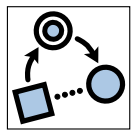
Komplexität und natürliche Sprache

Einordnung natürlicher Sprache in die Chomsky-Hierarchie

Timm Lichte & Christian Wurm

HHU Düsseldorf

SS 2017, 03.05.2015



Kursübersicht

Vier Formen der Komplexität:

- **Extensionale Komplexität**
- Algorithmische Komplexität / Verarbeitungskomplexität
- Beschreibungskomplexität
- Informatische Komplexität

Ontologie

Wir betrachten Stringsprachen:

$$a \triangleleft \overline{w} \in L_G \in \{L_G | G \in \hat{G}\}$$

Ontologie

Wir betrachten Stringsprachen:

$$a \triangleleft \overline{w} \in L_G \in \{L_G | G \in \hat{G}\}$$

a	ist ein **Buchstabe** und Element eines Alphabets Σ.
$\overline{w}$	ist ein **Wort** und Element einer Sprache L.
L_G	ist die **Sprache** der Grammatik G.
$\hat{G}$	ist eine Menge/Klasse von Grammatiken.

Ontologie

Wir betrachten Stringsprachen:

$$a \triangleleft \overline{w} \in L_G \in \{L_G | G \in \hat{G}\}$$

a	ist ein **Buchstabe** und Element eines Alphabets Σ.
$\overline{w}$	ist ein **Wort** und Element einer Sprache L.
L_G	ist die **Sprache** der Grammatik G.
$\hat{G}$	ist eine Menge/Klasse von Grammatiken.

Beispiele:

$\Sigma = \{a, b\}$

Ontologie

Wir betrachten Stringsprachen:

$$a \triangleleft \overline{w} \in L_G \in \{L_G | G \in \hat{G}\}$$

a	ist ein **Buchstabe** und Element eines Alphabets Σ.
$\overline{w}$	ist ein **Wort** und Element einer Sprache L.
L_G	ist die **Sprache** der Grammatik G.
$\hat{G}$	ist eine Menge/Klasse von Grammatiken.

Beispiele:

$$\Sigma = \{a, b\}$$
$$\overline{w} = abba$$

Ontologie

Wir betrachten Stringsprachen:

$$a \triangleleft \overline{w} \in L_G \in \{L_G | G \in \hat{G}\}$$

a	ist ein **Buchstabe** und Element eines Alphabets Σ.
$\overline{w}$	ist ein **Wort** und Element einer Sprache L.
L_G	ist die **Sprache** der Grammatik G.
$\hat{G}$	ist eine Menge/Klasse von Grammatiken.

Beispiele:

$$\begin{aligned} \Sigma &= \{a, b\} \\ \overline{w} &= abba \\ L &= \{abba, abab, bbb, \ldots\} \end{aligned}$$

Ontologie

Wir betrachten Stringsprachen:

$$a \triangleleft \overline{w} \in L_G \in \{L_G | G \in \hat{G}\}$$

a	ist ein **Buchstabe** und Element eines Alphabets Σ.
$\overline{w}$	ist ein **Wort** und Element einer Sprache L.
L_G	ist die **Sprache** der Grammatik G.
$\hat{G}$	ist eine Menge/Klasse von Grammatiken.

Beispiele:

Σ	$=$	$\{a, b\}$	
$\overline{w}$	$=$	$abba$	
L	$=$	$\{abba, abab, bbb, \ldots\}$	
G	$=$	$\langle \mathcal{N}, \Sigma, S, \mathcal{P} \rangle$	(z.B. eine CFG)

Ontologie

Wir betrachten Stringsprachen:

$$a \triangleleft \overline{w} \in L_G \in \{L_G | G \in \hat{G}\}$$

a	ist ein **Buchstabe** und Element eines Alphabets Σ.
$\overline{w}$	ist ein **Wort** und Element einer Sprache L.
L_G	ist die **Sprache** der Grammatik G.
$\hat{G}$	ist eine Menge/Klasse von Grammatiken.

Beispiele:

$$\begin{aligned}
\Sigma &= \{a, b\} \\
\overline{w} &= abba \\
L &= \{abba, abab, bbb, \ldots\} \\
G &= \langle \mathcal{N}, \Sigma, S, \mathcal{P} \rangle \qquad \text{(z.B. eine CFG)} \\
L_G &= \{\overline{w}' | S \vdash^*_{\mathcal{P}_G} \overline{w}'\}
\end{aligned}$$

Relative extensionale Komplexität

Intuition

Je komplexer desto umfangreicher.

Relative extensionale Komplexität

Intuition

Je komplexer desto umfangreicher.

Gegeben zwei Grammatikklassen $\hat{G}$ und $\hat{G}'$:

$K(\hat{G}) > K(\hat{G}') \quad \leftrightarrow \quad$ Für jede $G' \in \hat{G}'$ gibt es eine $G \in \hat{G}$ und $\mathcal{P}_{G'} = \mathcal{P}_G$

Relative extensionale Komplexität

Intuition

Je komplexer desto umfangreicher.

Gegeben zwei Grammatikklassen $\hat{G}$ und $\hat{G}'$:

$K(\hat{G}) > K(\hat{G}') \quad \leftrightarrow \quad$ Für jede $G' \in \hat{G}'$ gibt es eine $G \in \hat{G}$ und $\mathcal{P}_{G'} = \mathcal{P}_G$

$K(\hat{G}) > K(\hat{G}') \quad \leftrightarrow \quad$ Für jede $G' \in \hat{G}'$ gibt es eine $G \in \hat{G}$ und $L_{G'} = L_G$

Relative extensionale Komplexität

Intuition

Je komplexer desto umfangreicher.

Gegeben zwei Grammatikklassen $\hat{G}$ und $\hat{G}'$:

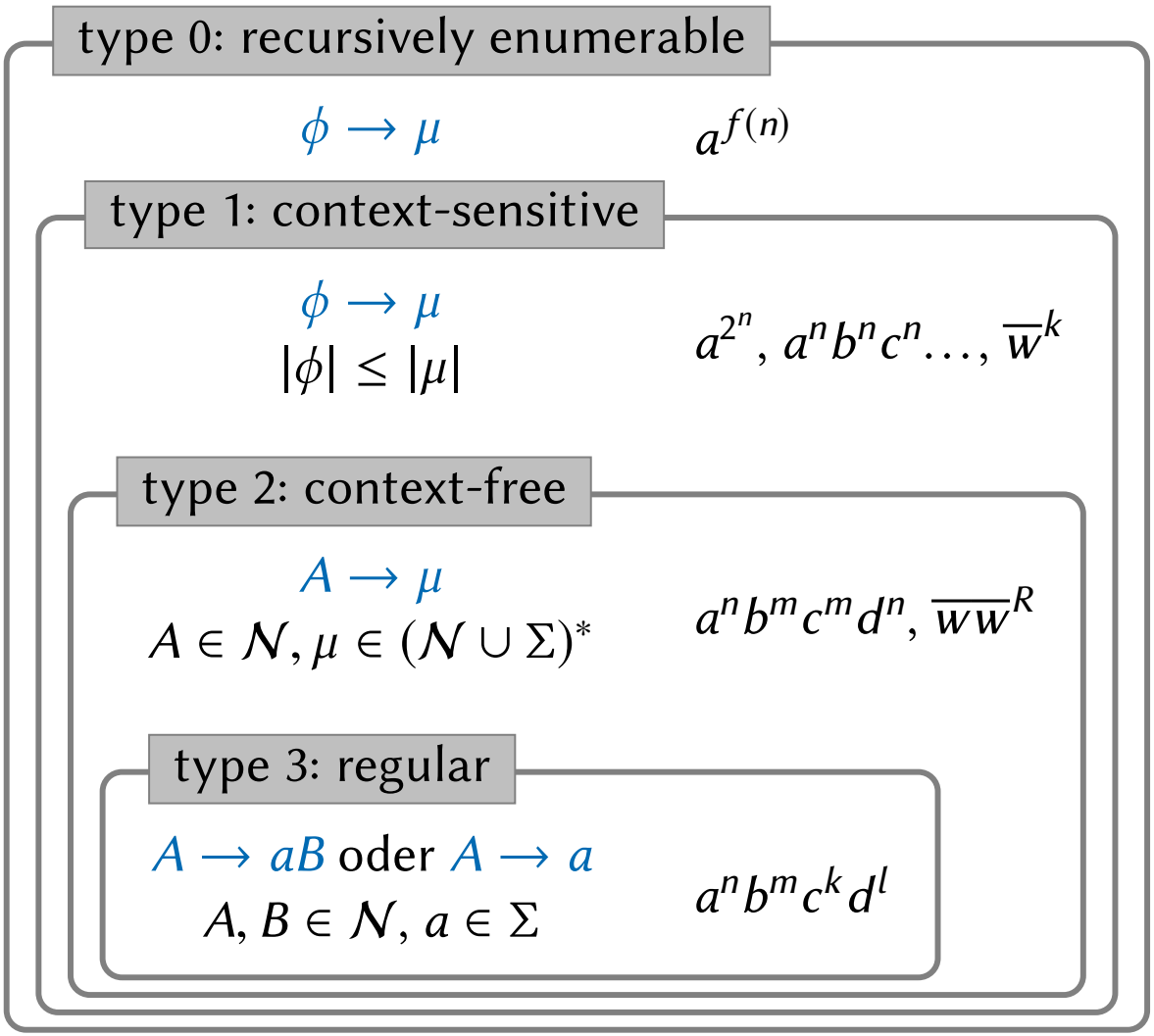
$K(\hat{G}) > K(\hat{G}') \quad \leftrightarrow \quad$ Für jede $G' \in \hat{G}'$ gibt es eine $G \in \hat{G}$ und $\mathcal{P}_{G'} = \mathcal{P}_G$

