Proseminar Sprachgebrauch M03,3: Einführung in die Computerlinguistik (CL)

Leitung:
Paul Rössler (Univ. Wien),
Dieter Merkl (TU Wien),
Gudrun Kellner (Univ.+TU Wien)

WiSe 2008

Mensch vs. Maschine

Maschinen arbeiten unfreier, beschränkter, eindeutiger und strenger als Menschen. Fehler beispielsweise, schlampiges Sprechen, nachdenkliche Redeweise, Ungenauigkeit, Doppeldeutigkeit, Hintersinn, Ironie, ‚zwischen den Zeilen etwas mitteilen' – all das kommt in menschlicher Kommunikation ständig vor.

Für den/die ComputerlinguistIn ist das nicht der Normalfall, sondern eine besondere Schwierigkeit.

Zentrale Fragen der CL in Bezug auf die menschl. Sprache u. Kommunikation

- Was muss ein Mensch können, der eine Sprache ‚kann‘? Welche Teile dieser Fähigkeit können so formalisiert werden, dass ein Computer sie ausführt?
- Woran erkennen Menschen, dass die Schallsignale, die sie wahrnehmen, Sprachlaute und nicht Geräusche sind? Wie kann das ein Computer erkennen?
- Beeinflussen Hören u. Lesen in je anderer Weise das Verstehen von Sprache? Sind beide für maschinelle Verfahren gleich aufwendig?
- Welche einzelnen Aktivitäten werden ausgelöst, wenn Texte verstanden werden? Was müssen Maschinen leisten, die diese Aktivitäten nachvollziehen/übernehmen sollen?
- Welches Vorwissen der Gesprächspartner spielt für eine Mitteilung eine Rolle? Wie kann das im Computer dargestellt werden?

Zentrale Fragen der CL in Bezug auf die menschl. Sprache u. Kommunikation

- Was wird auf welche Weise unmittelbar mitgeteilt? Was wird indirekt, z.B. durch Schlüsse, die der Hörer zieht, mitgeteilt? Wie kann man diese alltäglichen Vorgangsweisen in Regeln fassen?
- Wie erkennen Menschen, um welche sprachlichen Handlungen es geht: z.B. Voraussage, Scheinfrage, Argumentation, Drohung, Befehl, Bitte? Welche Absichten verfolgen Sprecher mit diesen Handlungen? Woran können Computer das ablesen?
- Wie erkennen wir, dass eine sprachliche Äußerung z.B. ein Vorschlag ist? Was wissen wir durch diese sprachl. Handlung vom Menschen, der sie vollzogen hat? Welche Folgen hat es für Menschen, wenn diese sprachl. Handlung (z.B. etw. vorschlagen) ein Computer vollzieht?

aus: Gesellsch. Rür Linguist. Datenverarbeitung GLDV

Forschungsaufgaben der CL

- Allg. Grundlagen der Interaktion zw. Mensch u. Maschine erarbeiten (Interaktionstheorie, Kommunikationsforschung, Kognitionswissenschaft) – Modelle für Dialog- u. Argumentationsformen kreieren, die jene der menschl. Benutzer nachbilden
- Formalismen konzipieren, um Sachgebietswissen maschinengerecht darzustellen – Prozeduren bauen, mit denen dieses Wissen effizient genutzt u. weiterverarbeitet wird
- Verfahren u. Strategien entwerfen, die grammatische Strukturen analysieren (Parsing) und die umgekehrt Texte erzeugen können
- Erforschung e. möglichst übersichtl. u. effizienten Aufbaus (Architektur) natürlich-sprachlicher Systeme
- Entwerfen einzelner Typen solcher natürlich-sprachl. Systeme (z.B. Dialogsysteme, natürl.-sprachl. Zugänge zu Experten-systemen, masch. Übersetzungssysteme, Textanalyse-, Textverstehens-, Textgenerierungssysteme)
- Evaluation der Systeme – Schnelligkeit, Fehler, Nutzen, Schaden

aus: Hahn 1990

„Ziel und Höhepunkt computerlinguistischer Arbeit ist der Bau funktionsfähiger Programmsysteme, die natürliche Sprache verarbeiten können. Im besten Fall kann ein solches ‚natürlich-sprachliches System‘ von Menschen eingegebene Texte so verstehen und so damit umgehen, wie es der Benutzer erwartet, und Antworten erzeugen, die den Benutzer zufriedenstellen.“

(Ulrich Schmitz: Computerlinguistik 1992, S. 34)

