Die Mächtigkeit natürlicher Sprachen

Was macht das Schweizerdeutsche so komplex?

Wiebke Petersen
Institut für Sprache und Information
Abteilung Computerlinguistik
Heinrich-Heine-Universität Düsseldorf

Computerlinguistik: Anwendungen

- maschinelle Übersetzung
- natürlichsprachliches Interface für Computeranwendungen
- automatische Spracherkennung
- automatische Zusammenfassung
- intelligente Suchmaschinen
- ...

In Anwendungen reichen i.d.R. Shallow-Prozesse

Theoretische Computerlinguistik

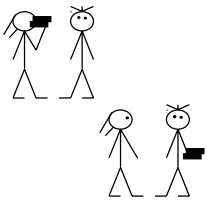
- **maschinelle Verarbeitung natürlicher Sprachen**
- **Grammatikformalismen:**
  - Formalismen zur Entwicklung von Modellen von Einzelsprachen (Grammatiken)
  - Generierung und Analyse von Sätzen (möglichst effizient)
- **Linguistik als empirische Wissenschaft**
  - Naturwissenschaft
- **Kognitionswissenschaft:**
  - starker kognitionswissenschaftlicher Ansatz:
    - maschinelle Sprachverarbeitung als Modell der menschlichen Sprachverarbeitung
  - vollständige, keine Shallow-Analysen

Was macht Sprachwissen aus?

- Phonetisches und Phonologisches Wissen
- Morphologisches Wissen
- Syntaktisches Wissen
- Semantisches Wissen
- Pragmatisches Wissen
- Diskurswissen
- Weltwissen

Was unterscheidet Natürliche Sprachen von z.B. Programmiersprachen?

- Ambiguitäten
  - lexikalische Ambiguitäten
    - (Ruf morgen an - Der Ruf der Möwen)
  - strukturelle Ambiguitäten
    - Die Frau sieht den Mann mit dem Fernrohr
      Die Frau sieht den Mann mit dem Fernrohr



- einzige Experten: Menschen
  - nur endlich viele Sätze abfragbar
  - nicht immer konsistente Antworten

Chomsky Hierarchie (1956)

- **reguläre Sprachen** (Typ 3, REG): $A \rightarrow bA$ a^*b^*
  - endliche Automaten (Wortproblem in linearer Zeit lösbar)
- **kontextfreie Sprachen** (Typ2, CF): $A \rightarrow \beta$ a^nb^n ww^{-1}
  - nichtdeterministische Kellerautomaten (WP in kubischer Zeit lösbar)
- **kontextsensitive Sprachen** (Typ1, CS):
  $\alpha A \eta \rightarrow \alpha\beta\eta$ ww $a^nb^nc^n$ a^{2^n}
  - linear beschränkte Automaten (WP in exponentieller Zeit lösbar)
- **rekursiv aufzählbare Sprachen** (Typ0, RE):
  $\alpha \rightarrow \alpha'$
  - Turingmaschinen (WP nicht entscheidbar)

$$REG \subset CF \subset CS \subset RE$$

Wo liegt die Klasse der natürlichen Sprachen?

Warum ist das eine interessante Frage?

- erlaubt Rückschlüsse auf Adäquatheit eines Grammatikformalismus für NL
- unter CL Aspekten sind möglichst effizient verarbeitbare Analysen gefragt
- erlaubt Rückschlüsse auf menschliche Sprachverarbeitung

Welche Idealisierungen über NL sind nötig?

1. Es gibt die Familie der natürlichen Sprachen
   - alle nat. Sprachen sind strukturell ähnlich
   - alle nat. Sprachen haben eine ähnliche generative Kapazität
2. nat. Sprache = Menge von Ketten über Alphabet
   - Muttersprachler haben volle Kompetenz
     - eindeutige Grammatikalitätsurteile
3. NL $\subset$ RE
4. nat. Sprache = unendliche Menge von Ketten

Zu den Idealisierungen (1)

Es gibt die Familie der natürlichen Sprachen:

- alle nat. Sprachen sind strukturell ähnlich
- alle nat. Sprachen haben eine ähnliche generative Kapazität

- Alle Sprachen haben gleiche Funktion
- Jeder Mensch kann jede Sprache im Erstspracherwerb erlernen (in ähnlicher Zeit)

➢ Kein Hinweis auf prinzipiell unterschiedliche Struktur natürlicher Sprachen

Zu den Idealisierungen (2)

nat. Sprache = Menge von Ketten

- Muttersprachler haben volle Kompetenz
  - eindeutige Grammatikalitätsurteile

- Kompetenz $\leftrightarrow$ Performanz

Grammatikalitätsurteile (Matthews 1979)

(1) The canoe floated down the river sank.

