Einführung in die Computerlinguistik

Endliche Automaten II/ Morphologiesysteme/ Syntaktische Strukturen

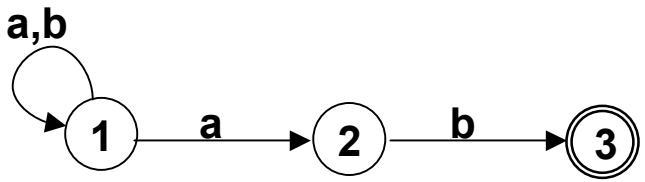
WS 2005/2006

Manfred Pinkal

Potenzautomatenkonstruktion, ein weiteres Beispiel

NEA A= < {1,2,3}, {a,b}, Δ,1, {3} >

Δ gegeben durch:



DEA

A‘ = < $\wp$ ({1,2,3}), {a,b}, δ, {1}, F‘>)

F´= {{3}, {1,3}, {2,3}, {1,2,3}}

Potenzautomatenkonstruktion, Beispiel 2: Die Übergangstabelle

q	δ(q, a)	δ(q, b)
{1}	{1,2}	{1}
{2}	Ø	{3}
{3}	Ø	Ø
{1,2}	{1,2}	{1,3}
{1,3}	{1,2}	{1}
{2,3}	Ø	{3}
{1,2,3}	{1,2}	{1,3}
Ø	Ø	Ø

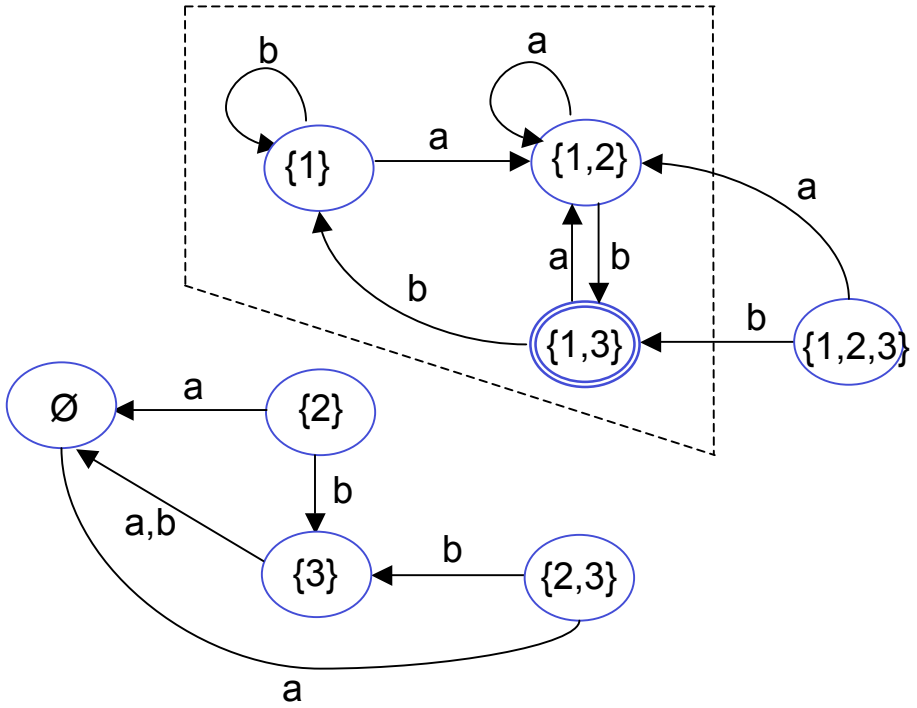
Potenzautomatenkonstruktion, Beispiel 2: Die Übergangstabelle

q	δ(q, a)	δ(q, b)
{1}	{1,2}	{1}
{2}	Ø	{3}
{3}	Ø	Ø
{1,2}	{1,2}	{1,3}
{1,3}	{1,2}	{1}
{2,3}	Ø	{3}
{1,2,3}	{1,2}	{1,3}
Ø	Ø	Ø