$K(\hat{G}) > K(\hat{G}') \quad \leftrightarrow \quad$ Für jede $G' \in \hat{G}'$ gibt es eine $G \in \hat{G}$ und $L_{G'} = L_G$

$\Rightarrow$ berühmtestes Beispiel: Chomsky(-Schützenberger)-Hierarchie[6]

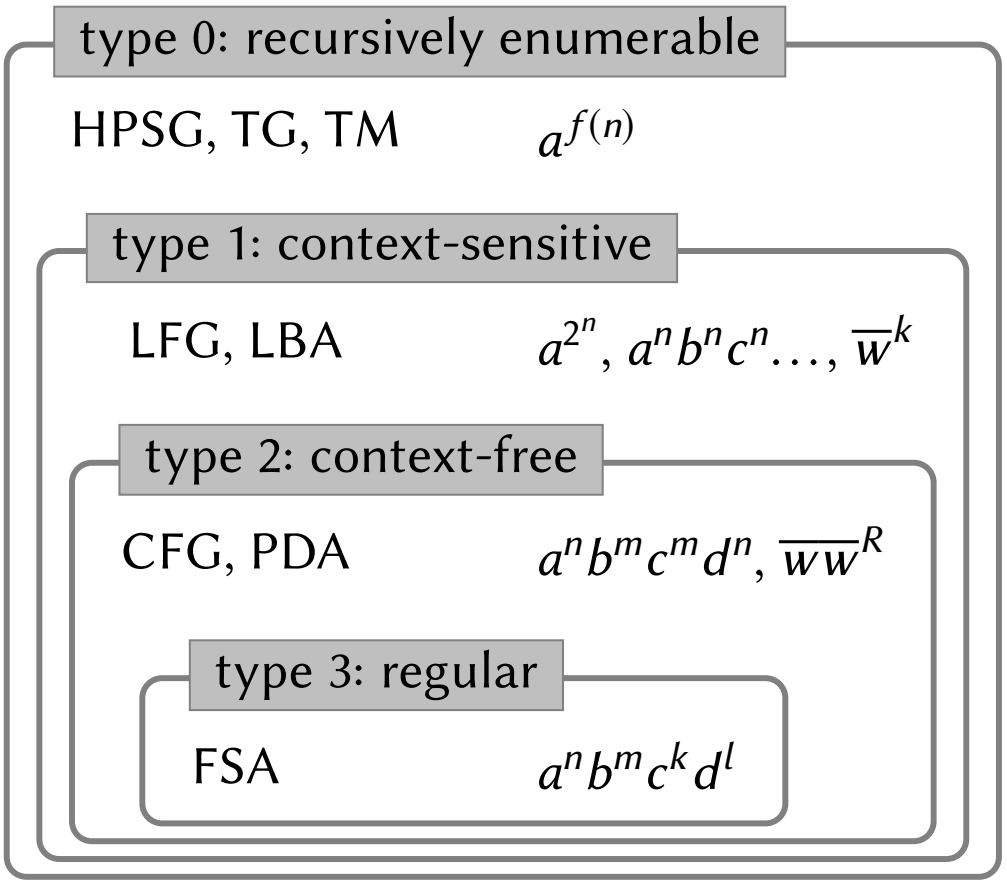
Relative extensionale Komplexität von Grammatikklassen

Chomsky(-Schützenberger)-Hierarchie[6]



Relative extensionale Komplexität von Grammatikklassen

Chomsky(-Schützenberger)-Hierarchie[6]



Wo liegt die Klasse der natürlichen Sprachen?

(Einzelne natürliche Sprachen können in unterschiedlichen Klassen sein.)

NL $\subset$ Type 3 / regular?

Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

NL ⊂ Type 3 / regular?

Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

- Q: endliche Menge von Zuständen (inklusive einem Startzustand und mindestens einem Endzustand)

NL $\subset$ Type 3 / regular?

Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

- Q: endliche Menge von Zuständen (inklusive einem Startzustand und mindestens einem Endzustand)
- δ: endliche Menge von Übergängen zwischen Zuständen

NL ⊂ Type 3 / regular?

Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

- Q: endliche Menge von Zuständen (inklusive einem Startzustand und mindestens einem Endzustand)
- δ: endliche Menge von Übergängen zwischen Zuständen
- Bei jedem Übergang wird ein Symbol von der Eingabe gelesen.

NL ⊂ Type 3 / regular?

Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

- Q: endliche Menge von Zuständen (inklusive einem Startzustand und mindestens einem Endzustand)
- δ: endliche Menge von Übergängen zwischen Zuständen
  - Bei jedem Übergang wird ein Symbol von der Eingabe gelesen.
  - kein Speicher (außer den Zuständen)!

NL ⊂ Type 3 / regular?

Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

- Q: endliche Menge von Zuständen (inklusive einem Startzustand und mindestens einem Endzustand)
- δ: endliche Menge von Übergängen zwischen Zuständen
  - Bei jedem Übergang wird ein Symbol von der Eingabe gelesen.
  - kein Speicher (außer den Zuständen)!

NL ⊂ Type 3 / regular?

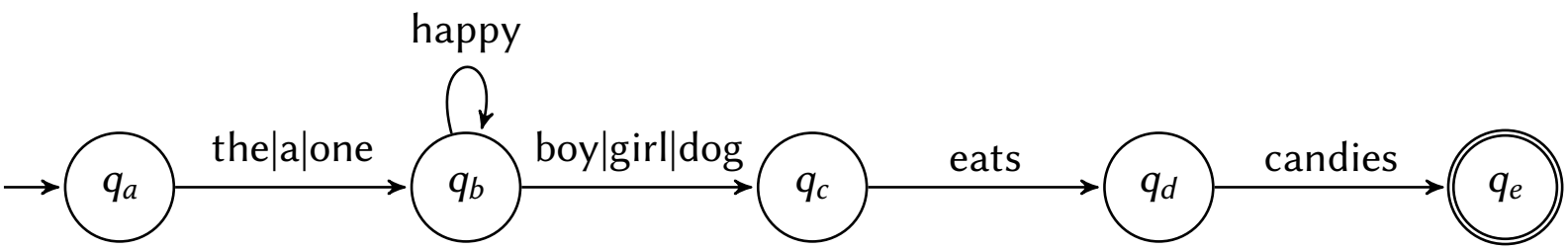
Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

Q: endliche Menge von Zuständen (inklusive einem Startzustand und mindestens einem Endzustand)

δ: endliche Menge von Übergängen zwischen Zuständen

- Bei jedem Übergang wird ein Symbol von der Eingabe gelesen.
- kein Speicher (außer den Zuständen)!



NL ⊂ Type 3 / regular?

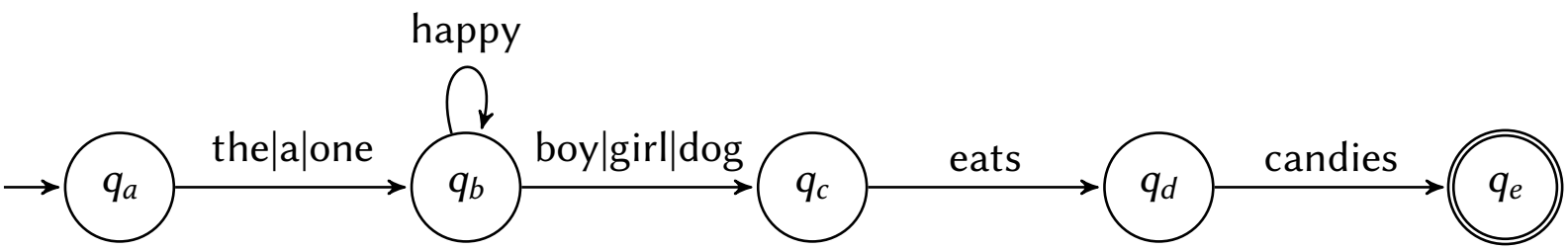
Hypothese

Für jede natürliche Sprache L gibt es einen **endlichen Automaten**, der genau die Wörter aus L akzeptiert.

