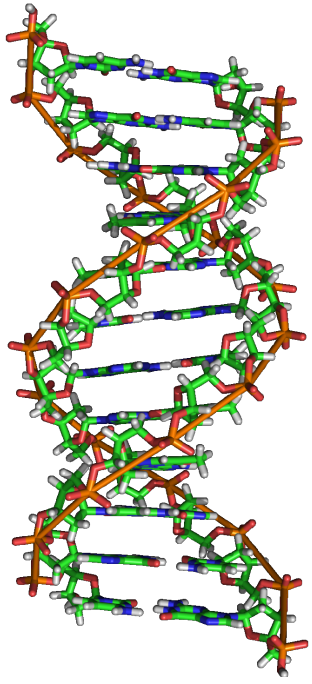
Grammatik der Genome

Andreas de Vries

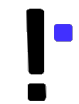
Fachhochschule Südwestfalen, Campus Hagen

SS 2011



Inhalt

Übersicht



Formale Sprachen

- Die Theorie der formalen Sprachen behandelt die systematische Analyse, Klassifizierung und Konstruktion von Wortmengen über endlichen Alphabeten.
- Entstanden in den 1950er Jahren in der Linguistik.
- Formale Sprachen (Prädikatenkalkül, Java, HTML, …) haben wohldefinierte und eindeutige ***Grammatiken***.
- Natürliche Sprachen haben dagegen **Mehrdeutigkeiten** auf jeder Beschreibungsebene, von der phonetischen bis zur soziologischen.
- Sprachliche **Witze** verwenden diese Mehrdeutigkeiten, insbesondere Wortspiele: *„Bis dann, wir sehen uns! – Wir sind ja nicht blind.“*

Formale Sprachen

- Die Theorie der formalen Sprachen behandelt die systematische Analyse, Klassifizierung und Konstruktion von Wortmengen über endlichen Alphabeten.
- Entstanden in den 1950er Jahren in der Linguistik.
- Formale Sprachen (Prädikatenkalkül, Java, HTML, …) haben wohldefinierte und eindeutige ***Grammatiken***.
- Natürliche Sprachen haben dagegen **Mehrdeutigkeiten** auf jeder Beschreibungsebene, von der phonetischen bis zur soziologischen.
- Sprachliche **Witze** verwenden diese Mehrdeutigkeiten, insbesondere Wortspiele: *„Bis dann, wir sehen uns! – Wir sind ja nicht blind.“*
- Merke: ***Formale Sprachen eignen sich nicht für Späße!***

Formale Grammatiken

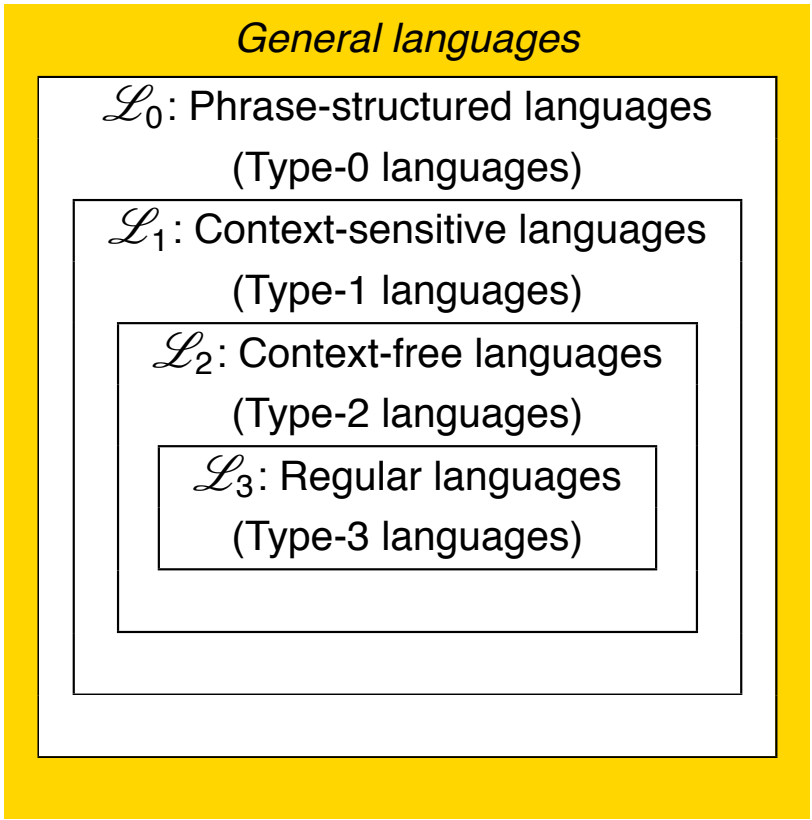
Example (*Simple English*)

