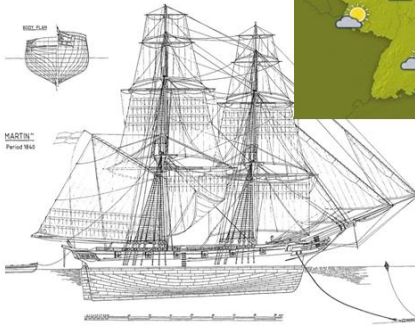
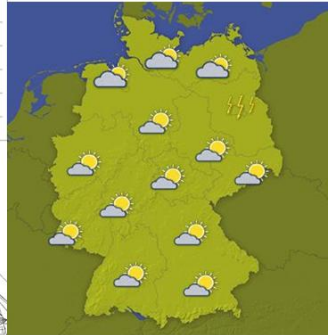
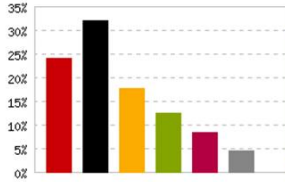
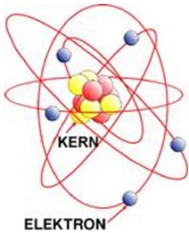
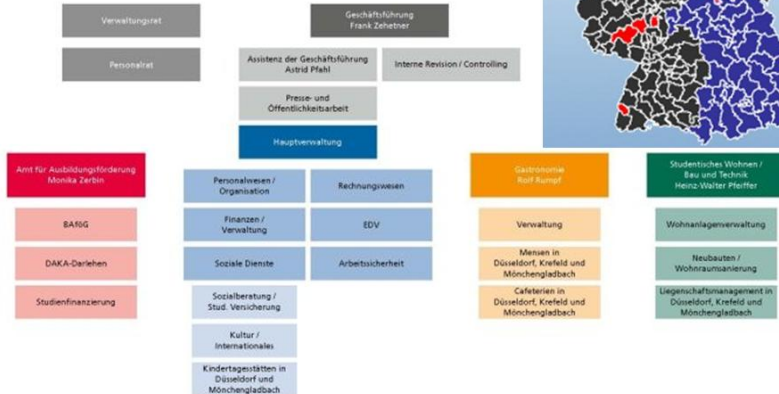
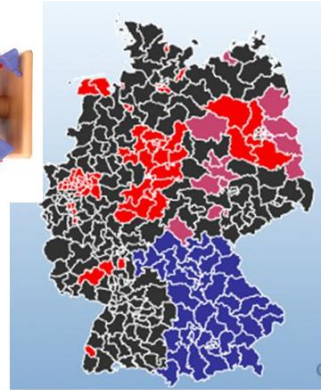
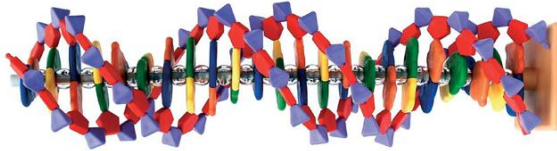


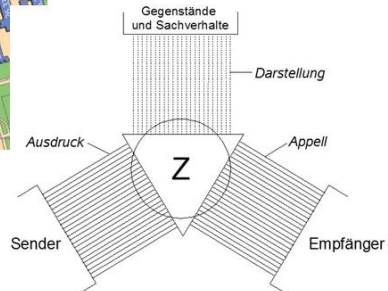
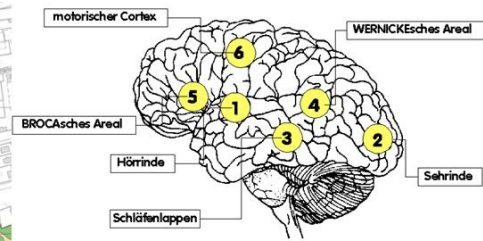
Einführung in die Computerlinguistik – formale Sprachen

Dozentin: Wiebke Petersen

26.4.2010







Modell

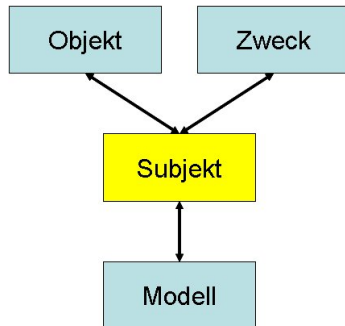
- künstlich geschaffen
- materiell oder immateriell
- vereinfachtes Abbild
- zweckgerichtet
- Abstraktion
- Repräsentation
- Modellierungsannahmen

Modellierung

Ein **Subjekt** entwirft zu einem **Original** ein **Modell** zu einem bestimmten **Zweck**.

Stachowiak:

- Abbildsmerkmal
- Vereinfachungsmerkmal
- Pragmatisches Merkmal



Modellierung natürlicher Sprachen

Formale Sprachen

Formale Sprachen sind Mengen von **Wörtern** (entspricht in natürlichen Sprachen den **Sätzen**), die ihrerseits aus **Zeichen/Symbolen** (in natürlichen Sprachen **Wörtern**) aufgebaut sind. Was in der Menge ist, ist ein “grammatisch korrektes Wort”, alles andere nicht.

Für “strukturierte” formale Sprachen lassen sich endliche Mengen von Regeln/Grammatiken angeben, die diese beschreiben.

Sprachmodell

Formale Sprachen dienen als Modell für natürliche Sprachen.

Wir gehen davon aus, daß alle natürlichen Sprachen durch endlich viele Regeln beschreibbar sind, da wir sie ansonsten nicht sprechen / verstehen könnten.

Welche Modellannahmen werden hier implizit gemacht?

Mengen

Georg Cantor (1845-1918)

Eine **Menge** ist eine Zusammenfassung beliebiger Objekte, genannt Elemente, zu einer Gesamtheit, wobei keines der Objekte die Menge selbst sein darf. Zwei Mengen sind **gleich**, g.d.w. sie die gleichen Elemente enthalten. Es gibt genau eine Menge, die keine Elemente enthält, die **leere Menge** \emptyset .



Mengenbeschreibungen

explizite Mengendarstellung $\{a_1, a_2, \dots, a_n\}$ ist die Menge, die genau die Elemente a_1, a_2, \dots, a_n enthält.

Beispiel: $\{2, 3, 4, 5, 6, 7\}$

implizite Mengendarstellung $\{x|A\}$ ist die Menge, die genau die Objekte x enthält, auf die die Aussage A zutrifft.

Beispiel: $\{x|x \in \mathbb{N} \text{ und } x < 8 \text{ und } 1 < x\}$,
 $\{x|x \in \mathbb{N} \text{ und } x \text{ ist eine gerade Zahl}\}$

Notation

$x \in M$: x ist ein **Element** der Menge M ($2 \in \{1, 2, 3\}$, $2 \notin \{1, 3, 5\}$)

$N \subseteq M$: die Menge N ist eine **Teilmenge** der Menge M

($\{2, 3\} \subseteq \{1, 2, 3, 4\}$)

Hinweise: Die leere Menge ist eine Teilmenge jeder Menge

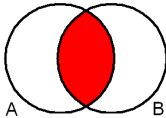
($\emptyset \subseteq \{1, 2, 3, 4\}$)

$N \subset M$: die Menge N ist eine **echte Teilmenge** der Menge M

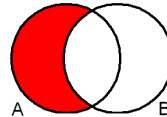
($\{1, 2\} \subseteq \{1, 2\}$ aber $\{1, 2\} \not\subset \{1, 2\}$, Bsp. $\{1\} \subset \{1, 2, \}$)

Mengenoperationen

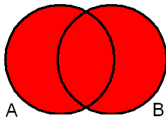
Schnitt: $A \cap B$



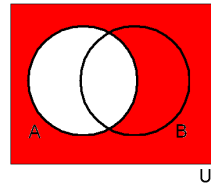
Differenz: $A \setminus B$



Vereinigung: $A \cup B$



Komplement (in U): $C_U(A)$



Wenn U feststeht, dann auch \bar{A}

Potenzmenge

Die **Potenzmenge** einer Menge M ist die Menge aller Teilmengen von M , also $\mathcal{POT}(M) = \{N \mid N \subseteq M\}$.

Für endliche Mengen gilt: ist M eine n -elementige Menge, so ist $\mathcal{POT}(M)$ eine 2^n -elementige Menge.

$$\begin{array}{l} \{1 \quad 2 \quad 3\} \\ \{1 \quad 2 \quad \quad\} \\ \{1 \quad \quad \quad 3\} \\ \{ \quad 2 \quad 3\} \\ \{1 \quad \quad \quad \quad\} \\ \{ \quad 2 \quad \quad \quad\} \\ \{ \quad \quad \quad 3\} \\ \{ \quad \quad \quad \quad \quad\} \end{array}$$

Alphabete und Wörter

Definition

- **Alphabet Σ** : nichtleere endliche Menge von **Symbolen / Zeichen**.
- **Wort**: eine endliche Kette/Folge $x_1 \dots x_n$ von Symbolen/Zeichen eines Alphabets ($n \geq 0$). Das Wort, das aus null Zeichen besteht heißt **leeres Wort** und wird mit ϵ bezeichnet.

Die Menge aller Wörter über einem Alphabet Σ bezeichnen wir mit Σ^* .

$\Sigma^+ = \Sigma^* \setminus \{\epsilon\}$ ist die Menge der nichtleeren Wörter.

- **Länge** eines Wortes $|w|$: Gesamtzahl der Zeichen eines Wortes w ($|abbaca| = 6$, $|\epsilon| = 0$)

Leersymbol, leeres Wort und leere Menge

Vorsicht Verwechslungsgefahr!

Das **Leersymbol** \sqcup ist ein *Zeichen* des Alphabets, also auch ein Wort der Länge 1.

Das **leere Wort** ϵ ist ein *Wort* der Länge 0.

Die **leere Menge** \emptyset ist eine *Menge*.

Übung: Alphonete und Wörtern

Sei $\Sigma = \{a, b, c\}$ ein Alphabet:

- Gib ein Wort der Länge 4 über Σ an.
- Welche der folgenden Ausdrücke sind Wörter über Σ und welche Länge haben sie:
'aa', 'caab', 'da'
- Was ist der Unterschied zwischen Σ^* , Σ^+ und Σ ?
- Wieviele Elemente haben Σ^* und Σ^+ ?

Operationen auf Wörtern

Definition

Verkettung / Konkatenation Die **Konkatenation / Verkettung** zweier Wörter $u = a_1 a_2 \dots a_n$ und $v = b_1 b_2 \dots b_m$ mit $n, m \geq 0$ ist

$$u \circ v = a_1 \dots a_n b_1 \dots b_m$$

Häufig schreiben wir uv statt $u \circ v$.

$$w \circ \epsilon = \epsilon \circ w = w \quad \text{Neutrales Element}$$

$$u \circ (v \circ w) = (u \circ v) \circ w \quad \text{Assoziativität}$$

Operationen auf Wörtern

Exponenten

- w^n : w wird n -mal mit sich selbst verkettet.
- $w^0 = \epsilon$: w wird '0-mal' mit sich selbst verkettet.

Umkehrung

- Die **Umkehrung** eines Wortes w wird mit w^R bezeichnet.
 $(abcd)^R = dcba$.
- Ein Wort w , für das $w = w^R$ gilt, heißt **Palindrom**.

(madam, reliefpfeiler, otto, anna, ...)

Übung: Operationen auf Wörtern

Seien $w = abc$ und $v = bcc$ Wörter, berechne:

- $w \circ v$
- $((w^R \circ v)^R)^2$
- $w \circ (v^R \circ w^3)^0$

Formale Sprache

Definition

Eine *formale Sprache* L ist eine Menge von Wörtern über einem Alphabet Σ , also $L \subseteq \Sigma^*$.

Beispiele:

- Sprache L_{rom} der gültigen römischen Zahldarstellungen über dem Alphabet $\Sigma_{rom} = \{I, V, X, L, C, D, M\}$.
- Sprache L_{Mors} der Buchstaben des lateinischen Alphabets dargestellt im Morsecode. $L_{Mors} = \{\cdot-, -\cdots, \dots, -\cdots\}$
- Sprache L_{pal} der Palindrome im deutschen Duden
 $L_{pal} = \{\text{Madam, reliefpfeiler, \dots}\}$
- Leere Menge
- Menge der Wörter der Länge 13 über dem Alphabet $\{a, b, c\}$
- Sprache der syntaktisch wohlgeformten Java-Programme
- Deutsch?

Operationen auf Sprachen

Seien $L \subseteq \Sigma^*$ und $K \subseteq \Sigma^*$ zwei Sprachen über dem Alphabet Σ , dann entstehen durch die Verknüpfung mit Mengenoperatoren neue Sprachen über Σ :

$$K \cup L, K \cap L, K \setminus L$$

Die Verkettung von Wörtern kann ausgedehnt werden auf die Verkettung von Sprachen:

$$K \circ L := \{v \circ w \in \Sigma^* \mid v \in K, w \in L\}$$

Beispiel: Sei $K = \{abb, a\}$ und $L = \{bbb, ab\}$

- $K \circ L = \{abbbbb, abbab, abbb, aab\}$ und
 $L \circ K = \{bbbabb, bbba, ababb, aba\}$
- $K \circ \emptyset = \emptyset$
- $K \circ \{\epsilon\} = K$
- $K^2 = K \circ K = \{abbabb, abba, aabb, aa\}$

Hausaufgaben

(Abgabe bis zum 6.5.2010; für den BN: 2 aus 5)

Sei $K = \{aa, aaa, ba\}$, $L = \{bb, aa\}$

- 1 Geben sie die Sprachen $L \circ L$, $L \circ K$, $\{\epsilon\} \circ L$, $\{\epsilon\} \circ \emptyset$ und $K \circ \emptyset$ an.
- 2 Geben sie die Sprache L^3 an.
- 3 Geben sie die Sprache $L \setminus K$ an.
- 4 Geben sie eine implizite Mengendarstellung der Sprache $K \circ K$ an.
- 5 Wie unterscheiden sich die Sprachen L^* und L^+ ?