

Grundkurs Linguistik mathematische Linguistik

Zur formalen Komplexität natürlicher Sprachen

Wiebke Petersen

Prolog: Was ist 'mathematische Linguistik'?

Ziel:

Konstruktion mathematischer Modelle der (phonologischen, grammatischen, semantischen) Struktur von Sprache.

Beschreibung der "Association for Mathematics of Language"

Mathematics of Language is the study of mathematical structures and methods that are of importance to the study of language. This work is generally more theoretical than other sub-fields of linguistics. It includes work at every level of linguistic structure, and involves applications of both discrete and continuous maths, as well as statistics.

<http://www.molweb.org/>

Prolog: Was ist 'mathematische Linguistik'?

Ziel:

Konstruktion mathematischer **Modelle** der (phonologischen, grammatischen, semantischen) **Struktur** von Sprache.

Beschreibung der "Association for Mathematics of Language"

Mathematics of Language is the study of mathematical structures and methods that are of importance to the study of language. This work is generally more theoretical than other sub-fields of linguistics. It includes work at every level of linguistic structure, and involves applications of both discrete and continuous maths, as well as statistics.

<http://www.molweb.org/>

Formale Komplexität natürlicher Sprachen

Formale Komplexität natürlicher Sprachen

- Deutsch, Englisch, Chinesisch, Finnisch, ...

Formale Komplexität natürlicher Sprachen

- Deutsch, Englisch, Chinesisch, Finnisch, ...
- Prolog, Pascal, ...

Formale Komplexität natürlicher Sprachen

- Deutsch, Englisch, Chinesisch, Finnisch, ...
- Prolog, Pascal, ...
- Esperanto, Volapük, Interlingua, ...

Formale Komplexität natürlicher Sprachen

- Deutsch, Englisch, Chinesisch, Finnisch, ...
- Prolog, Pascal, ...
- Esperanto, Volapük, Interlingua, ...
- Aussagenlogik, Prädikatenlogik, ...
- ...

Formale Komplexität **natürlicher Sprachen**

- Deutsch, Englisch, Chinesisch, Finnisch, ...

Formale Komplexität **natürlicher Sprachen**

- Deutsch, Englisch, Chinesisch, Finnisch, ...
- vage, ambig,

Formale Komplexität **natürlicher Sprachen**

- Deutsch, Englisch, Chinesisch, Finnisch, ...
- vage, ambig,
- Ambiguitäten
 - lexikalische Ambiguitäten (Ruf morgen an - Der Ruf der Möwen)

Formale Komplexität **natürlicher Sprachen**

- Deutsch, Englisch, Chinesisch, Finnisch, ...
- vage, ambig,
- Ambiguitäten
 - lexikalische Ambiguitäten (Ruf morgen an - Der Ruf der Möwen)
 - strukturelle Ambiguitäten:

- Die Frau sieht \lceil den Mann \rceil mit dem Fernrohr



- Die Frau sieht \lceil den Mann mit dem Fernrohr \rceil



Formale Komplexität natürlicher Sprachen

- Deutsch, Englisch, Chinesisch, Finnisch, ...
- vage, ambig,
- Ambiguitäten
 - lexikalische Ambiguitäten (Ruf morgen an - Der Ruf der Möwen)
 - strukturelle Ambiguitäten:

- Die Frau sieht \lceil den Mann \rceil mit dem Fernrohr



- Die Frau sieht \lceil den Mann mit dem Fernrohr \rceil



- einzige Experten: Menschen
- dynamische Sprachentwicklung

Formale **Komplexität natürlicher Sprachen**

- schwierig zu erlernen im Erstspracherwerb / Zweitspracherwerb

Formale **Komplexität natürlicher Sprachen**

- schwierig zu erlernen im Erstspracherwerb / Zweitspracherwerb
- komplexe Phonologie / Morphologie / Syntax / ...

Formale **Komplexität natürlicher Sprachen**

- schwierig zu erlernen im Erstspracherwerb / Zweitspracherwerb
- komplexe Phonologie / Morphologie / Syntax / ...
- schwierig zu parsen (die Struktur eines Ausdrucks zu erkennen)

Formale Komplexität natürlicher Sprachen

- Komplexität der Berechnung / Verarbeitungskomplexität (computational complexity)

Formale Komplexität natürlicher Sprachen

- Komplexität der Berechnung / Verarbeitungskomplexität (computational complexity)
- Komplexität der Struktur

Formale Komplexität natürlicher Sprachen

- Komplexität der Berechnung / Verarbeitungskomplexität (computational complexity)
- Komplexität der Struktur

Komplexität der Struktur:

- Natürliche Sprachen werden als abstrakte Symbolsysteme betrachtet, bestehend aus elementaren Zeichen und Kombinationsvorschriften.
- Fragen nach der Grammatikalität natürlichsprachlicher Sätze entsprechen Fragen nach der syntaktischen Korrektheit von Programmen oder der Wohlgeformtheit logischer Ausdrücke.