(2) The editor authors the newspaper hired liked laughed.

(3) The man (that was) thrown down the stairs died.

(4) The editor (whom) the authors the newspaper hired liked laughed.

Grammatikalitäsurteile: abhängig von der Situation?

Zu den Idealisierungen (2)

nat. Sprache = Menge von Ketten

- Muttersprachler haben volle Kompetenz
  - eindeutige Grammatikalitätsurteile

- aber: inkonsistente Grammatikalitätsurteile
  - Annahme: Grade von Grammatikalität
    - in Experimenten sind graduelle Urteile nicht konsistent
    - ja/nein-Grammatikalitätsurteile sind innerhalb eines Experiments erstaunlich konsistent

- Lösung: Jede Testsituation definiert eine eigene Sprache

Zu den Idealisierungen (3)

NL $\subseteq$ RE

- Matthews 1979: Mensch hat kein effizientes Verfahren zur Sprachanalyse sondern nur ein Bündel von heuristischen Methoden
  ($\Rightarrow$ inkonsistente Urteile)
- Rogers 1967: Naturgesetze sind universell
  $\Rightarrow$ Churchs These ist universell
- Testprozedur
  + heuristische Strategien (menschliches Orakel)
  + Churchs These
  $\Rightarrow$ NL ist Typ 0

Zu den Idealisierungen (4)

nat. Sprache = unendliche Menge von Ketten

- endliches Alphabet
  - Donaudampfschiffskapitänsmützenschirm...
- Rekursion in natürlichen Sprachen
  - 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.

Zu den Idealisierungen (4)

nat. Sprache = unendliche Menge von Ketten

- beliebig lange Sätze entstehen z.B. durch Konjunktion
  - Sie sah den Hasen und den Hund und die Katze und die Maus und den Käse und ...

konsistente Grammatikalitätsurteile!

Kornai (1985): NL sind regulär

Selbsteinbettende Strukturen in NL sind nicht iterativ

- Dies ist derJunge, 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
  - **Simplizitätsgebot** => Wenn es keine nichtregulären Strukturen in NL gibt, dann beschreibe NL mit regulären Methoden

NL und Chomsky Hierarchie

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 analyses"

Chomsky (1965):

- Transformationsgrammatik
- Peters and Ritchies Theorem (1973)

NL $\subseteq$ CF? fehlerhafte Argumente

- die Frau sieht den Hund
  - die Frauen sehen den Hund
- die Frau, die gestern auf den Baum, der letztes Jahr, als es soviel geregnet hat, gepflanzt wurde, geklettert ist, sieht den Hund

  - die Frauen, die gestern auf den Baum, der letztes Jahr, als es soviel geregnet hat, gepflanzt wurde, geklettert sind, sehen den Hund

NL ⊆ CF? fehlerhafte Argumente

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

- (description of auxiliary-initial interrogatives) "Since **there seems to be no way** of using such PS rules to represent an obviously signigicant generalization about one language, namely, English, we can be sure that phrase structure grammars cannot possibly represent all the significant aspects of language structure."

NL ⊆ CF? fehlerhafte Argumente

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 ... this type of syntactic dependency can extend as memory or patience permits ...

  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**."

NL $\subseteq$ CF? fehlerhafte Argumente

Transformational grammar (Grinder & Elgin, 1973):

- the defining characteristic of a context-free rule is that the symbol to be rewritten **is to be rewritten without reference to the context** in which it occurs ... Thus by definition, one cannot write a context-free rule that will expand the symbol **V** into *kiss* in the context of being immediately preceded by the sequence *the girls* and that will expand the symbol **V** into *kisses* in the context of being immediately preceded by the sequence *the girl*.