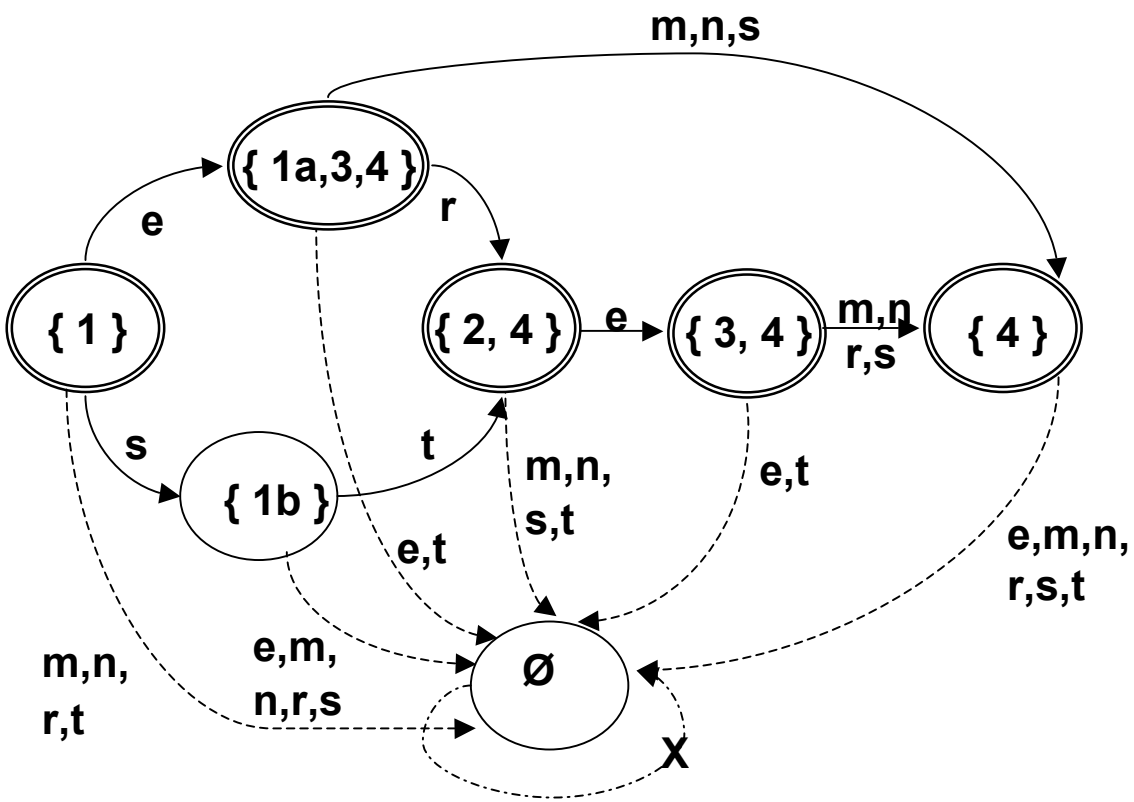
Potenzautomatenkonstruktion, Beispiel 2: Das Zustandsdiagramm

Nur ein Teil der Zustände ist vom Startzustand aus erreichbar.

Die übrigen Zustände sind funktionslos.



Noch einmal der Adjektivendungs-DEA



Noch einmal: Der DEA für Adjektiv-Endungen

- Die Zustandsmenge des Potenzautomaten A‘ ist eigentlich K' = $\wp$(K), er hat in unserem Beispiel also $2^6 = 64$ Zustände. Wie die Übergangstabelle zeigt, sind vom Startzustand {1} aus aber nur 7, ohne den „trap state“ ∅ 6 echte Zustände erreichbar. Die übrigen Zustände können ignoriert werden.

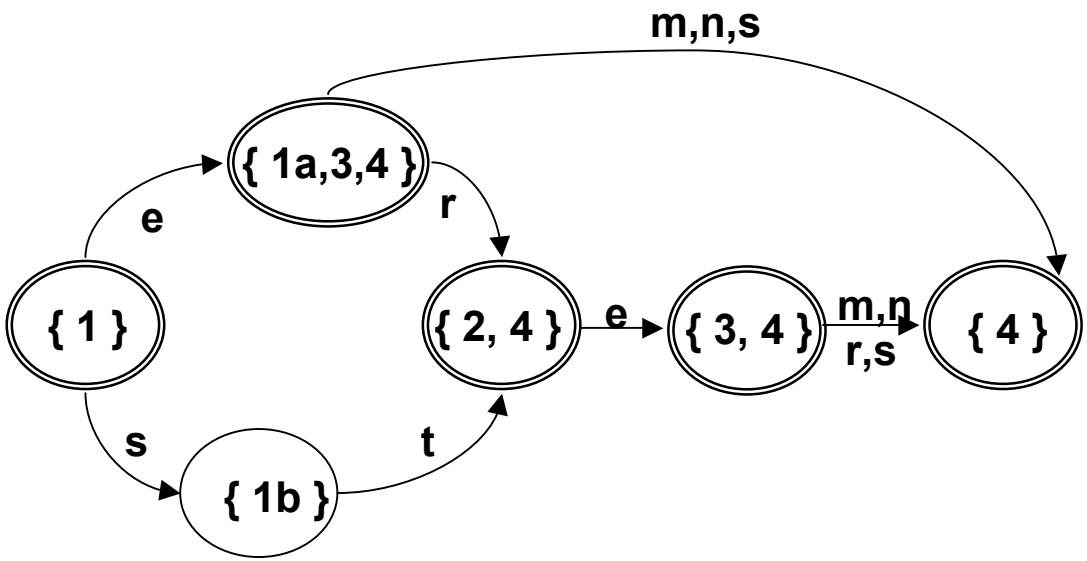
  Wir können also, wie im Diagramm, ausgehen von:

  K‘ = {{1}, {1a,3,4}, {1b}, {2,4}, {3,4}, {4}, ∅} und

  F‘ = {{1}, {1a,3,4}, {2,4}, {3,4}, {4}}

- Im Diagramm können wir außerdem noch, per Konvention, den Zustand ∅ und alle hinführenden Kanten unterschlagen, und erhalten dann das vereinfachte Diagramm auf der folgenden Folie.

Das Diagramm, vereinfacht



Morphologiesysteme

- Flexionsmorphologie: Lemmatisierung/Stemming
  - *veranstalt+et, Veranstaltung+en*
- Ableitungs-/Derivationsmorphologie
  - *Veranstalt+ung, un+glaubwürdig*
- Komposita-Zerlegung
  - *Fach+veranstaltung, glaub+würdig*

Morphologiesysteme

- Flexionsmorphologie: Lemmatisierung/Stemming
- Ableitungs-/Derivationsmorphologie
- Komposita-Zerlegung

Lemmatisierung/Stemming

- Rückführung flektierter Formen auf Stämme, erlaubt je nach morphologischer Struktur der Sprache Reduktion der Lexikongröße (Verhältnis Wortstämme/Lemmata : Wortformen) von 1:2 (Englisch), 1:5 (Deutsch), 1:200 und mehr (Türkisch, Finnisch).
- Ermittlung von grammatischen Merkmalen, die für die syntaktische Analyse nötig sind.
- Methode: Endliche Automaten + Flexionsklassen-Information im Lexikon + Regeln für vorhersagbare morphophonologische Prozesse
- Morphologische Analyse (und damit auch deren Umkehrung, die Wortformengenerierung) ist für das Stemming gelöst. Verbleibendes Problem für das Deutsche: Trennbare Präfixverben. Kommerzielle Systeme haben eine gute Abdeckung (bis ca. 98%). Abdeckungsprobleme sind meist auf Abdeckungsprobleme bei der Kompositazerlegung zurückzuführen.

Wortbildungsmorphologie: Ableitung/Derivationsmorphologie

- Analyse liefert über den Wortstamm Information zur Ableitung der Bedeutung und über das Suffix Wortart- und Flexionsklasseninformation: Vervielfältig+ung bezeichnet den Vorgang des Vervielfältigens, ist feminines Substantiv und wird schwach flektiert.
- Zwei Probleme, die zusammenhängen:
  - viele Ableitungsprä- und -suffixe sind semiproduktiv
  - viele Ableitungen sind semantisch "nicht transparent": Sie haben eine konventionelle, lexikalisierte Bedeutung, die sich nicht aus den Bestandteilen erschließen lässt.
- Derivationsanalyse führt aus diesen Gründen zur Übergenerierung:
  - die *Lesung* bezeichnet den Akt des Vorlesens,
  - die *Vorlesung* aber nicht den Akt des Vorlesens,
  - die *Schreibung* nicht den Akt des Schreibens, und
  - die *Singung* ist ganz unmöglich

Wortbildungsmorphologie: Ableitung/Derivationsmorphologie

- Die Analyse von Derivativen ist in Morphologie-systeme meist mehr oder weniger ausführlich mit integriert. Sie kann grundsätzlich mit endlichen Automaten realisiert werden.
- Wegen Semiproduktivität und fehlender semantischer Transparenz ist es sinnvoll, Wissen über Ableitungen im großen Umfang im Lexikon zu kodieren.

Morphologiesysteme

- Lemmatisierung/Stemming
- Ableitungs-/Derivationsmorphologie
- Komposita-Zerlegung

Kompositazerlegung

- Die Analyse von Komposita ist besonders in Sprachen wie dem Deutschen unerlässlich (potienziell beliebige Länge von Komposita und deshalb unbegrenzt großer Wortschatz).
- Rechtschreibprüfung ohne oder mit begrenzter Kompositabehandlung führt zu vielen korrekten, aber unerkannten Wörtern (MS Word bis vor einigen Jahren).
- Vollständigere Suche beim Informationszugriff (z.B. IR): Suche nach *Rentenversicherung* soll auch *Angestelltenrentenversicherung* finden.

Kompositazerlegung: Probleme

- Fugenelemente:
  - sind keine Flexionssuffixe, vgl.: "*Absichts"* in *A*bsichtserklärung
  - sind nicht vollständig aus der phonologischen/ orthografischen Umgebung vorhersagbar
  - gehören nicht zu der Information, die in Einträgen konventioneller Wörterbücher mit augeführt wird.
- Übergenerierung:
  - Beispiel zur Konversion alte-neue Rechtschreibung (falsch angewandte Drei-Konsonsanten-Regel)
    - Hufeisenniere → Huf|ei|senn|niere
  - Beispiel aus einem (älteren) Maschinellen Übersetzungssystem:
    - Bar|bar|ei → nightclub nightclub egg

Komposita-Zerlegung in Morphologiesystemen

- Gute Morphologien haben eine Kompositazerlegung mit guter Abdeckung und akzeptablem Übergenerierungsverhalten.
- Zur Vermeidung von Übergenerierung werden Blockierungslisten mit kurzen und seltenen Wörtern angelegt. Komposita, die diese Wörter enthalten, werden explizit im Lexikon aufgeführt.

Morphologiesysteme insgesamt

- Abdeckung (und Präzision)
- Zeitbedarf
- Preis

Anwendungen für endliche Automaten

- Morphologische Analyse
- Template-basierte Suche in Textdokumenten (PERL: Sprache zur Stringsuche)
- Dialogmodellierung
  - Beschreibung von Dialogmustern mit Automaten
- Syntax ???

Syntax

- Gegenstand der Morphologie ist die Struktur des Wortes: der Aufbau von Wörtern aus Morphemen, den kleinsten funktionalen oder bedeutungstragenden Einheiten der Sprache.
- Gegenstand der Syntax ist die Struktur des Satzes: der Aufbau von Sätzen aus Wörtern.
- Morphologie beschreibt die grammatischen Eigenschaften von Wörtern, die durch Wortform oder Flexionsmorpheme kodiert werden.
- Syntax beschreibt die Interaktion der grammatischen Eigenschaften unterschiedlicher Wörter im Satz.

Eigenschaften der syntaktischen Struktur [1]

- Sätze setzen sich aus Satzteilen (Konstituenten) zusammen, die wiederum aus einfachen oder ihrerseits komplexen Satzteilen bestehen können. Sätze können deshalb beliebig lang und beliebig tief geschachtelt sein.

Konstituenten

- *Er hat die Übungen gemacht.*
- *Der Student hat die Übungen gemacht.*
- *Der interessierte Student hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach Informatik studiert, hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert, hat die Übungen gemacht.*

Syntaktische Kategorien [1]

Konstituenten-Typen werden „syntaktische Kategorien“ genannt;

Beispiele:

- Nominalphrasen (Nominalausdrücke): *er – der Student – der interessierte Student – die Übungen – computerlinguistischen Fragestellungen*
- Präpositionalphrasen (Präpositionalausdrücke): *an computerlinguistischen Fragestellungen – im ersten Semester, – nach langer Überlegung*
- Adjektivphrasen: *interessierte – an computerlinguistischen Fragestellungen interessierte*
- Satz: Haupt- und Nebensätze unterschiedlicher Art

Syntaktische Kategorien [2]

Konstituenten /syntaktische Kategorien können beliebig ineinander verschachtelt sein:

- Der Nominalausdruck *„der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert“* enthält
- den (Relativ-)Satz „ *der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert*“; der enthält
- den Nominalausdruck „*Hauptfach, für das er sich nach langer Überlegung entschieden hat“;* der enthält
- den (Relativ-)Satz „ *für das er sich nach langer Überlegung entschieden hat*“; der enthält
- unter anderem den Nominalausdruck *„er“.*

Eigenschaften der syntaktischen Struktur

- Sätze setzen sich aus Satzteilen (Konstituenten) zusammen, die wiederum aus einfachen oder ihrerseits komplexen Satzteilen bestehen können. Sätze können deshalb beliebig lang und beliebig tief geschachtelt sein.
- Die Syntax natürlicher Sprachen erlaubt variable Wortstellung: Wörter und Konstituenten mit der gleichen Funktion können an unterschiedlichen Stellen eines Satzes stehen. Unterschiedliche Sprachen erlauben sehr unterschiedliche Freiheitsgrade.

Variable Wortstellung

Peter hat der Dozentin das Übungsblatt heute ins Büro gebracht.

Das Übungsblatt hat Peter der Dozentin heute ins Büro gebracht.

Der Dozentin hat Peter heute das Übungsblatt ins Büro gebracht.

Ins Büro hat heute Peter der Dozentin das Übungsblatt gebracht.

Heute hat Peter das Übungsblatt der Dozentin ins Büro gebracht.

Ins Büro hat das Übungsblatt der Dozentin Peter heute gebracht.

** Ins Büro heute Peter das Übungsblatt hat gebracht der Dozentin.*

** Ins heute Büro der Peter Dozentin das hat Übungsblatt gebracht.*

Eigenschaften der syntaktischen Struktur [3]

- Sätze setzen sich aus Satzteilen (Konstituenten) zusammen, die wiederum aus einfachen oder ihrerseits komplexen Satzteilen bestehen können. Sätze können deshalb beliebig lang und beliebig tief geschachtelt sein.
- Die Syntax natürlicher Sprachen erlaubt variable Wortstellung: Wörter und Konstituenten mit der gleichen Funktion können an unterschiedlichen Stellen eines Satzes stehen. Unterschiedliche Sprachen erlauben sehr unterschiedliche Freiheitsgrade.
- Die grammatischen Eigenschaften unterschiedlicher Wörter und Konstituenten im Satz hängen voneinander ab – zum Teil auch in Fällen, in denen die Wörter und Konstituenten im Satz weit auseinander liegen.

Eigenschaften der syntaktischen Struktur [3]

- *Der [m,sg, nom]an computerlinguistischen Fragestellungen interessierte [m,sg, nom] Student [m,sg, nom] im ersten Semester, der [m,sg, nom] im Hauptfach, für das er [m,sg, nom] sich nach langer Überlegung entschieden hat [sg], Informatik studiert [sg], hat die Übungen gemacht.*