Linguistik für Kognitionswissenschaften 1:

Formale Komplexität natürlicher Sprachen

Gerhard Jäger

Universität Tübingen

21. Oktober 2010

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Die Chomsky-Hierarchie

Ersetzungssysteme

$$G = \langle N, T, S, R \rangle$$

N ... Nichtterminal-Symbole
T ... Terminal-Symbole
S ... Startsymbol ($S \in N$)
R ... Regeln

Regeln haben die Form

$$\alpha \rightarrow \beta$$

wobei α, β Ketten über $T \cup N$ sind und β nicht leer ist.

Die Chomsky-Hierarchie

- $$L(G) = \{w \in T^* | S \rightarrow^* w\}$$

  "$\rightarrow^*$" ist der reflexive und transitive Abschluss von $\rightarrow$.

- Jede rekursiv aufzählbare Sprache kann durch ein Ersetzungssystem beschrieben werden.
- (Unbeschränkte) Ersetzungssysteme sind in ihrere Ausdrucksstärke zu Turing-Maschinen äquivalent.
- (Chomsky-) Typ-0-Grammatiken = unbeschränkte Ersetzungssysteme
- Zugehörigkeit zu einer Typ-0-Sprache ist **unentscheidbar**

Die Chomsky-Hierarchie

Kontext-sensitive Grammatiken

- Unterklasse der Typ-0-Grammatiken
- Einschränkung:
  Alle Regeln haben die Form

$$\alpha \rightarrow \beta$$

  wobei

$$\mathit{length}(\alpha) \leq \mathit{length}(\beta)$$

- Effekt: Zugehörigkeit zu einer kontext-sensitiven Sprache ist entscheidbar

Die Chomsky-Hierarchie

Kontext-sensitive Grammatiken

- alternative (ursprüngliche) Formulierung:

  Alle Regeln haben die Form

  $$\alpha A \beta \rightarrow \alpha \gamma \beta$$

  wobei $A \in N$, $\alpha, \beta, \gamma \in (T \cup N)^*, \gamma \neq \varepsilon$
- Beide Formulierungen beschreiben die selbe Klasse von Sprachen.
- Nicht alle entscheidbaren Sprachen sind kontext-sensitiv (aber „fast alle“)
- Zugehörigkeit für kontext-sensitive Sprachen ist PSPACE-vollständig
- Kontext-sensitive Grammatiken sind äquivalent zu **linear beschränkten Automaten**.

Die Chomsky-Hierarchie

Kontext-freie Grammatiken

- Unterklasse der kontext-sensitiven Grammatiken[1]
- zusätzliche Beschränkung:

Regeln haben die Form

$$A \rightarrow \alpha$$

wobei

$$A \in N, \alpha \in (T \cup N)^+$$

- Zugehörigkeit zu kontext-freien Sprachen ist entscheidbar in **polynomialer Zeit** ($O(n^3)$).
- Kontext-freie Grammatiken sind expressiv äquivalent zu **Kellerspeicher-Automaten**.

[1] Wenn man mal von dem eher notationellem Problem der ε-Regeln absieht.

Die Chomsky-Hierarchie

Reguläre Grammatiken

- Unterklasse der kontext-freien Grammatiken
- Zusätzliche Einschränkung:

