

Wintersemester 98/99

Vorlesung: Einführung in die Computerlinguistik

Teil 1: Forschungsgegenstand und Nachbarwissenschaften

Hans Uszkoreit

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

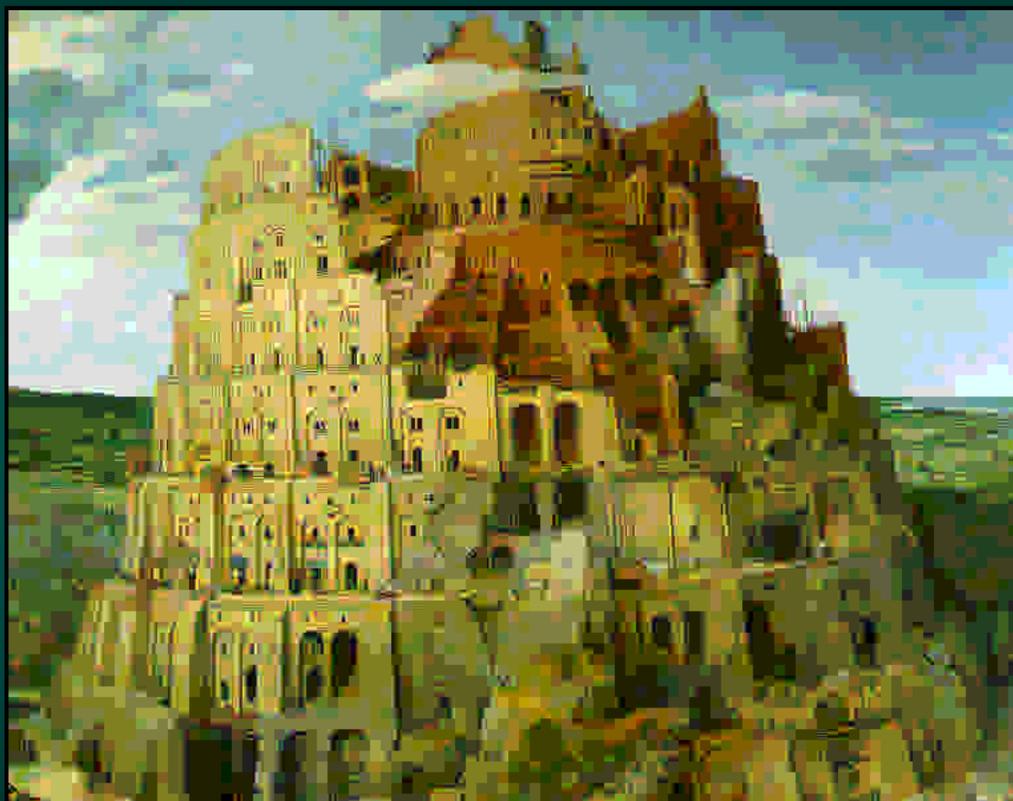
☐ **Faszination**

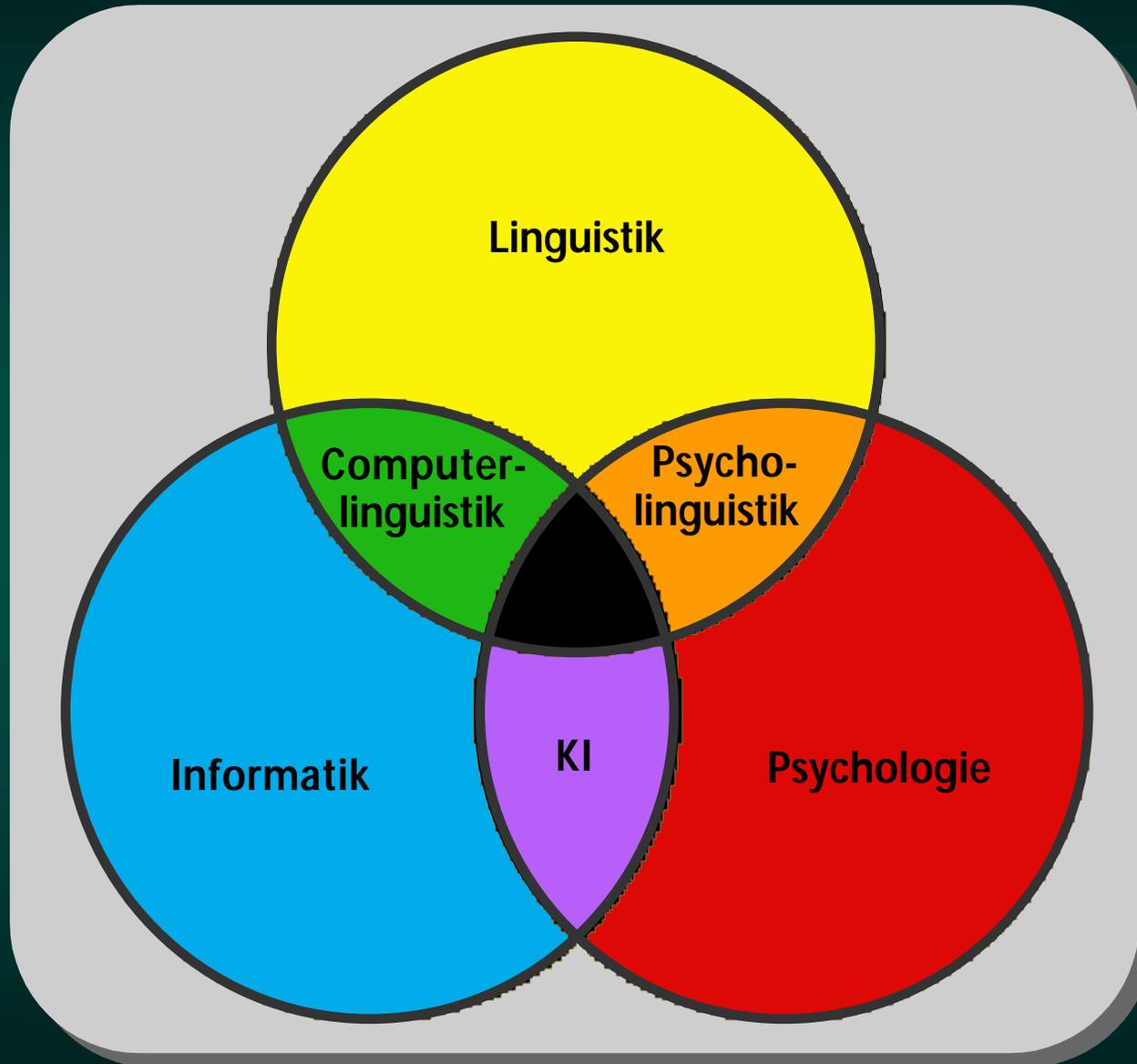
☐ **Wissenschaft**

☐ **Technologie**

- ❑ 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 Wörtern und eine komplexe Grammatik lernt.

WELT DER SPRACHE (Film)





- ❑ **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.

