Vorlesung: Einführung in die Computerlinguistik

Hans Uszkoreit

Übersicht des ersten Teils

- Aufgaben und Einordnung des Faches
- Motivationen für die Modellierung menschlicher Sprache
- Computerlinguistik als eine moderne Sprachwissenschaft
- Repräsentationen und Verarbeitungskomponenten

Faszination

Wissenschaft

Technologie

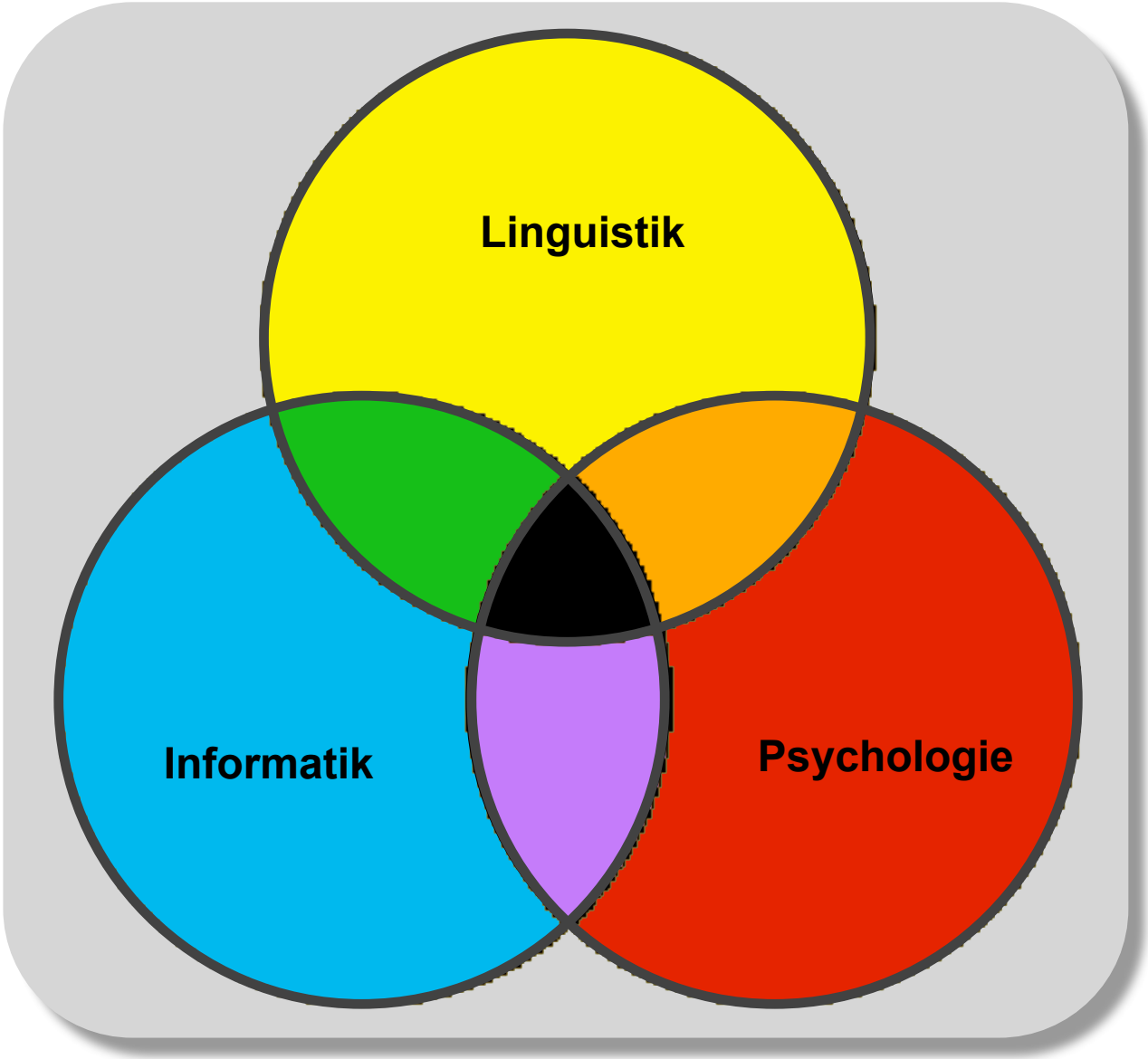
Faszination

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

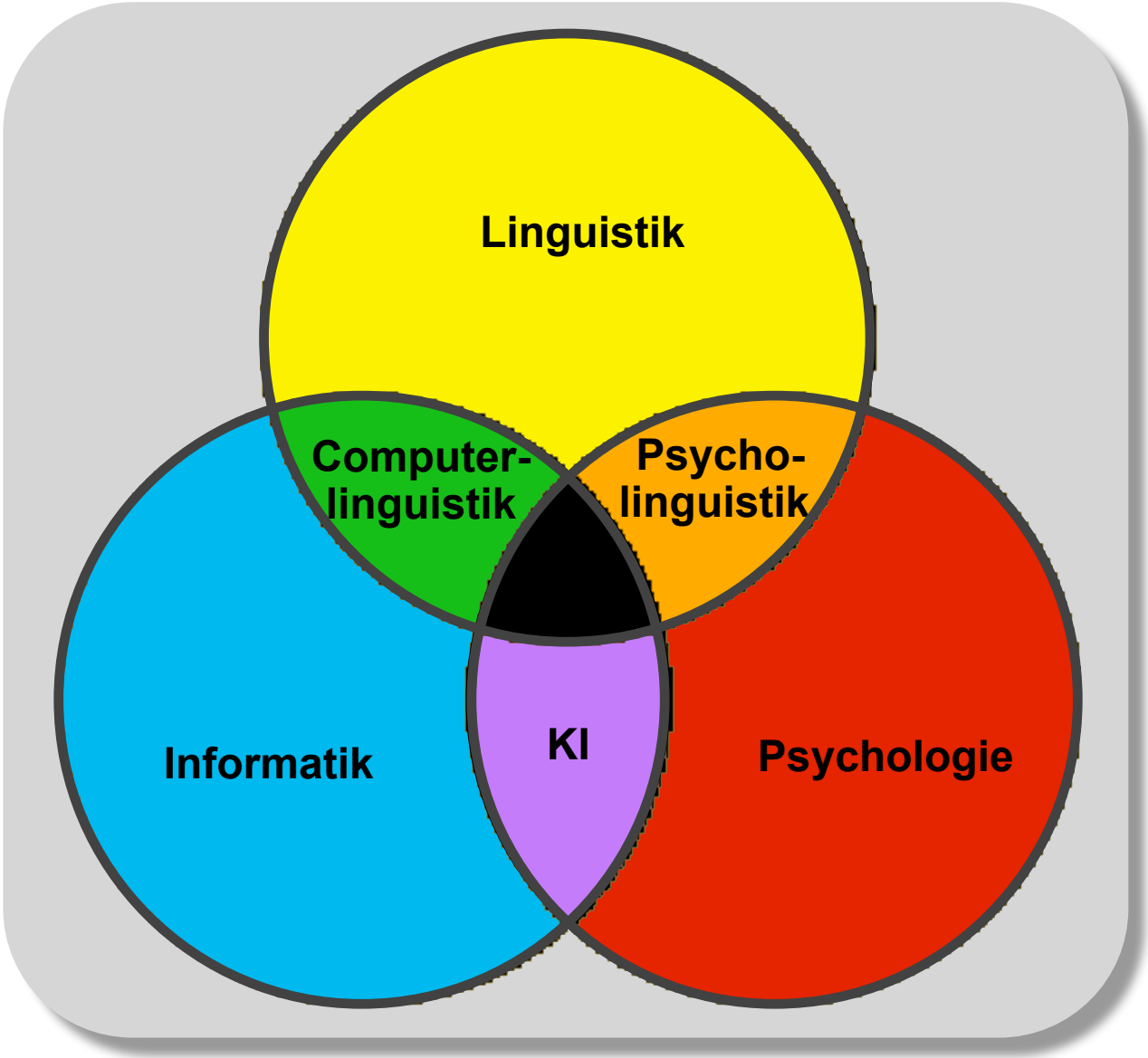
Es ist ein Wunder, wie wir in Sekundenschnelle komplexe Gedanken in einem Satz ausdrücken können.

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

NACHBARWISSENSCHAFTEN



NACHBARWISSENSCHAFTEN



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.

Die Disziplin

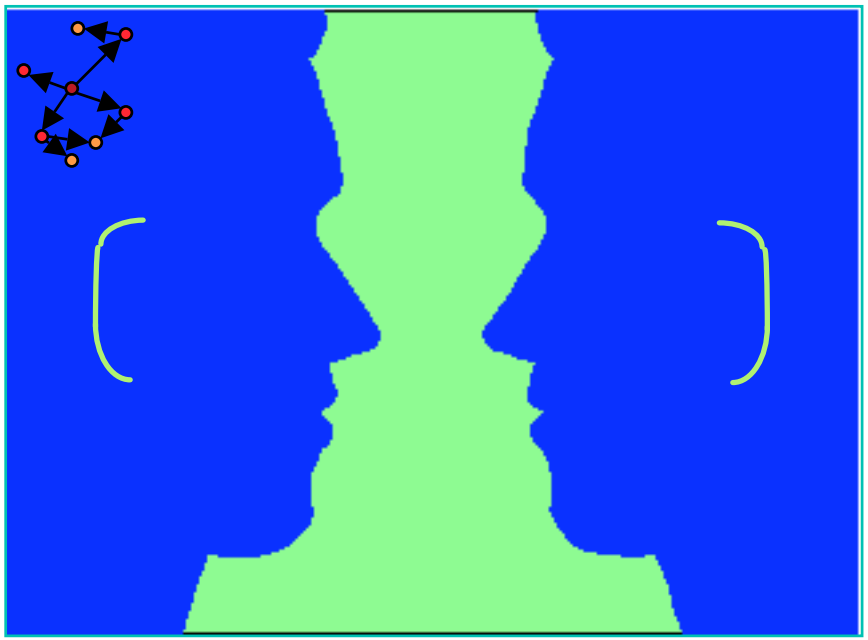
Theoretische Computerlinguistik

entwirft, implementiert und untersucht die Modelle mit dem Ziel, zum Verständnis, zur Verifikation und zur Verbesserung der zugrundeliegenden linguistischen und psychologischen Theorien beizutragen.