Falsche Ansätze

- unzulässige Induktionen
  - keine bekannte CFG beschreibt Englisch adäquat, also gibt es keine adäquate Beschreibung mit CFG's
- Kontextfreiheit wird intuitiv interpretiert

Gazdar & Pullum (1982 & 1985)

- These: "Alle bis dato präsentierten Argumente für die Nichtkontexfreiheit von NL sind nicht zwingend!"
  1. Folklore
  2. falsche Daten
  3. 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
  $NP_1\ NP_2\ NP_3\ VP_1\ VP_2\ VP_3$ "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
  $NP_1\ NP_2\ NP_3\ VP_3\ VP_2\ VP_1$ "nested dependencies"

NL $\not\subset$ CF: Beweis Shieber 1985

Homomorphismus:	f("laa") = c	f("es huus haend wele") = x
f("d'chind") = a	f("hälfe") = d	f("Jan säit das mer") = w
f("em Hans") = b	f("aastriiche") = y	f(s) = z otherwise

- f(Schweizerdeutsch) $\cap$ wa*b*xc*d*y = $wa^mb^nxc^md^ny$
- $wa^mb^nxc^md^ny$ ist nicht kontextfrei ($\rightarrow$Pumping Lemma)
- wa*b*xc*d*y ist regulär
- kontextfreie Sprachen sind abgeschlossen unter
  - Homomorphismen
  - Schnitt mit regulären Sprachen

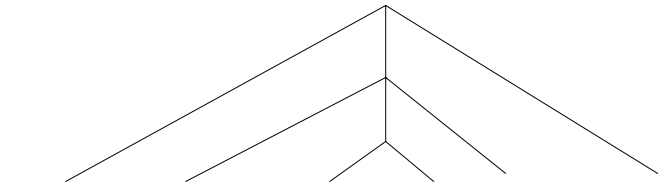
➢ Das Schweizerdeutsch ist nicht kontextfrei

potentielle Angriffspunkte des Beweis

- **falsche Daten**
  - Grammatikalitätsurteile
  - andere Konstituentenstrukturen sind auch möglich
- **Kasus ist nicht syntaktisch**
  - dann wäre Kasus bestimmt durch Semantik
- **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.

Dutch (Huybregts 1976)

dat Jan [Marie Pieter Japaans laat zien schrijven]
dass Jan Marie Pieter Japanisch schreiben sehen lässt



Marie Pieter Japaans laat zien schrijven

starke und schwache generative Kapazität

- Die **schwache generative Kapazität** eines linguistischen Formalismus ist die Eigenschaft alle grammatischen Sätze einer Sprache zu generieren
- Die **starke generative Kapazität** eines linguistischen Formalismus ist die Fähigkeit allen grammatischen Sätzen *ihre Struktur* zuzuweisen
- CFG's ????

Schwach kontextsensitive Sprachen (MCSL)

Schwach kontextsensitive Sprachen
= Teilmenge der kontextsensitiven Sprachen

- beschränktes Wachstum:
  ex. k, so dass für alle $w \in L$ ex. $w' \in L$ mit $|w'| \leq |w| + k$
- Wortproblem $w \in L$ in polynomialer Zeit entscheidbar
- MCSL enthält die folgenden nicht kontextfreien Sprachen:
  - $L_1 = \{a^n b^n c^n \mid n \geq 0\}$ (mehrfache Kongruenz),
  - $L_2 = \{a^n b^m c^n d^m \mid m,n \geq 0\}$ (gekreuzte Abhängigkeiten),
  - $L_3 = \{ww \mid w \in \{a,b\}^*\}$ (Duplikationen)

RL ⊂ CFL ⊂ MCSL ⊂ CSL ⊂ RE

L	CFL	MCSL	CSL
$a^n b^n$	√	√	√
$a^n b^n c^n$, ww	–	√	√
a^{2^n}	–	–	√

These: natürliche Sprachen sind schwach kontextsensitiv

beschränkte mathematische Formalismen

erster Ansatz: erweitere CFG's

- Transformationsgrammatik: CFG + Transformationen
- HPSG: CFG-Gerüst + Merkmalsstrukturen