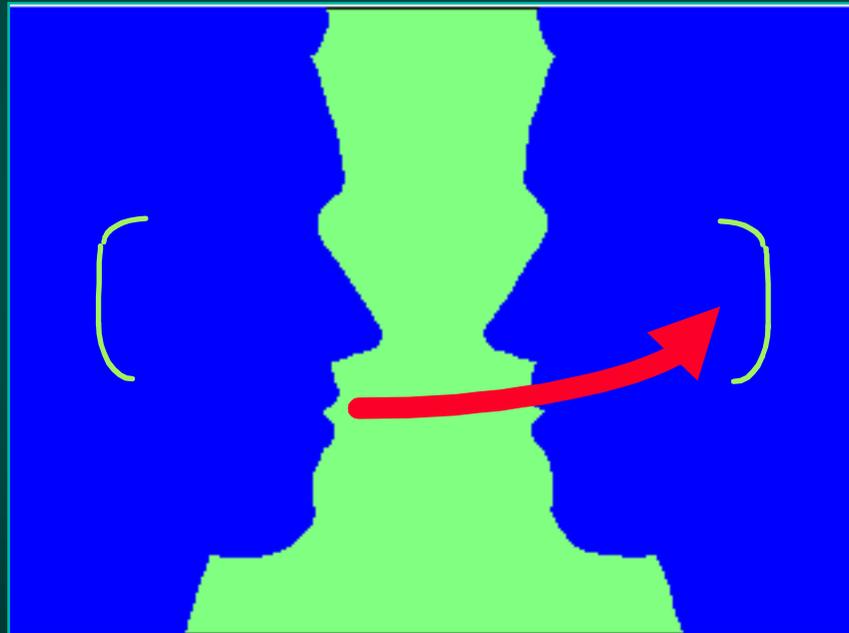
□ **Theoretische Computerlinguistik**

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

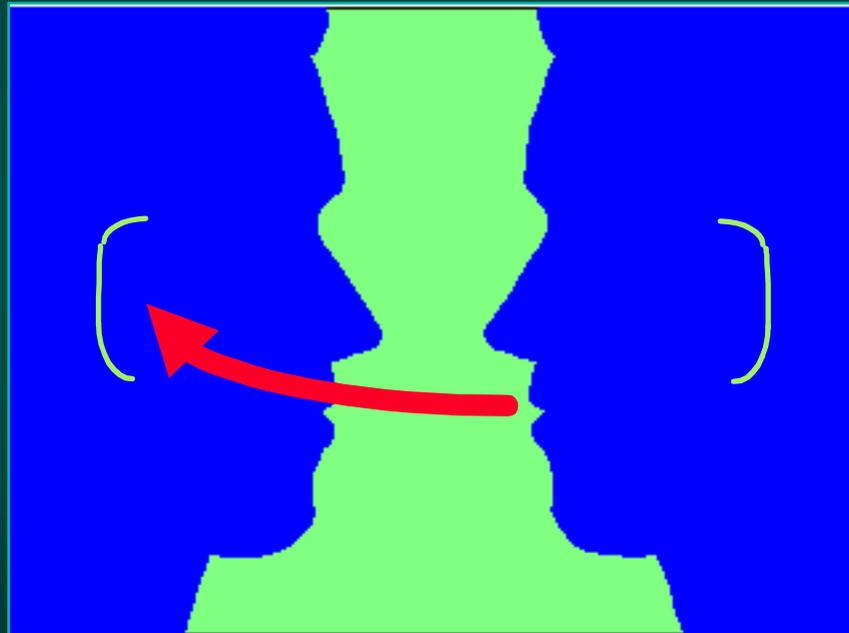
□ **Angewandte Computerlinguistik**

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