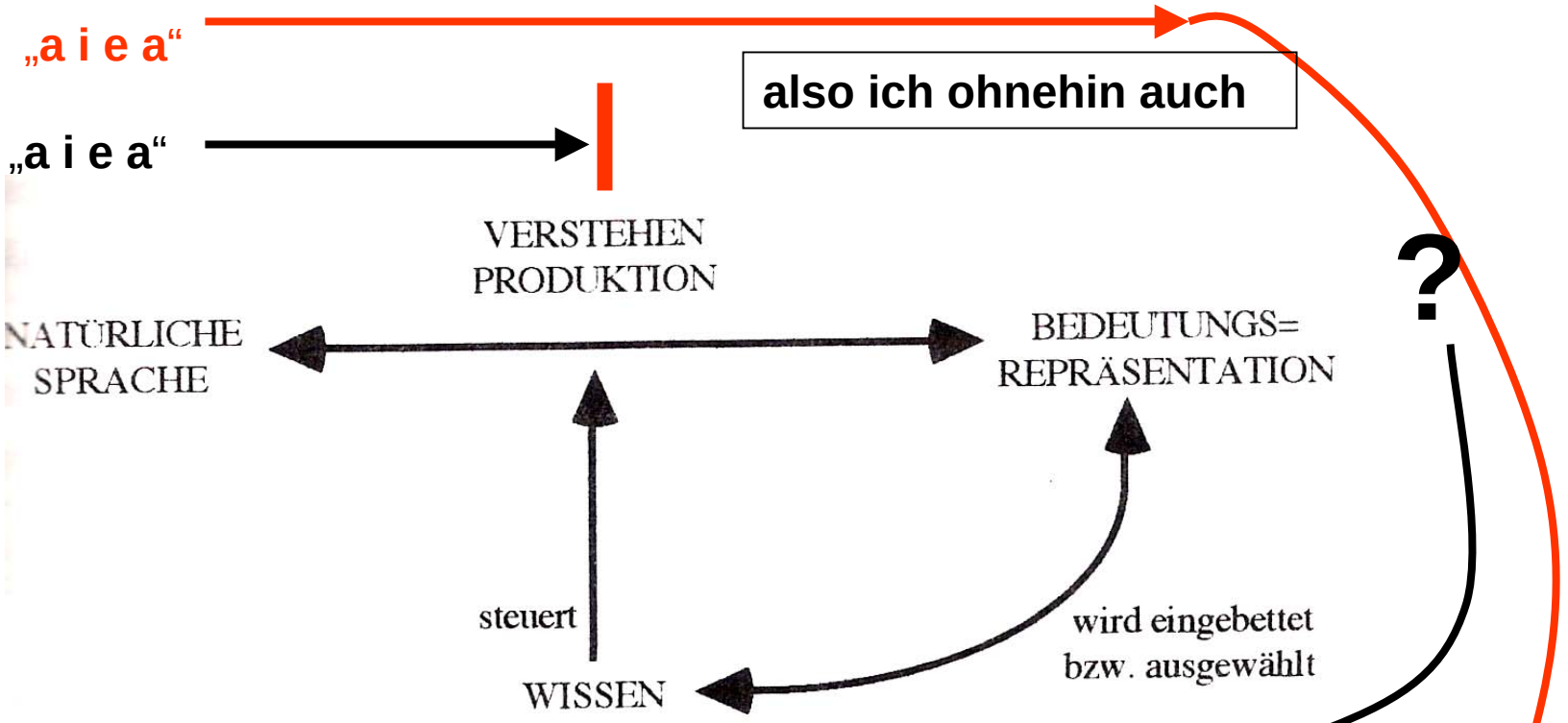
Linguistischer Ansatz vs. CL-Ansatz

- Linguisten wollen sprachl. Phänomene beschreiben u. erklären
- wissenschaftl. Theorie ist wichtig
- Vollständigkeit, Klarheit, Konsistenz sind wichtig
- ‚innersprachl.' (Grammatik) u. ‚außersprachl.' (z.B. Pragmatik) sind getrennt
- Ziel: Wissen über Sprache

- CLinguisten wollen menschl. Sprachfähigkeit techn. rekonstruieren
- auf Theorie kann teilw. verzichtet werden
- Funktionsfähigkeit ist wichtig
- Elemente aus Grammatik u. Pragmatik werden oft bewusst vermengt
- Ziel: Verhältnis von Sprache und Wissen

Grundriss eines natürlich-sprachlichen Systems I

(in folgenden Folien auf Basis von U. Schmitz (1992))



