Prof. Dr. Hans-Jörg Kreowski **Juli 2007**

Formale Sprachen: DNA Computing

Splicing mit linearen Sprachen

Eins der interessantesten Ergebnisse für das Splicing ist die Tatsache, dass sich jede rekursiv aufzählbare Sprache durch 1-Splicing aus einer linearen Sprache mit einer linearen Regelmenge erzeugen lässt.

Theorem 1: $S_1(LIN, LIN) = RE$.

Der Beweis ergibt sich aus einem bekannten Ergebnis der Theorie formaler Sprachen, das besagt, dass jede rekursiv aufzählbare Sprache Linksquotient zweier linearer Sprachen ist.

Definition

Seien $L, L' \subseteq V^*$ zwei Sprachen. Dann heißt die Sprache

$$L \backslash L' = \{x' \mid xx' \in L' \ \& \ x \in L\}$$

der *Linksquotient* von L und L'.

Der Linksquotient enthält also alle Endstücke von Konkatenationen aus L', deren Anfangsstücke aus L sind.

Theorem 2: Sei $L_0 \in RE$. Dann existieren $L, L' \in LIN$ mit $L_0 = L \backslash L'$.

Bezeichnet man die Menge aller Linksquotienten linearer Sprachen mit $LIN \backslash LIN$, kann man Theorem 2 auch so formulieren:

Theorem 2′: $RE = LIN \backslash LIN$.

Theorem 2 sagt explizit nur $RE \subseteq LIN \backslash LIN$. Da aber Linksquotienten von rekursiv aufzählbaren Sprachen selbst rekursiv aufzählbar sind, ergibt sich mit $LIN \backslash LIN \subseteq RE \backslash RE \subseteq RE$ die obige Gleichheit.

Der Beweis von Theorem 1 wird durch die Beobachtung vervollständigt, dass Linksquotienten linearer Sprachen sich auch als 1-Splicing erzeugen lassen.

Beobachtung 1

Seien $L, L' \in LIN$ mit $L, L' \subseteq V^*$. Sei $c \notin V$. Sei $\sigma(L) = (V \cup \{c\}, \{\#c\$cx\# \mid x \in L\})$. Dann gilt:

$$L \backslash L' = \sigma(L)_1(\{c\}L').$$

Gemäß folgender Beobachtung ist mit L auch die Regelmenge linear und mit L' auch $\{c\}L'$.

Beobachtung 2

Sei $L \subseteq T^*$ linear, und seien $u_0, v_0 \in T^*$. Dann ist auch $\{u_0\}L\{v_0\}$ linear.

Somit gilt nach Beobachtung 2 $LIN \backslash LIN \subseteq S_1(LIN, LIN)$, was selbst wieder in RE liegt. Alle Inklusionen zusammen beweisen Theorem 1.

Es bleibt Theorem 2 und die beiden Beobachtungen zu beweisen, was in umgekehrter Reihenfolge geschieht.

Beweis von Beobachtung 2: Sei $G = (N, T, P, S)$ eine lineare Grammatik und $L = L(G)$. Dann erzeugt die lineare Grammatik $G_0 = (N \cup \{S_0\}, T, P \cup \{S_0 ::= u_0 S v_0\}, S_0)$ offensichtlich $\{u_0\}L\{v_0\}$.

Beweis von Beobachtung 1: Sei $x' \in \sigma(L)_1(\{c\}L')$. Das bedeutet, dass eine Regel $\#c\$cx\#$ mit $x \in L$ und $cy, cz \in \{c\}L'$ existieren mit

$$(cy, cz) \vdash_{\#c\$cx\#} x'.$$

Da c weder in y noch in z vorkommt, geht das nur, wenn $cz = cxx'$ und damit $z = xx' \in L'$. Es folgt $x' \in L \backslash L'$.

Sei umgekehrt $x' \in L \backslash L'$, was nach Definition bedeutet, dass ein $x' \in L$ existiert mit $xx' \in L'$. Dann gilt aber auch:

$$(cxx', cxx') \vdash_{\#c\$cx\#} x'.$$

Somit erweist sich $x' \in \sigma(L)_1(\{c\}L')$.

Beides zusammen beweist die behauptete Gleichheit der Sprachen.

Beweis von Theorem 2: Sei $L_0 \in RE$. Dann ist auch $L_0^t \in RE$. Und es existiert eine Typ0-Grammatik $G = (N, T, P, S)$ mit $L(G) = L_0^t$. Nach Definition gibt es dann für $w \in L(G) \subseteq T^*$ eine Ableitung der Form

$$S = w_1 \rightarrow w_2 \rightarrow \ldots \rightarrow w_{n-1} \rightarrow w_n \rightarrow w_{n+1} = w$$

und Regeln $u_i ::= v_i \in P$ für $i = 1, \ldots, n$ mit $w_i = x_i u_i y_i$ und $w_{i+1} = x_i v_i y_i$. Der Aufbau dieser Wörter kann in einem Wort zusammengefasst werden, indem man sie hintereinanderhängt. Um sie unterscheiden zu können, werden sie dabei jeweils durch c getrennt bzw. cc in der Mitte. Aus technischen Gründen werden die Wörter ab der Mitte transponiert. Das letzte Wort $(w_{n+1} = w)$ wird sogar durch ccc abgetrennt. Das ergibt

$$(*)\ x_n u_n y_n c \ldots c x_1 u_1 y_1 cc\ y_1^t v_1^t x_1^t c \ldots c\ y_{n-1}^t v_{n-1}^t x_{n-1}^t ccc\, y_n^t v_n^t x_n^t.$$

Der Teil bis ccc ist identisch zu

$$(**)\ w_n c \ldots c w_2\, cScc\, w_2^t\, c \ldots c\, w_n^t\, ccc$$

Wenn alle Wörter der Form $(*)$ die Sprache L_P bilden und alle Wörter der Form $(**)$ die Sprache $L_=$, bedeutet das gerade:

$$w^t = w_{n+1}^t = y_n^t v_n^t x_n^t \in L_= \backslash L_P.$$

Somit gilt $L_0 = (L_0^t)^t = L(G)^t = L_= \backslash L_P$. Bleibt also nur zu zeigen, dass die beiden Sprachen L_P und $L_=$ linear sind. Dazu werden für beide Sprachen lineare Grammatiken konstruiert, die sie erzeugen:

1. $G_P = (\{X_0, X_1, X_2, X_3\}, N \cup T \cup \{c\}, P_P, X_0)$ mit den Regeln

   $X_0 ::= aX_0\, a$ für $a \in T$

   $X_0 ::= uX_1\, v^t$ für $u ::= v \in P$ und $v \in T^*$

   $X_1 ::= aX_1\, a$ für $a \in T$

   $X_1 ::= cX_2\, ccc$

   $X_2 ::= aX_2\, a$ für $a \in N \cup T$

   $X_2 ::= uX_3\, v^t$ für $u ::= v \in P$

   $X_3 ::= aX_3\, a$ für $a \in N \cup T$

   $X_3 ::= cX_2\, c \mid cc$

2. $G_= = (\{Y_0, Y_1\}, N \cup T \cup \{c\}, P_=, Y_0)$ mit den Regeln

   $Y_0 ::= Y_1\, ccc$

   $X_1 ::= aY_1\, a$ für $a \in N \cup T \cup \{c\}$

   $Y_1 ::= cScc$

Es ist nicht allzu schwer, einzusehen, dass $L(G_P) = L_P$ und $L(G_=) = L_=$ gilt. Damit ist Theorem 2 bewiesen.