Hörer-Sprecher Modell



Hörer-Sprecher Modell

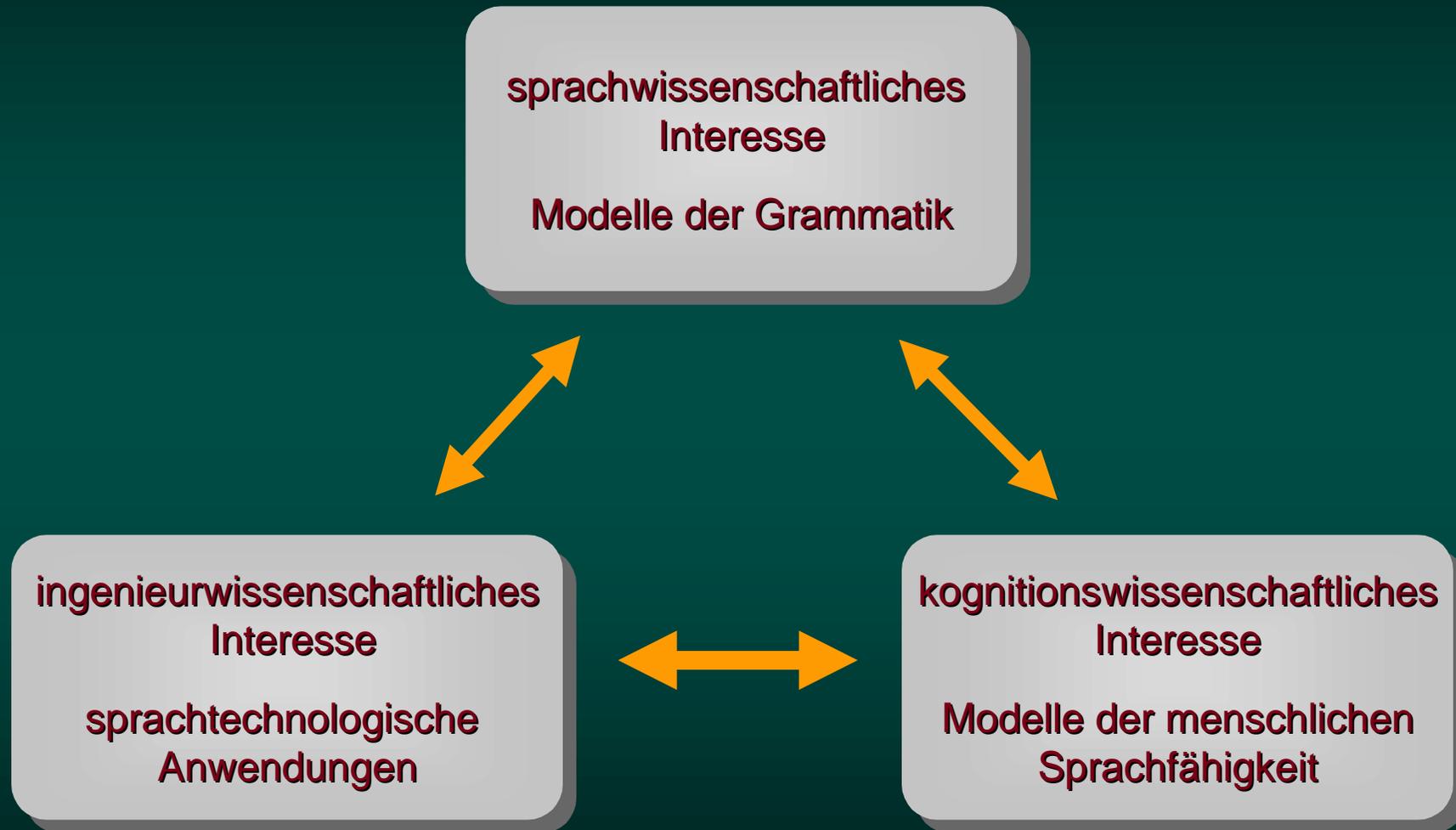


Grammatik



- ❑ **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



- ❑ 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"*)

☐ 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?