Mit dem Lateinischen Alphabet mit den 26 Kleinbuchstaben, dem Leerzeichen und dem Punkt ist durch die $\langle$Variablen$\rangle$ und Regeln für die Wortbildung **(„Produktionen“)** eine ***Grammatik*** gegeben:

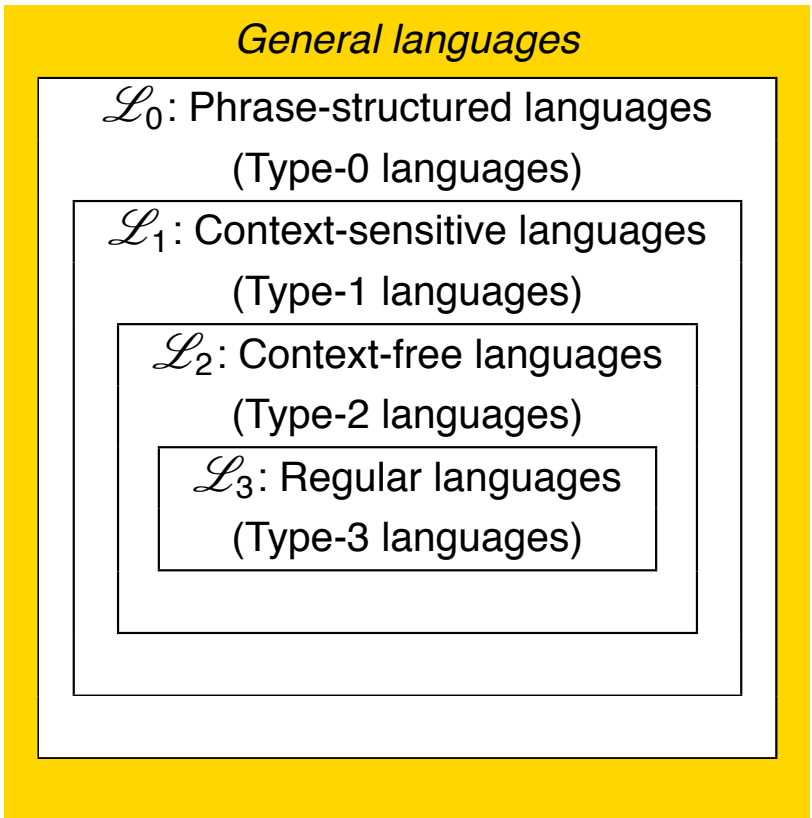
$$
\begin{aligned}
\langle \text{sentence} \rangle &\longrightarrow \langle \text{subject} \rangle\ \langle \text{predicate} \rangle\ \langle \text{object} \rangle . \\
\langle \text{subject} \rangle &\longrightarrow \langle \text{article} \rangle\ \langle \text{adjective} \rangle\ \langle \text{noun} \rangle \\
\langle \text{article} \rangle &\longrightarrow \texttt{a} \mid \texttt{the} \\
\langle \text{adjective} \rangle &\longrightarrow \texttt{sweet} \mid \texttt{quick} \mid \texttt{small} \\
\langle \text{noun} \rangle &\longrightarrow \texttt{duck} \mid \texttt{frog} \mid \texttt{mouse} \mid \texttt{hippopotamus} \\
\langle \text{predicate} \rangle &\longrightarrow \texttt{likes} \mid \texttt{catches} \mid \texttt{eats} \\
\langle \text{object} \rangle &\longrightarrow \texttt{cookies} \mid \texttt{chocolate} \mid \texttt{pizza}
\end{aligned}
$$