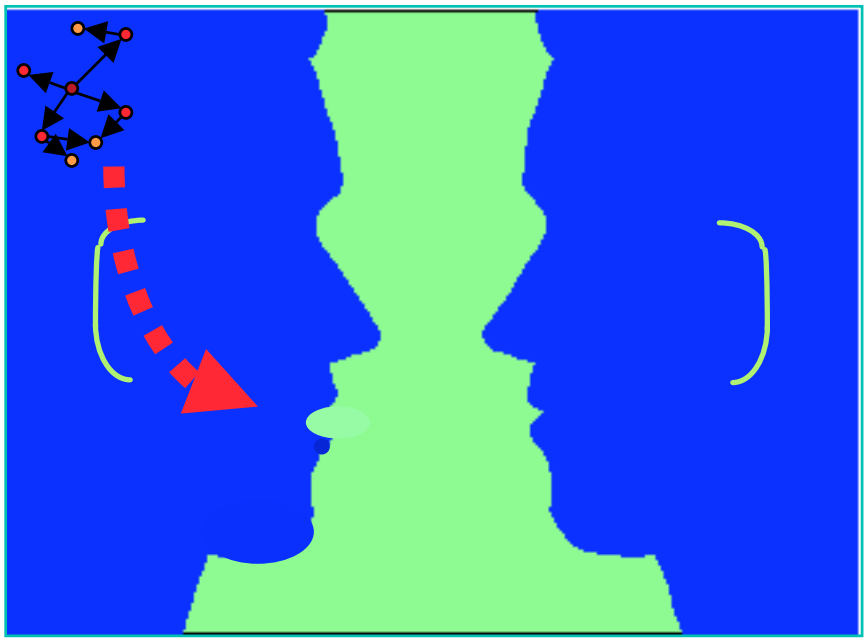
Angewandte Computerlinguistik

entwirft, implementiert und untersucht die Modelle mit dem Ziel, Softwareanwendungen zu ermöglichen, die über eine (eingeschränkte) Beherrschung menschlicher Sprache verfügen.

