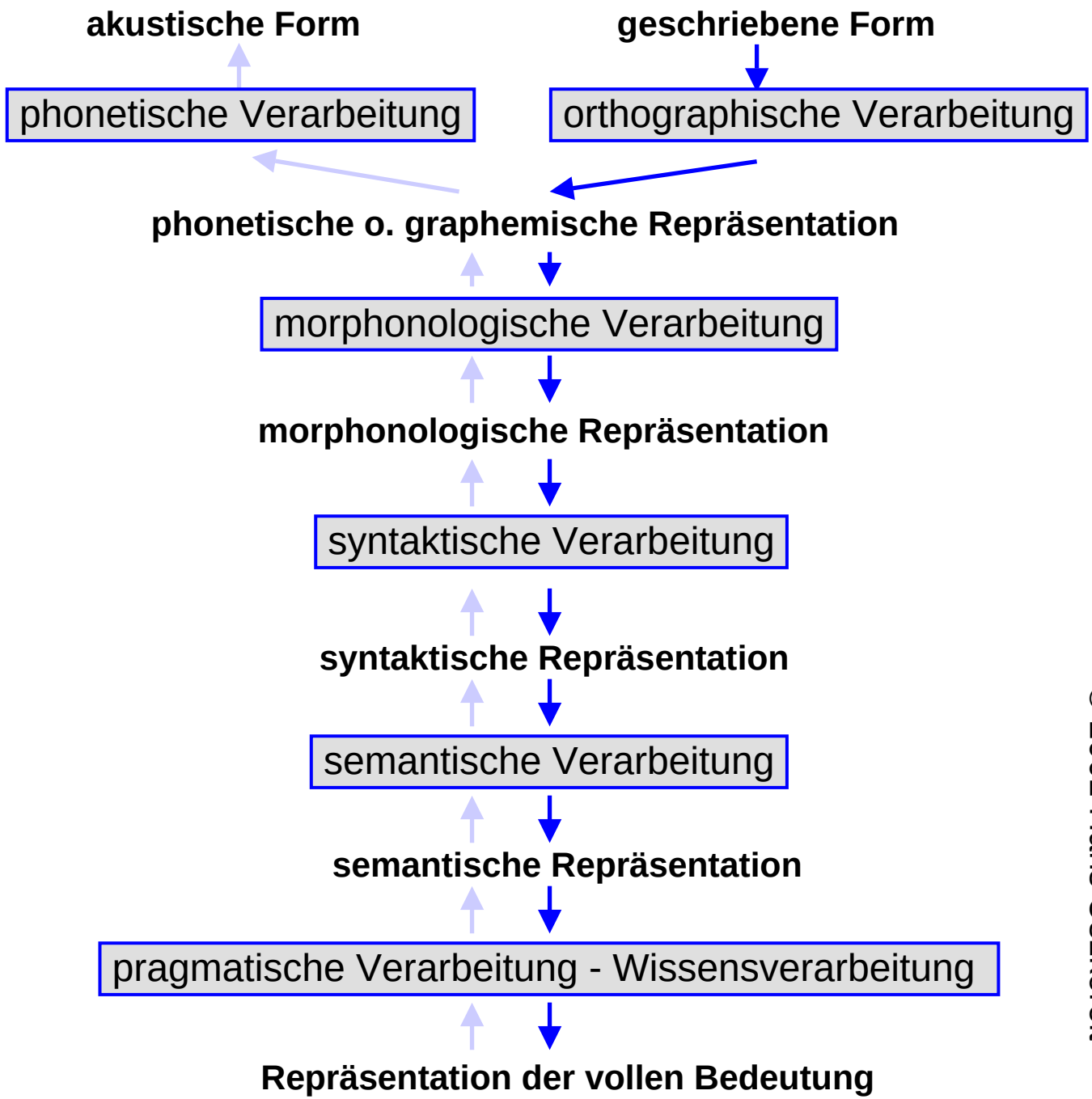
das Boot auf dem Main

oder

daß bot auf dem mein



Maschinelle Übersetzung



Anwendungen (1)

Korrekturprogramme: Rechtschreibkorrektur, Grammatikkorrektur Korrekturvorschläge, Verbesserung von Texterfassung mittels OCR.

Computergestützte Lexikographie: Hilfe bei der Erstellung und Pflege von Lexika; Akquisition lexikalischer Information, Repräsentation lexikalischer Information, Bereitstellung der lexikalischen Information für Anwendungen.

Volltextsuche (Information Retrieval): Indexkonstruktion, Auswertung von Suchanfragen, Retrievalmodell

Textmining: Strukturierung großer Textkollektionen, Textklassifikation, Schlüsselwortextraktion, Aufbau einer Taxonomie

Textklassifikation: Erlernen von Klassenprofilen anhand von Trainingsdaten, Klassifikationsalgorithmus

Informationsextraktion: Identifizierung relevanter Information in Texten, Instantiierung von Templates

Textzusammenfassung: Reduktion / Verdichtung, Textproduktion

Anwendungen (2)

Sprachsynthesesysteme: Produktion gesprochener Sprache aus geschriebener Sprache, Computerarbeitsplatz für Blinde, telefonische Auskunftsysteme, Navigationsysteme

Spracherkennungssysteme: Diktiersysteme, telefonische Auskunftsysteme; Signalanalyse, Geräuschfilterung, Adaption an verschiedene Sprecher

Natürlichsprachliche Retrieval-Schnittstellen: z.B. natürlichsprachliche Anfragen an Bibliothekskataloge

Dialogsysteme: ELIZA, automatische Auskunftsysteme, natürlichsprachliche Benutzerschnittstellen

Sprachlehr- und -lernsysteme: Hilfe bei dem Erwerb von Fremdsprachen; Anpassung an das individuelle Arbeitstempo und den Kenntnisstand, ortsungebunden, zeitlich flexibel, objektiv, nicht ermüdend

Anwendungen (3)

Elektronische Kommunikationshilfen: Wort- und Satzvervollständigung (SMS), Texttelephone, Textvereinfachungswerkzeuge, ...

Angewandte natürlichsprachliche Generierungs- und Auskunftssysteme: Wettervorhersagen, Gesundheitswesen, technische Dokumentationen, Computerspiele, ...

Maschinelle Übersetzung: Vollautomatische Übersetzung, Computergestützte Übersetzung