Was eine Grammatiktheorie erklären muss

- die Katze frisst den Hund
- ~~frisst die Katze Hund den~~
- ~~den Hund frisst die Katze~~
- ~~den die Hund Katze frisst~~
- ~~frisst die Katze den Hund~~
- ~~den Katze die Hund frisst~~
- ~~Katze frisst die den Hund~~
- ~~den Katze Hund die frisst~~
- ~~die Katze frisst Hund den~~
- ~~die Katze Hund den frisst~~
- ~~den Hund frisst Katze die~~
- ~~die den Hund Katze frisst~~
- ~~frisst den Katze die Hund~~
- ~~den Katze Hund frisst die~~
- ~~Katze frisst den Hund die~~
- ...

Auch wenn das Deutsche eine relativ freie Wortstellung hat, so wird doch deutlich, dass die Zahl der grammatisch korrekten Sätze verschwindend klein ist im Vergleich zu den ungrammatischen Wortketten.

Nur **3 von 120** möglichen Wortketten ergeben einen grammatischen Satz.

Wie komplex sind denn nun Sätze des Deutschen?

- Anne sieht Peter
- Anne sieht Peter am Rathaus mit dem Fernrohr
- Anne sieht Peter, den sie vorgestern kennengelernt hat, am Rathaus mit dem Fernrohr

Wie komplex sind denn nun Sätze des Deutschen?

- Anne sieht Peter
- Anne sieht Peter am Rathaus mit dem Fernrohr
- Anne sieht Peter, den sie vorgestern kennengelernt hat, am Rathaus mit dem Fernrohr
- Anne sieht Peter und Hans und Sabine und Joachim und Elfriede und Johanna und Maria und Jochen und Thomas und Andrea

Wie komplex sind denn nun Sätze des Deutschen?

- Anne sieht Peter
- Anne sieht Peter am Rathaus mit dem Fernrohr
- Anne sieht Peter, den sie vorgestern kennengelernt hat, am Rathaus mit dem Fernrohr
- Anne sieht Peter und Hans und Sabine und Joachim und Elfriede und Johanna und Maria und Jochen und Thomas und Andrea

Satzlänge spielt zwar bei der Verarbeitungskomplexität eine Rolle, ist aber kein strukturelles Komplexitätsmerkmal!

Grammatiktheorie versus Theorie formaler Sprachen

Grammatiktheorien

- sollen sprachliche Daten erklären
- sind einzelsprachspezifisch (Deutsch, Englisch, ...)

Grammatiktheorie versus Theorie formaler Sprachen

Grammatiktheorien

- sollen sprachliche Daten erklären
- sind einzelsprachspezifisch (Deutsch, Englisch, ...)

Theorie formaler Sprachen

- ist eine Theorie über den Aufbau und die Struktur von Symbolkettenmengen
- nicht einzelsprachspezifisch
- erlaubt Aussagen über die Mechanismen der Erzeugung und Erkennung von Symbolkettenmengen

Theorie formaler Sprachen

Definition (Definition: formale Sprache)

Eine *formale Sprache* L ist eine Menge von Symbolketten (Wörtern) über einem Alphabet Σ .

- Sprache L_{rom} der gültigen römischen Zahldarstellungen über dem Alphabet $\Sigma_{rom} = \{\mathbf{I}, \mathbf{V}, \mathbf{X}, \mathbf{L}, \mathbf{C}, \mathbf{D}, \mathbf{M}\}$.

Theorie formaler Sprachen

Definition (Definition: formale Sprache)

Eine *formale Sprache* L ist eine Menge von Symbolketten (Wörtern) über einem Alphabet Σ .

- Sprache L_{rom} der gültigen römischen Zahldarstellungen über dem Alphabet $\Sigma_{rom} = \{\mathbf{I}, \mathbf{V}, \mathbf{X}, \mathbf{L}, \mathbf{C}, \mathbf{D}, \mathbf{M}\}$.
- Sprache L_{pal} der Palindrome im deutschen Duden über dem Alphabet $\{a, b, c, \dots, z\}$ $L_{pal} = \{\text{neben, reliefpfeiler, gnutötung, retsinakanister, \dots}\}$

Theorie formaler Sprachen

Definition (Definition: formale Sprache)

Eine *formale Sprache* L ist eine Menge von Symbolketten (Wörtern) über einem Alphabet Σ .

- Sprache L_{rom} der gültigen römischen Zahldarstellungen über dem Alphabet $\Sigma_{rom} = \{\mathbf{I, V, X, L, C, D, M}\}$.
- Sprache L_{pal} der Palindrome im deutschen Duden über dem Alphabet $\{a, b, c, \dots, z\}$ $L_{pal} = \{\text{neben, reliefpfeiler, gnutötung, retsinakanister, \dots}\}$
- Menge der Wörter der Länge 13 über dem Alphabet $\{a, b, c\}$

Theorie formaler Sprachen

Definition (Definition: formale Sprache)

Eine *formale Sprache* L ist eine Menge von Symbolketten (Wörtern) über einem Alphabet Σ .