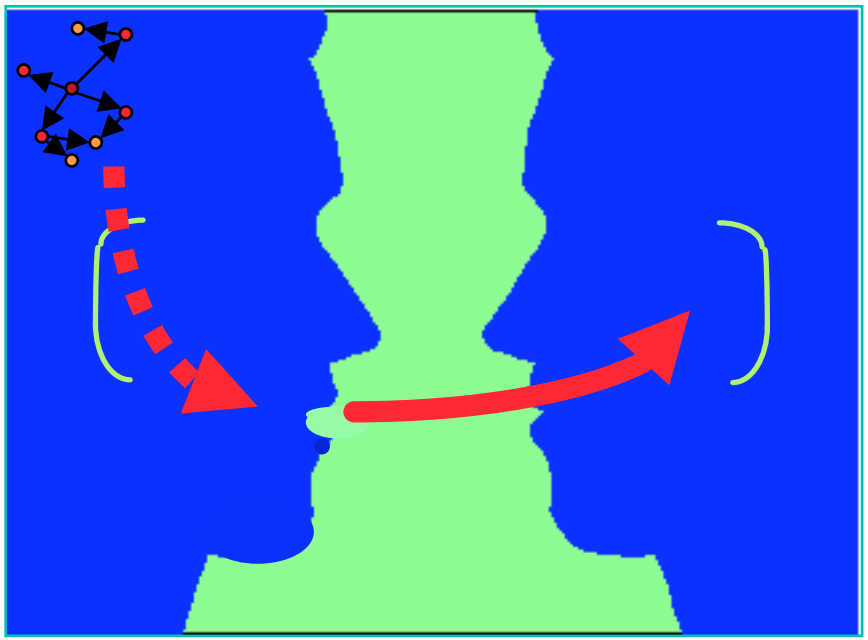
Sprachliche Kommunikation



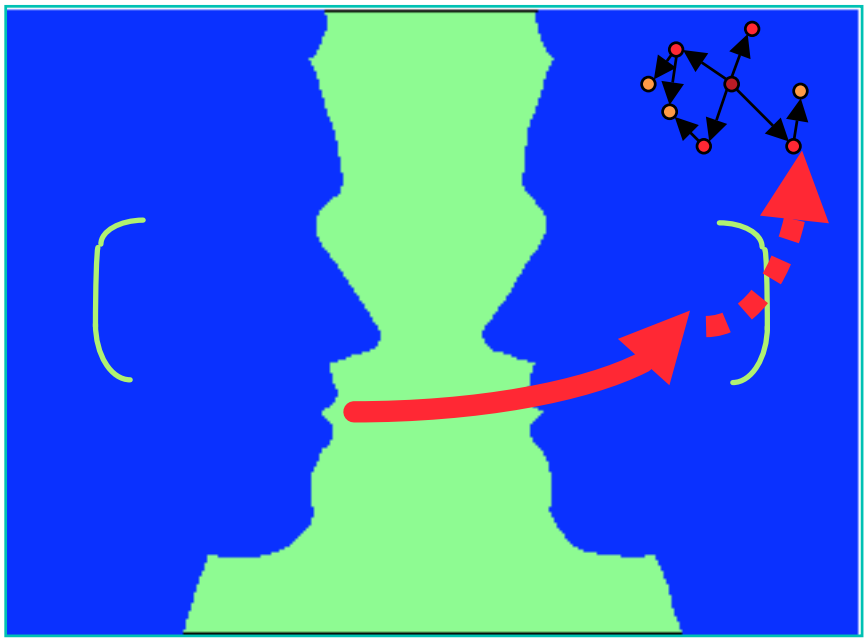
Sprachliche Kommunikation



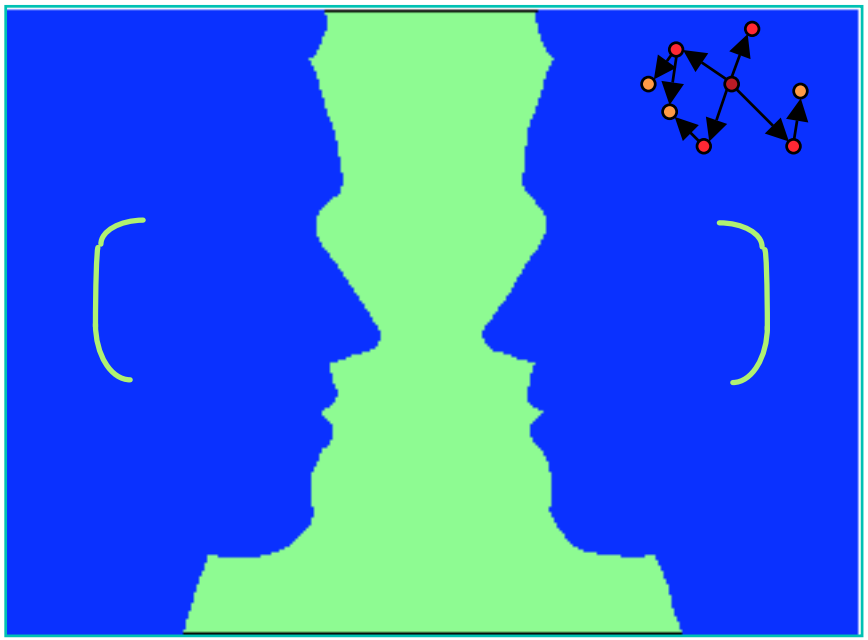
Sprachliche Kommunikation



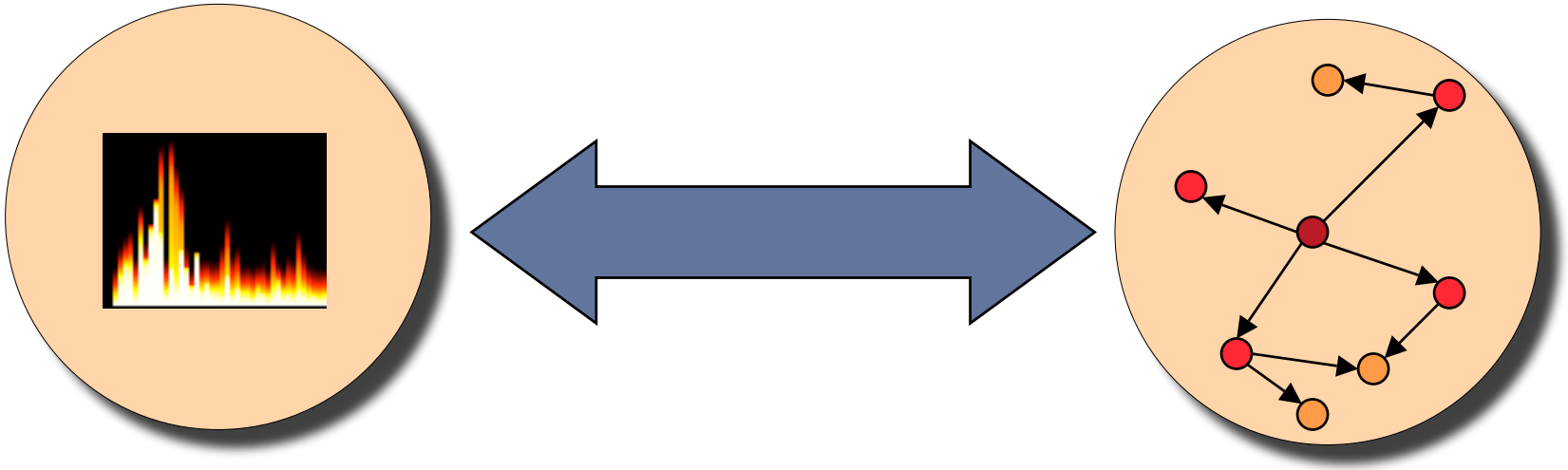
Sprachliche Kommunikation



Sprachliche Kommunikation



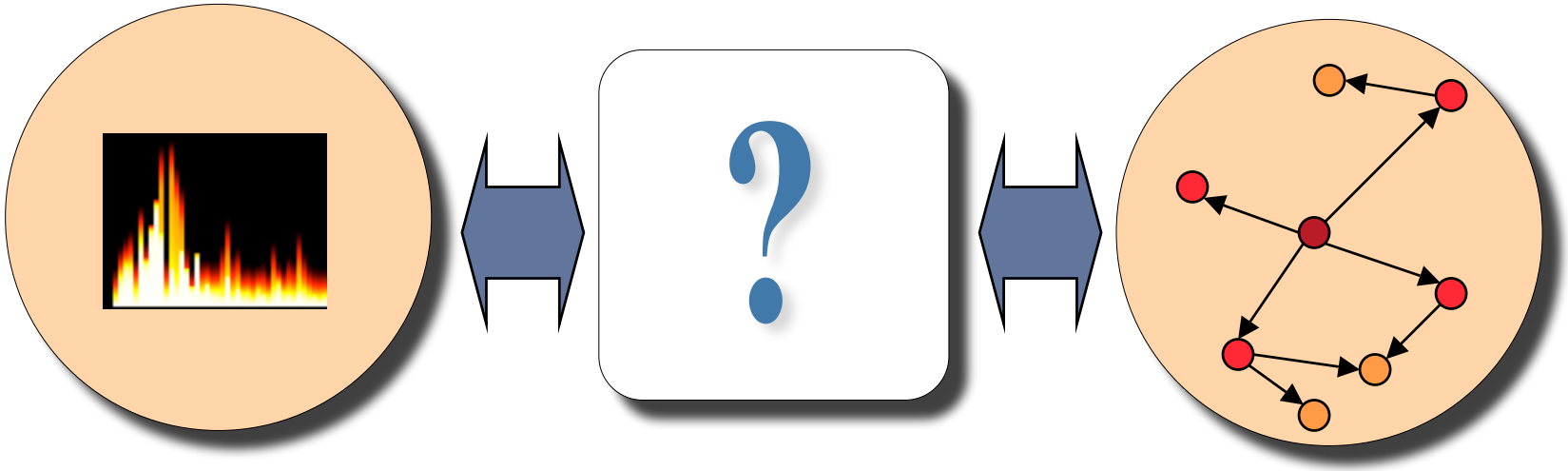
Grammatik



Schallwellen

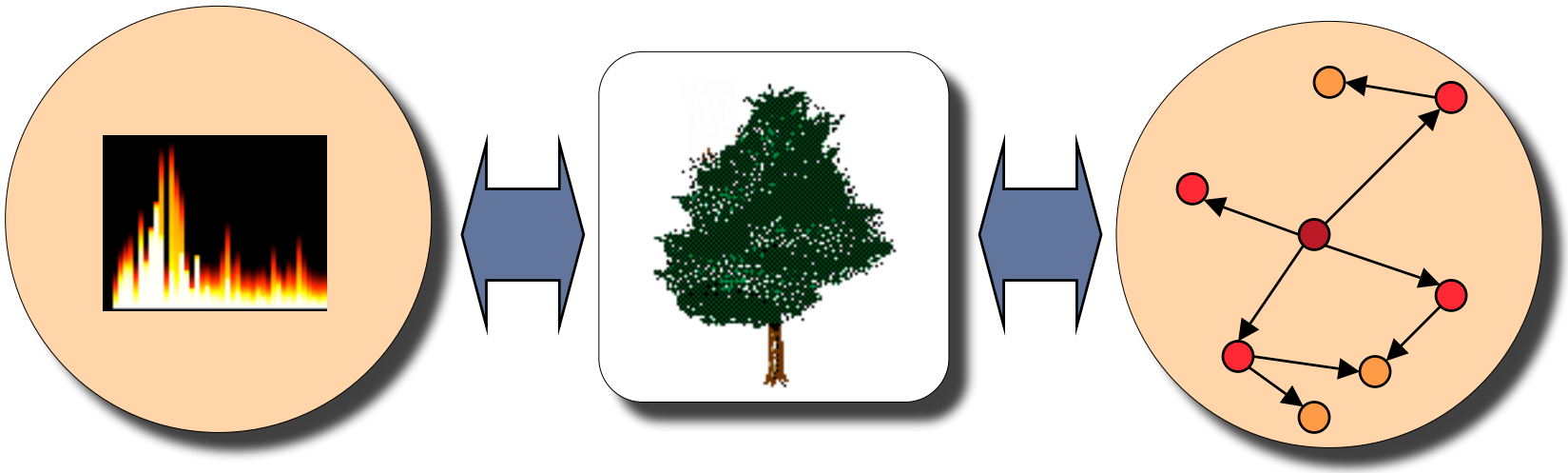
Aktivation von Konzepten

Grammatik



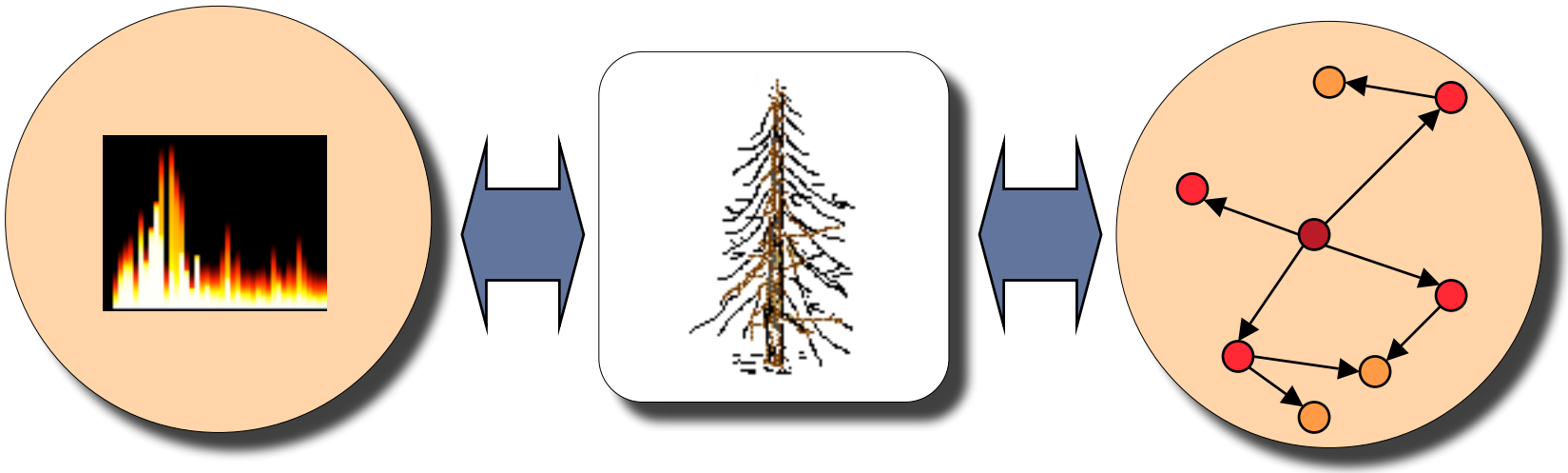
Schallwellen Grammatik Aktivation von Konzepten