Q: endliche Menge von Zuständen (inklusive einem Startzustand und mindestens einem Endzustand)

δ: endliche Menge von Übergängen zwischen Zuständen

- Bei jedem Übergang wird ein Symbol von der Eingabe gelesen.
- kein Speicher (außer den Zuständen)!



Wie **beweist** man die Inadäquatheit von endlichen Automaten auf der Ebene der Stringsprache?

NL $\subset$ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

NL $\subset$ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman
c. a woman whom another woman whom another woman hired hired hired another woman

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman
c. a woman whom another woman whom another woman hired hired hired another woman
d. ...

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman
c. a woman whom another woman whom another woman hired hired hired another woman
d. ...

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman
c. a woman whom another woman whom another woman hired hired hired another woman
d. ...

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman
c. a woman whom another woman whom another woman hired hired hired another woman
d. ...

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman
c. a woman whom another woman whom another woman hired hired hired another woman
d. ...

NL ⊂ Type 3 / regular?

Formaler Beweis durch Widerspruch (mittels **Abschlusseigenschaften**):

- Jede reguläre Sprache L erfüllt das **Pump-Lemma** für reguläre Sprachen, d.h. für jedes Wort ab einer bestimmten Wortlänge gilt: Es gibt eine Dekomposition $\overline{xay}$, so dass $\overline{xa}^n\overline{y} \in L$.

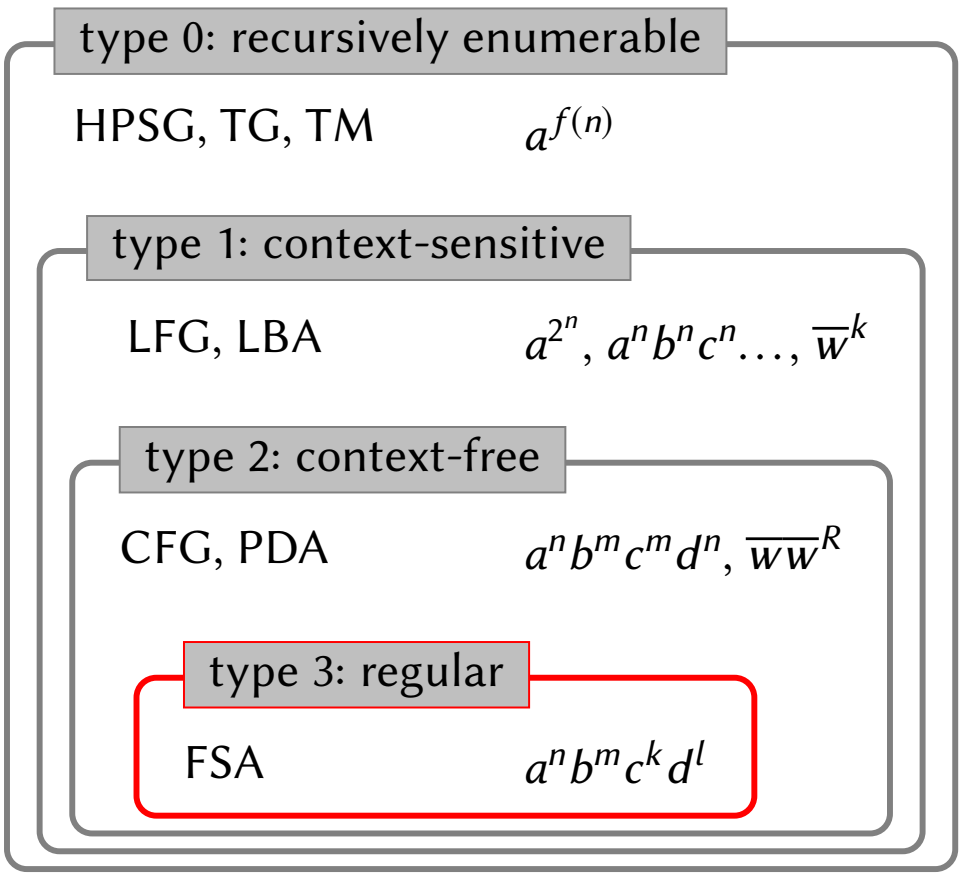
Einbettende Abhängigkeiten (Chomsky 1957)

(1) a. a woman hired another woman
b. a woman whom another woman hired hired another woman
c. a woman whom another woman whom another woman hired hired hired another woman
d. ...

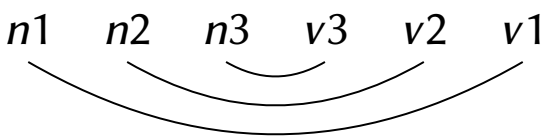
- Homomorphismus f: f(a woman) = x, f(whom another woman) = a, f(hired) = b, f(hired another woman) = y
  - xa^*b^*y ist eine reguläre Sprache; und
  - f(Englisch) $\cap$ $xa^*b^*y = xa^nb^ny$ sollte auch regulär sein.
- xa^nb^ny widerspricht dem Pump-Lemma für reguläre Sprachen.
- $\Rightarrow$ Englisch ist nicht regulär!

NL ⊄ Type 3 / regular!

Chomsky(-Schützenberger)-Hierarchie[6]



NL ist nicht regulär![3,4]
einbettende Abhängigkeiten
bei Relativsätzen



NL $\subset$ Type 2 / context-free?

- eine langwierige Diskussion

NL ⊂ Type 2 / context-free?

- eine langwierige Diskussion

Chomsky (1957):34

"Of course there are languages (in our general sense) that cannot be described in terms of phrase structure, but I do not know whether or not English is itself literally outside the range of such analysis."

NL ⊂ Type 2 / context-free?

- eine langwierige Diskussion

Chomsky (1957):34

"Of course there are languages (in our general sense) that cannot be described in terms of phrase structure, but I do not know whether or not English is itself literally outside the range of such analysis."

- mehrere fehlerhafte Versuche (siehe Pullum & Gazdar [17]), e.g.:

NL ⊂ Type 2 / context-free?

- eine langwierige Diskussion

Chomsky (1957):34

"Of course there are languages (in our general sense) that cannot be described in terms of phrase structure, but I do not know whether or not English is itself literally outside the range of such analysis."

- mehrere fehlerhafte Versuche (siehe Pullum & Gazdar [17]), e.g.:

Bresnan (1978):37–38

"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 ⊂ Type 2 / context-free?

- eine langwierige Diskussion

Chomsky (1957):34

"Of course there are languages (in our general sense) that cannot be described in terms of phrase structure, but I do not know whether or not English is itself literally outside the range of such analysis."

- mehrere fehlerhafte Versuche (siehe Pullum & Gazdar [17]), e.g.:

Bresnan (1978):37–38

"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."

- richtige Beweistechniken: Pump-Lemma und Abschlusseigenschaften
- Was ist ein nicht-kontextfreies Phänomen in natürlicher Sprache?

NL ⊂ Type 2 / context-free?

Nahe dran: Niederländisch (Bresnan et al. 1982)

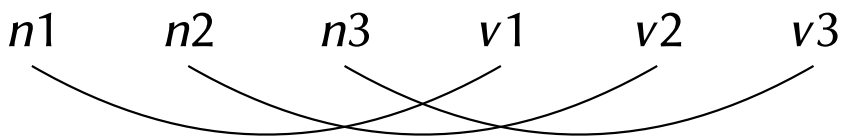
(2) dat Jan Piet de kinderen zag helpen zwemmen.
that Jan Piet the children saw help swim
'that Jan saw Piet helping the children to swim.'

NL ⊂ Type 2 / context-free?

Nahe dran: Niederländisch (Bresnan et al. 1982)

(2) dat Jan Piet de kinderen zag helpen zwemmen.
that Jan Piet the children saw help swim
'that Jan saw Piet helping the children to swim.'

Linguistische Abhängigkeiten sind kreuzend:

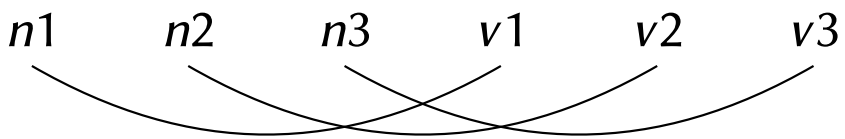


NL ⊂ Type 2 / context-free?