Standardsprecher

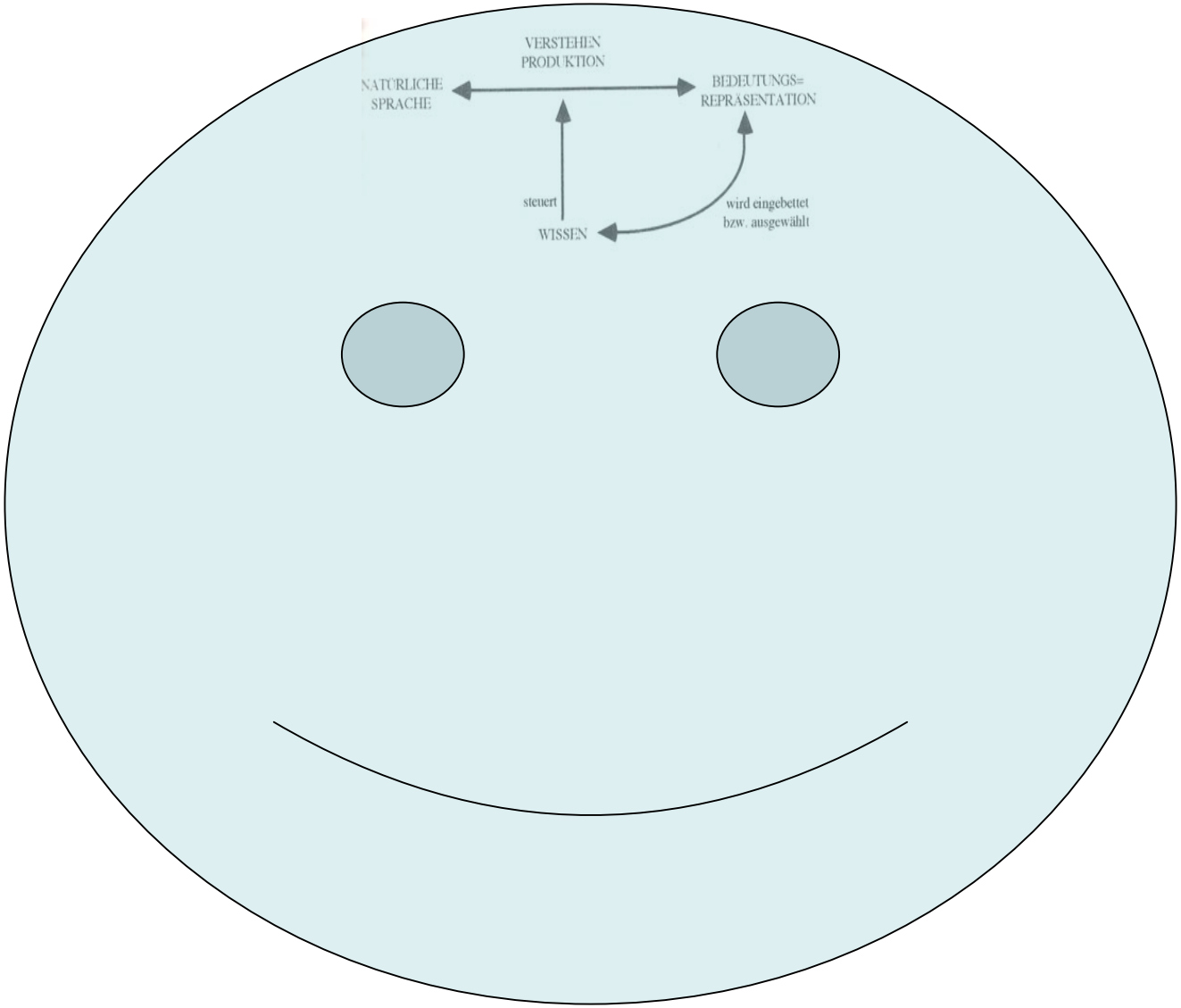
Dialektsprecher →

„a“ ... *auch/also*

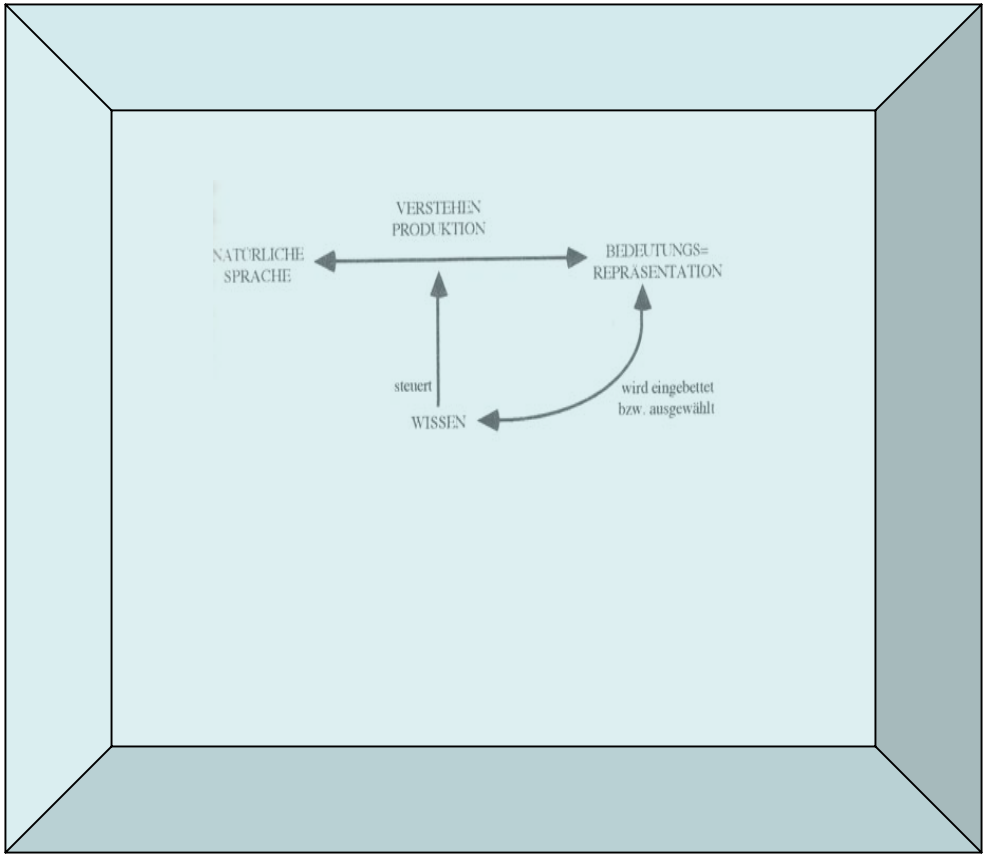
„i“ ... *ich*

„e“ ... ohnehin

Grundriss eines natürlich-sprachlichen Systems Ia

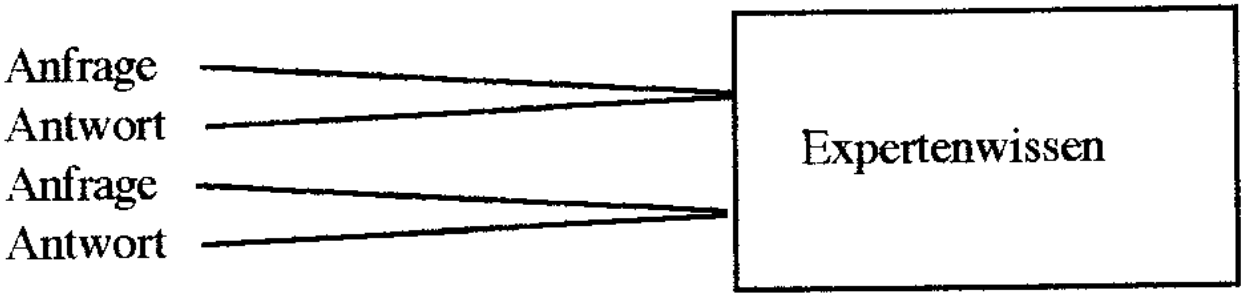


Grundriss eines natürlich-sprachlichen Systems Ib



Grundriss eines natürlich-sprachlichen Systems II

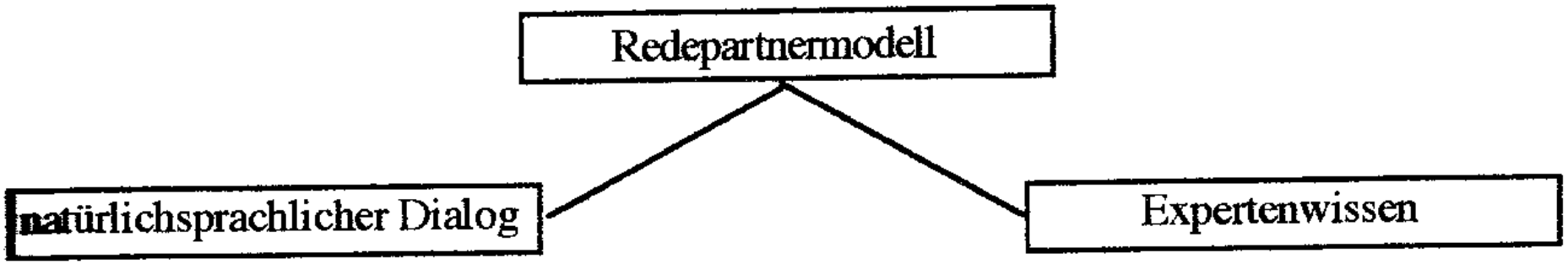
Natürlichsprachlicher Dialog:



Grundriss eines natürlich-sprachlichen Systems III: 3 Typen des Redepartnermodells

- Expertenwissen bleibt während des Dialogs unverändert
- Expertenwissen wird verändert: Maschine nimmt Wissen von Benutzer entgegen u. speichert es
- Expertenwissen + Gesprächsführung zw. Mensch u. Maschine werden verändert: Art u. Inhalt der Anfragen des Benutzers werden v. Maschine analysiert und für weitere Interaktion eingesetzt

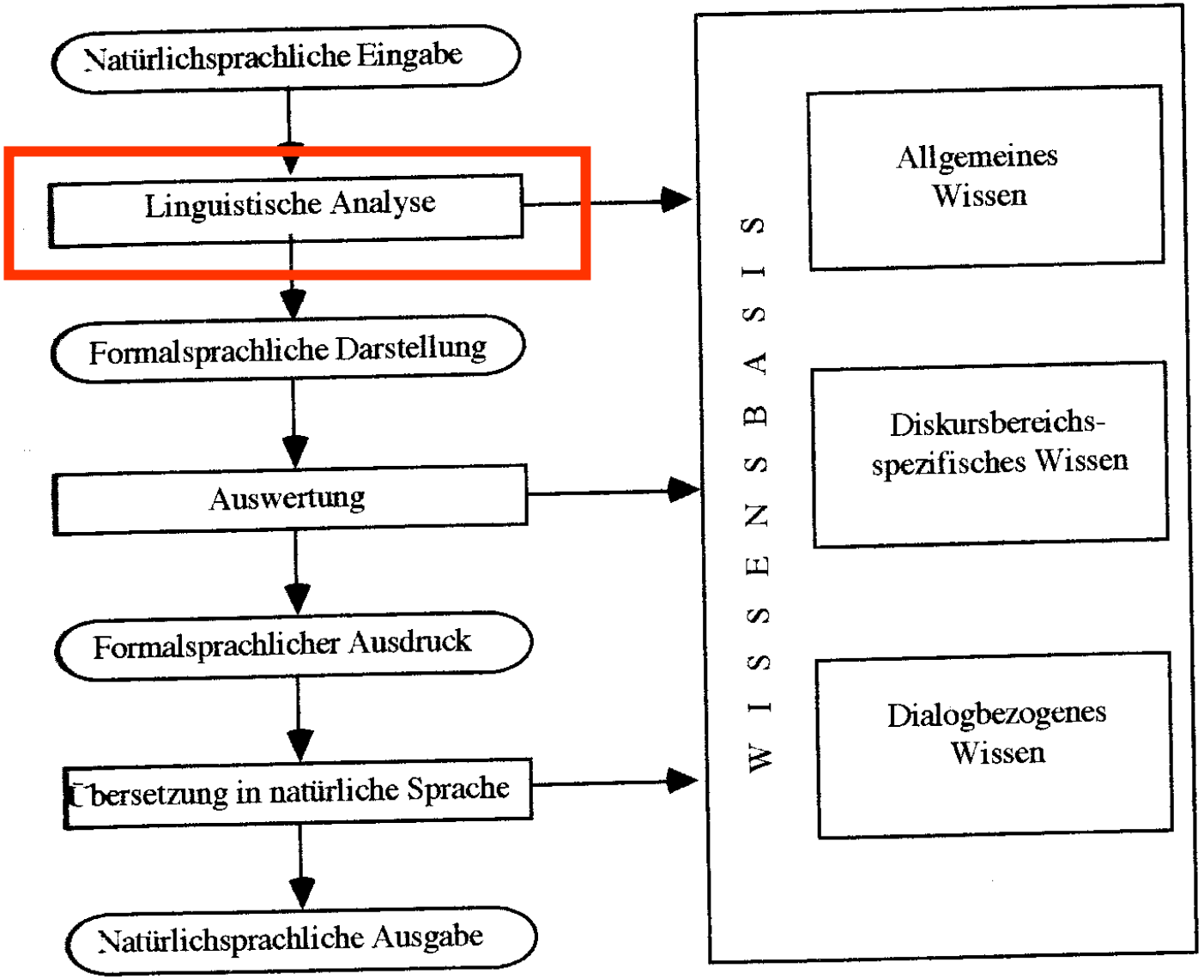
Grundriss eines natürlich-sprachlichen Systems IV