Grammatik



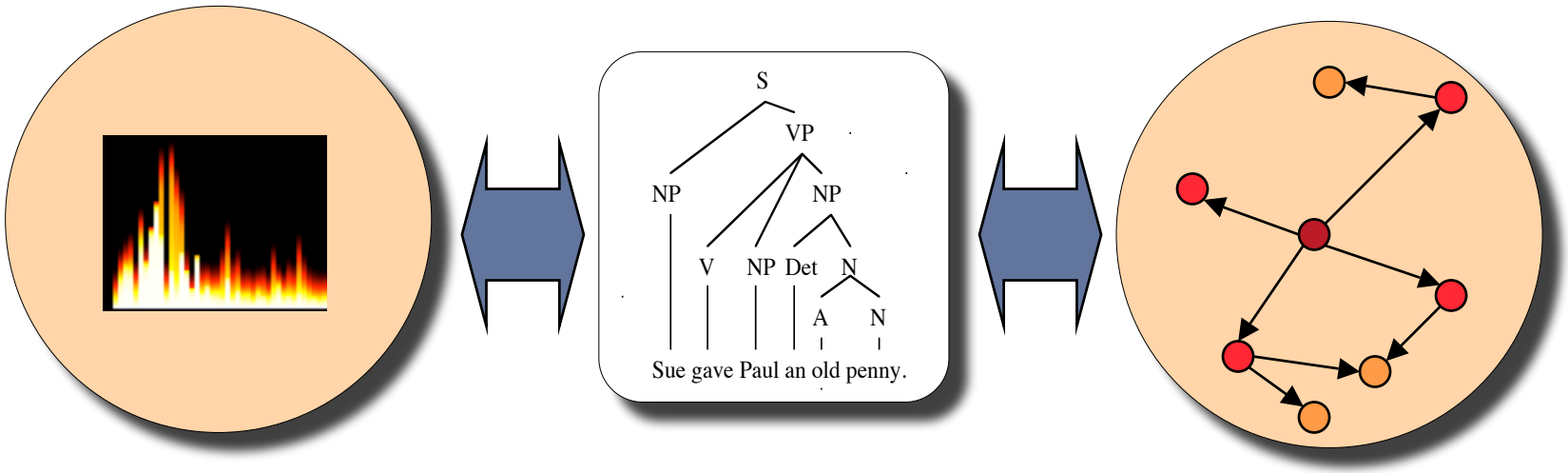
Schallwellen Grammatik Aktivation von Konzepten