- Sprache L_{rom} der gültigen römischen Zahldarstellungen über dem Alphabet $\Sigma_{rom} = \{\mathbf{I}, \mathbf{V}, \mathbf{X}, \mathbf{L}, \mathbf{C}, \mathbf{D}, \mathbf{M}\}$.
- Sprache L_{pal} der Palindrome im deutschen Duden über dem Alphabet $\{a, b, c, \dots, z\}$ $L_{pal} = \{\text{neben, reliefpfeiler, gnutötung, retsinakanister, \dots}\}$
- Menge der Wörter der Länge 13 über dem Alphabet $\{a, b, c\}$
- Sprache der syntaktisch wohlgeformten Java-Programme

Unmöglichkeit der Sprachbeschreibung durch Aufzählung

- Maria sagt, dass Otto vom Baum gefallen ist.
- Peter sagt, dass Maria sagt, dass Otto vom Baum gefallen ist.
- Lisa sagt, dass Peter sagt, dass Maria sagt, dass Otto vom Baum gefallen ist.
- Anna sagt, dass Lisa sagt, dass Peter sagt, dass Maria sagt, dass Otto vom Baum gefallen ist.
- ...

Im Deutschen gibt es unendliche viele Sätze.

Sprachbeschreibung durch Angabe einer Grammatik

Grammar

- Grammatiken sind Systeme zur Generierung von Wortketten.
- Eine Grammatik besteht aus endlich vielen Regeln.
- Die Menge aller Wortketten, die von einer Grammatik generiert werden, bilden die von der Grammatik beschriebene formale Sprache.

S	→	NP VP	VP	→	V	NP	→	D N
D	→	the	N	→	cat	V	→	sleeps
N	→	dog	V	→	runs			

Generiert: {the cat sleeps, the cat runs, the dog sleeps, the dog runs}

Chomskyhierarchie

$$\begin{aligned} S &\rightarrow NP VP \\ N &\rightarrow \text{dog} \end{aligned}$$
$$\begin{aligned} S &\rightarrow S \text{ und } S \\ D N &\rightarrow EN \end{aligned}$$

- Wenn man die Form der Regeln einschränkt erhält man Teilmengen der Menge aller durch eine Grammatik erzeugten Sprachen.

Chomskyhierarchie

$$\begin{aligned} S &\rightarrow NP VP \\ N &\rightarrow \text{dog} \end{aligned}$$
$$\begin{aligned} S &\rightarrow S \text{ und } S \\ D N &\rightarrow EN \end{aligned}$$

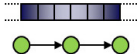
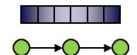








- Wenn man die Form der Regeln einschränkt erhält man Teilmengen der Menge aller durch eine Grammatik erzeugten Sprachen.
- Die Chomskyhierarchie ist eine Hierarchie über die Regelbedingungen (den verschiedenen Sprachklassen entsprechen Einschränkungen über die rechten und linken Regelseiten).

Chomskyhierarchie

$$\begin{aligned} S &\rightarrow NP VP \\ N &\rightarrow \text{dog} \end{aligned}$$
$$\begin{aligned} S &\rightarrow S \text{ und } S \\ D N &\rightarrow EN \end{aligned}$$

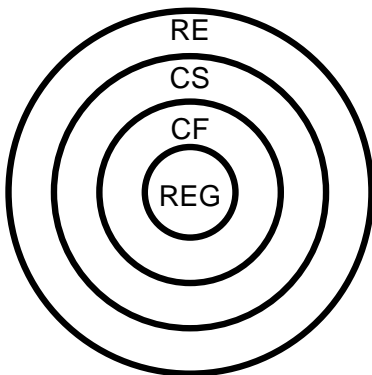
- Wenn man die Form der Regeln einschränkt erhält man Teilmengen der Menge aller durch eine Grammatik erzeugten Sprachen.
- Die Chomskyhierarchie ist eine Hierarchie über die Regelbedingungen (den verschiedenen Sprachklassen entsprechen Einschränkungen über die rechten und linken Regelseiten).
- Für Linguisten ist die Chomsky Hierarchie besonders interessant, da sie die Form der Regeln zentral stellt, und somit Aussagen über Grammatikformalismen zuläßt.

Chomsky-Hierarchie (grober Überblick)

<i>Sprache</i>	<i>Automat</i>	<i>Grammatik</i>	<i>Erkennung</i>	<i>Abhängigkeit</i>
rekursiv aufzählbar	Turing Maschine 	unbeschränkt $Baa \rightarrow \varepsilon$	unentscheidbar	beliebig
kontext- sensitiv	linear gebunden 	kontext- sensitiv $\gamma A \delta \rightarrow \gamma \beta \delta$	NP-vollständig 	überkreuzt 
kontext- frei	Kellerautomat (Stapel) 	kontextfrei $C \rightarrow bABa$	polynomiell 	eingebettet 
regulär	endlicher Automat 	regulär $A \rightarrow bA$	linear 	strikt lokal 

Chomskyhierarchie: Hauptsatz

$$\text{REG} \subset \text{CF} \subset \text{CS} \subset \text{RE}$$



REG: reguläre Sprachen, CF: kontextfreie Sprachen, CS: kontextsensitive Sprachen,
RE: rekursiv-aufzählbare Sprachen

Chomskyhierarchie: Beispiele

Die Sprache der Wörter mit fixem Präfix ist regulär.