Nahe dran: Niederländisch (Bresnan et al. 1982)

(2) dat Jan Piet de kinderen zag helpen zwemmen.
that Jan Piet the children saw help swim
'that Jan saw Piet helping the children to swim.'

Linguistische Abhängigkeiten sind kreuzend:



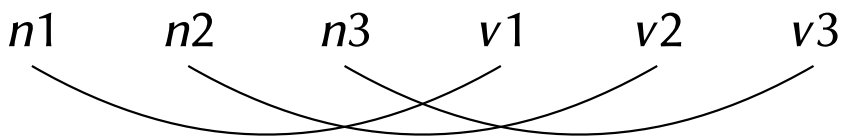
Aber: Kein Reflex an der Oberfläche, d.h. in der Stringsprache!

NL ⊂ Type 2 / context-free?

Nahe dran: Niederländisch (Bresnan et al. 1982)

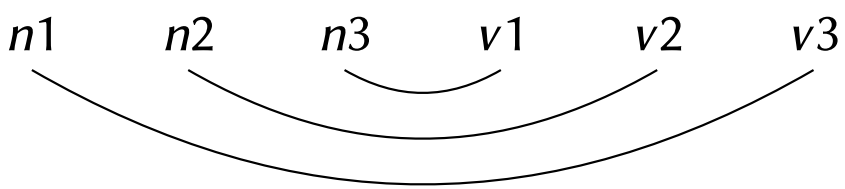
(2) dat Jan Piet de kinderen zag helpen zwemmen.
that Jan Piet the children saw help swim
'that Jan saw Piet helping the children to swim.'

Linguistische Abhängigkeiten sind kreuzend:



Aber: Kein Reflex an der Oberfläche, d.h. in der Stringsprache!

⇒ Der String kann durch eine CFG erzeugt werden, auch wenn die Abhängigkeiten verloren gehen.



NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3) a. wulu *'dog'*

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3) a. wulu *'dog'*
b. wulu-lela *'dog watcher'*

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3) a. wulu — *'dog'*
b. wulu-lela — *'dog watcher'*
c. wulu-lela-nyinila — *'dog watcher hunter'*

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3)	a.	wulu	*'dog'*
	b.	wulu-lela	*'dog watcher'*
	c.	wulu-lela-nyinila	*'dog watcher hunter'*
	d.	wulu-o-wulu	*'whatever dog'*

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3)	a.	wulu	*'dog'*
	b.	wulu-lela	*'dog watcher'*
	c.	wulu-lela-nyinila	*'dog watcher hunter'*
	d.	wulu-o-wulu	*'whatever dog'*
	e.	wulu-lela-o-wulu-lela	

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3)	a.	wulu	*'dog'*
	b.	wulu-lela	*'dog watcher'*
	c.	wulu-lela-nyinila	*'dog watcher hunter'*
	d.	wulu-o-wulu	*'whatever dog'*
	e.	wulu-lela-o-wulu-lela	*'whatever dog watcher'*

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3)	a.	wulu	*'dog'*
	b.	wulu-lela	*'dog watcher'*
	c.	wulu-lela-nyinila	*'dog watcher hunter'*
	d.	wulu-o-wulu	*'whatever dog'*
	e.	wulu-lela-o-wulu-lela	*'whatever dog watcher'*
	f.	wulu-lela-nyinila-o-wulu-lela-nyinila	*'whatever dog watcher hunter'*

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3)	a.	wulu	*'dog'*
	b.	wulu-lela	*'dog watcher'*
	c.	wulu-lela-nyinila	*'dog watcher hunter'*
	d.	wulu-o-wulu	*'whatever dog'*
	e.	wulu-lela-o-wulu-lela	*'whatever dog watcher'*
	f.	wulu-lela-nyinila-o-wulu-lela-nyinila	*'whatever dog watcher hunter'*

Muster: $a^n b^m a^n b^m$ oder $\overline{ww}$ (copy language)

$\Rightarrow$ nicht kontextfrei!

NL ⊂ Type 2 / context-free?

Ein weiterer Versuch von Culy (1985): Duplikation in der Morphologie des Bambara

(3)	a.	wulu	*'dog'*
	b.	wulu-lela	*'dog watcher'*
	c.	wulu-lela-nyinila	*'dog watcher hunter'*
	d.	wulu-o-wulu	*'whatever dog'*
	e.	wulu-lela-o-wulu-lela	*'whatever dog watcher'*
	f.	wulu-lela-nyinila-o-wulu-lela-nyinila	*'whatever dog watcher hunter'*

Muster: $a^n b^m a^n b^m$ oder $\overline{ww}$ (copy language)

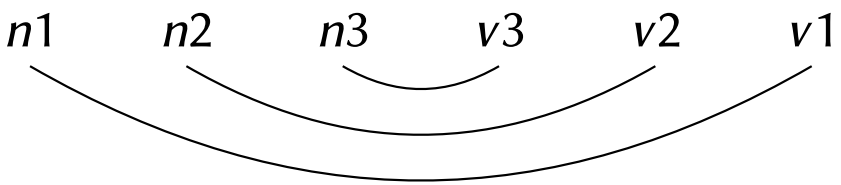
$\Rightarrow$ nicht kontextfrei!

Aber: Beweis für die Morphologie, nicht für die Syntax (Stringsprache)!

NL ⊂ Type 2 / context-free?

- Deutsch: **einbettende Abhängigkeiten** (VE-Sätze)

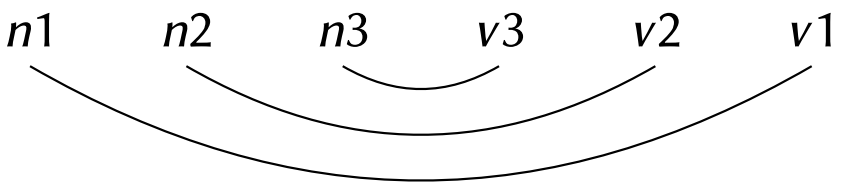
(4) (dass) er die Kinder dem Hans das Haus streichen helfen ließ
(that) he the children the Hans the house paint help let
‘(that) he let the children help Hans paint the house’



NL ⊂ Type 2 / context-free?

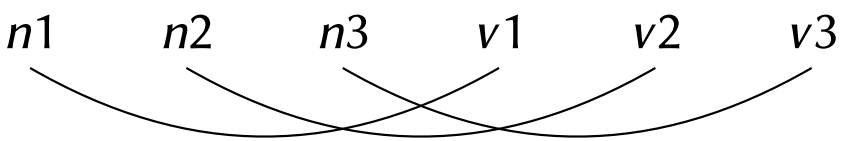
- Deutsch: **einbettende Abhängigkeiten** (VE-Sätze)

(4) (dass) er die Kinder dem Hans das Haus streichen helfen ließ
(that) he the children the Hans the house paint help let
'(that) he let the children help Hans paint the house'



- Schwyzerdütsch: **kreuzende Abhängigkeiten**

(5) …mer d'chind em Hans es huus lönd hälfe aastriiche
…we children.ACC the Hans.DAT the house.ACC let help paint
'…we let the children help Hans paint the house'



NL ⊂ Type 2 / context-free?

Beweis von Shieber (1985):

(6) Jan säit das mer d'chind em Hans es huus lönd hälfe aastriiche.

NL ⊂ Type 2 / context-free?

Beweis von Shieber (1985):

(6) Jan säit das mer d'chind em Hans es huus lönd hälfe aastriiche.

- Homomorphismus f:

$f(\text{Jan säit das mer}) = x$ $f(\text{d'chind}) = a$ $f(\text{em Hans}) = b$ $f(\text{es huus}) = y$
$f(\text{lönd}) = c$ $f(\text{hälfe}) = d$ $f(\text{aastriiche}) = z$

NL ⊂ Type 2 / context-free?

Beweis von Shieber (1985):

(6) Jan säit das mer d'chind em Hans es huus lönd hälfe aastriiche.

- Homomorphismus f:

  $f(\text{Jan säit das mer}) = x$ $f(\text{d'chind}) = a$ $f(\text{em Hans}) = b$ $f(\text{es huus}) = y$
  $f(\text{lönd}) = c$ $f(\text{hälfe}) = d$ $f(\text{aastriiche}) = z$

- $f(\text{Schwyzerdütsch}) \cap xa^*b^*yc^*d^*z = xa^mb^nyc^md^nz$

NL ⊂ Type 2 / context-free?

Beweis von Shieber (1985):

(6) Jan säit das mer d'chind em Hans es huus lönd hälfe aastriiche.