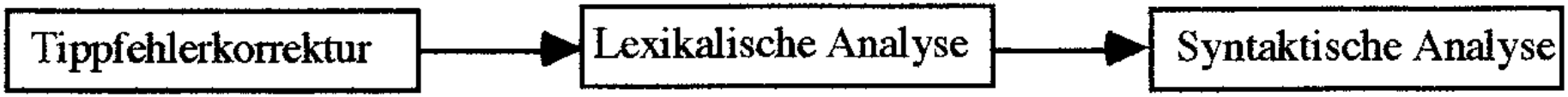
Grundriss eines natürlich-sprachlichen Systems V

Die Maschine greift zum Verständnis einer vom Benutzer eingegebenen Frage und zur Erzeugung einer adäquaten Antwort auf ein mehrdimensional strukturiertes, meist netzartig angeordnetes Universum von Wissen zurück, welches ihr in einer sehr kompakten, logisch-formalisierten Darstellungsform eingegeben wurde. Das Anfrage-Ergebnis muss dann erst noch in einen natürlich-sprachlichen Text (rück-) übersetzt werden

Grundriss eines natürlich-sprachlichen Systems VI



Grundriss eines natürlich-sprachlichen Systems VII
Anwendungsbeispiel: linguistische Analyse



Grundriss eines natürlich-sprachlichen Systems VIII

Eingabe per eintippen - Frage:

„Wie arbeitet eigentlich ein natürlich-sprachliches Dialogsystem?“

Grundriss eines natürlich-sprachlichen Systems IX
Lexikalische Analyse des Eingabesatzes

- **Tippfehlerkorrektur**
- **Vergleich jedes Wortes mit eingebautem Wortlexikon**:
- a) dt. Grundwortschatz
- b) wissensspezif. Spezialwortschatz (z.B. *„Dialogsystem“*)
- Analyse grammat. u. semant. Informationen jedes (z.B. *„Substantiv“, „Diskursbereich xy“)*
- Zuordnung flektierter Wortformen (z.B. „arbeitet“) zu Grundformen (>“arbeiten“) (=Lemmatisierung)
- Registrierung der grammat. Information (3. Pers. Sg. Präs. Ind.)
- Standardisierung von Sonderformen (z.B. kontrahierte Formen wie „im“ > expandiert zu „in dem“
- diskontinuierl. Konstituenten (z.B. trennbare Verben) > zusammengefasst
- Tilgung von für die Anwendung irrelevanten Füllwörtern (z.B. „eigentlich“)

Grundriss eines natürlich-sprachlichen Systems X
Syntaktische Analyse des Eingabesatzes

- „Wie“ , „?“ → Fragesatz
- Fragesatz → an 2. Stelle: Adjektiv / Adverb / Verb
- „arbeitet“ → hat best. Kasusrahmen / Valenzen: obligator. Aktant = Subjekt, fakultativ = Ergänzungen
- „eigentlich“ ... bereits in lexikal. Analyse getilgt
- „ein“ → leitet als Artikel eine NP ein: Art+N oder Art+Adj+N
- „natürlich-sprachliches Dialogsystem“ → NP + Subjekt

Grundriss eines natürlich-sprachlichen Systems XI
Auswertung der lexikal.+syntakt. Analyse

- zwei Ergebnisse aus lexikal. u. syntakt. Analyse für den Bezug zum Wissenssystem des Computers:
- a) eine Frage soll beantwortet werden
  - speziell: das Fragepronomen soll durch längeren Text ersetzt werden
- b) Suche nur nach begrifflichem Wissen, nicht nach situativem, referenziellem od. visuellem Wissen – nur diese Komponente muss in Datenbank gesucht werden (durch Initiator „ein“ ... unbest. Art.)