- Grammatik: $S \rightarrow aT$, $S \rightarrow a$, $T \rightarrow aT$, $T \rightarrow bT$, $T \rightarrow a$, $T \rightarrow b$
generiert die Menge aller Wörter über dem Alphabet $\{a, b\}$, die mit dem Präfix 'a' beginnen.

Chomskyhierarchie: Beispiele

Die Sprache der Wörter mit fixem Präfix ist regulär.

- Grammatik: $S \rightarrow aT$, $S \rightarrow a$, $T \rightarrow aT$, $T \rightarrow bT$, $T \rightarrow a$, $T \rightarrow b$
generiert die Menge aller Wörter über dem Alphabet $\{a, b\}$, die mit dem Präfix 'a' beginnen.

Die Sprache der Palindrome über einem Alphabet ist kontextfrei, aber nicht regulär.

- Ein Palindrom ist eine Zeichenkette, die von hinten und von vorne gelesen gleich bleibt (z.B. 'Reliefpfeiler').
- Grammatik: $S \rightarrow aSa$, $S \rightarrow bSb$, $S \rightarrow a$, $S \rightarrow b$
generiert die Menge aller Palindrome ungerader Länge über dem Alphabet $\{a, b\}$

Chomskyhierarchie: Beispiele

Die Sprache der Wörter mit fixem Präfix ist regulär.

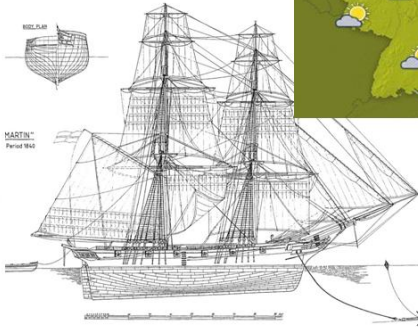
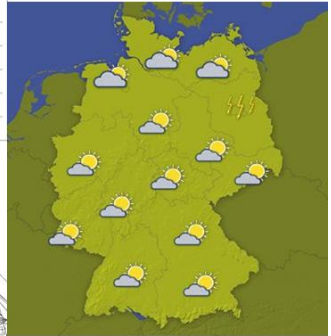
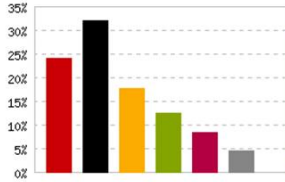
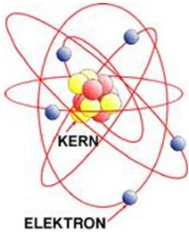
- Grammatik: $S \rightarrow aT$, $S \rightarrow a$, $T \rightarrow aT$, $T \rightarrow bT$, $T \rightarrow a$, $T \rightarrow b$
generiert die Menge aller Wörter über dem Alphabet $\{a, b\}$, die mit dem Präfix 'a' beginnen.

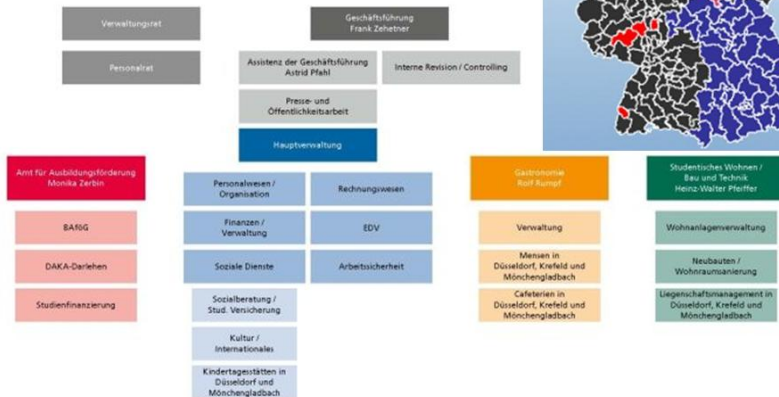
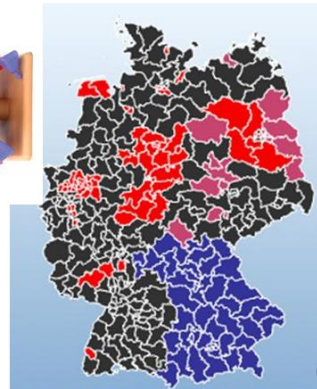
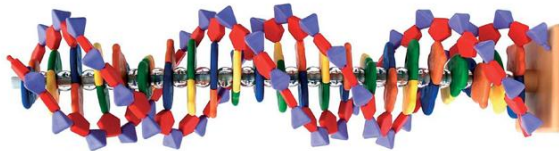
Die Sprache der Palindrome über einem Alphabet ist kontextfrei, aber nicht regulär.