- Homomorphismus f:

  $f(\text{Jan säit das mer}) = x$ $f(\text{d'chind}) = a$ $f(\text{em Hans}) = b$ $f(\text{es huus}) = y$
  $f(\text{lönd}) = c$ $f(\text{hälfe}) = d$ $f(\text{aastriiche}) = z$

- $f(\text{Schwyzerdütsch}) \cap xa^*b^*yc^*d^*z = xa^mb^nyc^md^nz$
  - Kontextfreie Sprachen sind **abgeschlossen** unter Schnitt mit regulären Sprachen.

NL $\subset$ Type 2 / context-free?

Beweis von Shieber (1985):

(6) Jan säit das mer d'chind em Hans es huus lönd hälfe aastriiche.

- Homomorphismus f:

  $f(\text{Jan säit das mer}) = x$ $f(\text{d'chind}) = a$ $f(\text{em Hans}) = b$ $f(\text{es huus}) = y$
  $f(\text{lönd}) = c$ $f(\text{hälfe}) = d$ $f(\text{aastriiche}) = z$

- $f(\text{Schwyzerdütsch}) \cap xa^*b^*yc^*d^*z = xa^mb^nyc^md^nz$
  - Kontextfreie Sprachen sind **abgeschlossen** unter Schnitt mit regulären Sprachen.
  - $xa^*b^*yc^*d^*z$ ist regulär, also sollte $xa^mb^nyc^md^nz$ kontextfrei sein.

NL $\subset$ Type 2 / context-free?

Beweis von Shieber (1985):

(6) Jan säit das mer d'chind em Hans es huus lönd hälfe aastriiche.

- Homomorphismus f:

  $f(\text{Jan säit das mer}) = x$ $\quad f(\text{d'chind}) = a$ $\quad f(\text{em Hans}) = b$ $\quad f(\text{es huus}) = y$
  $f(\text{lönd}) = c$ $\quad f(\text{hälfe}) = d$ $\quad f(\text{aastriiche}) = z$

- $f(\text{Schwyzerdütsch}) \cap xa^*b^*yc^*d^*z = xa^mb^nyc^md^nz$
  - Kontextfreie Sprachen sind **abgeschlossen** unter Schnitt mit regulären Sprachen.
  - $xa^*b^*yc^*d^*z$ ist regulär, also sollte $xa^mb^nyc^md^nz$ kontextfrei sein.
  - Aber mittels **Pump-Lemma für kontextfreie Sprachen**: $xa^mb^nxy^md^nz$ ist nicht kontextfrei.

NL $\subset$ Type 2 / context-free?

Beweis von Shieber (1985):

(6) Jan säit das mer d'chind em Hans es huus lönd hälfe aastriiche.

- Homomorphismus f:

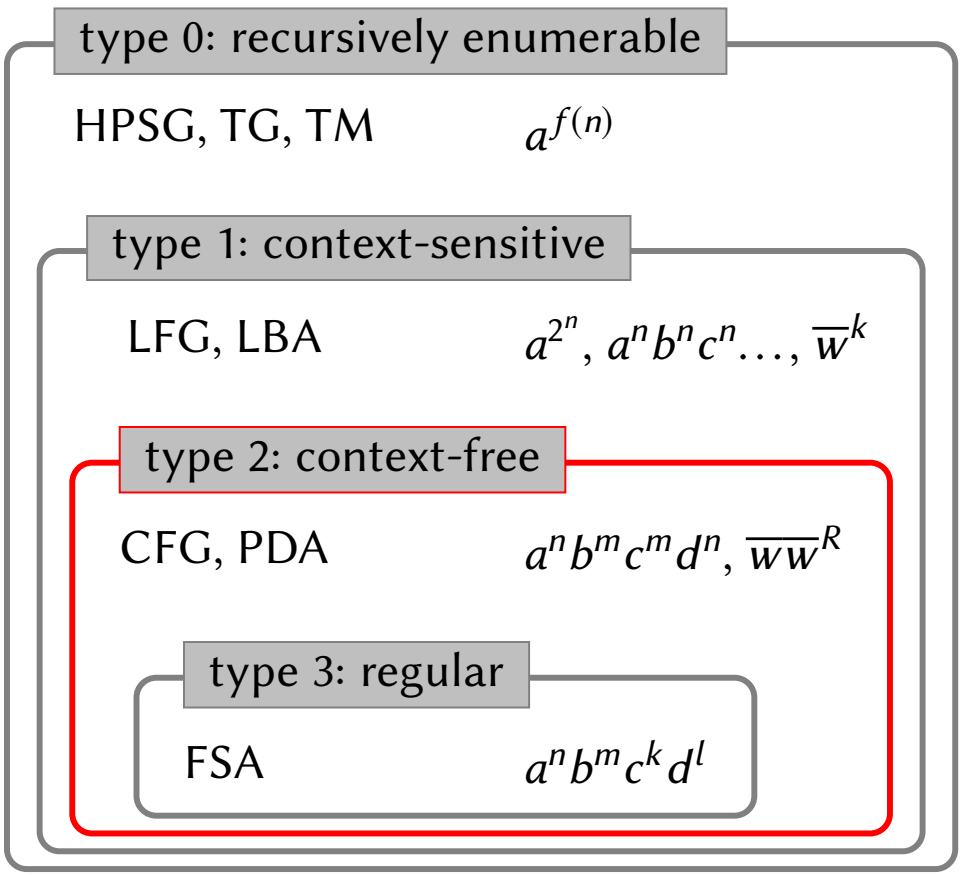
  $f(\text{Jan säit das mer}) = x$ $\quad f(\text{d'chind}) = a$ $\quad f(\text{em Hans}) = b$ $\quad f(\text{es huus}) = y$
  $f(\text{lönd}) = c$ $\quad f(\text{hälfe}) = d$ $\quad f(\text{aastriiche}) = z$

- $f(\text{Schwyzerdütsch}) \cap xa^*b^*yc^*d^*z = xa^mb^nyc^md^nz$
  - Kontextfreie Sprachen sind **abgeschlossen** unter Schnitt mit regulären Sprachen.
  - $xa^*b^*yc^*d^*z$ ist regulär, also sollte $xa^mb^nyc^md^nz$ kontextfrei sein.
  - Aber mittels **Pump-Lemma für kontextfreie Sprachen**: $xa^mb^nxy^md^nz$ ist nicht kontextfrei.

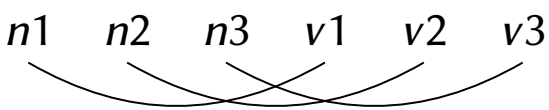
$\Rightarrow$ Schwyzerdütsch ist nicht kontextfrei!

NL ⊄ Type 2 / context-free!

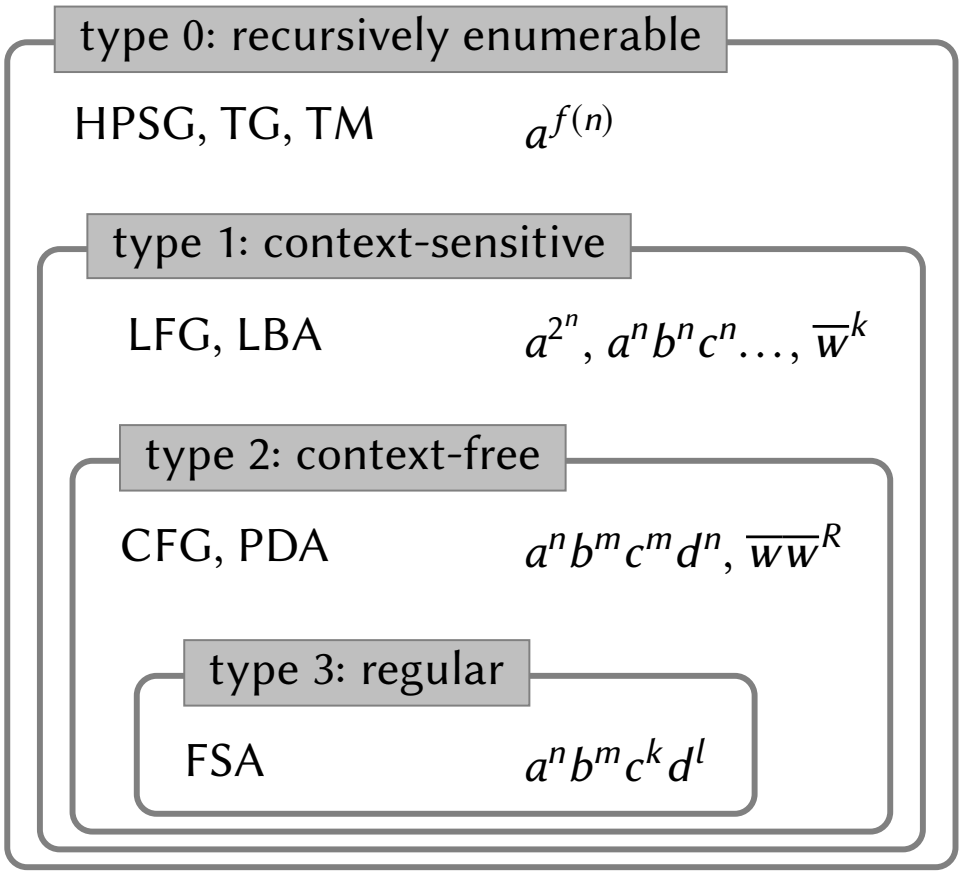
Chomsky(-Schützenberger)-Hierarchie[6]