Grammatik

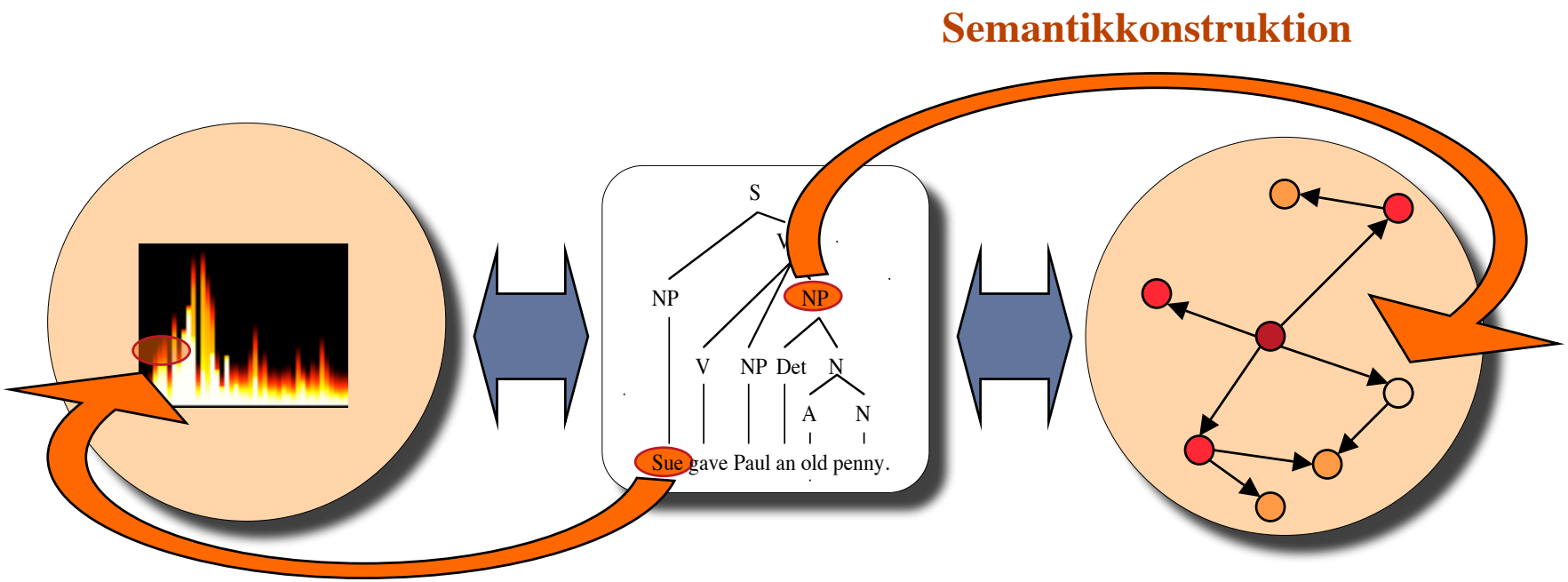


Schallwellen Grammatik Aktivation von Konzepten

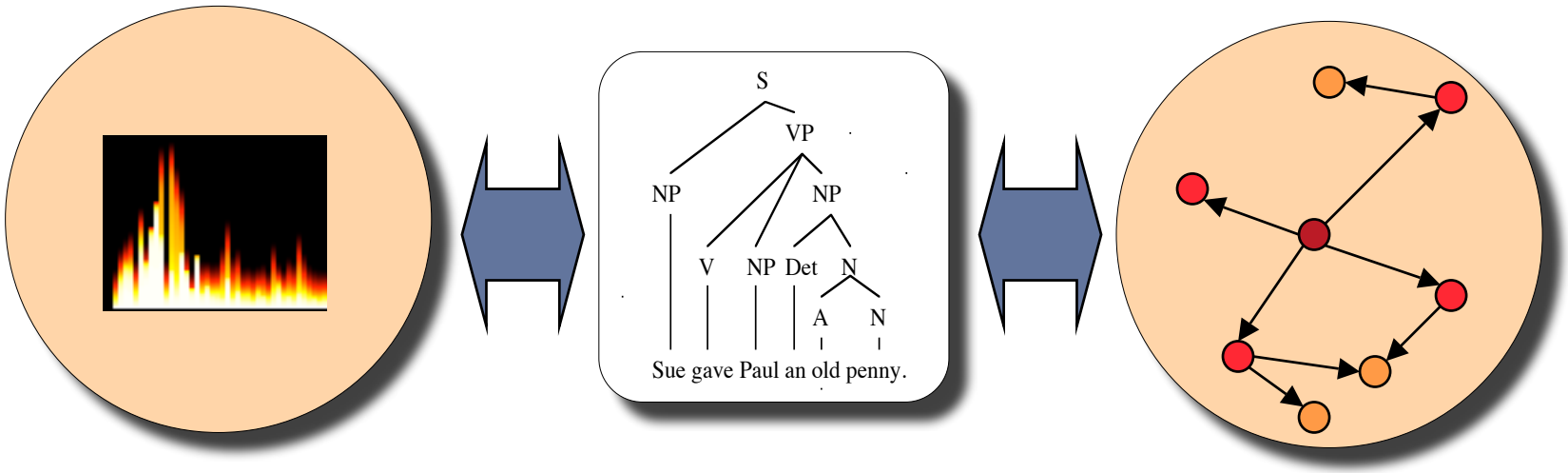
Grammatik



Schallwellen Grammatik Aktivation von Konzepten

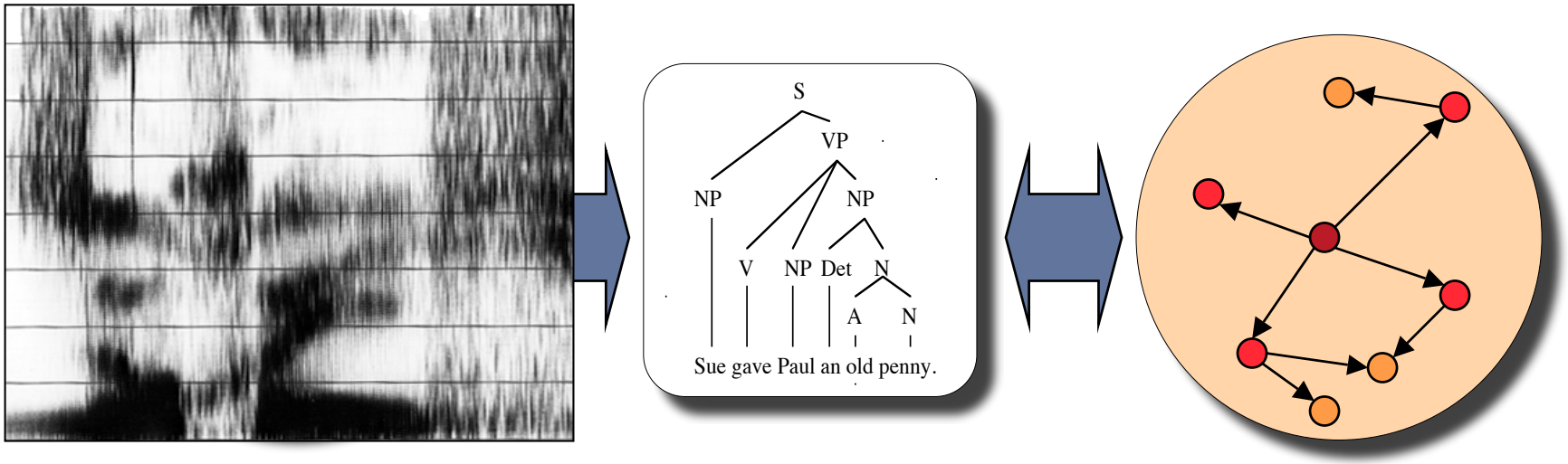


Grammatik



Schallwellen Grammatik Aktivation von Konzepten

Grammatik



Schallwellen Grammatik Aktivation von Konzepten

Verwandte Begriffe

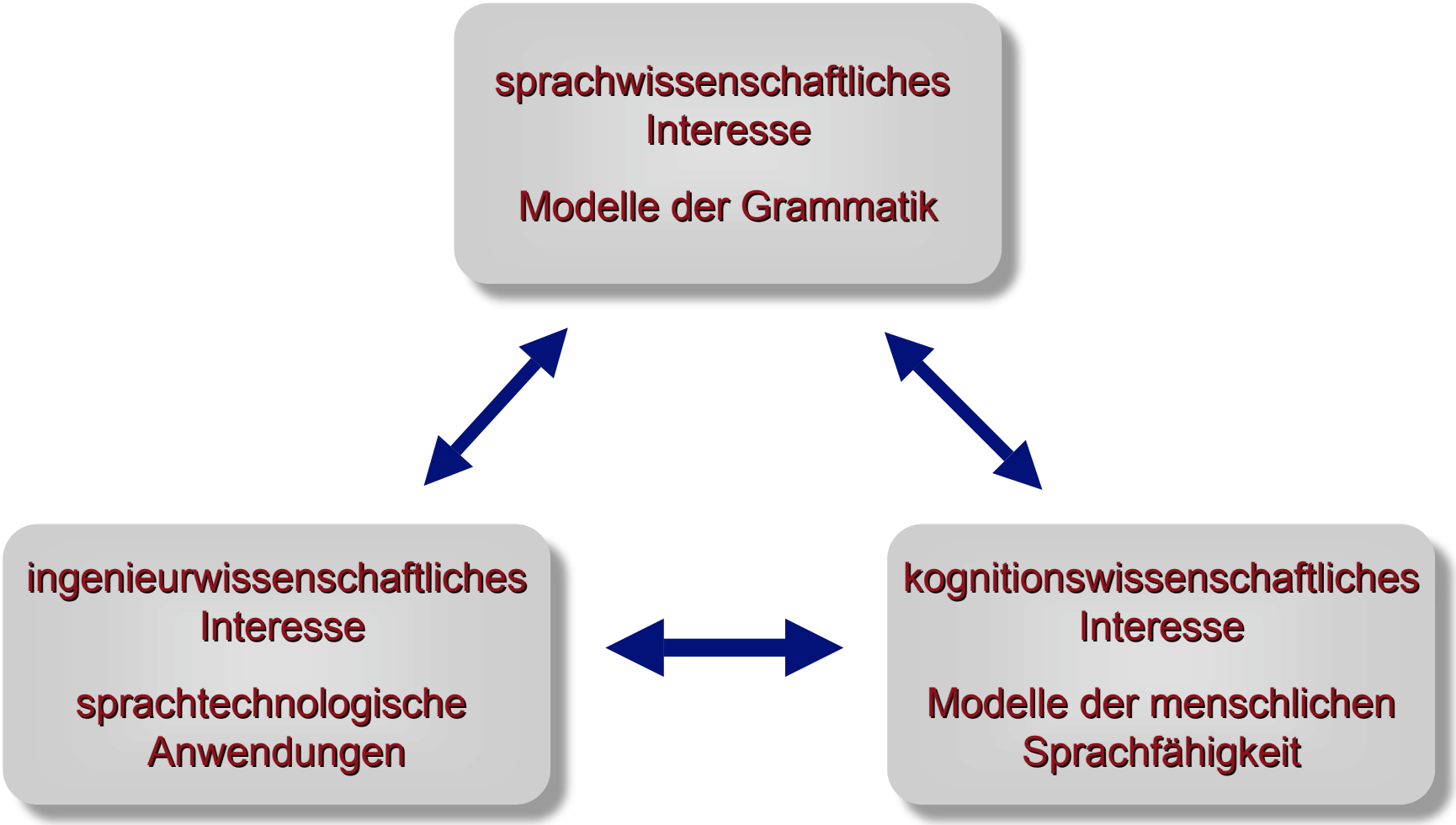
Maschinelle Sprachverarbeitung
: Analyse und Generierung von natürlicher Sprache mit dem Computer. Englisch: Natural Language Processing (NLP).

Sprachtechnologie(n)
: Übergriff für die Technologien sprachbeherrschender Systeme. Ingenieurwissenschaftliches Forschungsgebiet, in dem die Sprachtechnologien entwickelt werden.

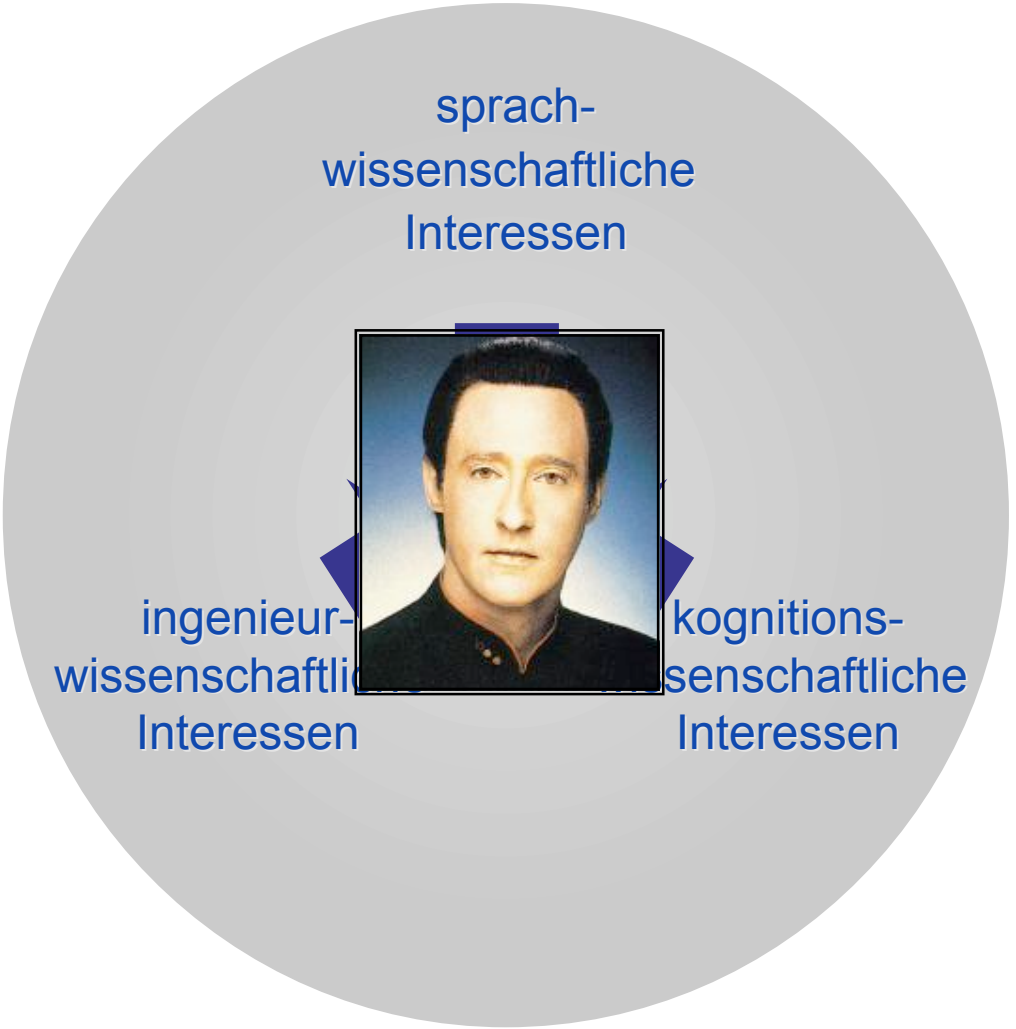
Linguistische Datenverarbeitung (LDV)
: Traditionell ein Teilgebiet der elektronischen Datenverarbeitung, das sich sowohl mit der Anwendung von Methoden der Datenverarbeitung für die linguistische Forschung als auch mit maschineller Sprachverarbeitung beschäftigt. Die LDV versteht sich heute als ein Gebiet, das die Computerlinguistik einschließt.

Sprachdatenverarbeitung
: Verarbeitung von sprachlichen Daten mit dem Computer. Schließt ein: mono- und multilinguale Textverarbeitung, elektronische Wörterbücher, Konkordanzen, Terminologiebanken, maschinelle und maschinengestützte Übersetzung.