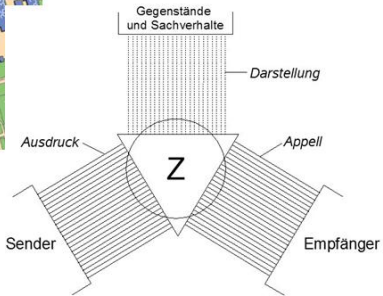
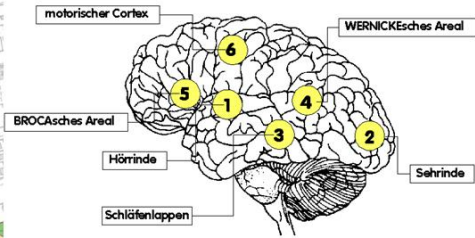
- Ein Palindrom ist eine Zeichenkette, die von hinten und von vorne gelesen gleich bleibt (z.B. 'Reliefpfeiler').
- Grammatik: $S \rightarrow aSa$, $S \rightarrow bSb$, $S \rightarrow a$, $S \rightarrow b$
generiert die Menge aller Palindrome ungerader Länge über dem Alphabet $\{a, b\}$

Die Sprache der Wiederholwörter über einem Alphabet ist kontextsensitiv, aber nicht kontextfrei.

- Eine Zeichenkette ist ein Wiederholwort, wenn sie sich in zwei gleiche Teile zerlegen lässt (z.B.: 'Papa', 'Momo')







Modell

- künstlich geschaffen
- materiell oder immateriell
- vereinfachtes Abbild
- zweckgerichtet
- Abstraktion
- Repräsentation
- Modellierungsannahmen

Modell

- künstlich geschaffen
- materiell oder immateriell
- vereinfachtes Abbild
- zweckgerichtet
- Abstraktion
- Repräsentation
- Modellierungsannahmen

Modellierung

Ein **Subjekt** entwirft zu einem **Original** ein **Modell** zu einem bestimmten **Zweck**.

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Warum ist die formale Komplexität natürlicher Sprachen von Interesse?

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Warum ist die formale Komplexität natürlicher Sprachen von Interesse?

- gibt Information über die Struktur von natürlichen Sprachen (NL)

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Warum ist die formale Komplexität natürlicher Sprachen von Interesse?

- gibt Information über die Struktur von natürlichen Sprachen (NL)
- erlaubt Rückschlüsse auf Adäquatheit eines Grammatikformalismus für NL

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Warum ist die formale Komplexität natürlicher Sprachen von Interesse?

- gibt Information über die Struktur von natürlichen Sprachen (NL)
- erlaubt Rückschlüsse auf Adäquatheit eines Grammatikformalismus für NL
- unter computerlinguistischen Aspekten sind möglichst effizient verarbeitbare Analysen gefragt

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Warum ist die formale Komplexität natürlicher Sprachen von Interesse?

- gibt Information über die Struktur von natürlichen Sprachen (NL)
- erlaubt Rückschlüsse auf Adäquatheit eines Grammatikformalismus für NL
- unter computerlinguistischen Aspekten sind möglichst effizient verarbeitbare Analysen gefragt
- erlaubt Rückschlüsse auf menschliche Sprachverarbeitung

Notwendige Modellierungsannahmen

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Warum ist die formale Komplexität natürlicher Sprachen von Interesse?

- gibt Information über die Struktur von natürlichen Sprachen (NL)
- erlaubt Rückschlüsse auf Adäquatheit eines Grammatikformalismus für NL
- unter computerlinguistischen Aspekten sind möglichst effizient verarbeitbare Analysen gefragt
- erlaubt Rückschlüsse auf menschliche Sprachverarbeitung

Notwendige Modellierungsannahmen

- 1 Es gibt die Familie der natürlichen Sprachen
- 2 Natürliche Sprachen lassen sich als Mengen von Zeichenketten modellieren
- 3 Natürliche Sprachen lassen sich durch endliche Regelsysteme beschreiben

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Warum ist die formale Komplexität natürlicher Sprachen von Interesse?

- gibt Information über die Struktur von natürlichen Sprachen (NL)
- erlaubt Rückschlüsse auf Adäquatheit eines Grammatikformalismus für NL
- unter computerlinguistischen Aspekten sind möglichst effizient verarbeitbare Analysen gefragt
- erlaubt Rückschlüsse auf menschliche Sprachverarbeitung

Notwendige Modellierungsannahmen

- 1 Es gibt die Familie der natürlichen Sprachen
- 2 Natürliche Sprachen lassen sich als Mengen von Zeichenketten modellieren
- 3 Natürliche Sprachen lassen sich durch endliche Regelsysteme beschreiben
- 4 Damit die Frage spannend ist: Natürliche Sprachen sind unendlich

Annahme 1: Familie der natürlichen Sprachen

Es gibt die Familie der natürlichen Sprachen:

- alle natürlichen Sprachen sind strukturell ähnlich
- alle natürlichen Sprachen haben eine ähnliche Komplexität

Argumente:

- Alle Sprachen haben gleiche Funktion
- In allen Sprachen lassen sich ähnliche Elemente identifizieren (Phoneme, Morpheme, Wortarten, . . .)
- Jeder Mensch kann jede Sprache im Erstspracherwerb erlernen (in ähnlicher Zeit)

Annahme 2: nat. Spr. = Mengen von Zeichenketten

Natürliche Sprachen lassen sich als Mengen von Zeichenketten modellieren:

- Muttersprachler haben volle Kompetenz (können also entscheiden, ob sie einen Satz als grammatisch oder nicht beurteilen)
- konsistente Grammatikalitätsurteile

Argumente:

- Fehler beruhen auf Performanz- und nicht Kompetenzproblemen

Annahme 2: nat. Spr. = Mengen von Zeichenketten

Natürliche Sprachen lassen sich als Mengen von Zeichenketten modellieren:

- Muttersprachler haben volle Kompetenz (können also entscheiden, ob sie einen Satz als grammatisch oder nicht beurteilen)
- konsistente Grammatikalitätsurteile

Argumente:

- Fehler beruhen auf Performanz- und nicht Kompetenzproblemen
- Mathews (1979) Gegenbeispiel:
 - The canoe floated down the river sank.

Annahme 2: nat. Spr. = Mengen von Zeichenketten

Natürliche Sprachen lassen sich als Mengen von Zeichenketten modellieren:

- Muttersprachler haben volle Kompetenz (können also entscheiden, ob sie einen Satz als grammatisch oder nicht beurteilen)
- konsistente Grammatikalitätsurteile

Argumente:

- Fehler beruhen auf Performanz- und nicht Kompetenzproblemen
- Mathews (1979) Gegenbeispiel:
 - The canoe floated down the river sank.
 - The man thrown down the stairs died.

Annahme 2: nat. Spr. = Mengen von Zeichenketten

Natürliche Sprachen lassen sich als Mengen von Zeichenketten modellieren:

- Muttersprachler haben volle Kompetenz (können also entscheiden, ob sie einen Satz als grammatisch oder nicht beurteilen)
- konsistente Grammatikalitätsurteile

Argumente:

- Fehler beruhen auf Performanz- und nicht Kompetenzproblemen
- Mathews (1979) Gegenbeispiel:
 - The canoe floated down the river sank.
 - The man thrown down the stairs died.
 - The man that was thrown down the stairs died.

Reihenfolge beeinträchtigt Grammatikalitätsurteile

Annahme 3: nat. Spr. = beschreibbar durch endliches Regelsystem

Natürliche Sprachen lassen sich durch endliche Regelsysteme beschreiben:

Argumente:

- Regeln wie 'S → NP VP' erfassen Generalisierungen.
- Aufzählung erscheint nicht möglich.

Annahme 4: Natürliche Sprachen sind unendlich

Jede natürliche Sprache besteht aus einer unendlichen Zahl von Zeichenketten
Rekursion:

- Dies ist der Hund, der die Katze ärgerte, die die Maus tötete, die den Käse fraß, der in dem Haus lag, das Maja gebaut hat.

Annahme 4: Natürliche Sprachen sind unendlich

Jede natürliche Sprache besteht aus einer unendlichen Zahl von Zeichenketten
Rekursion:

- Dies ist der Hund, der die Katze ärgerte, die die Maus tötete, die den Käse fraß, der in dem Haus lag, das Maja gebaut hat.
- Dies ist der Hund, der die Katze, die die Maus, die den Käse, der in dem Haus, das Maja gebaut hat, lag, fraß, tötete, ärgerte.

Annahme 4: Natürliche Sprachen sind unendlich

Jede natürliche Sprache besteht aus einer unendlichen Zahl von Zeichenketten
Rekursion:

- Dies ist der Hund, der die Katze ärgerte, die die Maus tötete, die den Käse fraß, der in dem Haus lag, das Maja gebaut hat.
- Dies ist der Hund, der die Katze, die die Maus, die den Käse, der in dem Haus, das Maja gebaut hat, lag, fraß, tötete, ärgerte.
- Donaudampfschiffskapitänsmützenschirmi

Annahme 4: Natürliche Sprachen sind unendlich

Jede natürliche Sprache besteht aus einer unendlichen Zahl von Zeichenketten
Rekursion:

- Dies ist der Hund, der die Katze ärgerte, die die Maus tötete, die den Käse fraß, der in dem Haus lag, das Maja gebaut hat.
- Dies ist der Hund, der die Katze, die die Maus, die den Käse, der in dem Haus, das Maja gebaut hat, lag, fraß, tötete, ärgerte.
- Donaudampfschiffskapitänsmützenschirmi

Aber ist Deutsch wirklich unendlich?