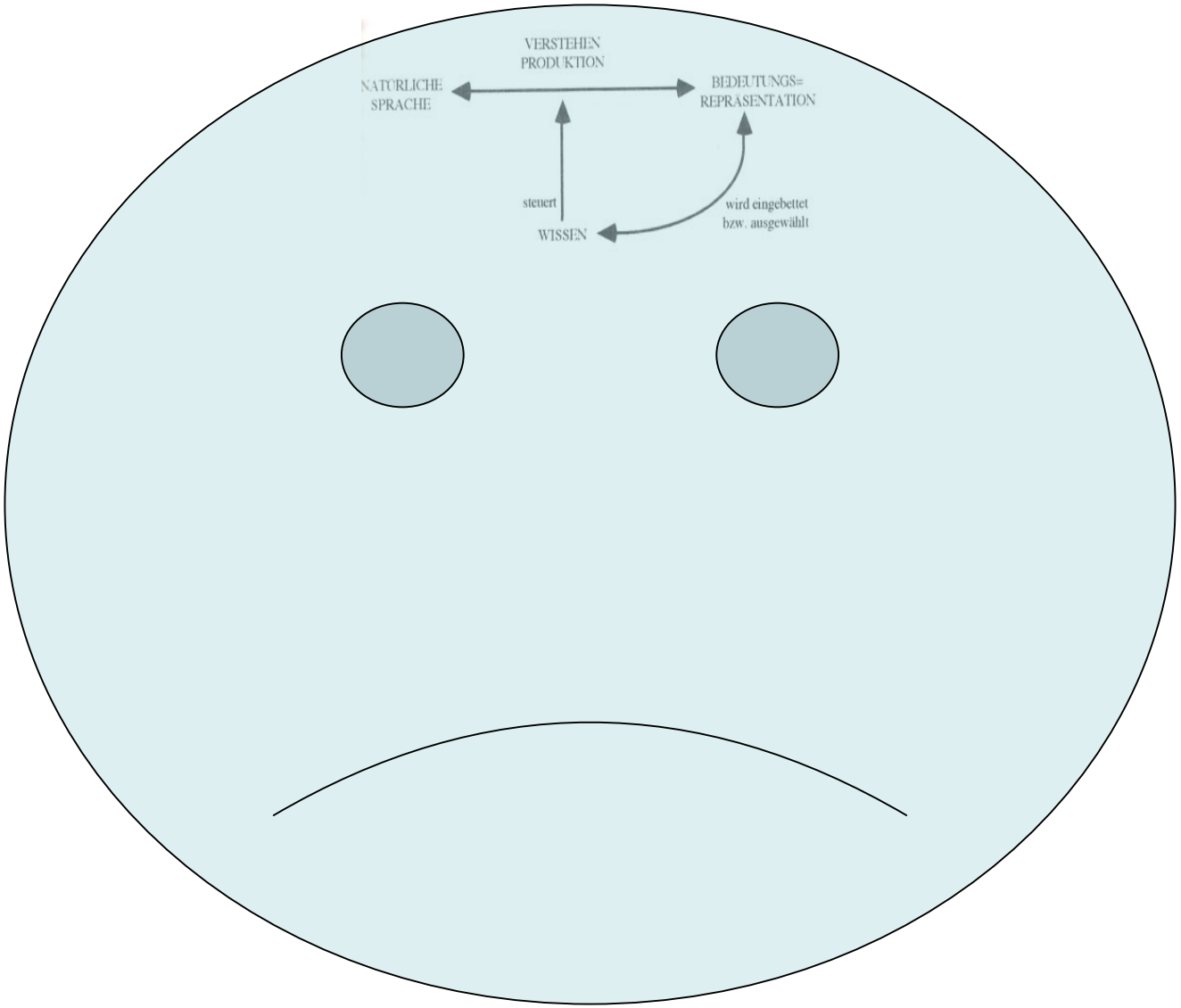
Grundriss eines natürlich-sprachlichen Systems XII
Auswertung der lexikal.+syntakt. Analyse

- Subjekt des Fragesatzes → Datenbanksuche
- Datenbanksuche-Möglichkeiten:
  - a) direkter Eintrag „natürlich-sprachliches Dialogsystem“ in Datenbank enthalten
  - b) Synonymierungspfade durchlaufen, z.B. „NSD“
  - c) an „NSD“-Knoten hängen Informationen differenten Typs:
    - Ober/Unterbegriffsrelationen,
    - Teil-von-Relationen
    - Eigenschaften etc.