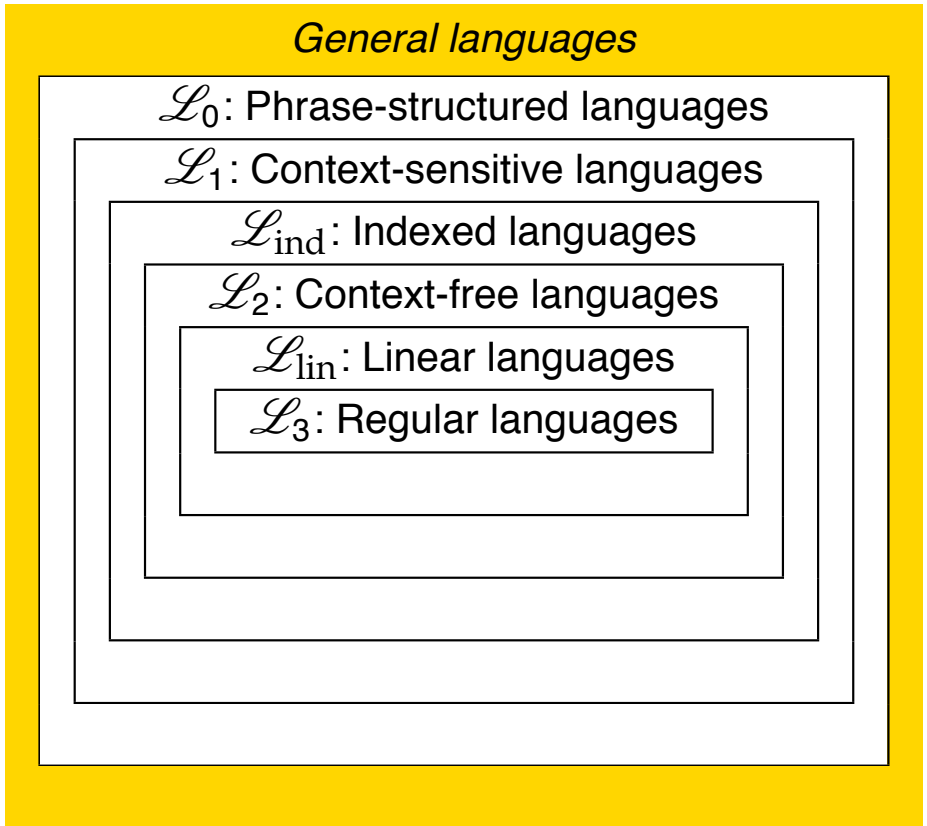
Die durch diese Grammatik erzeugte Sprache beinhaltet Sätze wie “`the small duck eats pizza.`” und “`a quick mouse catches cookies.`”

Chomsky-Hierarchie



Chomsky-Hierarchie



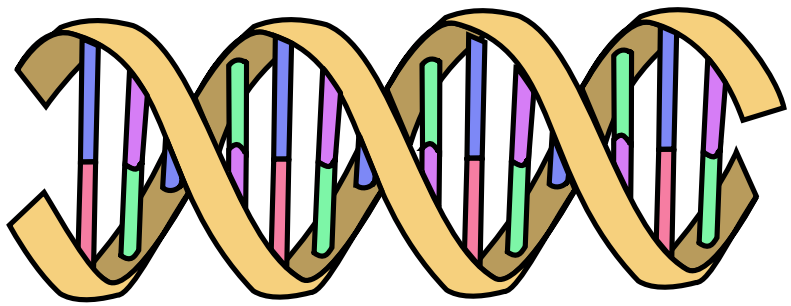


Übersicht

Genomsequenzen

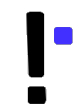
Genomsequenzen sind auffassbar als Strings über dem DNA Alphabet

$$\Sigma_{\mathrm{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}. \tag{1}$$



Phosphatrückgrat (–) Adenin (A) Thymin (T) Guanin (G) Cytosin (C)

Angelehnt an: *http://upload.wikimedia.org/wikipedia/commons/e/e7/DNA_simple.svg*



DNA-Komplement

Definition

Sei die Abbildung $\bar{\cdot} : \Sigma_{\mathrm{DNA}} \cup \{\varepsilon\} \to \Sigma_{\mathrm{DNA}} \cup \{\varepsilon\}$ mit

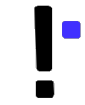
$$\bar{\varepsilon} = \varepsilon, \quad \bar{A} = T, \quad \bar{C} = G, \quad \bar{G} = C, \quad \bar{T} = A$$

gegeben. Für ein Wort $w \in \Sigma_{\mathrm{DNA}}^*$ ist dann das ***Komplement*** $\bar{w}$ gegeben durch

$$\overline{w_1 w_2 \cdots w_n} = \bar{w}_n \cdots \bar{w}_2 \bar{w}_1, \tag{2}$$

mit $w = w_1 w_2 \cdots w_n$, wobei $w_i \in \Sigma$. Die so definierte Abbildung $\bar{\cdot} : \Sigma_{\mathrm{DNA}}^* \to \Sigma_{\mathrm{DNA}}^*$ heißt ***DNA-Komplementierung***, und $\bar{w}$ das ***DNA-Komplement*** des Wortes w.

- Für ein DNA-Wort $w \in \Sigma_{\mathrm{DNA}}^*$ modelliert das Komplement $\bar{w}$ den gegenüberliegenden Helixstrang.
- Komplementäre Wörter können auch als Resultate der DNA-Replikation angesehen werden.



Sätze über komplementäre Genomsequenzen

Theorem (Watson und Crick 1953)

Die Replikation des Wortes des gegenüberliegenden DNA-Strangs ergibt wieder das originale Wort.

Lemma (Dyadische Symmetrie)

Für ein Wort $w \in \Sigma_{\mathrm{DNA}}^*$ *gilt* $w = \bar{w}$ *genau dann, wenn* $w = u\bar{u}$ *für ein* $u \in \Sigma_{\mathrm{DNA}}^*$. *Insbesondere hat w eine gerade Länge.*

Lemma (Kopiereigenschaft)

Für ein Wort $w \in \Sigma_{\mathrm{DNA}}^*$ *mit* $w = \bar{w}$ *folgt* $w^n = \overline{w^n}$ *für alle* $n \in \mathbb{N}$.

DNA-Grammatik: Haarnadel-Moleküle

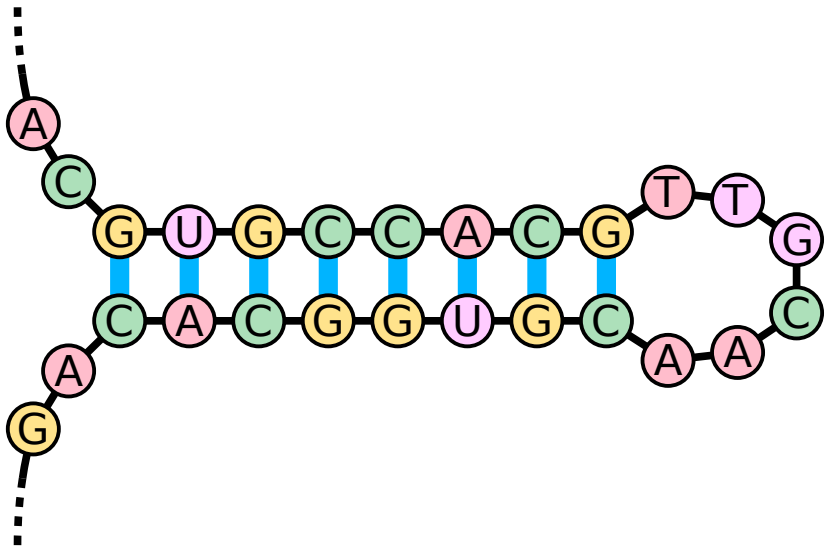
Haarnadel-Moleküle sind Einzelstränge, die sich auf sich selbst falten und sich wie getrennte Stränge verhalten. Sie bilden die formale Sprache

$$L_{\mathrm{h}} = \left\{ w \in \Sigma_{\mathrm{DNA}}^{*} \,\middle|\, w = \bar{w} \right\}. \tag{3}$$

L_{h} ähnelt einer Palindrom-Sprache. Sie wird erzeugt von der kontextfreien Grammatik

$$S \to x S \bar{x} \;\middle|\; \varepsilon \tag{4}$$

mit $x \in \Sigma_{\mathrm{DNA}}$.



Angelehnt an: *http://upload.wikimedia.org/wikipedia/commons/3/3f/Stem-loop.svg*

DNA-Grammatik: Haarnadel-Moleküle

Example

Die Wörter $v = \mathtt{GATC}$ und $w = \mathtt{TGGCCA}$,

$$v = \boxed{\begin{array}{lll} 5' & \longrightarrow \ \mathtt{GA} & \longrightarrow 3' \\ 3' & \longleftarrow \ \mathtt{CT} & \longleftarrow 5' \end{array}} \qquad w = \boxed{\begin{array}{lll} 5' & \longrightarrow \ \mathtt{TGG} & \longrightarrow 3' \\ 3' & \longleftarrow \ \mathtt{GGT} & \longleftarrow 5' \end{array}} \tag{5}$$

sind in L_{h}, da die Produktionen (4) ergeben:

$$S \Rightarrow \mathtt{G}S\mathtt{C} \Rightarrow \mathtt{GA}S\mathtt{TC} \Rightarrow \mathtt{GATC}$$

und

$$S \Rightarrow \mathtt{G}S\mathtt{C} \Rightarrow \mathtt{GG}S\mathtt{CC} \Rightarrow \mathtt{TGG}S\mathtt{CCA} \Rightarrow \mathtt{TGGCCA}.$$

Zu welcher Sprachfamilie gehört die DNA-Grammatik?

Theorem (Searl 1999)

L_{h} ist kontextfrei, aber nicht regulär.

Zu welcher Sprachfamilie gehört die DNA-Grammatik?

Theorem (Searl 1999)

L_{h} ist kontextfrei, aber nicht regulär.

Theorem (Head, Păun und Pixton 2010)

Falls die Sprache der DNA nicht kontextfrei ist, so ist sie auch nicht kontextsensitiv.

Zu welcher Sprachfamilie gehört die DNA-Grammatik?

Theorem (Searl 1999)

L_{h} ist kontextfrei, aber nicht regulär.

Theorem (Head, Păun und Pixton 2010)

Falls die Sprache der DNA nicht kontextfrei ist, so ist sie auch nicht kontextsensitiv.

Vermutung

Die Sprache der DNA ist nicht kontextsensitiv.

Diskussion

Noch Fragen