Dies ist der Junge, der den Hasen, der den Hund, der die Katze, die die Maus, die den Käse, der in dem Haus, das Maja gebaut hat, lag, fraß, tötete, ärgerte, jagte, sah, mochte.

Annahme 4: Natürliche Sprachen sind unendlich

Jede natürliche Sprache besteht aus einer unendlichen Zahl von Zeichenketten
Rekursion:

- Dies ist der Hund, der die Katze ärgerte, die die Maus tötete, die den Käse fraß, der in dem Haus lag, das Maja gebaut hat.
- Dies ist der Hund, der die Katze, die die Maus, die den Käse, der in dem Haus, das Maja gebaut hat, lag, fraß, tötete, ärgerte.
- Donaudampfschiffskapitänsmützenschirmi

Aber ist Deutsch wirklich unendlich?

Dies ist der Junge, der den Hasen, der den Hund, der die Katze, die die Maus, die den Käse, der in dem Haus, das Maja gebaut hat, lag, fraß, tötete, ärgerte, jagte, sah, .

Zurück zur Frage

Zu welcher Sprachklasse gehören die natürlichen Sprachen?

Chomsky (1957):

- “English is not a regular language”
- context-free languages: “I do not know whether or not English is itself literally outside the range of such analysis”

Sind natürliche Sprachen regulär? Nein!

- a woman hired another woman
- a woman whom another woman hired hired another woman
- a woman whom another woman whom another woman hired hired hired another woman
- a woman whom another woman whom another woman whom another woman hired hired hired hired another woman
- ...

Sind natürliche Sprachen regulär? Nein!

- a woman hired another woman
- a woman whom another woman hired hired another woman
- a woman whom another woman whom another woman hired hired hired another woman
- a woman whom another woman whom another woman whom another woman hired hired hired hired another woman
- ...
-
- a woman (whom another woman)ⁿ (hired)ⁿ hired another woman ($n > 0$)
- Struktur $a^n b^n$ nicht regulär.

Sind natürliche Sprachen kontextfrei? fehlerhafte Argumente (1)

unzulässige Induktion:

- wir kennen keine adäquate kontextfreie Grammatik des Englischen \Rightarrow es gibt keine adäquate kontextfreie Grammatik des Englischen

Sind natürliche Sprachen kontextfrei? fehlerhafte Argumente (1)

unzulässige Induktion:

- wir kennen keine adäquate kontextfreie Grammatik des Englischen ⇒ es gibt keine adäquate kontextfreie Grammatik des Englischen

An Introduction to the Principles of Transformational Syntax (Akmajian & Heny, 1975):

“**Since there seems to be no way** of using such PS rules to represent an obviously significant generalization about one language, namely, English, **we can be sure** that phrase structure grammars cannot possibly represent all the significant aspects of language structure.”

Sind natürliche Sprachen kontextfrei? fehlerhafte Argumente (2)

intuitive Auffassung von Kontextfreiheit:

Sind natürliche Sprachen kontextfrei? fehlerhafte Argumente (2)

intuitive Auffassung von Kontextfreiheit:

- the **girl sees** the dog / the **girls see** the dog

Sind natürliche Sprachen kontextfrei? fehlerhafte Argumente (2)

intuitive Auffassung von Kontextfreiheit:

- the **girl** **sees** the dog / the **girls** **see** the dog
- the **girl** who climbed the tree which was planted last year when it rained so much **sees** the dog
- the **girls** who climbed the tree which was planted last year when it rained so much **see** the dog

Sind natürliche Sprachen kontextfrei? fehlerhafte Argumente (2)

intuitive Auffassung von Kontextfreiheit:

- the **girl** **sees** the dog / the **girls** **see** the dog
- the **girl** who climbed the tree which was planted last year when it rained so much **sees** the dog
- the **girls** who climbed the tree which was planted last year when it rained so much **see** the dog

Der Numerus von 'girl' bestimmt den Numerus von 'see' ⇒ Kontextabhängigkeit.

Sind natürliche Sprachen kontextfrei? fehlerhafte Argumente (2)

intuitive Auffassung von Kontextfreiheit:

- the **girl** **sees** the dog / the **girls** **see** the dog
- the **girl** who climbed the tree which was planted last year when it rained so much **sees** the dog
- the **girls** who climbed the tree which was planted last year when it rained so much **see** the dog

Der Numerus von 'girl' bestimmt den Numerus von 'see' ⇒ Kontextabhängigkeit.

A realistic transformational grammar (Bresnan, 1987)

" [...] in many cases the number of a verb agrees with that of a noun phrase at some **distance** from it [...] the distant type of agreement [...] **cannot be** adequately **described** even **by context-sensitive** phrase-structure rules, for **the possible context is not correctly describable as a finite string of phrases.**"

Pullum & Gazdar (1982)

- These: Alle bis dato präsentierten Argumente für die Nichtkontextfreiheit von NL sind nicht zwingend!
 - Folklore
 - falsche Daten
 - formale Fehler
- 30 Jahre vergebliche Suche nach einer nichtkontextfreien natürlichen Sprache
- Menschen scheinen Sätze in linearer Zeit zu parsen
Probleme bereiten genau die Sätze, die beweisen, dass NL nicht regulär sind.