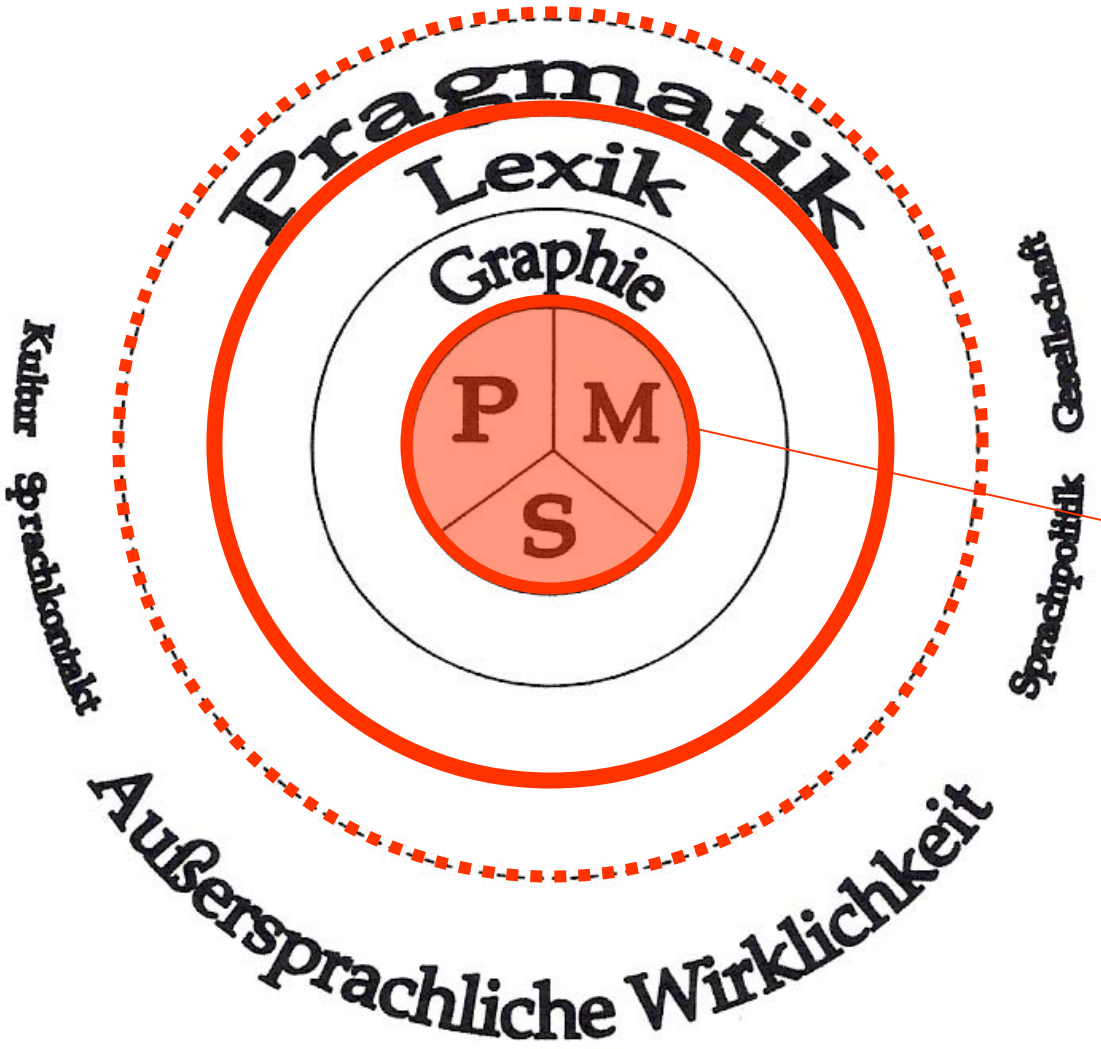
Grundriss eines natürlich-sprachlichen Systems XIII Auswertung>formalsprachl. Ausdruck>Übersetzung in natürliche Sprache

- Überführung von Gegenstandbezeichner in Datenbank (z.B. „NSD“) in eine Nominalphrase (z.B. „natürlich-sprachliche Dialogsysteme“)
- Ersetzen formalsprachl. Operatoren (z.B. BUT, NOT, IS-A) durch passende Lexikonwörter
- Erstellen semant.-syntakt. Grundgerüste (~ Kette von Satzbestandteilen)
- syntakt. u. morpholog. Bearbeitung dieses Grundgerüsts
- Überarbeitung (z.B. wiederkehrende NPs werden pronominalisiert, semant. Kohärenzen zw. mehreren Sätzen werden kohäsiert, d.h. sprachl. sichtbar gemacht (z.B. Satz 1: „NSD“ > Satz 2: „Solche Systeme“)

ENDE


VERSTEHEN
PRODUKTION
NATÜRLICHE
SPRACHE
BEDEUTUNGS=
REPRÄSENTATION
steuert
wird eingebettet
bzw. ausgewählt
WISSEN

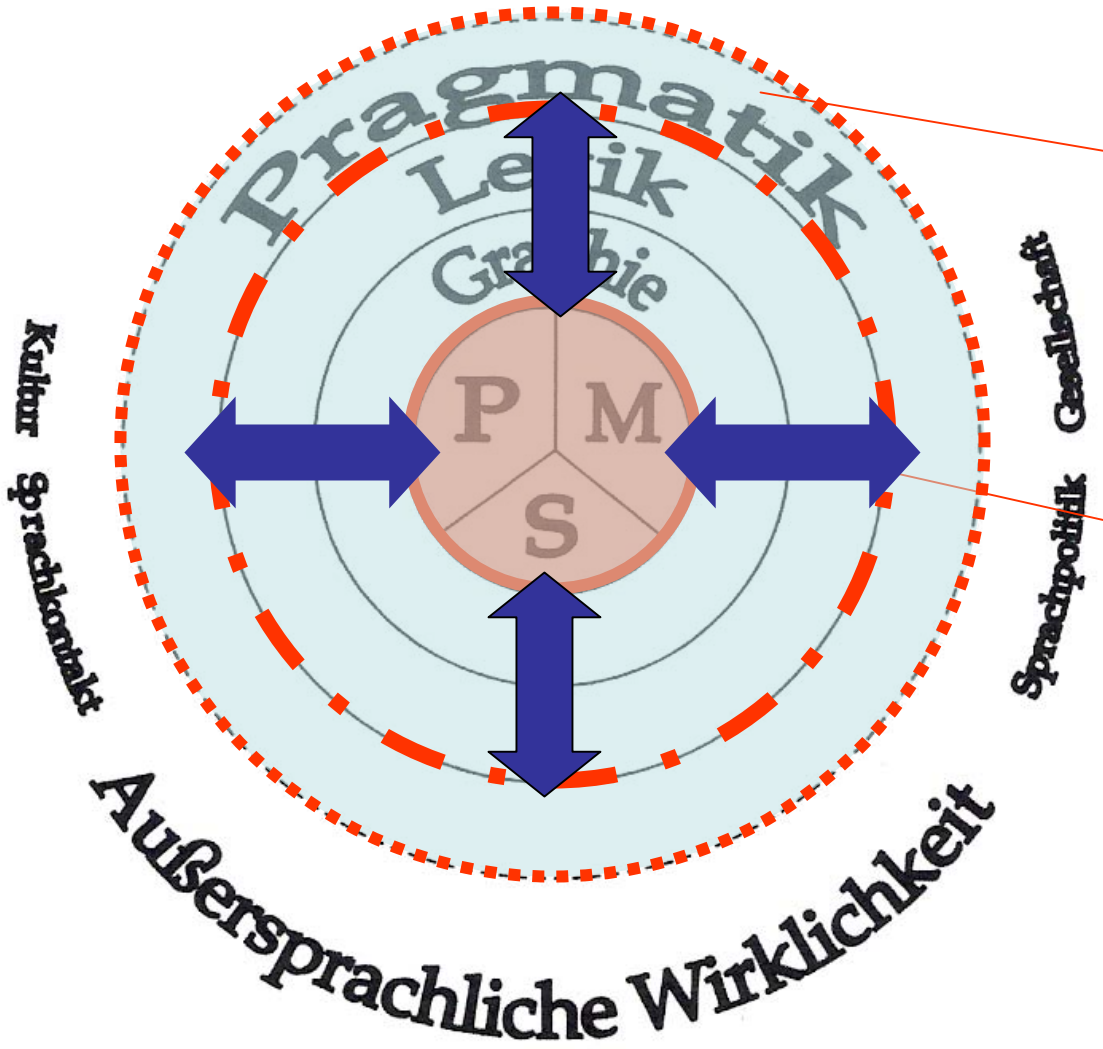
Linguist. Ansatz:
Trennung der sprachlichen Ebenen