  Regeln haben die Form

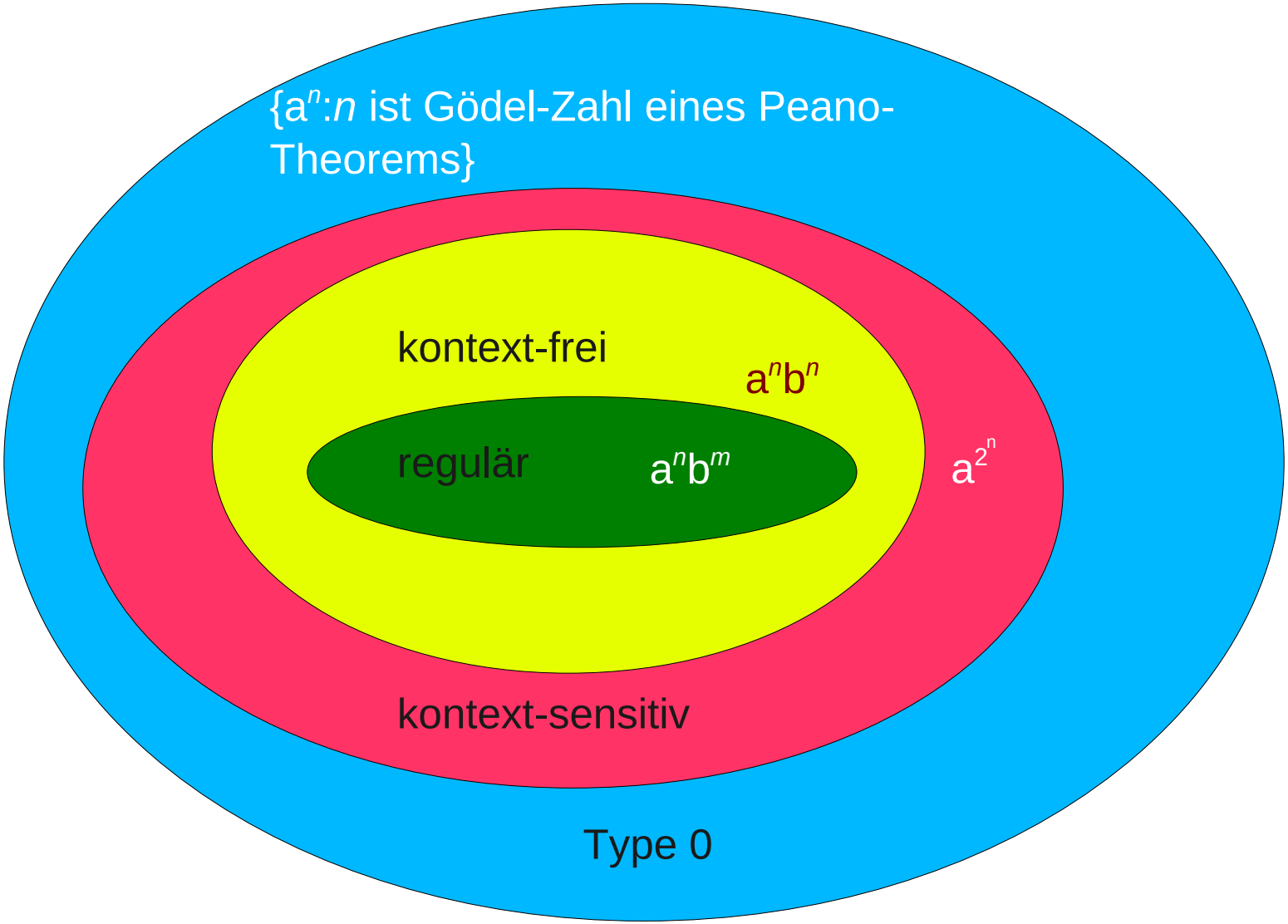
  $$A \to a$$

  oder

  $$A \to Ba$$

  wobei $A, B \in N$ *und* $a \in T$
- Zugehörigkeit ist entscheidbar in **linearer Zeit**.
- Reguläre Grammatiken sind expressiv äquivalent zu **endlichen Automaten**.

Die Chomsky-Hierarchie



Position der natürlichen Sprachen

Wo ordnen sich natürliche Sprachen in dieser Hierarchie ein?

- Frage war über mehrere Jahrzehnte hinweg heiß umstritten
- übliche Struktur eines Arguments:
  - finde eine rekursive Konstruktion C in einer natürlichen Sprache L
  - gib Gründe an, dass die Sprach-**Kompetenz** der Sprachbenutzer unbeschränkte Rekursion zulässt (während die Performanz natürlich *de facto* eine Obergrenze festlegt)
  - reduziere C mit Hilfe eines *Homomorphismus* auf eine formale Sprache L', deren Komplexität bekannt ist
  - zeige, dass L mindestens so komplex ist wie L'
  - generalisiere auf alle natürlichen Sprachen: Wenn es eine Sprache gibt, die mindestens so komplex ist wie ..., dann muss die menschliche Sprachfähigkeit diesen Komplexitätsgrad zulassen.

Position der natürlichen Sprachen

Sind natürliche Sprachen regulär?

Chomsky 1957: Natürliche Sprachen sind nicht regulär.
Struktur seiner Argumentation:

- Betrachte folgende drei hypothetischen Sprachen
  1. $ab, aabb, aaabbb$ $(a^n b^n)$
  2. $aa, bb, abba, baab, aaaa, bbbb, aabbaa, abbbba,$ ...
     (Palindrom-Sprache)
  3. $aa, bb, abab, baba, aaaa, bbbb, aabaab, abbabb, aababaabab$
     (Kopier-Sprache)
- Es kann leicht gezeigt werden, dass diese Sprachen nicht regulär sind.
- Also können Sprachen, die eine ähnliche Struktur haben (dabei aber statt a und b auch komplexe Einheiten zulassen), auch nicht regulär sein.
- Natürliche Sprachen lassen unbeschränkte Rekursion zu.