□ 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

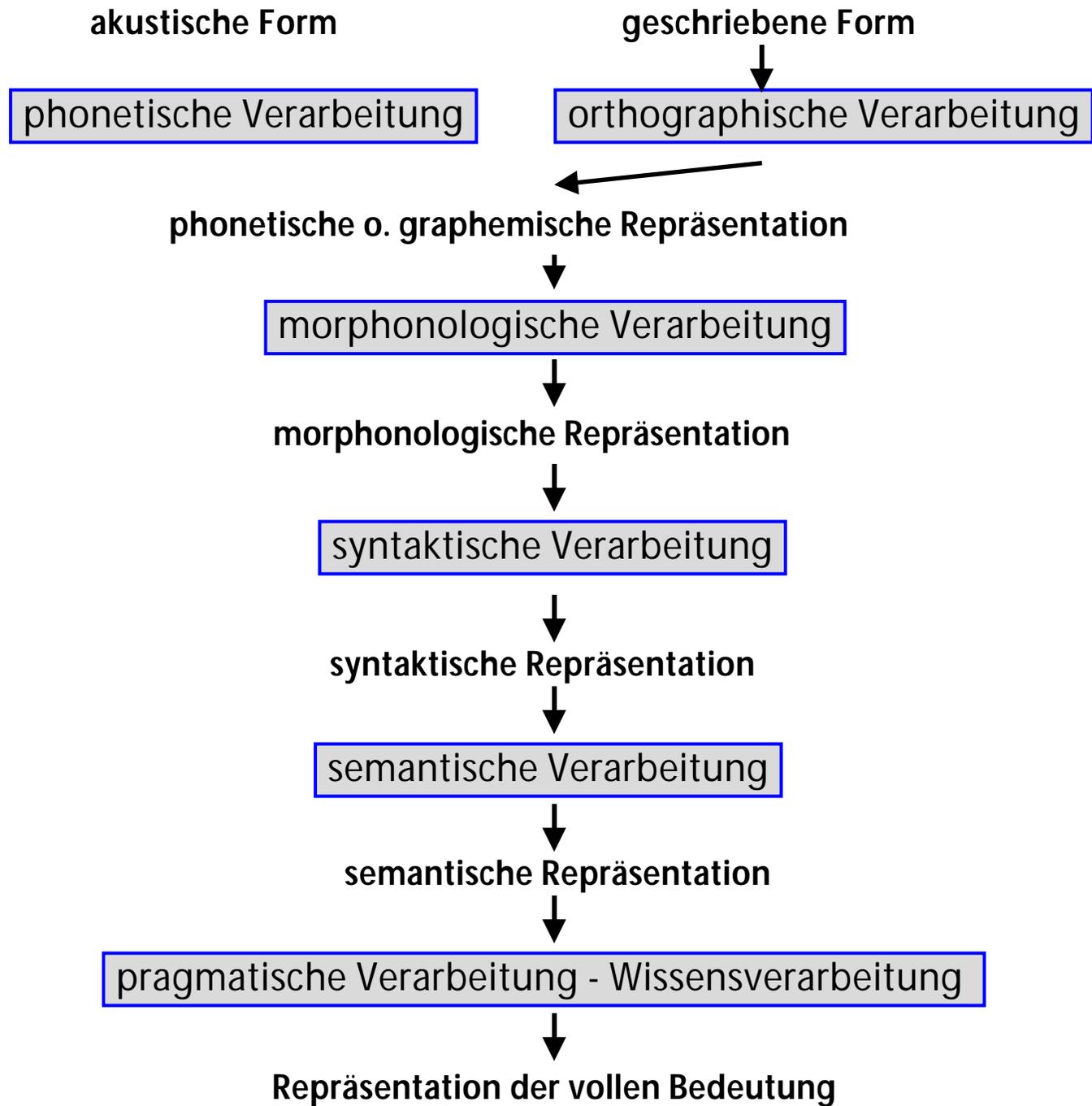
❑ **Grammatikfehler und Sprechfehler:**

Das Verfassen der Kinderbücher und der Reiseberichte haben dem Autor viel Ruhm eingebracht.
Die Poxen zum Backen...

