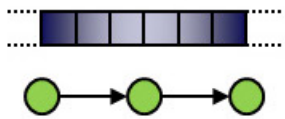
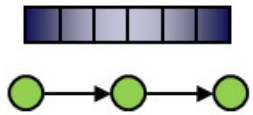
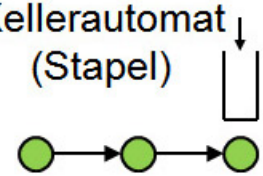
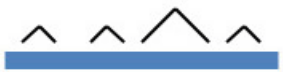
Automatentheorie und formale Sprachen
Entscheidbarkeitsprobleme
endliche Transduktoren
Komplexität natürlicher Sprachen

Dozentin: Wiebke Petersen

15.7.2009

Chomsky-Hierarchie & Automaten

Sprache	*Automat*	*Grammatik*	*Erkennung*	*Abhängigkeit*
rekursiv aufzählbar	Turing Maschine 	unbeschränkt Baa → ε	unentscheidbar	beliebig
kontext-sensitiv	linear gebunden 	kontext-sensitiv γAδ → γβδ	NP-vollständig 	überkreuzt 
kontext-frei	Kellerautomat (Stapel) 	kontextfrei C → bABa	polynomiell 	eingebettet 
regulär	endlicher Automat 	regulär A → bA	linear 	strikt lokal 

Vokabular zur Theorie der Entscheidbarkeit

Algorithmus: Eine aus endlich vielen Schritten bestehende Verarbeitungsvorschrift, die, mechanisch angewandt zur Lösung eines Problems führt.

Entscheidbarkeit: Ein Problem ist entscheidbar, wenn ein Algorithmus existiert, der bei Eingabe einer Instantiierung des Problems nach endlich vielen Schritten angibt, ob dieses lösbar ist oder nicht.

Entscheidbarkeitsprobleme

Gegeben: Grammatiken $G = (N, \Sigma, S, P)$, $G' = (N', \Sigma, S', P')$, Wort $w \in \Sigma^*$

Wortproblem Ist w in G ableitbar?

Leerheitsproblem Erzeugt G eine nichtleere Sprache?

Äquivalenzproblem Erzeugen G und G' die gleichen Sprachen $(L(G) = L(G'))$?

Ergebnisse zu Entscheidbarkeitsproblemen

	Typ3	Typ2	Typ1	Typ0
Wortproblem	E	E	E	U
Leerheitsproblem	E	E	U	U
Äquivalenzproblem	E	U	U	U

E steht für entscheidbar

U steht für unentscheidbar

Entscheidbarkeitsprobleme für kontextfreie Sprachen

Wortproblem: Argumentation über Wortlänge

Leerheitsproblem: Markiere die Symbole der Regeln aus denen ein Terminalwort ableitbar ist (wenn Startsymbol markiert, dann ist die Sprache nicht leer).

Hausaufgaben

1. Überlegen sie sich, warum das Wort- und das Leerheitsproblem für reguläre Sprachen entscheidbar ist (argumentieren sie mit endlichen Automaten).

Endliche Transduktoren

Parsing/Generierung vs. Erkennen

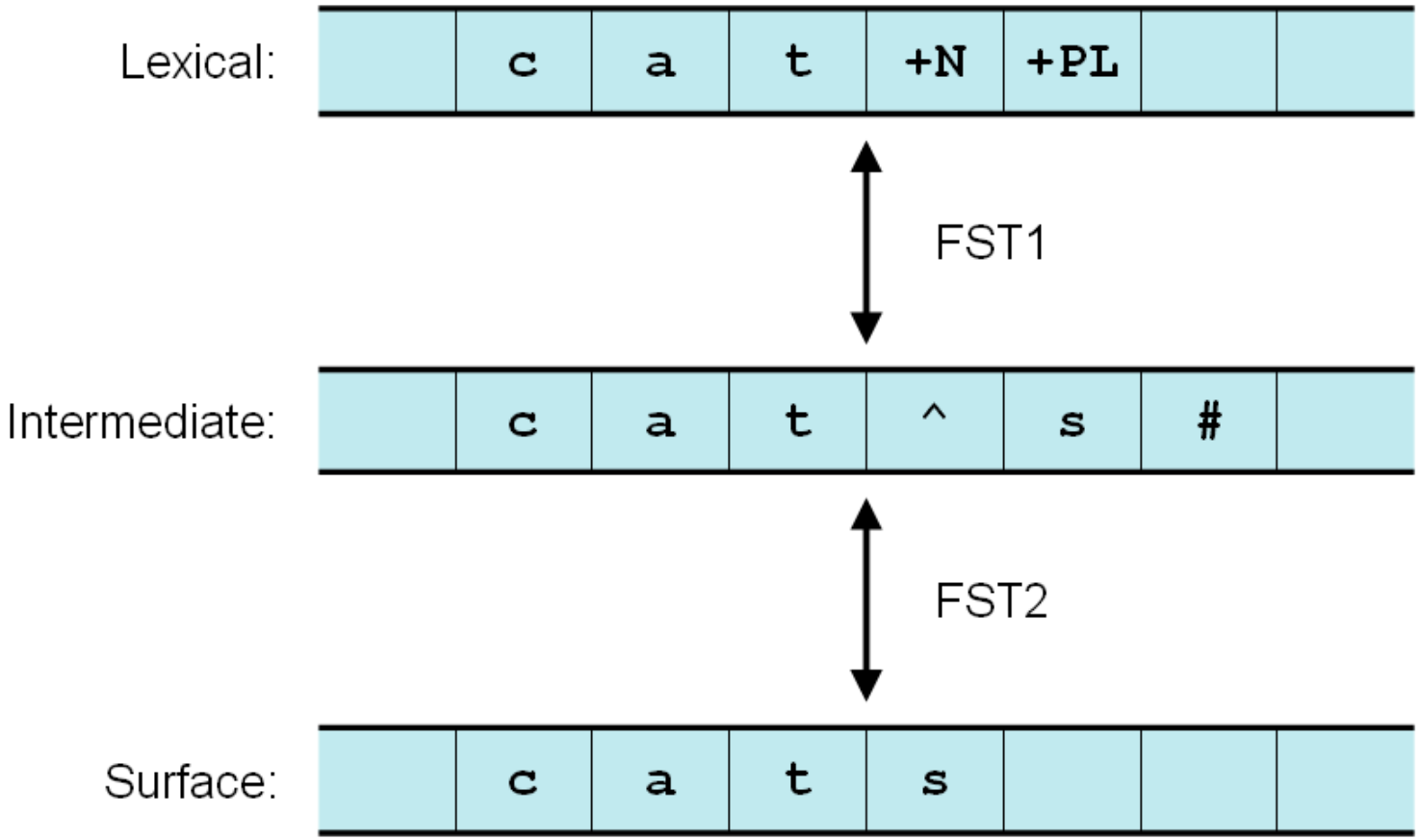
- Endliche Automaten sind lediglich in der Lage zu erkennen, ob ein Wort zur Sprache gehört oder nicht (Anwendungsgebiete: z. B. Rechtschreibprüfung).
- Häufig wird mehr gefordert:
  - Gegeben ein sprachlicher Ausdruck, gesucht wird seine Struktur (Parsing)
  - Gegeben eine Struktur, gesucht die Oberflächenform bzw. die sprachliche Realisierung (Generierung).

Endliche Transduktoren / Finite State Transducers (FST)

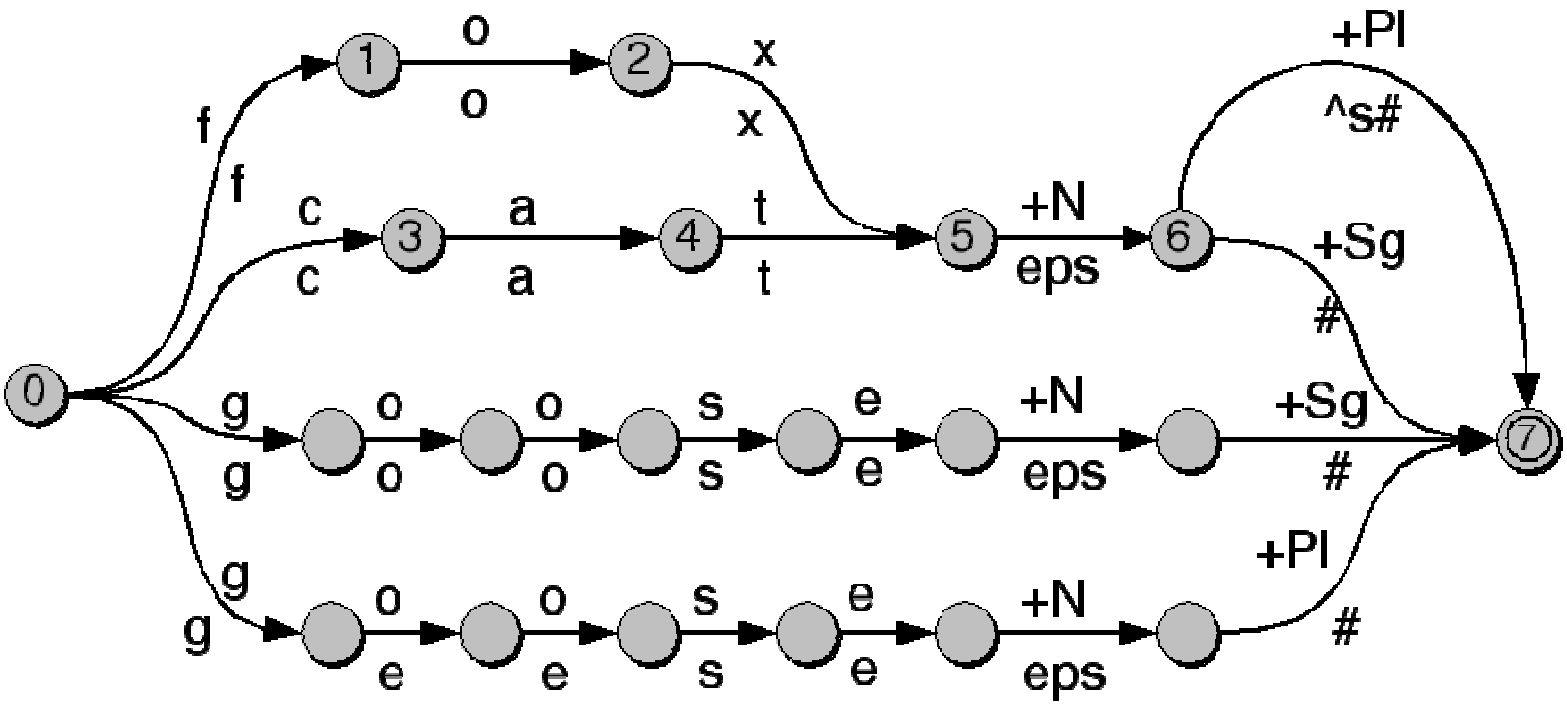
Einfache Idee:

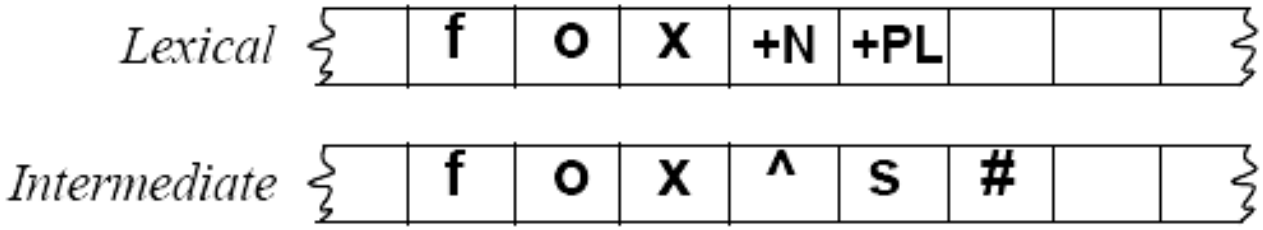
- Hinzufügen eines zweiten Bands (Schreibband) zu dem vorhandenen Leseband.
- Hinzufügen spezieller Symbole zu den Übergängen.
- Beispiel: auf dem einen Band wird "`cats`" gelesen auf dem anderen "`cat +N +PL`" geschrieben.

Endliche Transduktoren werden häufig hintereinander geschaltet:



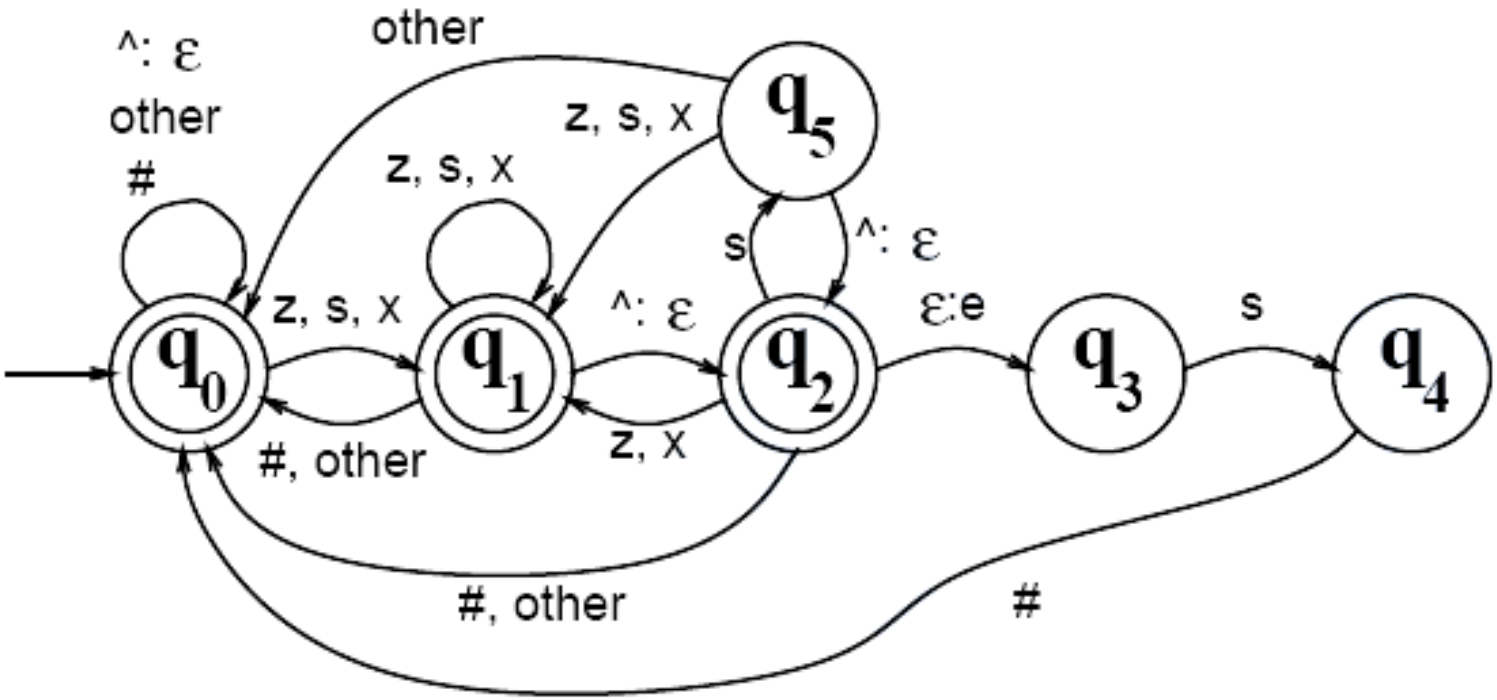
FST1: lexical ↔ intermediate

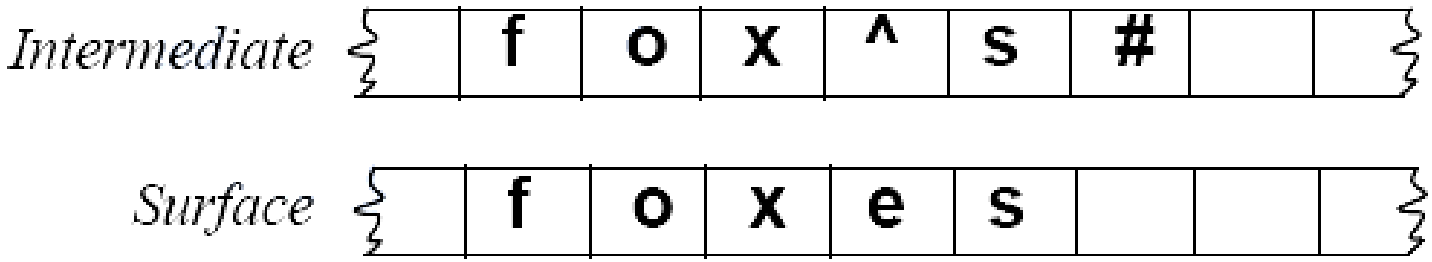




(aus Jurafsky & Martin 2008)

FST2: intermediate ↔ surface





(aus Jurafsky & Martin 2008)

Which is the class of natural languages?

Why is the formal complexity of natural languages interesting?

- It gives information about the general structure of natural language

Which is the class of natural languages?

Why is the formal complexity of natural languages interesting?

- It gives information about the general structure of natural language
- It allows to draw conclusions about the adequacy of grammar formalisms

Which is the class of natural languages?

Why is the formal complexity of natural languages interesting?

- It gives information about the general structure of natural language
- It allows to draw conclusions about the adequacy of grammar formalisms
- It determines a lower bound for the computational complexity of natural language processing tasks

Which is the class of natural languages?

Why is the formal complexity of natural languages interesting?

- It gives information about the general structure of natural language
- It allows to draw conclusions about the adequacy of grammar formalisms
- It determines a lower bound for the computational complexity of natural language processing tasks
- It allows to draw conclusions about human language processing

Which idealizations about NL are necessary?

1. The family of natural languages exists:
   - all natural languages are structurally similar
   - all natural languages have a similar generative capacity

Which idealizations about NL are necessary?

1. The family of natural languages exists:
   - all natural languages are structurally similar
   - all natural languages have a similar generative capacity
2. Language = set of strings over an alphabet:
   - native speakers have full competence
   - consistent grammaticality judgements

Which idealizations about NL are necessary?

1. The family of natural languages exists:
   - all natural languages are structurally similar
   - all natural languages have a similar generative capacity
2. Language = set of strings over an alphabet:
   - native speakers have full competence
   - consistent grammaticality judgements
3. $NL \subset RE$
   - each natural language is describable by a formal grammar (a finite rule system)

Which idealizations about NL are necessary?

1. The family of natural languages exists:
   - all natural languages are structurally similar
   - all natural languages have a similar generative capacity
2. Language = set of strings over an alphabet:
   - native speakers have full competence
   - consistent grammaticality judgements
3. $NL \subset RE$
   - each natural language is describable by a formal grammar (a finite rule system)
4. Each NL consists of an *infinite* set of strings

About the idealizations

The family of natural languages exists:

- all natural languages are structurally similar
- all natural languages have a similar generative capacity

About the idealizations

The family of natural languages exists:

- all natural languages are structurally similar
- all natural languages have a similar generative capacity

Arguments:

- all NLs serve for the same tasks
- children can learn each NL as their native language (within a similar period of time)

$\Rightarrow$ No evidence for a principal structural difference

About the idealizations (cont.)

Language = sets of strings over an alphabet:

- native speakers have full competence
- consistent grammaticality judgements

About the idealizations (cont.)

Language = sets of strings over an alphabet:

- native speakers have full competence
- consistent grammaticality judgements

Arguments:

- all mistakes are due to performance not to competence
- Mathews (1979) counter examples:
  - The canoe floated down the river sank.
  - The editor authors the newspaper hired liked laughed.
  - The man (that was) thrown down the stairs died.
  - The editor (whom) the authors the newspaper hired liked laughed.

About the idealizations (cont.)

$NL \subset RE$:

- each natural language is describable by a formal grammar (a finite rule system)

Arguments:

Rogers (1967)

- Laws of nature are universal
- Church's thesis is universal
- human oracle + Church's thesis $\Rightarrow$ NL is RE

About the idealizations (cont.)

Each NL consists of an *infinite* set of strings

About the idealizations (cont.)

Each NL consists of an *infinite* set of strings

Arguments:

- Recursion in NL:
  - john likes peter
  - john likes peter and mary
  - john likes peter and mary and sue
  - john likes peter and mary and sue and otto and ...
- (Donaudampfschiffskapitänsmützenschirm ...)

Are natural languages regular?

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"

Are natural languages regular? No!

- 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 ($n > 0$)

Kornai (1985): NL are regular

Self-embedding (nested) structures in NL are not iterative!

> This is the woman whom the man whom the girl whom the boy whom the teacher whom the doctor admired met called chased

Kornai (1985): NL are regular

Self-embedding (nested) structures in NL are not iterative!

> This is the woman whom the man whom the girl whom the boy whom the teacher whom the doctor admired met called chased liked

NL $\subseteq$ CFL? wrong arguments

inadmissible induction:

- no known CFG describes English adequately, thus no adequate description with CFG's exists.

NL $\subseteq$ CFL? wrong arguments

inadmissible induction:

- no known CFG describes English adequately, thus no adequate description with CFG's exists.

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 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."

NL $\subseteq$ CFL? wrong arguments

"context-freeness" intuitively understood:

NL $\subseteq$ CFL? wrong arguments

"context-freeness" intuitively understood:

- the girl sees the dog
- the girls see the dog

NL $\subseteq$ CFL? wrong arguments

"context-freeness" intuitively understood:

- 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

NL $\subseteq$ CFL? wrong arguments

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*."

NL $\subseteq$ CFL? wrong arguments

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**."

Gazdar & Pullum (1982 & 1985)

- thesis: all published arguments for the non-context-freeness of NL are not compelling
  1. folklore
  2. wrong data
  3. formal mistakes
- 30 years of fruitless search for a non-context- free language
- human seem able to parse sentences in linear time

NL $\not\subseteq$ CF: Shieber 1985

Nebensatzeinbettung im Schweizerdeutschen

- Jan säit das
  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"

potential attacks

- wrong data
  - grammaticality judgements

potential attacks

- wrong data
  - grammaticality judgements
- case is not a syntactic phenomenon
  - case is determined by semantics (unterstützen/helfen)

potential attacks

- wrong data
  - grammaticality judgements
- case is not a syntactic phenomenon
  - case is determined by semantics (unterstützen/helfen)
- the length of the sentences is restricted
  - 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."