Position der natürlichen Sprachen

- Die folgenden Konstruktionen können unbeschränkte geschachtelt werden:
  - If S_1, then S_2.
  - Either S_3 or S_4.
  - The man that said that S_5 is arriving today.
- Deshalb — so Chomsky — kann Englisch nicht regulär sein.

"It is clear, then that in English we can find a sequence a + S1 + b, where there is a dependency between a and b, and we can select as S1 another sequence c + S2 + d, where there is a dependency between c and d ... etc. A set of sentences that is constructed in this way...will have all of the mirror image properties of [2] which exclude [2] from the set of finite state languages."
(Chomsky 1957)

Position der natürlichen Sprachen

Abschluss-Eigenschaften regulärer Sprachen.

Theorem 1: Wenn L_1 und L_2 regulär sind, dann ist $L_1 \cap L_2$ auch regulär.

Theorem 2: Die Klasse der regulären Sprachen ist unter Homomorphismen abgeschlossen.

Theorem 3: Die Klasse der regulären Sprachen ist unter Inversion abgeschlossen.

Position der natürlichen Sprachen

- Homomorphismus:

$$\text{neither} \mapsto a$$

$$\text{nor} \mapsto b$$

$$\textit{alles andere} \mapsto \varepsilon$$

If it neither rains nor snows, then if it rains then it snows.
$\mapsto ab$

Position der natürlichen Sprachen

- Englisch wird dadurch nicht auf die Spiegelsprache abgebildet, sondern auf L_1:

$$\begin{aligned} S &\rightarrow aST \\ T &\rightarrow bST \\ T &\rightarrow bS \\ S &\rightarrow \varepsilon \end{aligned}$$

Position der natürlichen Sprachen

Das Pumping-Lemma für reguläre Sprachen

Sei L eine reguläre Sprache. Dann gibt es eine Konstante n so dass, wenn z eine beliebige Kette aus L ist und *length*$(z) \geq n$, wir $z = uvw$ schreiben können, so dass *length*$(uv) \leq n$, $v \neq \varepsilon$, und für alle $i \geq 0, uv^iw \in L$.

Position der natürlichen Sprachen

- Angenommen Englisch ist regulär.
- Wegen Abschluss unter Homomorphismus ist auch L_1 regulär.
- a^*b^* ist regulär. (Übungsaufgabe: Warum?)
- Daher ist $a^*b^* \cap L_1$ eine reguläre Sprache

$$L_2 = L_1 \cap a^*b^* = \{a^n b^m | n \leq m\}$$

aufgrund von Theorem 1.

Position der natürlichen Sprachen

- Aufgrund von Abschluss unter Inversion und Homomorphismus ist

$$L_3 = \{a^n b^m | n \geq m\}$$

  auch regulär.
- Daher ist L_4 regulär:

$$L_4 = L_2 \cap L_3 = a^n b^n$$

- L_4 jedoch kann nicht regulär sein aufgrund des Pumping-Lemmas.
- Daher kann Englisch keine reguläre Sprache sein.

Position der natürlichen Sprachen

Abweichende Sichtweisen:

- *Alle derartigen Argumentationen benutzen Zentral-Einbettung.*
- *Menschen sind sehr schlecht in der Verarbeitung von Zentral-Einbettung.*
- *Eine Auffassung von Sprach-Kompetenz die dieses Faktum ignoriert, ist fragwürdig.*
- *Daher sind natürliche Sprachen tatsächlich regulär.*

Position der natürlichen Sprachen

Übung:

Zeige, dass Chomsky $a^n b^n$, die Spiegel-Sprache und die Kopier-Sprache korrekt als nicht regulär klassifiziert hat!

Position der natürlichen Sprachen

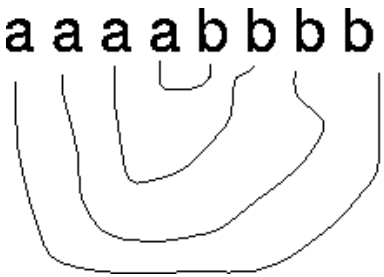
Sind natürliche Sprachen kontext-frei?

