An introduction to mildly context sensitive grammar formalisms

Gerhard Jäger & Jens Michaelis

University of Potsdam

{jaeger,michael}@ling.uni-potsdam.de

Rewriting systems

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

N ... nonterminal symbols
T ... terminal symbols
S ... start symbol ($S \in N$)
R ... rules

Rules take the form

$$\alpha \rightarrow \beta$$

where α, β *are strings over* $T \cup N$ *and* β *is non-empty.*

The Chomsky Hierarchy

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

  "$\rightarrow^*$" is the reflexive and transitive closure of $\rightarrow$.
- Every recursively enumerable language can be described by a rewriting system.
- (Unrestricted) Rewriting systems are equivalent to Turing machines in expressive power.
- "(Chomsky) Type-0 grammars" = unrestricted rewriting systems
- membership in a type-0 language is **undecidable**

Context-sensitive grammars

- subclass of type-0 grammars
- restriction:
  all rules take the form

  $$\alpha \rightarrow \beta$$

  where

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

- consequence: membership in a context-sensitive language (CSL) is decidable

Context-sensitive grammars

- alternative (original) formulation:

  All rules take the form

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

  where $A \in N$, $\alpha, \beta, \gamma \in (T \cup N)^*, \gamma \neq \varepsilon$
- The two formulations define the same class of languages.
- Not all decidable languages are context-sensitive (but most are).
- Membership problem for CSLs is PSPACE-complete.
- CSGs are expressively equivalent to **linear bounded automata**.

Context-free grammars

- subclass of context-sensitive grammars
- restriction:

  rules take the form

  $$A \rightarrow \alpha$$

  where

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

- Membership in context-free language (CFL) is decidable in **polynomial time** ($O(n^3)$).
- CFG are expressively equivalent ot **pushdown automata**.

Regular grammars

- subclass of context-free grammars
- restriction:

  rules take the form

$$A \to B$$

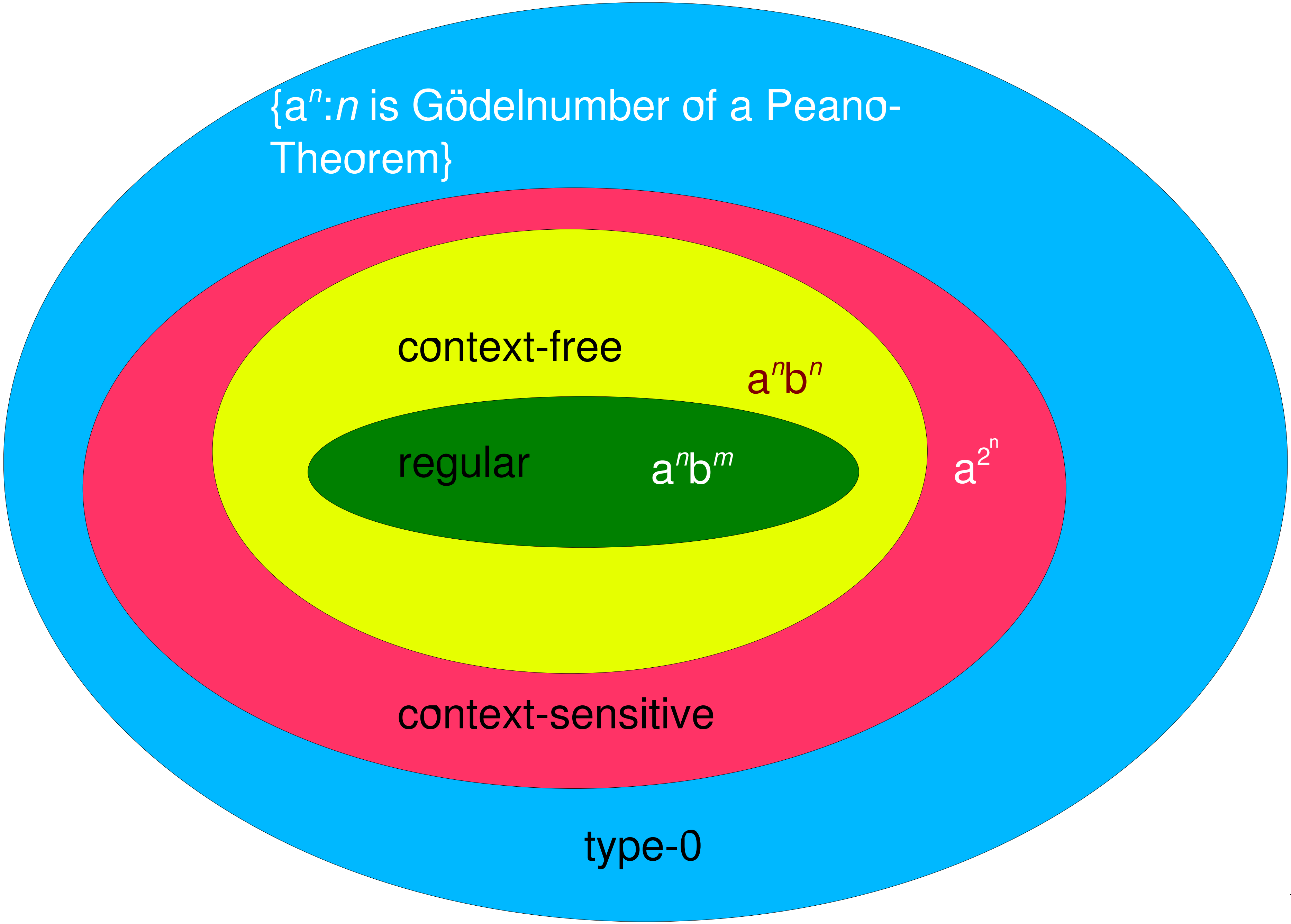
  or

$$A \to Ba$$

  where $A, B \in N$ *and* $a \in T$

- Membership is decidable in **linear time**.
- RGs are expressively equivalent to **finite state automata**.

The Chomsky Hierarchy



Where are natural languages located?

- hotly contested issue over several decades
- typical argument:
  - find a recursive construction C in a natural language L
  - argue that the **competence** of speakers admits unlimited recursion (while the performance certainly poses an upper limit)
  - reduce C to a formal language L' of known complexity via **homomorphisms**
  - make a case that L must be at least as complex as L'
  - extrapolate to all human languages: if there is one languages which is at least as complex as ..., then the human language faculty must allow it in general

Are natural languages regular?

Chomsky 1957: Natural languages are not regular.
Structure of his argument:

- Consider 3 hypothetical languages:
  1. $ab, aabb, aaabbb$ ($a^n b^n$)
  2. $aa, bb, abba, baab, aaaa, bbbb, aabbaa, abbbba,$ ... (palindromic)
  3. $aa, bb, abab, baba, aaaa, bbbb, aabaab, abbabb, aababaabab$ (copy language)
- can easily be shown that these are not regular languages
- also languages like 1, 2 and 3 except allowing for embeddings of as and bs are not regular
- natural language is infinitely recursive

- The following constructions can be arbitrarily embedded into each other:
  - If S_1, then S_2.
  - Either S_3 or S_4.
  - The man that said that S_5 is arriving today.
- Therefore—Chomsky says—English cannot be regular.

“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 languages.”

(Chomsky 1957)

Closure properties of regular languages

Theorem 1: If L_1 and L_2 are regular languages, then $L_1 \cap L_2$ is also a regular language.

Theorem 2: The class of regular languages is closed under homomorphism.

Theorem 3: The class of regular languages is closed under inversion.

- homomorphism:

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

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

$$\textit{everything else} \mapsto \varepsilon$$

If it neither rains nor snows, then if it rains then it snows.

$$\mapsto ab$$

- maps English not to the mirror language, but to the language L_1:

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

The pumping lemma for regular languages

Let L be a regular language. Then there is a constant n such that if z is any string in L, and *length*$(z) \geq n$, we may write $z = uvw$ in such a way that *length*$(uv) \leq n$, $v \neq \varepsilon$, and for all $i \geq 0, uv^i w \in L$.