P: Phonologie
M: Morphologie
S: Syntax

Grammatik

Computerlinguist. Ansatz: Vermengung der sprachlichen Ebenen



P: Phonologie
M: Morphologie
S: Syntax

Grammatik

Verbvalenz – das Konzept von Helbig/Schenkel

Wörterbuch zur Valenz und Distribution deutscher Verben

- Quantitative Valenz
- Qualitative Valenz
- Semantische Valenz

- I. beachten$_2$
- II. beachten →Sn, Sa
- III. Sn →1. Hum (*Die Mutter* beachtet die Anweisung)
  2. Abstr (als Hum) (*Die Polizei* beachtet den Hinweis)

  Sa → keine Selektionsbeschränkungen (Er beachtet *den Freund,*

  den Hund, den Betrieb, das Schild, die Warnung, das Pfeifen)

- I. enterben$_2$
- II. enterben → Sn, Sa
- III. Sn → Hum (*Der Vater* enterbt den Sohn)

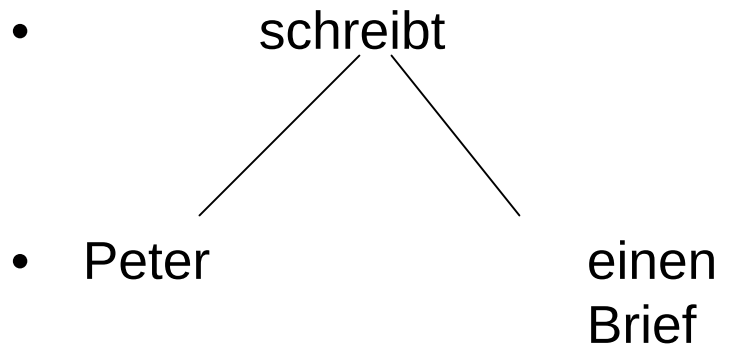
  Sa → Hum (Er enterbt *den Sohn*)

5 Typen der thematischen Progression nach Daneš

- Einfache lineare Progression: Rhema der ersten Äußerung wird zu Thema der zweiten.
- *Es war einmal **ein** **Linguist**. **Der** schrieb **einen** **Aufsatz**. Dieser Aufsatz hatte **Fehler**.*
- Progression mit durchlaufendem Thema: Thema der ersten Äußerung wird in der nächsten wieder aufgenommen.
- *Noam Chomsky ist Linguist. Er ist aber auch politisch aktiv. Er wirkt am MIT.*
- Progression mit abgeleiteten Themen: aus einem Hyperthema werden mehrere Themen abgeleitet.
- *1901 war die letzte orthografische Konferenz. Von 1998 bis 2005 dauert der Übergangszeitraum zur neun Rechtschreibung. Ab 1. August gilt nur noch die neue Rechtschreibung. (Hyperthema: Daten zur Orthografiegeschichte)*
- Entwickeln eines gespaltenen Rhemas: aus einem Rhema werden Teile als Thema für die folgende Äußerung genommen.
- *Es gibt verschiedene Morpheme. Lexikalische Morpheme liefern lexikalische Informationen. Grammatische Morpheme tragen grammatische Informationen.*
- Progression mit thematischem Sprung: ein Glied der thematischen Kette wird ausgelassen.
- *Gestern war Chomsky in Wien. Nicht nur die Generativisten wollten ihn hören.*

Unterschiede zw. IC-Grammatik bzw. Phrasenstrukturgrammatik und Dependenzgrammatik beim Verhältnis zw. Subjekt u. Prädikat

- Dependenzgrammatik:



- IC-Grammatik + Phrasenstrukturgr.: