Chomsky: Sprachen und Grammatiken

DEF. 1: Wir definieren eine Grammatik als Quadrupel $G = (\Sigma, \mathcal{N}, \mathcal{P}, \mathcal{S})$, wobei

- Σ die Menge der Terminale,
- $\mathcal{N}$ die Menge der Nichtterminale (wobei $\Sigma \cap \mathcal{N} = \varnothing$),
- $\mathcal{S} \in \mathcal{N}$ das Startsymbol oder auch Axiom oder Ziel,
- $(\mathcal{V}, \mathcal{P})$ ein Semi-Thue-System und
- $\mathcal{V} = \Sigma \cup \mathcal{N}$ das Vokabular von G ist.

DEF. 2: Sei G eine Grammatik. Dann bezeichnet $\mathcal{L}(G) = \{ w \in \Sigma^* \mid \mathcal{S} \Rightarrow^* w \}$ die Sprache, die von G erzeugt wird.

Wir wollen die Mengen der so erzeugten Sprachen je nach Komplexität der in ihnen enthaltenen Wörter einteilen. Dazu definiert man vier Grammatikklassen, aus denen man vier Sprachklassen ableitet. Im Folgenden definieren wir die vier Chomsky-Typ Grammatiken[1]:

DEF. 3:

CH-0 Grammatik: Produktionen beliebiger Form ($l \mapsto r$; $l,r \in \mathcal{V}^*$) sind zulässig

CH-1 Grammatik: kontextsensitive Produktionen ($uAv \mapsto urv$; $A \in \mathcal{N}$; $r \in \mathcal{V}^+$; $u,v \in \mathcal{V}^*$) sind zulässig; $\mathcal{S}\varepsilon$-Produktionen ($S \mapsto \varepsilon$) sind zulässig, wenn das Startsymbol $\mathcal{S}$ auf keiner rechten Seite vorkommt

CH-2 Grammatik: kontextfreie Produktionen ($A \mapsto r$; $A \in \mathcal{N}$; $r \in \mathcal{V}^*$) sind zulässig

CH-3 Grammatik: entweder linkslineare ($A \mapsto Bx$; $A,B \in \mathcal{N}$; $x \in \Sigma$) oder rechtslineare Produktionen ($A \mapsto xB$; $A,B \in \mathcal{N}$; $x \in \Sigma$) sind zulässig; terminierende Produktionen ($A \mapsto x$; $A \in \mathcal{N}$; $x \in \Sigma$) sind zulässig und ε-Produktionen ($A \mapsto \varepsilon$; $A \in \mathcal{N}$) sind zulässig.

DEF. 4: Eine Sprache L ist vom Typ Chomsky X genau dann, wenn es eine Grammatik G mit $\mathcal{L}(G) = L$ und $G \in$ CH-X gibt. Die dementsprechend erzeugten Sprachen bezeichnet man als

- Chomsky 0, die Klasse der allgemeinen oder berechenbaren Sprachen,
- Chomsky 1, die Klasse der kontextsensitiven oder entscheidbaren Sprachen,
- Chomsky 2, die Klasse der kontextfreien Sprachen und
- Chomsky 3, die Klasse der regulären Sprachen.

Über die Sprachklassen herrscht, wie über die Produktionstypen, weitgehende Einigkeit in der Fachwelt. Die wesentliche Ausnahme davon bildet das leere Wort ε, das, je nach Autor, in den Sprachklassen Chomsky 1 und Chomsky 3 nicht zugelassen wird. Davon abgesehen gilt in Einklang mit DEF. 4 für die Sprachklassen aber folgende Inklusion:

$$\text{Chomsky } 3 \subset \text{Chomsky } 2 \subset \text{Chomsky } 1 \subset \text{Chomsky } 0$$

In unserer Definition der Grammatiken erlauben wir aus praktischen Gründen das Enthaltensein des leeren Wortes ε in Chomsky 3 Sprachen, z.B. um den „*“ in regulären Ausdrücken zuzulassen. Ferner lassen wir $\mathcal{S}\varepsilon$-Produktionen ($S \mapsto \varepsilon$) unter der oben genannten Randbedingung in CH-1 Grammatiken zu, um die Inklusion sicherzustellen.

---

[1] Hinweis: Diese Definitionen sind in der Literatur nicht einheitlich!

Problem: Weniger Einigkeit als über die Sprachklassen herrscht darüber, welche Grammatiktypen welche Produktionstypen zulassen. Letztlich führen aber alle Regelsysteme zu den gleichen Sprachklassen. Damit können Produktionen einer CH-X Grammatik aus einem Regelsystem in CH-X Grammatiken aus einem anderen Regelsystem transformiert werden.

Alternative Regelsysteme fordern beispielsweise beschränkte Produktionen ($l \mapsto r$; $l,r \in \mathcal{V}^*$; $1 \leq |\, l \,| \leq |\, r \,|$) in CH-1 Grammatiken. Es ist offensichtlich, dass die von uns geforderten kontextsensitiven Produktionen beschränkt sind. Andererseits lässt sich aber auch zeigen, dass sich beschränkte Produktionen in kontextsensitive umwandeln lassen.[2] Die Sprachklasse ist also invariant gegenüber der Wahl dieser beiden Definitionen. Da man oft auch von „kontextsensitiven Grammatiken" (statt CH-1) spricht, ist die Definition über kontextsensitive Produktionen die übliche.

Ähnlich fordert die „historische" Definition, dass die Produktionen einer CH-3 Grammatik ausschließlich rechtslinear oder terminierend sind. Die Forderung nach ausschließlich linkslinearen und terminierenden Produktionen ist dazu äquivalent in Bezug auf die erzeugte Sprachklasse; es gibt Algorithmen, die die eine Form in die andere Form überführen. Wichtig ist jedoch, dass rechts- und linkslineare Produktionen nicht gemeinsam verwendet werden dürfen.

Betrachtet man weitere denkbare Produktionstypen, so stellt man fest, dass viele davon die Sprachklasse ebenfalls nicht beeinflussen. Obwohl sie nicht den gegebenen Regeln genügen um als CH-X Grammatik eingestuft zu werden, ändert dies nichts an der Tatsache, dass die erzeugte Sprache eine CH-X-Sprache ist. Die Grammatik könnte also diesbezüglich noch auf das jeweilige Regelsystem „optimiert" werden.

Beispiele für Produktionstypen und wie sie den Chomsky-Typ der Sprache beeinflussen:

- Kettenproduktionen ($A \mapsto B$; $A,B \in \mathcal{N}$) lassen sich vollständig entfernen, daher keine Beeinflussung des Sprachtyps.
- Eine nichtprimitive linkslineare Produktion ($A \mapsto Bx$; $A,B \in \mathcal{N}$; $x \in \Sigma^+$) kann auf primitive linkslineare Produktionen ($A \mapsto B'y$; $y \in \Sigma$) zurückgeführt werden.
- Eine nichtprimitive rechtslineare Produktion ($A \mapsto xB$; $A,B \in \mathcal{N}$; $x \in \Sigma^+$) kann auf primitive rechtslineare Produktionen ($A \mapsto yB'$; $y \in \Sigma$) zurückgeführt werden.
- Eine nichtprimitive terminierende Produktion ($A \mapsto x$; $A \in \mathcal{N}$; $x \in \Sigma^+$) kann auf primitive terminierende Produktionen ($A \mapsto y$; $y \in \Sigma$) zurückgeführt werden.
- ε-Produktion ($l \mapsto \varepsilon$, $l \in \mathcal{V}^*$). Zu Grammatiken mit ε-Produktion gibt es für jede Chomsky-1/2/3-Sprache immer eine äquivalente Grammatik ohne ε-Produktionen (abgesehen von $\mathcal{S}$ε-Produktionen). Daher beeinflussen ε-Produktionen den Sprachtyp nur, falls das leere Wort ε für Chomsky Typ 1 oder 3 Sprachen ausgeschlossen wird (hier nicht der Fall!).

Zusammenfassend lässt sich feststellen, dass es zu einer Chomsky-3-Sprache auch Grammatiken geben, die nicht regulär sind, sondern beispielsweise nur CH-0. Der umgekehrte Fall ist jedoch ausgeschlossen, da die CH-3 Grammatiken hierfür „nicht mächtig genug" sind.

Autoren: Tom Gelhausen (gelhausen@fzi.de), Rubino Geiß im November 2003

[2] siehe Goos, Vorlesungen über Informatik, Band 1, 3. Auflage, S. 36