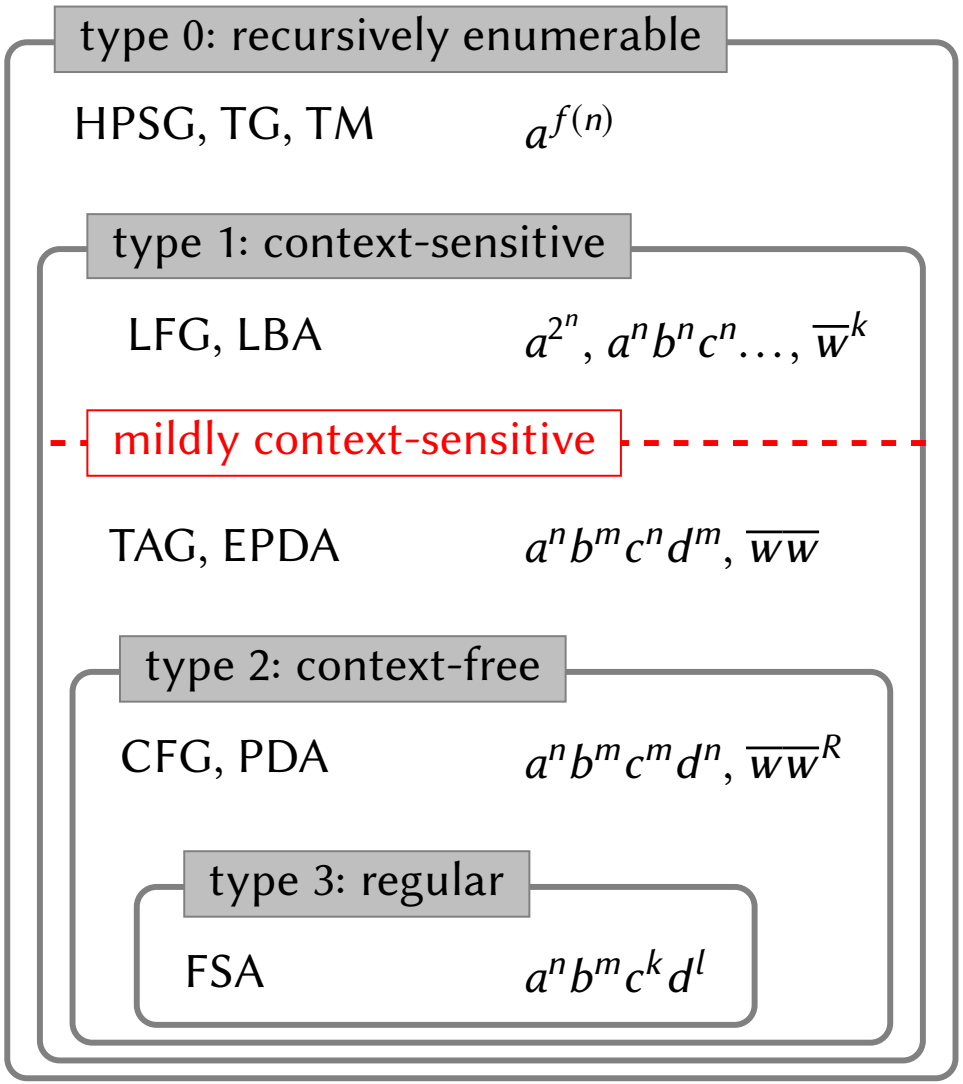
NL ist nicht kontextfrei![21]
kreuzende Abhängigkeiten
im Niederländischen und
Schwyzerdütsch



NL ⊂ Type 1–2 / mildly context-sensitive?



NL ⊂ Type 1–2 / mildly context-sensitive?



NL ist **schwach** kontextsensitiv? (Joshi [13])

- ⊃ CFL
- kreuzende Abh.
- semi-linear
- in PTIME

NL ⊂ Type 1–2 / mildly context-sensitive?

Argumente gegen Semilinearität:

- Kasusstapel (“Suffixaufnahme”) im Alt-Georgischen (Michaelis & Kracht 1997)

$$N_1\text{-NOM} \quad N_2\text{-GEN-NOM} \quad N_3\text{-GEN}^2\text{-NOM} \quad \ldots \quad N_k\text{-GEN}^{k-1}\text{-NOM}$$

> “The number of all genitive suffixes of all nouns within a complex NP consisting of k stacked NPs, where $k \in \mathbb{N}^+$, is bounded by $k^2/2 - k/2$.”

NL ⊂ Type 1–2 / mildly context-sensitive?

Argumente gegen Semilinearität:

- Kasusstapel (“Suffixaufnahme”) im Alt-Georgischen (Michaelis & Kracht 1997)

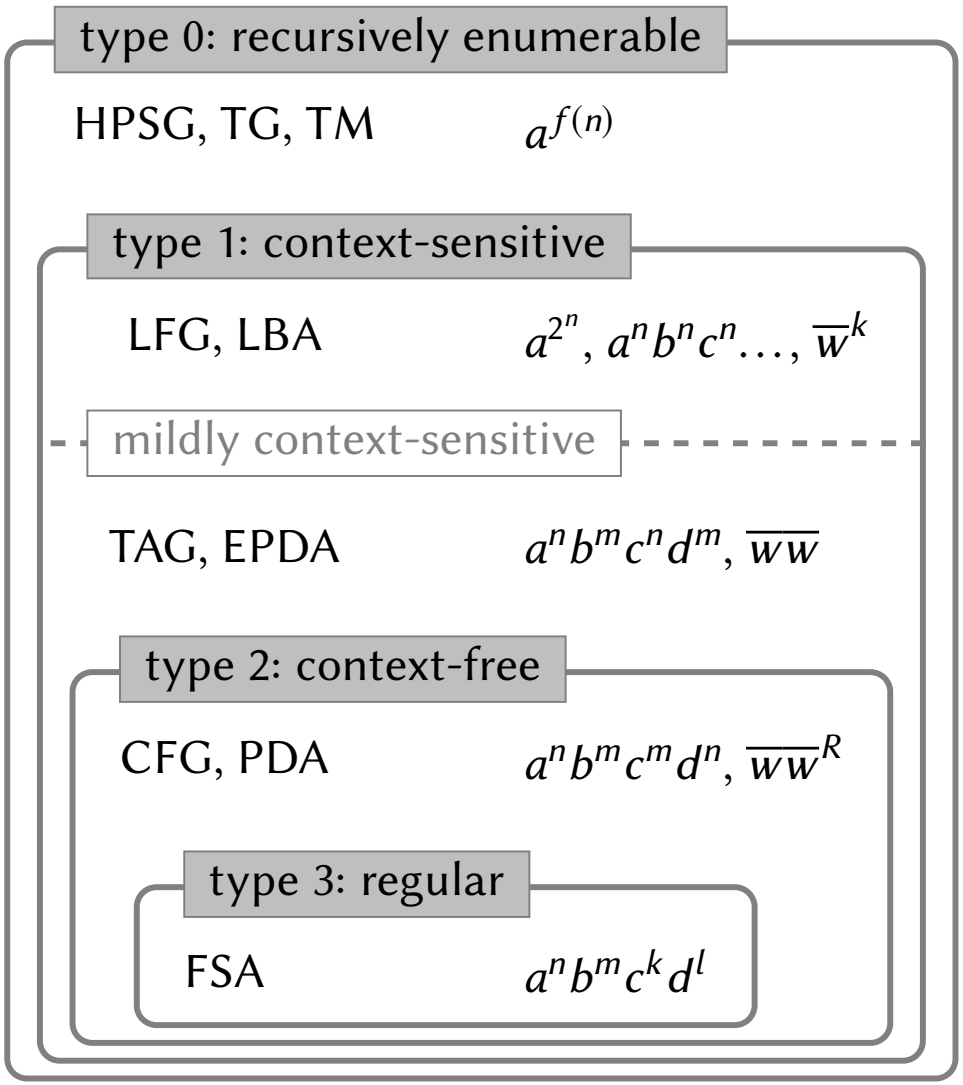
$$N_1\text{-NOM} \quad N_2\text{-GEN-NOM} \quad N_3\text{-GEN}^2\text{-NOM} \quad \ldots \quad N_k\text{-GEN}^{k-1}\text{-NOM}$$

> “The number of all genitive suffixes of all nouns within a complex NP consisting of k stacked NPs, where $k \in \mathbb{N}^+$, is bounded by $k^2/2 - k/2$.”

- Chinesische Zahlterme (Radzinski 1991)
- Koordination im Niederländischen (Groenink 1997)
- Relativsätze im Yoruba (Kobele 2006)

$\Rightarrow$ Im Allgemeinen in der Literatur nicht berücksichtigt.

NL ⊂ Type 1 / context-sensitive?



NL unentscheidbar?

- Hintikka (1974)
- Chomsky (1980)
- Pollard (1996)

Subreguläre Typen: Type 4, 5, ...

Wie Reguläre Sprachen weiter beschränken?

- Regelschema: Anzahl der Nichtterminale
- Automaten: Anzahl der Zustände, Struktur der Übergänge

Subreguläre Typen: Type 4, 5, ...

Wie Reguläre Sprachen weiter beschränken?

- Regelschema: Anzahl der Nichtterminale
- Automaten: Anzahl der Zustände, Struktur der Übergänge

Endliche Sprachen: Naheliegend aber formal eher uninteressant

- $S \to \Sigma^+$
- $S \to ab$
- $S \to a$

Subreguläre Typen: Type 4, 5, ...

Wie Reguläre Sprachen weiter beschränken?

- Regelschema: Anzahl der Nichtterminale
- Automaten: Anzahl der Zustände, Struktur der Übergänge

Endliche Sprachen: Naheliegend aber formal eher uninteressant

- $S \to \Sigma^+$
- $S \to ab$
- $S \to a$

Interessantere subreguläre Sprachen:

- Strictly Local Languages (SL_k)[12,19]

Subreguläre Typen: Type 4, 5, …

Wie Reguläre Sprachen weiter beschränken?

- Regelschema: Anzahl der Nichtterminale
- Automaten: Anzahl der Zustände, Struktur der Übergänge

Endliche Sprachen: Naheliegend aber formal eher uninteressant

- $S \to \Sigma^+$
- $S \to ab$
- $S \to a$

Interessantere subreguläre Sprachen:

- Strictly Local Languages (SL_k)[12,19]

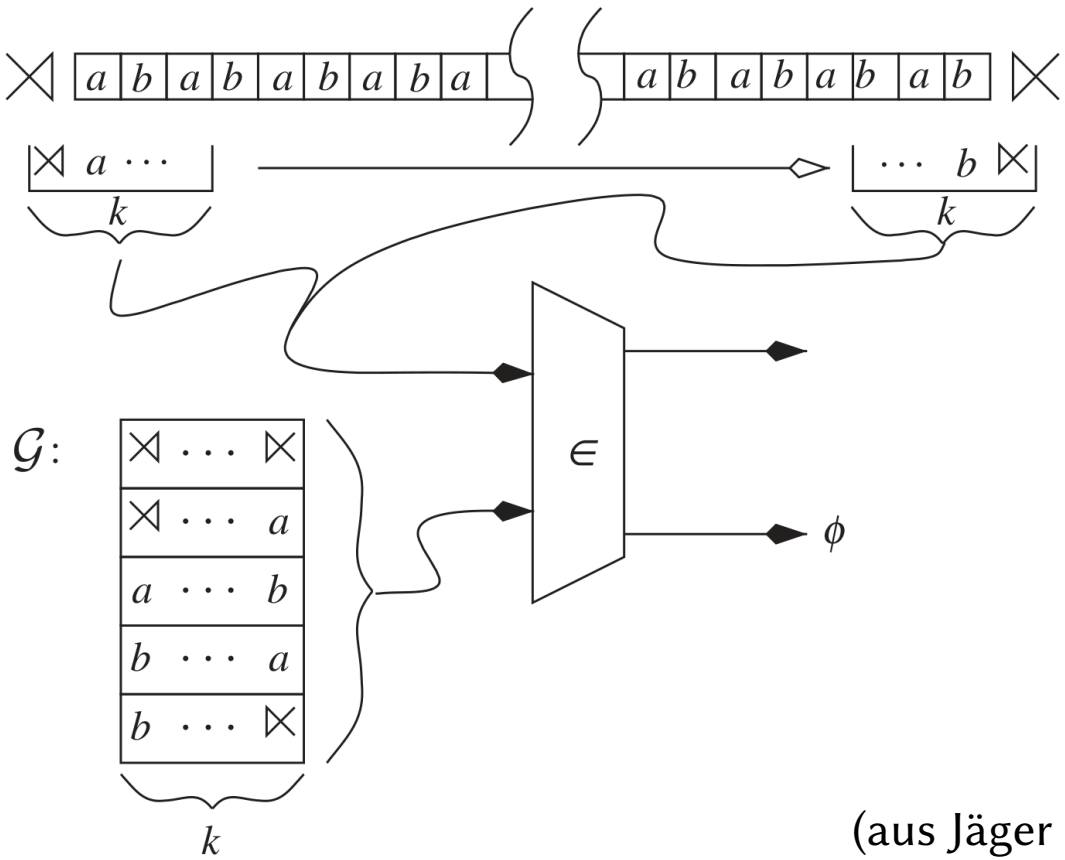
Aber was hat das mit natürlicher Sprache zu tun?

$\Rightarrow$ Syntax/Typologie[11], Phonologie

Strictly Local Languages (SL_k)

Die Sprache wird nur anhand zusammenhängender “Blöcke” der Größe k beschränkt.

Automatenmodell: “scanner” mit “look-up list”



(aus Jäger & Rogers [12])

(siehe z.B. Rogers & Pullum [19])

Strictly Local Languages (SL_k)

Die Sprache wird nur anhand zusammenhängender "Blöcke" der Größe k beschränkt.

Automatenmodell: "scanner" mit "look-up list"

- Die Grammatik G besteht aus einer Menge zulässiger Blöcke, z.B. $G = \{\rtimes A, AB, BA, B\ltimes\}$ mit $k = 2$.
- $F_k(\overline{w})$ ist die Menge der k-Faktoren in $\overline{w}$.

$$F_k(\overline{w}) = \begin{cases} \{y \mid \overline{w} = x \cdot y \cdot z, \quad x, y, z \in \Sigma^*, |y| = k\} & \text{if } |\overline{w}| > k, \\ \{\overline{w}\} & \text{otherwise.} \end{cases}$$

- $L(G) = \{\overline{w} | F_k(\rtimes \cdot \overline{w} \cdot \ltimes) \subseteq G, \overline{w} \text{ finite}\}$

(siehe z.B. Rogers & Pullum [19])

Strictly Local Languages (SL_k)

Die Sprache wird nur anhand zusammenhängender "Blöcke" der Größe k beschränkt.

Automatenmodell: "scanner" mit "look-up list"

- Die Grammatik G besteht aus einer Menge zulässiger Blöcke, z.B. $G = \{\rtimes A, AB, BA, B\ltimes\}$ mit $k = 2$.
- $F_k(\overline{w})$ ist die Menge der k-Faktoren in $\overline{w}$.

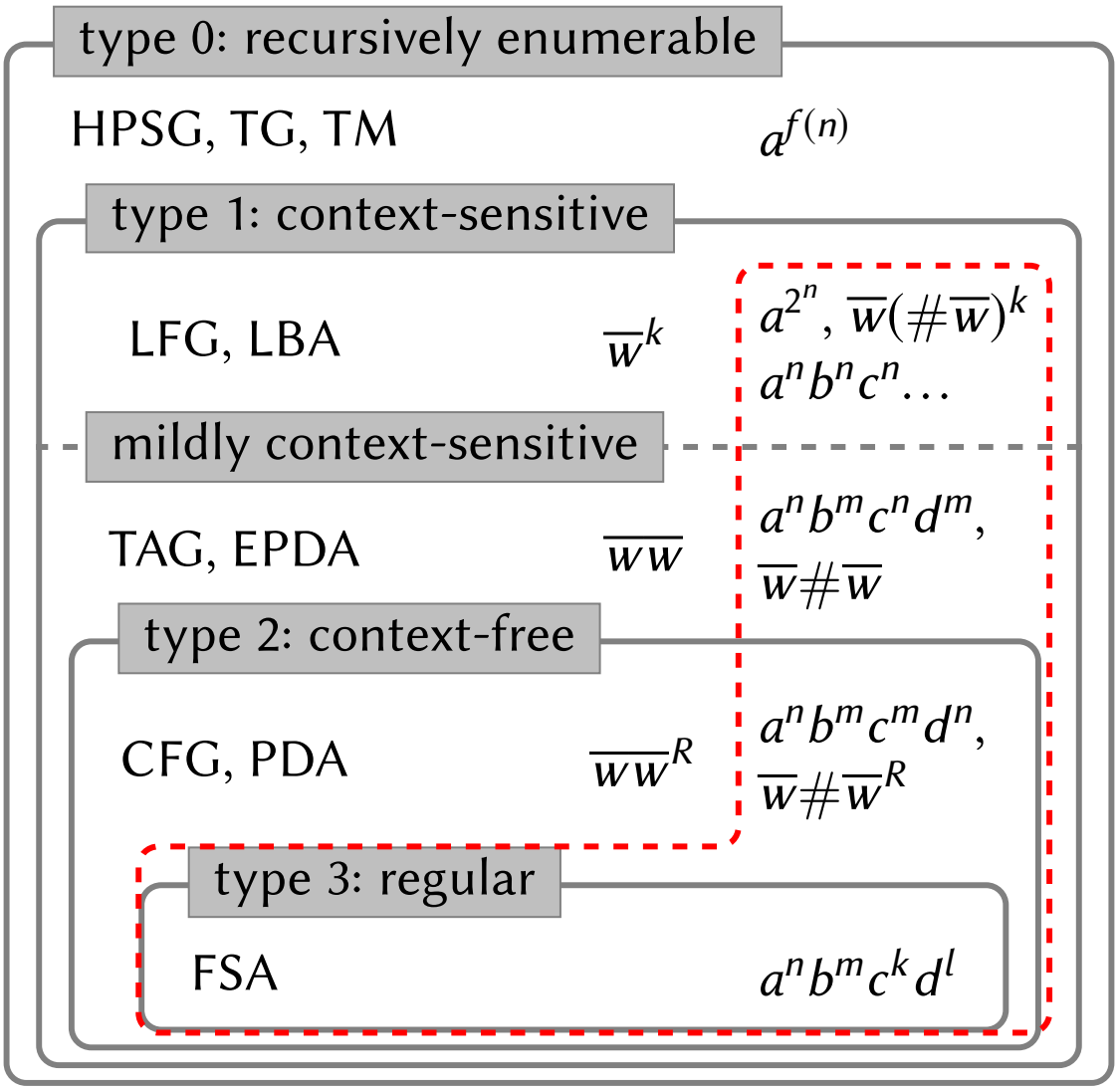
$$F_k(\overline{w}) = \begin{cases} \{y \mid \overline{w} = x \cdot y \cdot z, \quad x, y, z \in \Sigma^*, |y| = k\} & \text{if } |\overline{w}| > k, \\ \{\overline{w}\} & \text{otherwise.} \end{cases}$$

- $L(G) = \{\overline{w} | F_k(\rtimes \cdot \overline{w} \cdot \ltimes) \subseteq G, \overline{w} \text{ finite}\}$

- $(ab)^n \in SL_2$
- aber: $a^* ba^* \notin SL_k$

(siehe z.B. Rogers & Pullum [19])

Orthogonale Hierachien



orthogonal classes:[9,22]
deterministic real-time definable (Rosenberg 1967)

(siehe auch die Sitzung über Lernbarkeit)

Zusammenfassung

- extensionale Komplexität von (formalen) Sprachklassen
- Wo liegt natürliche Sprache in der Chomsky-Hierarchie?
- Die **Klasse** der natürlichen Sprachen is zumindest schwach kontextsensitiv.
- formale Beweise mittels Abschlusseigenschaften und Pump-Lemma

Appendix: Abschlusseigenschaften

	$L_1 \cap L_2$	$L_1 \cup L_2$	$L^{\complement}$	$L_1 \bullet L_2$	L^*
Type 0	+	+	−	+	+
Type 1	+	+	+	+	+
Type 2	−	+	−	+	+
Type 3	+	+	+	+	+
Type 2/3	+	+	+	+	+

Appendix: Pump-Lemma für kontextfreie Sprachen

Ist eine Sprache L kontextfrei, dann existiert eine Mindestwortlänge k für L, ab der ein Wort $\overline{w} \in L$ in Teilworte $\overline{xayb\overline{z}}$ dekomponiert werden kann, so dass auch $\overline{xa}^n\overline{y}\overline{b}^n\overline{z} \in L$ für $n \geq 0$, wobei gelten muss:

1. $|\overline{xyz}| \leq k$,
2. $|\overline{ab}| \geq 1$.

[1] Bresnan, Joan. 1978. A realistic transformational grammar. In Morris Halle, Joan Bresnan & George A. Miller (eds.), *Linguistic theory and psychological reality*, 1–59. Cambridge, MA: The MIT Press.

[2] Bresnan, Joan, Ronald M. Kaplan, Stanley Peters & Annie Zaenen. 1982. Cross-serial dependencies in dutch. *Linguistic Inquiry* 13(4). 613–634.

[3] Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2. 113–124.

[4] Chomsky, Noam. 1957. *Syntactic structures*. Den Haag: Mouton.

[5] Chomsky, Noam. 1980. *Rules and representations*. Oxford, UK: Basil Blackwell.

[6] Chomsky, Noam & Marcel-Paul Schützenberger. 1963. The algebraic theory of context-free languages. In P. Braffort & D. Hirschberg (eds.), *Computer programming and formal systems* (Studies in Logic and the Foundations of Mathematics 35), 118–161. Elsevier.

[7] Culy, Christopher. 1985. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy* 8(3). 345–351. http://www.jstor.org/stable/25001211.

[8] Groenink, Annius V. 1997. Mild context-sensitivity and tuple-based generalizations of context-grammar. *Linguistics and Philosophy* 20(6). 607–636. http://dx.doi.org/10.1023/A%3A1005376413354.

[9] Hausser, Roland. 1992. Complexity in Left-Associative Grammar. *Theoretical Computer Science* 106(2). 283–308. http://www.sciencedirect.com/science/article/pii/030439759290253C.

[10] Hintikka, Jaakko. 1974. Quantifiers vs. quantification theory. *Linguistic Inquiry* 5(2). 153–177. http://www.jstor.org/stable/4177815.

[11] Jackendoff, Ray & Eva Wittenberg. 2014. What you can say without syntax: A hierarchy of grammatical complexity. In Frederick J. Newmeyer & Laurel B. Preston (eds.), *Measuring grammatical complexity*, 65–82. Oxford: Oxford University Press.

[12] Jäger, Gerhard & James Rogers. 2012. Formal language theory: Refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1598). 1956–1970.

[13] Joshi, Aravind K. 1985. Tree adjoining grammars: how much context-sensitivity is required to provide reasonable structural descriptions. In David Dowty, Lauri Karttunen & Arnold Zwicky (eds.), *Natural language parsing*, 206–250. Cambridge University Press.

[14] Kobele, Gregory M. 2006. *Generating copies: an investigation into structural identity in language and grammar*. Los Angeles: University of California dissertation. `http://home.uchicago.edu/~gkobele/files/Kobele06GeneratingCopies.pdf`.

[15] Michaelis, Jens & Marcus Kracht. 1997. Semilinearity as a syntactic invariant. In Christian Retoré (ed.), *Logical aspects of computational linguistics* (Lecture Notes in Computer Science 1328), 329–345. Berlin: Springer. `http://dx.doi.org/10.1007/BFb0052165`.

[16] Pollard, Carl. 1996. The nature of constraint grammar. Paper presented at the 11th Pacific Conference of Language, Information and Computation (PACLIC).

[17] Pullum, Geoffrey K. & Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy* 4(4). 471–504. `http://www.jstor.org/stable/25001071`.

[18] Radzinski, Daniel. 1991. Chinese number-names, tree adjoining languages, and mild context-sensitivity. *Computational Linguistics* 17. 277–299.

[19] Rogers, James & Geoffrey K. Pullum. 2007. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Language, Logic and Information* 20. 329–342.

[20] Rosenberg, Arnold L. 1967. Real-time definable languages. *Journal of the Association for Computing Machinery* 14(4). 645–662.

[21] Shieber, Stuart. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8. 333–343.

[22] Wurm, Christian. 2012. Regular Growth Automata: properties of a class of finitely induced infinite machines. In Makoto Kanazawa, Markus Kracht, Hiroyuki Seki & Andras Kornai (eds.), *Proceedings of the 12th conference on the mathematics of language (MOL 2012)* (LNCS 6878), 192–208. Nara, Japan: Springer.