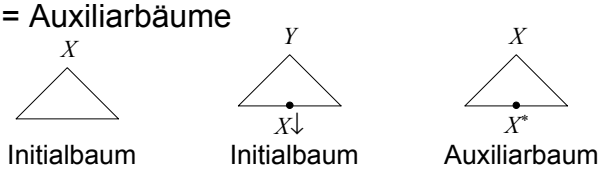
nicht mehr beschränkt!

zweiter Ansatz: ersetze CFG's

- Tree Adjoining Grammar (TAG)
  tree rewriting statt ***string rewriting***

Tree Adjoining Grammars

- TAG G=(Σ,N,I,A,S)
- Σ = Terminale, N = Nichtterminale, S∈N Startsymbol
- I = Initialbäume
- A = Auxiliarbäume

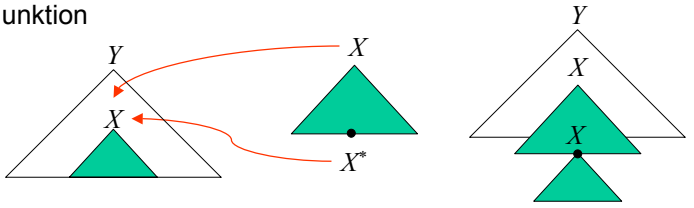


Initialbaum Initialbaum Auxiliarbaum

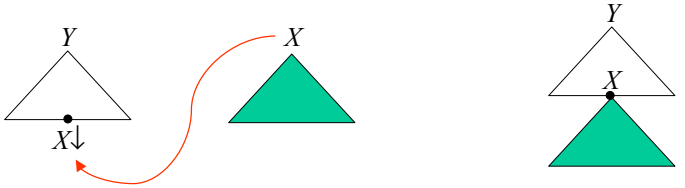
- $X\downarrow$ = Substitutionsknoten
- X^* = Adjunktionsknoten
- I∪A = Elementarbäume

TAG-Operationen

- Adjunktion



- Substitution



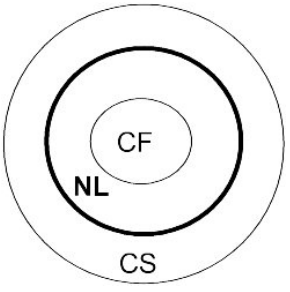
Eigenschaften von TAG-Sprachen

$T(G)=\{t \mid \exists \alpha \in I : \alpha \Rightarrow^* t, \text{root}(\alpha)=S, \text{yield}(\alpha) \in \Sigma^*\}$

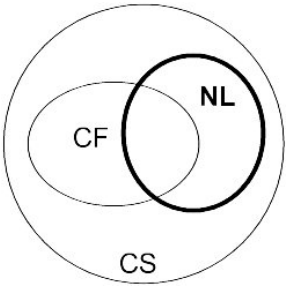
$L(G)=\{\text{yield}(t) \mid t \in T(G)\}$

- $RL \subset CFL \subset TAL \subset MCSL \subset CSL \subset RE$
- Das Wortproblem für TAL ist in polynomialer Zeit entscheidbar ($\mathcal{O}(n^6)$)
- L_1, L_2, L_3 sind in TAL
  - $L_1 = \{a^n b^n c^n \mid n \geq 0\}$ (mehrfache Kongruenz),
  - $L_2 = \{a^n b^m c^n d^m \mid m,n \geq 0\}$ (gekreuzte Abhängigkeiten),
  - $L_3 = \{ww \mid w \in \{a,b\}^*\}$ (Duplikationen)

Was ist, wenn NL unvergleichbar ist zur Chomsky Hierarchie?



?



