Mathematische Grundlagen:
Formale Sprachen und Automaten

Literatur

Dieses Skript basiert den Skripten von Werner Saurer und Espen Vestre und orientiert sich an:

Harry R. Lewis and Christos H. Papadimitriou. Elements of the Theory of Computation (2nd Edition). Prentice Hall 1998.

Weitere Literatur:

Barbara H. Partee, Alice ter Meulen and Robert E. Wall. Mathematical Methods in Linguistics. Kluwer 1990.

Hopcroft and Ullmann: Introduction to Automata Theory, Languages and Computation. Addison-Wesley 1979.

Programm der Vorlesung

Die Vorlesung gliedert sich in drei Teile:

1. Reguläre Sprachen und endliche Automaten:
   L&P Kapitel 1.8, 1.9, 2
2. Kontextfreie Sprachen und Kellerautomaten:
   L&P Kapitel 3
3. Typ-0 Sprachen und Turing-Maschinen:
   L&P Kapitel 4 und 5.2

Einführung

Formale Sprachen sind Mengen von Wörtern über einem Alphabet von Zeichen. Zum Beispiel ist

$abba$

ein *Wort über dem Alphabet* $\Sigma = \{a, b\}$ und

$L = \{ acb, aacbb, aaacbbb, \ldots \}$

$= \{ a^n cb^n : n \geq 1 \}$

eine *Sprache über* Σ.

(a^n bezeichnet eine Folge von n a's.)

Im nicht-trivialen Fall sind Sprachen immer unendlich, daher benötigen wir eine *endliche* Charakterisierung einer *unendlichen* Sprache.

Charakterisierungen von Sprachen

In dieser Vorlesung betrachten wir drei verschiedene Charakterisierungen von Sprachen:

- Automaten
- Grammatiken
- Reguläre Ausdrücke

Was haben diese Charakterisierungen miteinander zu tun?

- Automaten = Sprachversteher. Automaten *akzeptieren* Wörter einer Sprache.
- Grammatiken = Spracherzeuger. Grammatiken *generieren* die Wörter einer Sprache.

aber in einem gewissen Sinne sind Automaten mit Grammatiken äquivalent.

Wir werden eine Hierarchie von Sprachen, Grammatiken und Automaten kennenlernen, zwischen denen bestimmte Beziehungen bestehen.

Ein Beispiel

Die Sprache $L = \{a^n cb^n : n \geq 0\}$ ist eine *kontextfreie* Sprache. Sie kann einerseits durch eine kontextfreie Grammatik mit den Regeln

R_1: $S \to c$

R_2: $S \to aSb$

generiert werden. Das zu L gehörige Wort *aaacbbb* kann z.B. mit folgender Ableitung erzeugt werden:

1. S	Start
2. aSb	R_2, aus 1
3. $aaSbb$	R_2, aus 2
4. $aaaSbbb$	R_2, aus 3
5. $aaacbbb$	R_1, aus 4

Auf der anderen Seite gibt es einen Automaten, einen sogenannten *Kellerautomaten*, der genau die Wörter in L akzeptiert. Seine Arbeitsweise kann grob so beschrieben werden: Er fängt an beim ersten Zeichen der Eingabe. Er zählt zunächst die a's, falls vorhanden. Nachdem er ein c gelesen hat, zählt er die b's „rückwärts." Hat er genau so viele b's gezählt wie a's, akzeptiert er das Wort. Fehlen b's oder sind zu viele b's da, weist er die Eingabe als nicht zu L gehörig zurück.

Für jede kontextfreie Sprache gibt es eine kontextfreie Grammatik und einen Kellerautomaten, die genau diese Sprache generieren bzw. akzeptieren. Grammatik und Automat sind also in einem gewissen Sinne äquivalent, obwohl sie ganz verschiedene formale Mechanismen sind. Sie beleuchten aber verschiedene Aspekte ein und derselben Sache, nämlich einer formalen Sprache eines bestimmten Typs.

Ähnliches gilt für andere Sprachen in dieser Hierarchie.

Fragen der Komplexität von Sprachen

- Mit welchen minimalen Mitteln kann eine Sprache erzeugt und verstanden werden? Die Antwort sagt etwas über die Komplexität von Sprachen aus. So sind z.B. kontextfreie Sprachen komplexer als reguläre Sprachen, da sie zu ihrer Charakterisierung kontextfreie Grammatiken brauchen, die nicht regulär sind, und zu ihrer Erkennung Kellerautomaten und nicht nur Endliche Automaten.
- Als Linguisten sind wir speziell an der Komplexität von natürlichen Sprachen interessiert. Wie komplex sind natürliche Sprachen? D.h. welche Typen von Grammatiken und Automaten benötigt man zu ihrer Charakterisierung?
- Die Menge aller endlich langen Zeichenketten über einem Alphabet L, üblicherweise mit L^* bezeichnet, ist abzählbar unendlich. Damit gibt es überabzählbar viele Teilmengen von L^*, also überabzählbar viele Sprachen über L. Es kann aber höchstens abzählbar viele endliche Charakterisierungen – formale Grammatiken, Automaten, etc. – von Sprachen geben. Damit ist klar, dass es über-

abzählbar viele Sprachen über L gibt, die nicht charakterisiert werden können, die also nicht durch eine Grammatik oder einen sonstigen Formalismus erfasst werden können. Das sind also höchst „unregelmäßige“ Sprachen.