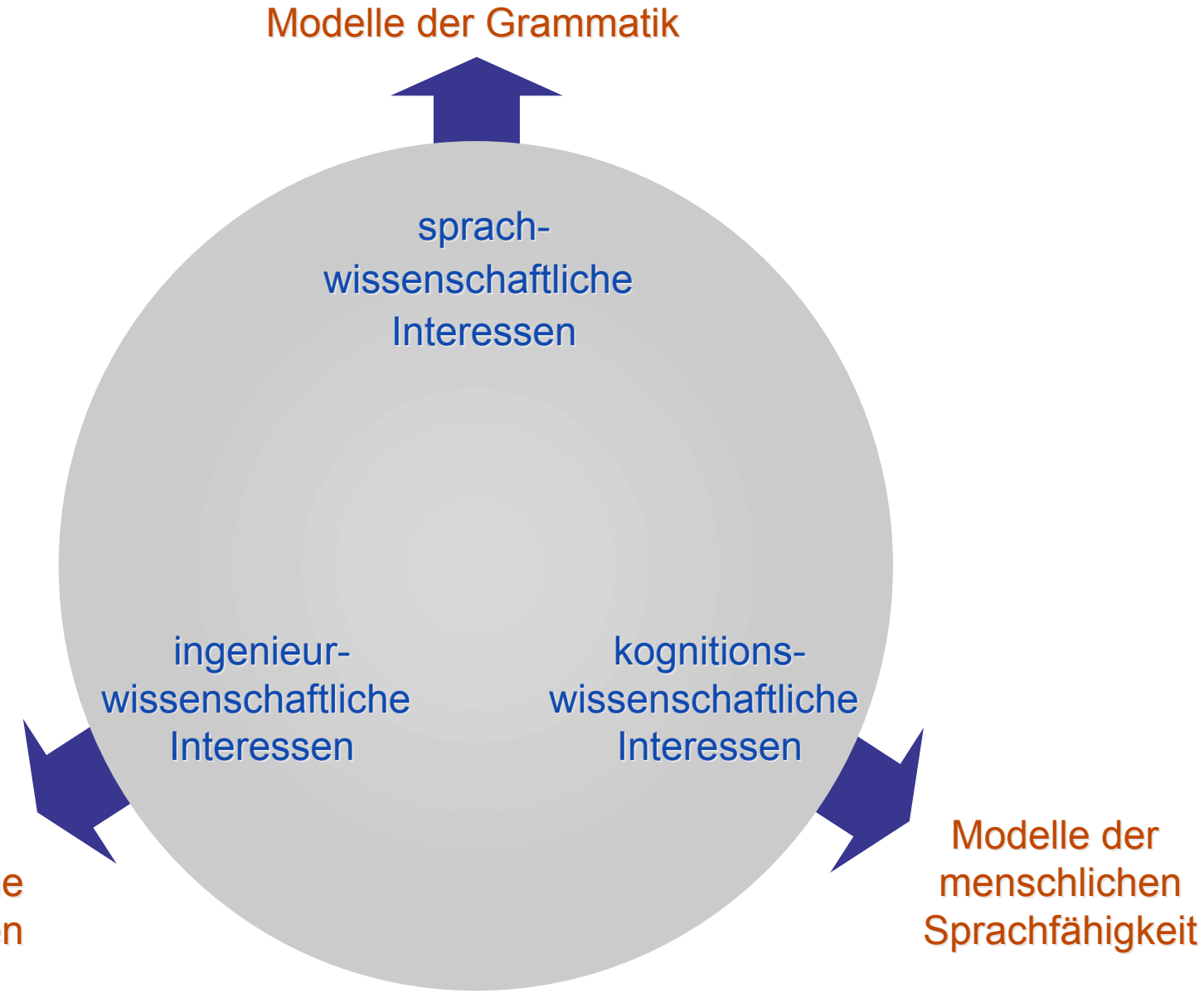
Motivationen



Motivationen



Motivationen



Linguistik

- Die Linguistik ist eine "moderne", synchron orientierte, auf die interne Struktur der Sprache bezogene Wissenschaft, die sprachliche Regularitäten auf allen Beschreibungsebenen untersucht und ihre Ergebnisse in explizierter (formalisierter) Beschreibungssprache und in integrierten Modellen darlegt.

- (*H. Bußmann "Lexikon der Sprachwissenschaft"*)

Teilgebiete der Linguistik

- Nach Beschreibungsebenen
  - Phonetik
  - Phonologie
  - Morphologie
  - Syntax
  - Semantik
  - Pragmatik/Text/Diskurs
- Andere Teildisziplinen
  - Psycholinguistik
  - Neurolinguistik
  - Historische Linguistik
  - Sozio- und Ethnolinguistik,
  - Dialektologie
  - Mathematische Linguistik

SPRACHLICHES WISSEN

Was sind die Inhalte und Strukturen dieses unbewußten Wissens?

SPRACHVERARBEITUNG

Wie produzieren und verstehen wir sprachliche Äußerungen?

SPRACHERWERB

Wie lernt das Kind seine Muttersprache?

SPRACHWANDEL

Wie entstehen Sprachen, Dialekte, Soziolekte?

Kompetenz und Performanz

- Sprachliche Kompetenz:
  - die endliche strukturierte Wissensbasis, die es den Sprechern einer Sprache ermöglicht, die wohlgeformten Äußerungen der Sprache zu generieren und zu interpretieren.
- Sprachliche Performanz:

  die Generierung oder Interpretation realer Äußerungen, bzw. die Gesamtheit der Prozesse, die beteiligt sind, wenn der Mensch auf der Basis der sprachlichen Kompetenz reale Äußerungen generiert und interpretiert.

Ein Kompetenzmodell sollte beinhalten:

Regeln, Prinzipien, Beschränkungen auf jeder Beschreibungsebene, die in ihrem Zusammenwirken genau die wohlgeformten Sätze der Sprache charakterisieren.

Es bietet für jede Sprache eine formalisierte endliche Definition einer unendlichen Menge von Paaren <Satz, Bedeutung>.

(Dazu gehören: Grammatik, Lexikon, morphologische Regeln, semantische Regeln.)

Performanzmodell

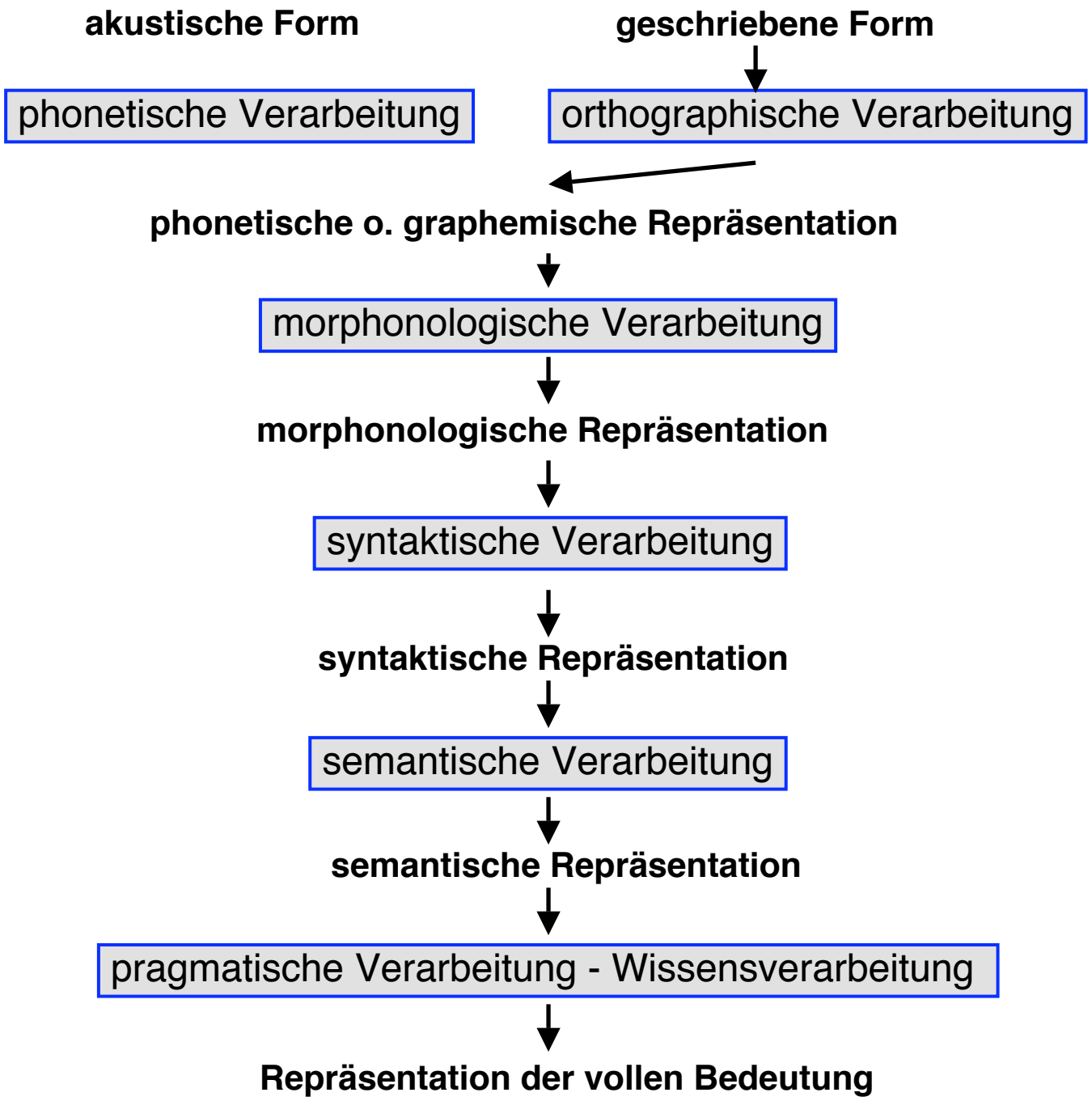
Ein Performanzmodell sollte erklären:

- warum viele ungrammatische Sätze erzeugt werden
  - ➥ z.B. Sprechfehler, Grammatikfehler
- warum viele ungrammatische Sätze verstanden werden
  - ➥ z.B. in der der Kommunikation mit Kindern oder Ausländern
- warum viele grammatische Sätze nicht erzeugt werden
  - ➥ z.B. durch Präferenzen in der Generierung
- warum viele grammatische Sätze nicht verstanden werden
  - ➥ z.B. Holzwegsätze
- wie die Verarbeitung zeitlich strukturiert ist
  - ➥ z.B. Effizienz, Abfolge der Verarbeitungsschritte
- welchen Aufwand die Verarbeitungsschritte erfordern
  - ➥ z.B. Abhängigkeiten von anderen kognitiven Belastungen

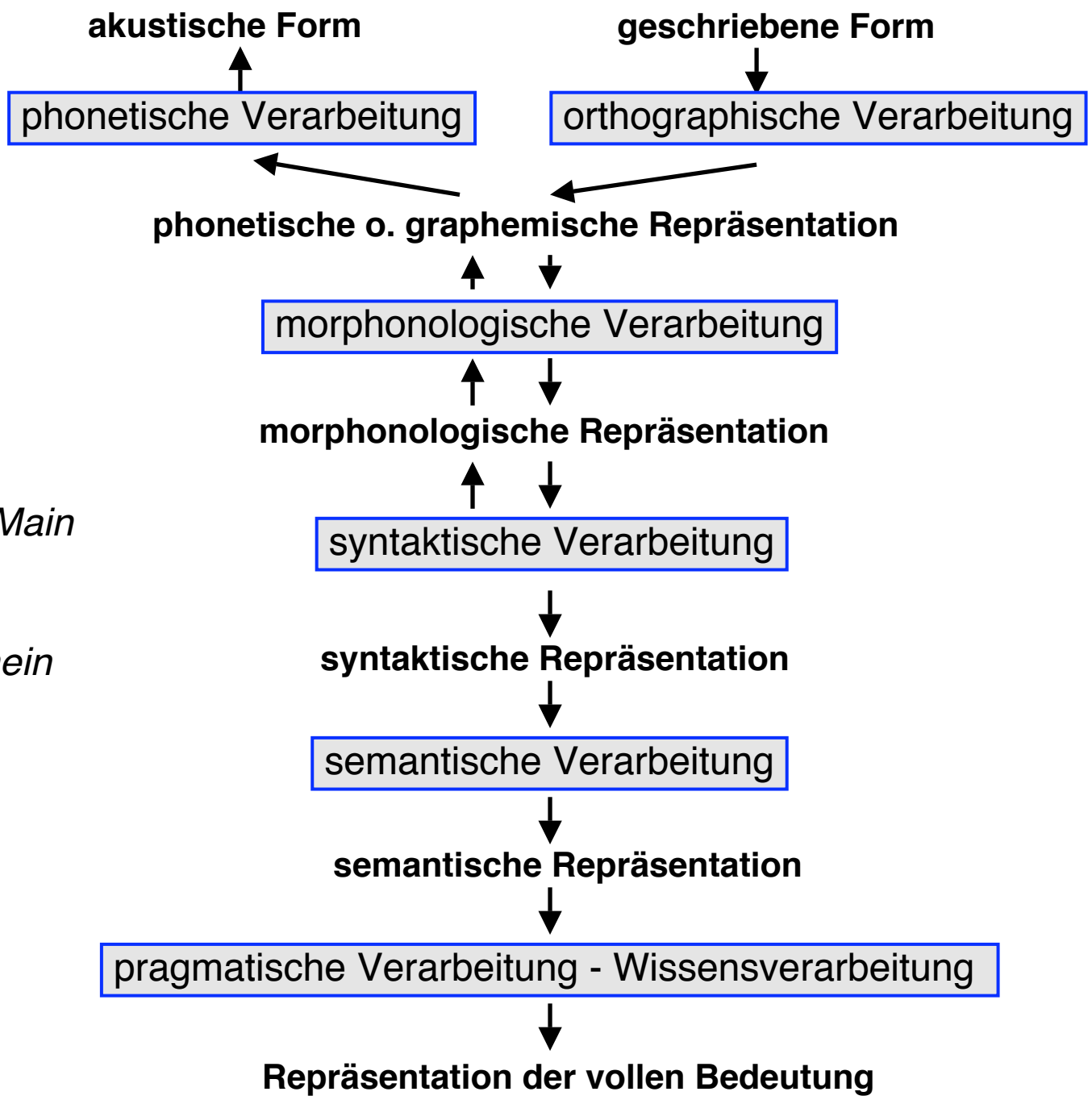
Performanzkriterien

efficiency	Fähigkeit, Lösungen mit geringem Zeit- und Speicherbedarf zu liefern
accuracy	Fähigkeit, linguistisch korrekte Lösungen zu liefern
robustness	Fähigkeit, mit allen möglichen Eingaben fertigzuwerden
coverage	größtmögliche Abdeckung der Grammatik
specificity	Fähigkeit, die intendierte Analyse zu selegieren