?

nichtklassische Ansätze

Beispiele für Grammatiken mit regulierte Produktionen (Martín-Vide)

- matrix grammars: (N,T,S,M)
  - die Reihen von M sind Ketten von Produktionen
- context grammars: (N,T,S,P)
  - die Produktionen sind Tripel ($\propto \rightarrow \beta$, Q,R). Q ist der positive und R der negative Kontext der Regel.
- additive valence grammars: (N,T,S,P,v)
  - v: P$\rightarrow\mathbb{Z}$
- ordered grammars: (N,T,S,P,<)
  - < ist eine strikte partielle Ordnung auf P
- ...

abschließend

- endliche Automaten in vielen Anwendungen
  - Phonologie
  - Morphologie
  - ...
- Menschen parsen sehr schnell => niedrige Komplexitätsklasse
- Lernbarkeit von NL muss erklärt werden

Literatur (1)

- **Bach**, Emmon und **Marsh**, William 1987: *An Elementary Proof of the Peters-Ritchie Theorem*. In Savitch et al. 1987, 41-55.
- **Chomsky**, Noam 1956: *Three models for the description of language.* In IRE Transactions on Information Theory 2(3), 113-124.
- **Chomsky**, Noam 1965: *Aspects of a Theory of Syntax*. MIT Press, Cambridge, Massachusetts.
- **Gazdar**, Gerald und **Pullum**, Geoffrey K. 1987[1985]: *Computationally Relevant Properties of Natural Languages and Their Grammars.* In Savitch et al. 1987, 41-55. (1985 erschienen in Technical Report CSLI-85-24.)
- **Joshi**, Aravind K. and **Schabes**, Yves 1997: *Tree-Adjoining Grammars*. In Handbook of Formal Languages, G. Rozenberg and A. Salomaa (Hrg.), Vol. 3, Springer, Berlin, New York, 1997, 69 - 124.
- **Kornai**, Andras 1985: *Natural languages and the Chomsky hierarchy.* In M. King (Hrg.): Proceedings of the 2nd European Conference of the Association for Computational Linguistics (1985),1-7.

Literatur (2)

- **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 1987 [1982]: *Natural Languages and Context-Free Languages.* In Savitch et al. 1987, 138-183. (1982 erschienen in Linguistics and Philosophy, 4:471--504.)
- **Rogers**, Hartley 1967: *Theory of Recursive Functions and Effective Computability*. McGraw-Hill Book Company, New York, 1967.
- **Savitch** et. al. 1987: *The Formal Complexity of Natural Language.* Reidel, Dordrecht.
- **Shieber**, Stuart M. 1987 [1985]. *Evidence against the context-freeness of natural language*. In Savitch et al. 1987, 320-335. (1985 erschienen in Linguistics and Philosophy, 8:333--343.)

Elster 1978 (Pi-Argument)

- The first two million numbers in the decimal expansion of π are $a_1 a_2 a_{3...} a_{2.000.000}$.
  The first two million million numbers in the decimal expansion of π are $a_1 a_{2...} a_{2.000.000.000.000}$.
  ...
  The first two (million)^k numbers in the decimal expansion of π are $a_1 a_2 a_{3...} a_{2 \cdot 10^{6k}}$.
- Pumping Lemma
- The first two (million)$^{k+q}$ numbers in the decimal expansion of π are $a_1 a_{2 ...} a_t a_{r ...} a_{t...} a_{2 \cdot 10^{6k}}$.
- In "The A are B" muss die Zahl der Entitäten in B mit der Zahl A übereinstimmen:
  - * The two largest animals in a zoo are a mouse
  - The two largest animals in a zoo are a mouse and a horse
- aber: The two largest animals in the zoo are Mickey, Minnie and Donald.

respectively-Konstruktion (Langendoen 1977)

- S = (the woman + the men)$^+$ and (the woman + the men) (smokes and drink)$^+$ and (smokes and drink) respectively.
- S ∩ Englisch = S'
- S'= {*xx' respectively* | x ∈ *L*= ((the woman + the men)$^+$ and (the woman + the men)) und x' ist die entsprechende Kette, in der *the woman* mit *smokes* und *the men* mit *drink* ersetzt sind}
- **aber:**
  * the woman and the men smokes and drink respecitvely
  * the man and the woman smokes and drinks respectively
  the man and the woman smoke and drink respectively
  Ira, Walt, and Louise have been dating Frank, Edith, Cedric, and Bruce, respectively