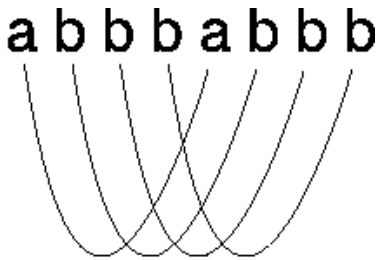
- Geschichte des Problems:
  - Chomsky 1957: Vermutung, dass natürliche Sprachen nicht kontext-frei sind
  - 1960er und 197er Jahre: viele Versuche, die Konjektur zu beweisen.
  - Pullum and Gazdar 1982:
    - Alle diese Versuche sind fehlgeschlagen.
    - Soweit wir wissen, sind tatsächlich alle natürlichen Sprachen (aufgefasst als Mengen von Zeichenketten) kontext-frei.
  - Huybregts 1984, Shieber 1985: Beweis, dass Schweizerdeutsch nicht kontext-frei ist.
  - Culy 1985: Beweis, dass die westafrikanische Sprache Bambara nicht kontext-frei ist.

Position der natürlichen Sprachen

Geschachtelte und überkreuzende Abhängigkeiten

- Kontext-freie Sprachen können — im Unterschied zu regulären Sprachen — unbeschränkte Abhängigkeiten aufweisen
- Diese Abhängigkeiten müssen allerdings **geschachtelt** sein, nicht überkreuzend
- Beispiel:
  - $a^n b^n$ hat unbeschränkte geschachtelte Abhängigkeiten $\rightarrow$ kontext-frei
  - Die Kopiersprache hat unbeschränkte überkreuzende Abhängigkeiten $\rightarrow$ nicht kontext-frei





Position der natürlichen Sprachen

Wichtige Eigenschaften kontext-freier Sprachen

Theorem 4: Kontext-freie Sprachen sind abgeschlossen unter Schnitt mit regulären Sprachen: Wenn L_1 regulär ist und L_2 kontext-frei, dann ist $L_1 \cap L_2$ auch kontext-frei.

Position der natürlichen Sprachen

Wichtige Eigenschaften kontext-freier Sprachen

Theorem 5: Die Klasse der kontext-freien Sprachen ist unter Homomorphismen abgeschlossen.

Position der natürlichen Sprachen

Das Pumping-Lemma für kontext-freie Sprachen

Sei L eine kontext-freie Sprache. Dann gibt es eine Konstante n so dass, wenn $z \in L$ und *length*$(z) \geq n$, wir z als $z = uvwxy$ darstellen können, so dass

1. *length*$(vx) \geq 1$
2. *length*$(vwx) \leq n$
3. für alle $i \geq 0 : uv^iwx^iy \in L$.

Position der natürlichen Sprachen

Das respectively-Argument

- Bar-Hillel und Shamir (1960):
  - Englisch enthält die Kopiersprache.
  - Daher kann Englisch nicht kontext-frei sein.
- Betrachte den Satz

  John, Mary, David, ... are a widower, a widow, a widower, ..., respectively.
- Behauptung: der Satz ist nur unter der Bedingung grammatisch, dass, wenn der n-te Name männlich (weiblich) ist, dann die n-te Phrase nach dem Kopulaverb *a widower (a widow)* ist.

Position der natürlichen Sprachen

- Angenommen die Behauptung stimmt.
- Schnittmenge von Englisch mit einer regulären Sprache:

$$L_1 = (\textit{Paul}|\textit{Paula})^+ \textit{are} (\textit{a widower}|\textit{a widow})^+ \textit{respectively}$$

  Englisch $\cap L_1 = L_2$
- Homomorphismus $L_2 \rightsquigarrow L_3$:

$$\begin{aligned} \textit{John, David, Paul, ...} &\mapsto a \\ \textit{Mary, Paula, Betty, ...} &\mapsto b \\ \textit{a widower} &\mapsto a \\ \textit{a widow} &\mapsto b \\ \textit{are, respectively} &\mapsto \varepsilon \end{aligned}$$

Position der natürlichen Sprachen

- Resultat: Kopiersprache L_3

$$\{ww|w \in (a|b)^+\}$$

- Aufgrund des Pumping-Lemmas kann die Kopiersprache nicht kontext-frei sein. (Aufgabe: Warum?)
- Daher ist L_2 nicht kontext-frei.
- Daher ist Englisch nicht kontext-frei.

Position der natürlichen Sprachen

Gegenargument

- Die überkreuzenden Abhängigkeiten im Zusammenhang mit *respectively* sind semantischer Natur, nicht syntaktischer.
- Vergleiche o.g. Beispiel mit

  (Here are John, Mary and David.) They are a widower, a widow and a widower, respectively.