Textverstehen



Text-to-Speech

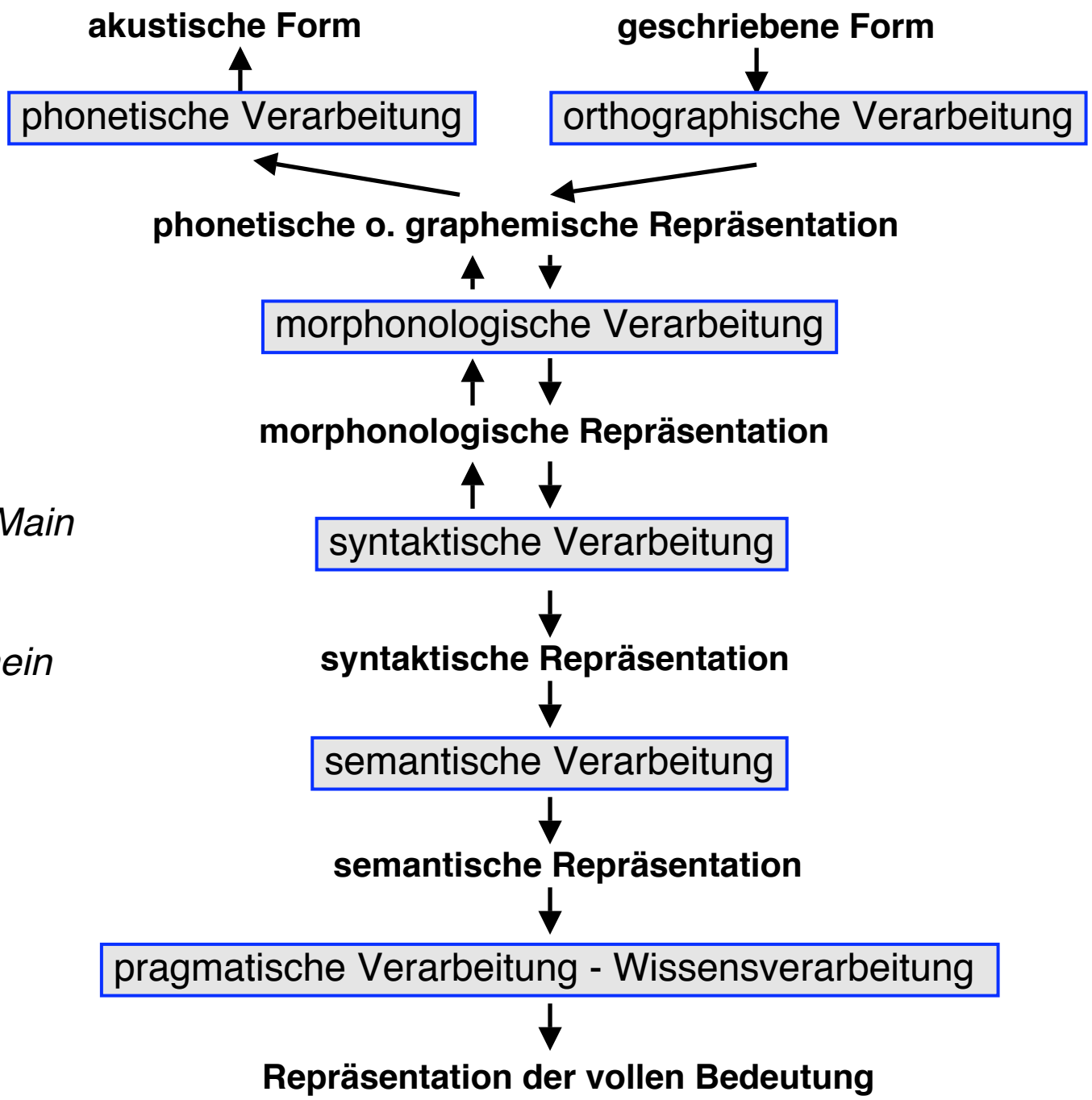


das Boot auf dem Main

oder

daß bot auf dem mein

Text-to-Speech

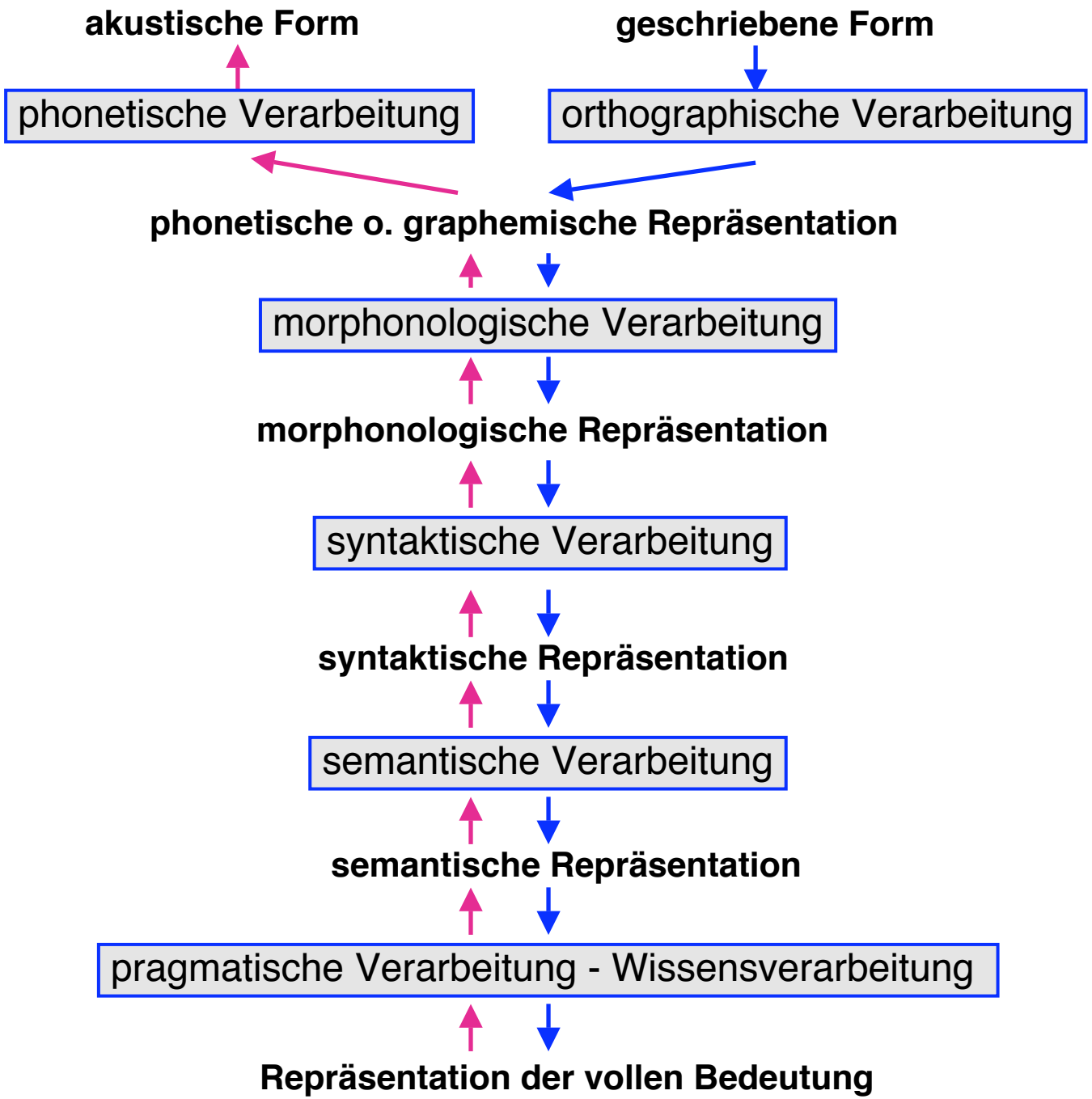


das Boot auf dem Main

oder

daß bot auf dem mein

Maschinelle Übersetzung



Beispiele

- Grammatikfehler und Sprechfehler:
  - Das Verfassen der Kinderbücher und der Reiseberichte haben dem Autor viel Ruhm eingebracht.
  - Die Poxen zum Backen...
- Holzwegsätze:
  - The canoe floated down the river sank.
  - Er bezichtigte den Vater des Schreibens unkundiger Kinder.
  - Peter beschuldigte sie der Geheimniskrämerei ähnlichen Verhaltens.

Ambiguität

phonetische Ambiguität (Homophone)

Miene - Mine

orthographische Ambiguität (Homographen)

übersetzen - übersetzen

lexikalische Ambiguität (Homonyme)

Ball - Ball

morphologische Ambiguität

Staubecken - Staubecken

Hauptpostsekretär

syntaktische Ambiguität

Peter fuhr seinen Freund sturzbetrunken nach Hause.

Visiting relatives can be boring.

Ich traf den Sohn des Nachbarn mit dem Gewehr.

kompositionell-semantische Ambiguität

Die zwei Mitarbeiter müssen vier Sprachen beherrschen.

pragmatische Ambiguität

Könnten Sie die Aufgabe lösen.

Ambiguität I

phonetische Ambiguität (Homophone)

Miene - Mine

orthographische Ambiguität (Homographen)

übersetzen - übersetzen

lexikalische Ambiguität (Homonyme)

Ball - Ball

morphologische Ambiguität

Staubecken - Staubecken

Hauptpostsekretär

Certain Readings are less preferred

*Auf dem Tisch lag ein **Heft**.*	*Auf der Werkbank lag ein **Heft**.*
*Ich habe einen **Stift** gefunden.*	*Ich habe einen jungen **Stift** gesucht.*

The preference can be influenced by concept.

*The goal keeper opened the **Ball**.*	**vs.**	*The President opened the ball*
*The astronomer married a **star**.*	**vs.**	*The movie director married a **star**.*

Ambiguität II

- **syntaktische Ambiguität**
- Peter fuhr seinen Freund sturzbetrunken nach Hause.
- Visiting relatives can be boring.
- Ich traf den Sohn des Nachbarn mit dem Gewehr.

- **kompositionell-semantische Ambiguität**
- Die zwei Mitarbeiter müssen vier Sprachen beherrschen.

- **pragmatische Ambiguität**
- Könnten Sie die Aufgabe lösen.

Ambiguität beim Parsing

In fast allen realen Situationen sind Sätze hochgradig ambig.

Beispiel:

Grammatik:	deutsche LFG-Grammatik von Christian Rohrer
Parser:	XLE Parser von XEROX PARC (Kaplan, Maxwell, Shemtov,...)
Korpus:	Teilmenge des NEGRA Korpus Frankfurter Rundschau (Saarbrücken)
∅Satzlänge:	ca. 16 Wörter

∅Ambiguität: >3000 Lesarten pro Satz
(durch heuristische Präferenzen reduziert auf 7 Lesarten)

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Der Satz weist lexikalische (L), syntaktische (S) und anaphorische (A) Ambiguitäten auf, die uns nicht auffallen.

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Der Satz weist lexikalische (L), syntaktische (S) und anaphorische (A) Ambiguitäten auf, die uns nicht auffallen.

Wieviele Lesarten besitzt dieser Satz?

258.048

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Das berechnet sich so:

L ***Früher*** **kann sowohl eigenständiges Adverb als auch Komparativ von** ***früh*** **sein (2);**

L **die Verbform** ***stellten*** **is ambig zwischen Präteritum und Konjunktiv (2);**

S **die Nominalphrase** ***die Frauen*** **kann sowohl Subjekt als auch Objekt des Satzes sein (2);**

S ***am Wochenende*** **kann die Insel, die Frauen oder das Verb modifizieren (3);**

S ***mit Blumenmotiven*** **kann sich auf die Kopftücher beziehen, ein Instrument der Herstellung**
sein oder ein Adjunkt im Sinne von ***gemeinsam mit Blumenmotiven*** **(3);**

L ***her*** **hat auch eine direktionale Bedeutung (2);**

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Und weiter:

- **S** der Relativsatz könnte jede der vier Nominalphrasen im Plural modifizieren (4);
- **S** sowohl *die* als auch *ihre Männer* kann Subjekt des Relativsatzes sein (2);
- **A** das Possessivpronomen *ihr*e kann auf jede der Nominalphrasen referieren (4);
- **L** *Montagen* hat eine zweite Lesart als Nominalisierung von *montieren* (2);
- **S** *die Hauptinsel* kann im Genitiv zu der vorangegangenen NP gehören oder im Dativ die Käuferin bezeichnen (2);
- **S** die drei Präpositionalphrasen des Relativsatzes können sich in insgesamt sieben Kombinationen mit den jeweils vorhergehenden NPs oder mit dem Verb verbinden (7);
- **L** *verkauften* zeigt wieder die Ambiguität zwischen Präteritum und Konjunktiv auf (2).

Ambiguität (Beispiel)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“

Durch Multiplikation ergibt sich die Gesamtambiguität:

2 x 2 x 2 x 3 x 3 x 2 x 4 x 2 x 4 x 2 x 2 x 7 x 2 = 258.048

Wir wollen wissen...

- warum viele ungrammatische Sätze erzeugt werden
  - z.B. Sprechfehler, Grammatikfehler
- warum viele ungrammatische Sätze verstanden werden
  - z.B. in der der Kommunikation mit Kindern oder Ausländern
- warum viele grammatische Sätze nicht erzeugt werden
  - z.B. durch Präferenzen in der Generierung
- warum viele grammatische Sätze nicht verstanden werden
  - z.B. Holzwegsätze
- wie die Verarbeitung zeitlich strukturiert ist
  - z.B. Effizienz, Abfolge der Verarbeitungsschritte
- welchen Aufwand die Verarbeitungsschritte erfordern
  - z.B. Abhängigkeiten von anderen kognitiven Belastungen

Der Wissenschaftler schrieb zwei Bücher über den Ursprung der menschlichen Sprache, die in vielen Fernsehsendungen diskutiert wurden, ab.