5. Maschinelle Verarbeitung natürlicher Sprache (Computerlinguistik) (1)

5.1 Vorbemerkungen

Wissen, das als Ergebnis menschlichen Denkens vorliegt, wird durch Sprache mitgeteilt. Unterscheidung von Sprachen:

a) **Natürliche** Sprachen (z.B. Deutsch, Englisch, Französisch, Türkisch). Natürliche Sprachen dienen der Mitteilung von Mensch zu Mensch.

b) **Formale** Sprachen (z. B. Programmiersprachen (z.B. Java, C, SQL), Auszeichnungssprachen (z.B. HTML, XML)). Formale Sprachen dienen der Mitteilung von Mensch zur Maschine oder von Maschine zur Maschine.

Natürliche Sprachen sind Vorbilder für formale Sprachen. Formale Sprachen versuchen den Formenreichtum **syntaktischer Konstrukte einzuschränken** (Reduktion von Komplexität) und Wörter in ihrer **Bedeutung eindeutig zu machen** (Heteronymie und Monosemie statt Homonymie und Polysemie).

5.1 Vorbemerkungen (Forts.) (2)

Anm.: **Homonymie** := gleichgeschriebene Wörter mit unterschiedlichen Bedeutungen und unterschiedlichen grammatischen Eigenschaften (z.B. das „Tor“: a) begehbare Öffnung in einem Gebäude, b) aus zwei Pfosten und einer Latte gebautes Konstrukt bei Ballspielen; der „Tor“: Mensch, dessen Handlungsweise als unklug angesehen wird. Weiteres Beispiel: die „Heide“ / der „Heide“).

Polysemie := gleichgeschriebene Wörter mit unterschiedlichen Bedeutungen (z.B. das Futter: a) Nahrungsmittel für Tiere, b) Stoff auf der Innenseite von Kleidungsstücken).

Im Ausschluss von Polysemie sind **semiformale** Sprachen zwischen natürlichen Sprachen und formalen Sprachen angesiedelt.

Bsp.: Die semiformale Sprache der **Geometrie** in den Büchern Euklids. Diese Sprache kennt **monoseme** Wörter (z.B.: Punkt, Linie, Winkel, Kreis, Dreieck), deren Bedeutung in Euklids Definitionen (vgl. Definitionen 1,2,8,15,20 im 1. Buch Euklids) mit Mitteln einer natürlichen Sprache (Altgriechisch) festgelegt werden:

5.1 Vorbemerkungen (Forts.) (3)

Definition 8: „Ein ebener **Winkel** ist die Neigung zweier Linien in einer Ebene gegeneinander, die einander treffen, ohne einander gerade fortzusetzen.“ [Euklid: „Die Elemente“, Buch I., übersetzt von C. Thaer, Leipzig (Akademische Verlagsgesellschaft) 1984, S.1]

Maschinelle Sprachverarbeitung ist zweierlei:

(1) Sie ist Anwendung der Informatik zur Verarbeitung natürlicher Sprache, die in Form maschinenlesbarer Texte oder in Form digitalisierter Aufzeichnung gesprochener Sprache gegeben ist.

(2) Sie ist Studium der Syntax und der Semantik natürlicher Sprache, um Konzepte, Methoden und Verfahren zu finden, wie formale Sprachen verbessert werden können und wie man dem Rechner Funktionen des natürlichen Sprachverhaltens antrainieren kann (z.B. Sprachsynthese, Vorübersetzen, Vokabeltraining).

5.2 Syntaktische Analyse (4)

In der Grammatik einer natürlichen Sprache (z.B. des Deutschen) werden Wörter zu **Wortarten** zusammengefasst. Wörter einer Wortart weisen gemeinsames syntaktisches und semantisches Verhalten auf.

Zum syntaktischen Verhalten gehört, ob die Wörter flektiert („gebeugt“) werden können oder nicht. Falls sie flektiert werden können, haben sie ein allgemeines Flektionsschema, das durch die Wortart gegeben ist. Zum semantischen Verhalten gehört, ob Wörter einer Wortart generell Bedeutungen tragen oder ob sie nur syntaktische Funktionen ausüben.

In der deutschen Sprache können zehn Wortarten unterschieden werden:

1. Substantive (S): Z.B. der Tisch, die Zeit, das Haus. Substantive werden in Satzsubjekten oder -objekten, in Attributen oder adverbialen Bestimmungen verwendet. Allgemeines Flektionsschema: Deklination nach Kasus und Numerus. Grundform: Nominativ / (i. d. R.: Singular).

5.2 Syntaktische Analyse (Forts.) (5)

2. Verben (V): Z.B. fahren, lachen, werden. Verben bilden das Prädikat eines Satzes. Flektionsschema: Konjugation. Grundform: Infinitiv.

3. Adjektive (A): Z.B. blau, groß, reich. Adjektive können u. a. als Attribute von Substantiven verwendet werden. Flektionsschema: Komparation („Steigerung“); in Verbindung mit Substativen: Deklination. Grundform: Positiv.

4. Adverbien (AV): Z.B. bald, dort, vielleicht, weit. Adverbien dienen der Spezifikation der Verbverwendung in einem Prädikat.