NL and the Chomsky Hierarchy

- Suppose English is regular.
- Due to closure under homomorphism, L_1 is reglar.
- a^*b^* is a regular language. (exercise: why?)
- Thus $a^*b^* \cap L_1$ is a regular language

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

due to Theorem 1

NL and the Chomsky Hierarchy

- Due to closure under inversion and homomorphism,

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

  is also regular.
- Hence L_4 is regular:

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

- L_4 cannot be regular due to the pumping lemma
- Therefore English cannot be a regular language.

Dissenting view:

- *all arguments to this effect use center-embedding*
- *humans are extremely bad at processing center-embedding*
- *notion of* competence *that ignores this is dubious*
- *natural languages are regular after all*

Exercises:

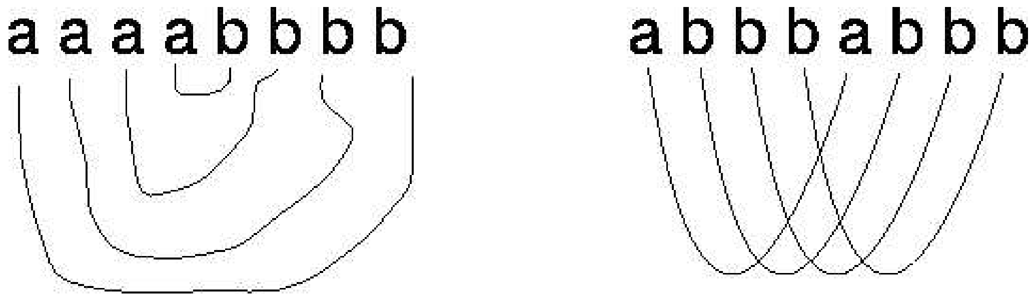
Show that Chomsky correctly classified $a^n b^n$, the mirror language, and the copy language as non-regular!

Are natural languages context-free?

- history of the problem:
  - Chomsky 1957: conjecture that natural languages are not cf
  - sixties, seventies: many attempts to prove this conjecture
  - Pullum and Gazdar 1982:
    - all these attempts have failed
    - for all we know, natural languages (conceived as string sets) might be context-free
  - Huybregts 1984, Shieber 1985: proof that Swiss German is not context-free
  - Culy 1985: proof that Bambara is not context-free

Nested and crossing dependencies

- CFLs—unlike regular languages—can have unbounded dependencies
- however, these dependencies can only be **nested**, not **crossing**
- example:
  - $a^n b^n$ has unlimited nested dependencies $\rightarrow$ context-free
  - the copy language has unlimited crossing dependencies $\rightarrow$ not context-free



Important properties of CFLs

Theorem 4: CFLs are closed under intersection with regular languages: If L_1 is a regular language and L_2 is context-free, then $L_1 \cap L_2$ is also context-free.

Important properties of CFLs

Theorem 5: The class of context-free languages is closed under homomorphism.

The pumping lemma for context-free languages

Let L be any CFL. Then there is a constant n, depending only on L, such that if z is in L and *length*$(z) \geq n$, then we may write $z = uvwxy$ such that

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

The *respectively* argument

- Bar-Hillel and Shamir (1960):
  - English contains copy-language
  - cannot be context-free
- Consider the sentence

  John, Mary, David, ... are a widower, a widow, a widower, ..., respectively.
- Claim: the sentence is only grammatical under the condition that if the nth name is male (female) then the nth phrase after the copula is *a widower (a widow)*

- suppose the claim is true
- intersect English with regular language

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

English $\cap L_1 = L_2$

- homomorphism $L_2 \rightsquigarrow L_3$:

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

- result: copy language L_3

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

- copy language is not cf due to pumping lemma (exercise: why is this so?)
- hence L_2 is not cf
- hence English is not cf

Counterargument

- crossing dependencies triggered by *respectively* are semantic rather than syntactic
- compare above example to

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

Cross-serial dependencies in Dutch

- Huybregt (1976):
  - Dutch has copy-language like structures
  - thus Dutch is not context-free

(1) dat Jan Marie Pieter Arabisch laat zien schrijven
THAT JAN MARIE PIETER ARABIC LET SEE WRITE
'that Jan let Marie see Pieter write Arabic'

Counterargument

- crossing dependencies only concern argument linking, i.e. semantics
- Dutch has no case distinctions
- as far as plain string are concerned, the relevant fragment of Dutch has the structure

$$NP^n V^n$$

  which is context-free