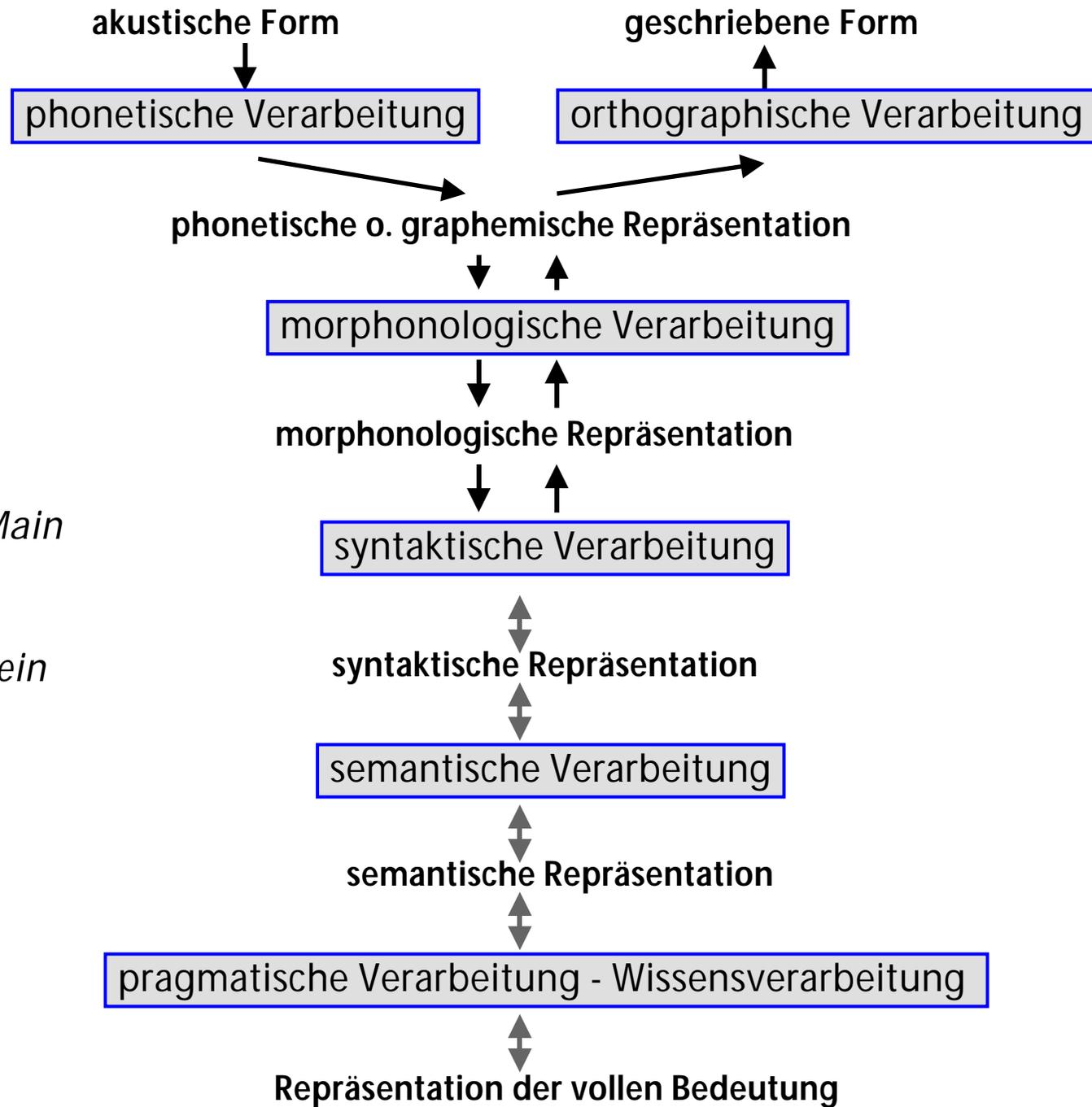
❑ **Holzwegsätze:**

The canoe floated down the river sank.
Er bezichtigte den Vater des Schreibens unkundiger Kinder.
Peter beschuldigte sie der Geheimniskrämerei ähnlichen Verhaltens.