Position der natürlichen Sprachen

Überkreuzende Abhängigkeiten im Niederländischen

- Huybregt (1976):
  - Niederländisch hat eine der Kopiersprache vergleichbare Struktur.
  - Daher ist Niederländisch nicht kontext-frei.

(1) dat Jan Marie Pieter Arabisch laat zien schrijven
DASS JAN MARIE PIETER ARABISCH LÄSST SEHEN SCHREIBEN
'dass Jan Marie Pieter Arabisch schreiben sehen lässt.'

Position der natürlichen Sprachen

Gegenargument

- Überkreuzende Abhängigkeiten betreffen nur die Zuordnung der Objekte zu den Verben, also die Semantik.
- NL hat keine Kasus-Unterscheidungen.
- Soweit die reinen Wortketten betroffen sind, hat das NL die Struktur

$$NP^n V^n,$$

die kontext-frei ist.

Beweis der Nicht-Kontextfreiheit

Deutsch

dass der Karl die Maria
dem Peter den Hans schwimmen lehren helfen lässt

Beweis der Nicht-Kontextfreiheit

Deutsch

dass der Karl die Maria
dem Peter den Hans schwimmen lehren helfen lässt

Niederländisch

dat Karel Marie Piet Jan laat helpen leren zwemmen

Beweis der Nicht-Kontextfreiheit

Deutsch

dass der Karl die Maria
dem Peter den Hans schwimmen lehren helfen lässt

Niederländisch

dat Karel Marie Piet Jan laat helpen leren zwemmen

Beweis der Nicht-Kontextfreiheit

Deutsch

dass der Karl die Maria
dem Peter den Hans schwimmen lehren helfen lässt

Niederländisch

dat Karel Marie Piet Jan laat helpen leren zwemmen

Beweis der Nicht-Kontextfreiheit

Deutsch

dass der Karl die Maria
dem Peter den Hans^m schwimmen lehren^m helfen lässt

Niederländisch

dat Karel Marie Piet Jan^m laat helpen leren^m zwemmen

Beweis der Nicht-Kontextfreiheit

Deutsch

dass der Karl die Maria
dem Peter^n den Hans^m schwimmen lehren^m helfen^n lässt

Niederländisch

dat Karel Marie Piet^n Jan^m laat helpen^n leren^m zwemmen

Beweis der Nicht-Kontextfreiheit

Deutsch

dass der Karl die Maria
dem Peter^n den Hans^m schwimmen lehren^m helfen^n lässt

- Deutsches Fragment entspricht formaler Sprache: $a^m b^n d^n c^m$ — kontext-frei

Niederländisch

dat Karel Marie Piet^n Jan^m laat helpen^n leren^m zwemmen

- NL Fragment entspricht formaler Sprache: $a^m b^n c^m d^n$ — nicht kontext-frei

Beweis der Nicht-Kontextfreiheit

Schweizerdeutsch

dass de Karl d'Maria en Peter^n de Hans^m laat hälfe^n lärne^m schwüme

Beweis der Nicht-Kontextfreiheit

Schweizerdeutsch

dass de Karl d'Maria en Peter^n de Hans^m laat hälfe^n lärne^m schwüme

Beweis der Nicht-Kontextfreiheit

Schweizerdeutsch

dass de Karl d'Maria en Peter^n de Hans^m laat hälfe^n lärne^m schwüme

- Scheizerdeutsches Fragment entspricht formaler Sprache: $a^m b^n c^m d^n$ — nicht kontext-frei

Beweis der Nicht-Kontextfreiheit

- Homomorphismus h:
  - em Peter $\mapsto$ a
  - de Hans $\mapsto$ b
  - hälfe $\mapsto$ c
  - lärne $\mapsto$ d
  - *alles andere* $\mapsto \varepsilon$

Beweis der Nicht-Kontextfreiheit

- Homomorphismus h:
  - em Peter $\mapsto$ a
  - de Hans $\mapsto$ b
  - hälfe $\mapsto$ c
  - lärne $\mapsto$ d
  - *alles andere* $\mapsto \varepsilon$
- Sei L das Schweizerdeutsche
- $h(L) \cap \{a^k b^l c^m d^n | k, l, m, n \geq 0\} = \{a^m b^n c^m d^n | m, n \geq 0\}$
- $\{a^m b^n c^m d^n | m, n \geq 0\}$ ist nicht kontext-frei
- Daher ist auch Schweizerdeutsch nicht kontext-frei!