Sind natürliche Sprachen kontextfrei?

Nebensatzeinbettung im Schweizerdeutschen

- mer d'chind em Hans es huus lönd hälfe aastriiche
wir die Kinder-AKK Hans-DAT das Haus-AKK ließen helfen anstreichen

NP₁ NP₂ NP₃ VP₁ VP₂ VP₃

"cross serial dependencies"



Sind natürliche Sprachen kontextfrei?

Nebensatzeinbettung im Schweizerdeutschen

- mer d'chind em Hans es huus lönd hälfe aastriiche
wir die Kinder-AKK Hans-DAT das Haus-AKK ließen helfen anstreichen

NP₁ NP₂ NP₃ VP₁ VP₂ VP₃

"cross serial dependencies"



- *mer d'chind de Hans es huus lönd hälfe aastriiche
wir die Kinder-AKK Hans-AKK das Haus-AKK ließen helfen anstreichen

Sind natürliche Sprachen kontextfrei?

Nebensatzeinbettung im Schweizerdeutschen

- mer d'chind em Hans es huus lönd hälfe aastriiche
wir die Kinder-AKK Hans-DAT das Haus-AKK ließen helfen anstreichen

NP₁ NP₂ NP₃ VP₁ VP₂ VP₃ "cross serial dependencies"



- *mer d'chind de Hans es huus lönd hälfe aastriiche
wir die Kinder-AKK Hans-AKK das Haus-AKK ließen helfen anstreichen

Nebensatzeinbettung im Deutschen

- weil er die Kinder dem Hans das Haus streichen helfen ließ

NP₁ NP₂ NP₃ VP₃ VP₂ VP₁ "nested dependencies"



Das Schweizerdeutsche ist nicht kontextfrei! (Shieber 1985)

Nebensatzeinbettung im Schweizerdeutschen:

em Hans es huus hälfe aastriche

Dat Akk Dat Akk

d'chind em Hans es huus lönd hälfe aastriche

Akk Dat Akk Akk Dat Akk

Nebensatzeinbettung im Deutschen:

dem Hans das Haus streichen helfen

Dat Akk Akk Dat

die Kinder dem Hans das Haus streichen helfen ließ

Akk Dat Akk Akk Dat Akk

Das Schweizerdeutsche ist nicht kontextfrei! (Shieber 1985)

Nebensatzeinbettung im Schweizerdeutschen:

em Hans es huus hälfe aastrüiche
Dat Akk Dat Akk

d'chind em Hans es huus lönd hälfe aastrüiche
Akk Dat Akk Akk Dat Akk

Die Sprache der Wiederholwörter über dem Alphabet {Akk, Dat} ist nicht kontextfrei!

Nebensatzeinbettung im Deutschen:

dem Hans das Haus streichen helfen
Dat Akk Akk Dat

die Kinder dem Hans das Haus streichen helfen ließ
Akk Dat Akk Akk Dat Akk

Die Sprache der Palindrome über dem Alphabet {Akk, Dat} ist kontextfrei!

Mögliche Angriffspunkte

- falsche Daten (unsichere Grammatikalitätsurteile)
- Kasus ist nicht syntaktisch sondern semantisch (aber ‚helfen‘ vs. ‚unterstützen‘)
- Die Länge der Sätze ist beschränkt

Shieber: "Down this path lies tyranny. Acceptance of this argument opens the way to proofs of natural languages as regular, nay, finite. The linguist proposing this counterargument to salvage the context-freeness of natural language may have won the battle, but has certainly lost the war.

Literatur

- Chomsky, Noam 1956: Three models for the description of language. In IRE Transactions on Information Theory 2(3), 113-124.
- Chomsky, Noam 1957: Syntactic structures. The Hague: Mouton
- Matthews, Robert J. 1979: Are the Grammatical Sentences of a Language a Recursive Set? In Synthese 40, 209–224.
- Pullum, Geoffrey K. und Gazdar, Gerald 1982: Natural Languages and Context-Free Languages. In Linguistics and Philosophy, 4:471–504.
- Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. In Linguistics and Philosophy, 8:333–343.

Hausaufgabe

Sie haben in der Sitzung ein Argument kennengelernt, das anhand des Schweizerdeutschen zeigt, dass natürliche Sprachen nicht kontextfrei sind. Dieses Argument basiert unter anderem auf der Annahme, dass sich eine natürliche Sprache adäquat als eine unendliche Menge von Zeichenketten über einem festgelegten Alphabet modellieren lässt.

Diskutieren Sie diese **Annahme** (nicht das Argument) und beziehen Sie zu ihr Stellung (300-400 Wörter).

Die beiden Texte von Geoffrey K. Pullum liefern wertvolle Hintergrundinformationen zu der Diskussion um die formale Mächtigkeit natürlicher Sprachen und geben einen Eindruck in die linguistische Methodologie.