Textverstehen

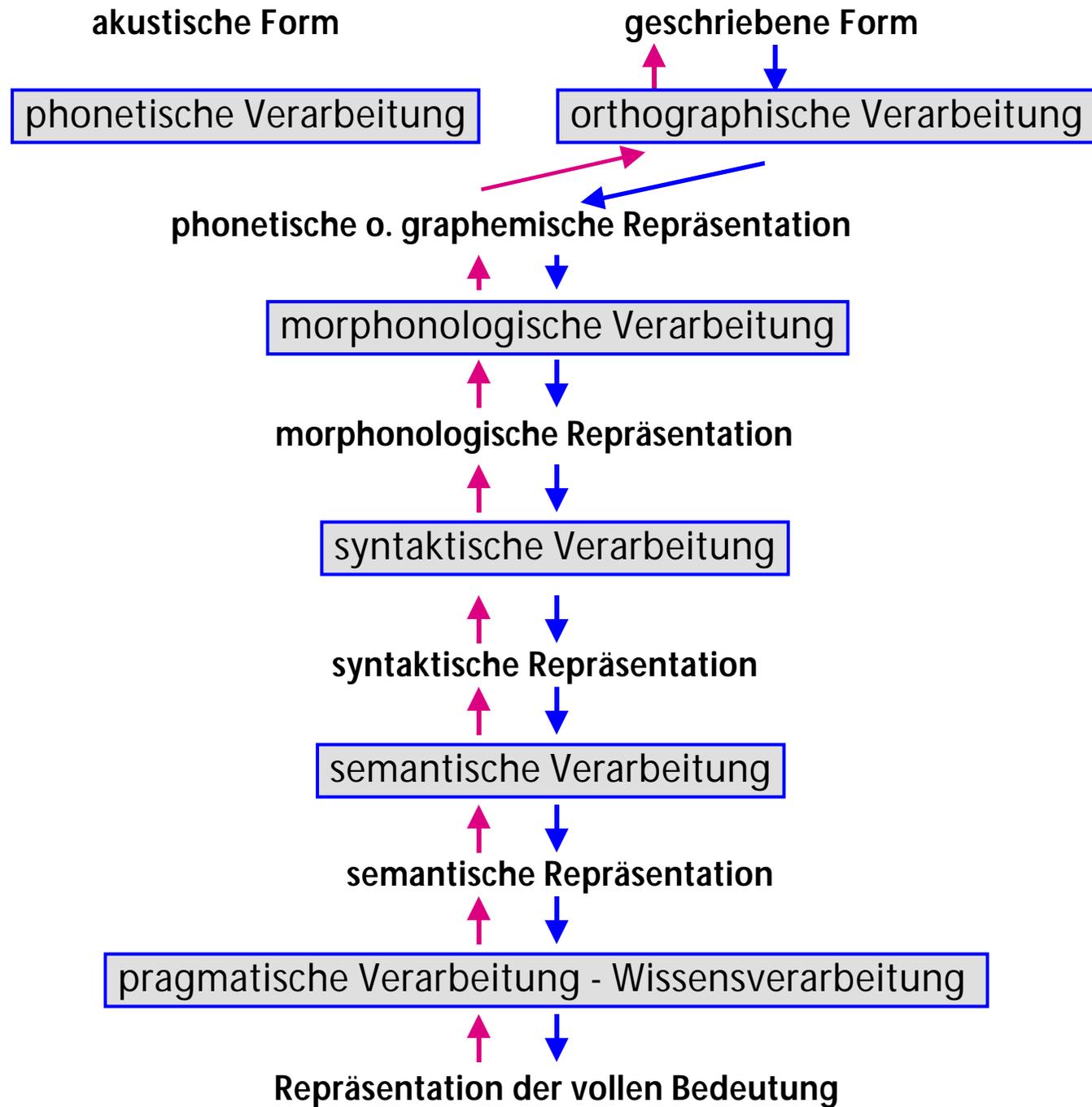


Diktat



das Boot auf dem Main
oder
daß bot auf dem mein

Maschinelle Übersetzung



Ambiguität

phonetische Ambiguität (Homophone)

Miene - Mine

orthographische Ambiguität (Homographen)

übersetzen - übersetzen

lexikalische Ambiguität (Homonyme)

Ball - Ball

morphologische Ambiguität

Staubecken - Staubecken

Hauptpostsekretär

Ambiguität

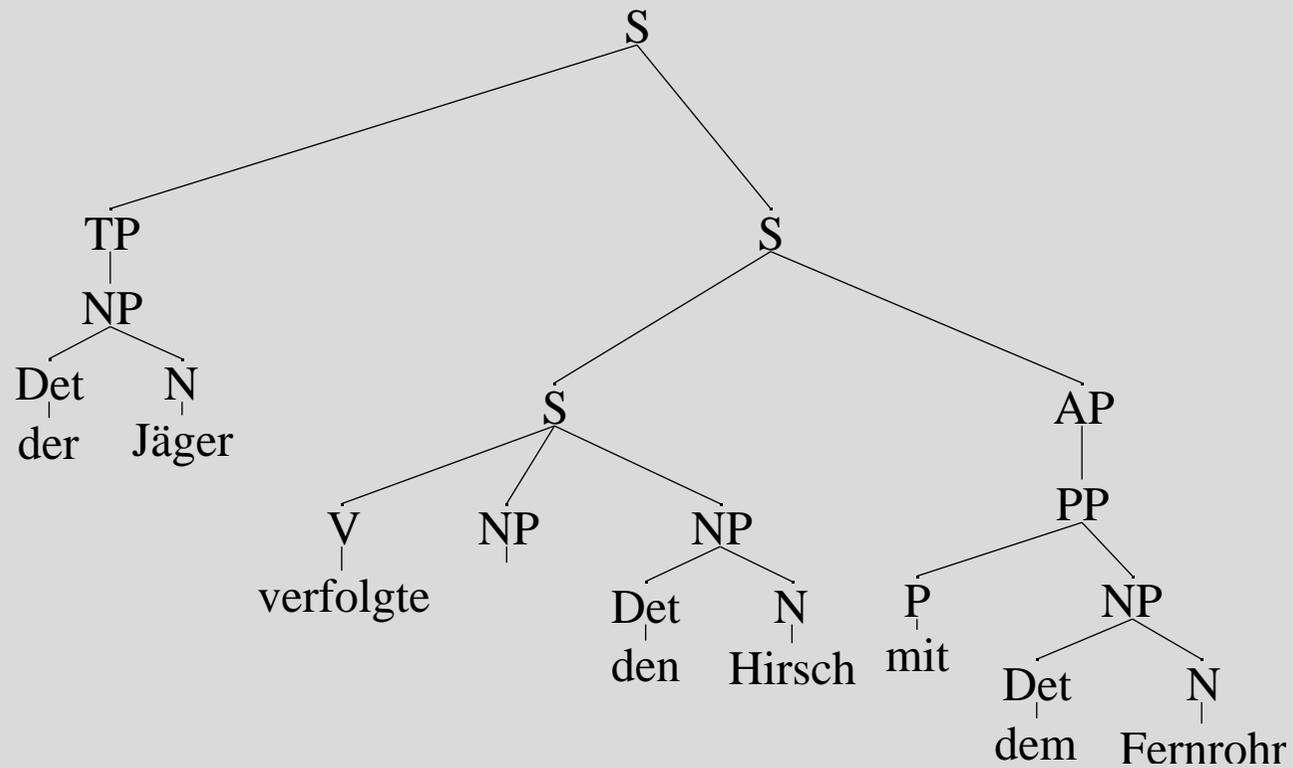
Der Jäger verfolgte den Hirsch mit dem Fernrohr.

Der Hirsch verfolgte den Jäger mit dem Fernrohr.

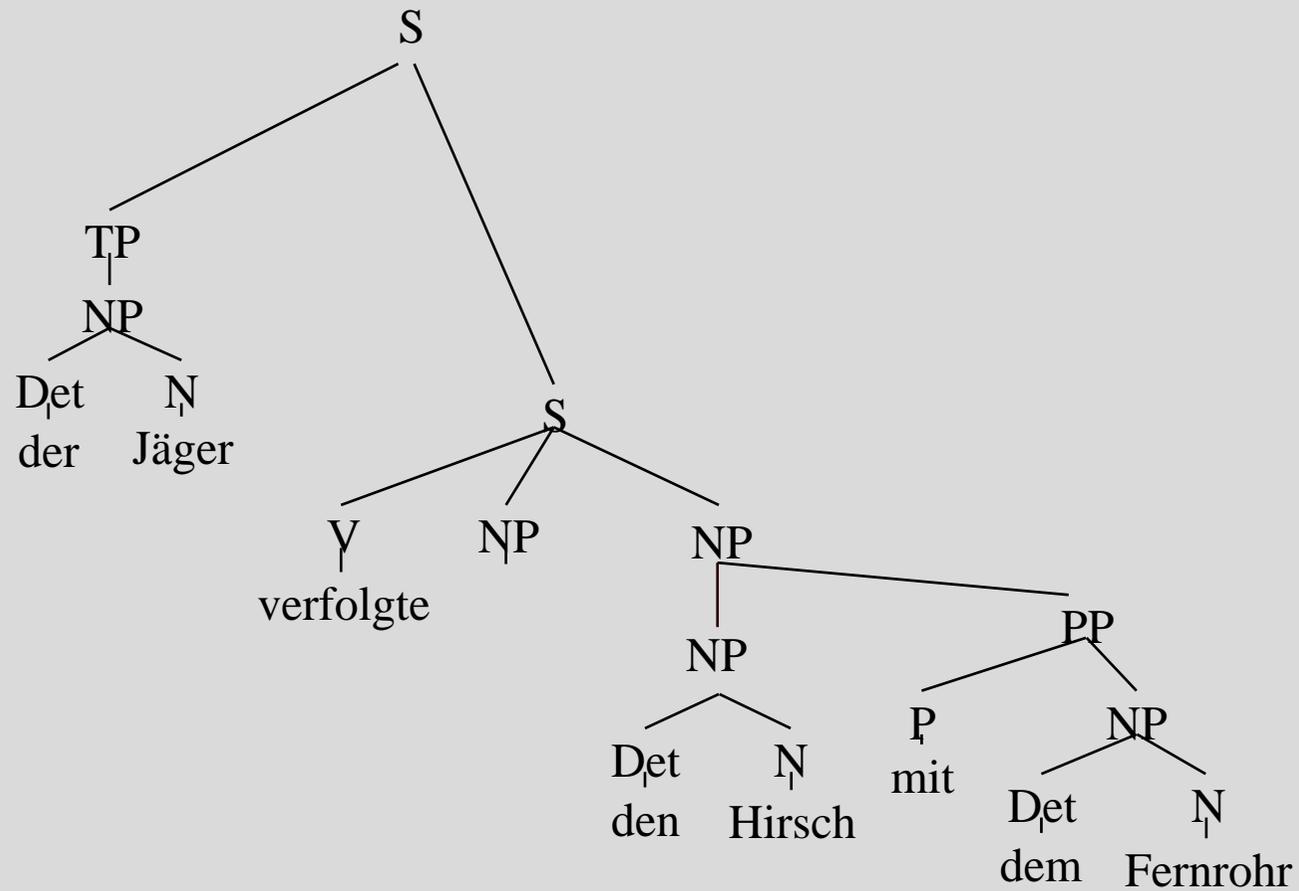
Der Frau verfolgte den Mann mit dem Fernrohr.

Die Frau verfolgte die Amöbe mit dem Mikroskop.

Ambiguität



Ambiguität



Ambiguität

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 (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.“

Wieviele Lesarten (lexikalische, syntaktische und anaphorische Ambiguitäten) besitzt dieser Satz?

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.“

Wieviele Lesarten (lexikalische, syntaktische und anaphorische Ambiguitäten) besitzt dieser Satz?

Tatsächlich sind es:

258.048!

Ohne Anaphernambiguität:

64.512!

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:

- ⇒ *Früher* kann sowohl eigenständiges Adverb als auch Komparativ von *früh* sein (2);
- ⇒ die Verbform *stellten* ist ambig zwischen Präteritum und Konjunktiv (2);
- ⇒ die Nominalphrase *die Frauen* kann sowohl Subjekt als auch Objekt des Satzes sein (2);
- ⇒ *am Wochenende* kann die Insel, die Frauen oder das Verb modifizieren (3);
- ⇒ *mit Blumenmotiven* kann sich auf die Kopftücher beziehen, ein Instrument der Herstellung sein oder ein Adjunkt im Sinne von *gemeinsam mit Blumenmotiven* (3);
- ⇒ *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.“

Das berechnet sich so:

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

Das berechnet sich so:

$$2 \times 2 \times 2 \times 3 \times 3 \times 2 \times 4 \times 2 \times 4 \times 2 \times 2 \times 7 \times 2 = \underline{258.048}$$

Paraphrasen (1)

phonetisch:

Mótor

Motór

orthographisch:

Photographie

Fotografie

lexikalisch:

Orange

Apfelsine

syntaktisch:

Diesen Mann hatte sie gestern
auf der Straße getroffen.

Gestern hatte sie diesen
Mann auf der Straße getroffen.

Er wurde von allen erkannt.

Alle erkannten ihn.

Paraphrasen (2)

Semantisch:

Mindestens einer war nicht erschienen.

Mindestens einer fehlte.

Mindestens einer war nicht erschienen.

Nicht alle waren erschienen.

pragmatisch:

Könnten Sie diesen Text
übersetzen?

Übersetzen Sie diesen Text!