5. Präpositionen (PP): Z.B. auf, bei, nach. Präpositionen sind in der Regel das einleitende Wort einer Umstandsbestimmung, eines Attributs oder eines Präpositionalobjekts.

6. Konjunktionen (PK): Z.B. dann, oder, und. Konjunktionen verbinden Satzteile oder Sätze.

7. Numerale (Zahlwörter) (Z): Z.B. achtel, drei, fünfte.

8. Artikel (PA): Z.B. der, die, ein, eine. Artikel werden in Verbindung mit Substantiven gebraucht.

9. Pronomen (PN): Z.B. du, mein, dieser, welcher. Pronomen sind Vertreter von Substantiven. Pronomen werden dekliniert.

5.2 Syntaktische Analyse (Forts.) (6)

10. Interjektionen (I): Z.B. ach, au, oh, pst. Interjektionen sind Ausrufe-, Ausdrucks- oder Empfindungswörter.

Die Wortarten Substantiv, Verben, Adjektiv und Adverb sind bedeutungstragend. Die übrigen sechs Wortarten enthalten Funktionswörter und sind in ihrer Wortmenge begrenzt und können daher einfach in einer relationalen Datenbank gespeichert werden (WortDB).

5.2 Syntaktische Analyse (Forts.) (7)

Die maschinelle **syntaktische Analyse** von maschinenlesbaren Sätzen einer natürlichen Sprache hat folgende **Aufgaben**:

- Die **Wortart** jedes Worts des gegebenen Satzes wird bestimmt.
- Bei flektierbaren Wörtern wird die **Grundform** ermittelt.
- Zu jeder Folge von Wörtern, die eine bestimmte Phrase (ein Satzglied) bilden, soll die **Phrasenart** ermittelt werden.

I

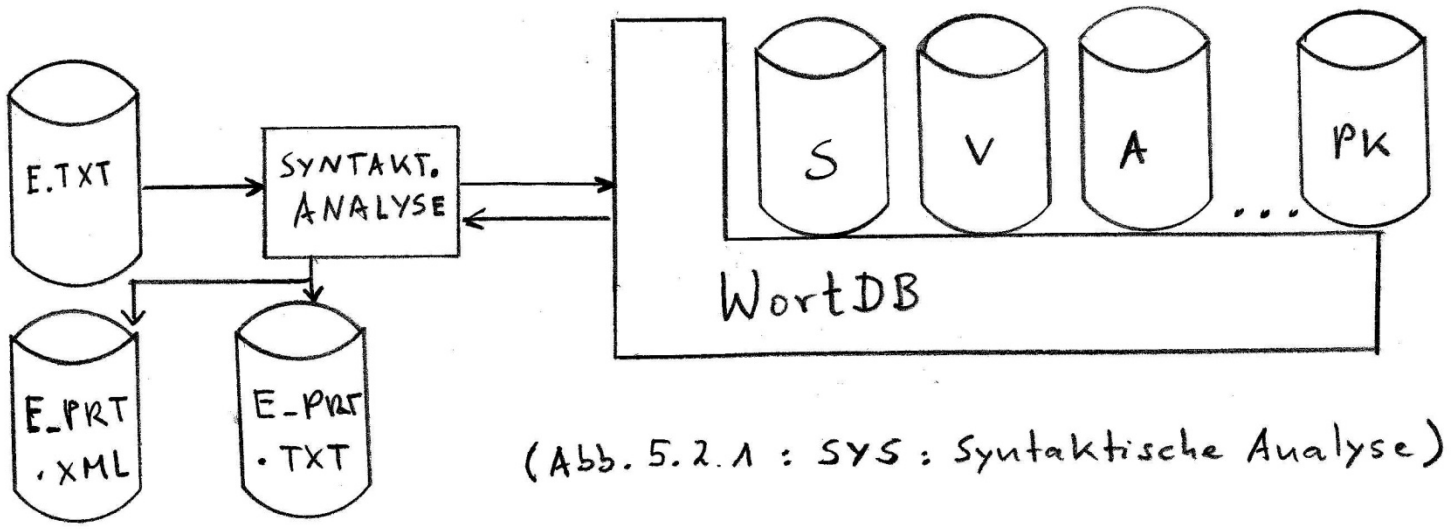
5.2 Syntaktische Analyse (Forts.) (8)

Nachfolgend ist eine einfache Systemübersicht einer syntaktischen Analyse gegeben: Legende:

E.TXT: Eingabetext

E_PRT.TXT : Ergebnisprotokoll der syntaktischen Analyse mit Angabe der Wortarten, der Grundformen und mit Kennzeichnung von Satzphrasen.

E_PRT.XML : Ergebnisprotokoll in XML-Format.



(Abb. 5.2.1 : SYS : Syntaktische Analyse)

5.2 Syntaktische Analyse (Forts.) (9)

Für den Eingabesatz: „Hohe Ströme erfordern starke Isolatoren“ (E.TXT) aus dem Kontext der Elektrotechnik ist nachfolgend das Ergebnisprotokoll der syntaktischen Analyse aufgeführt (E_PRT.TXT):

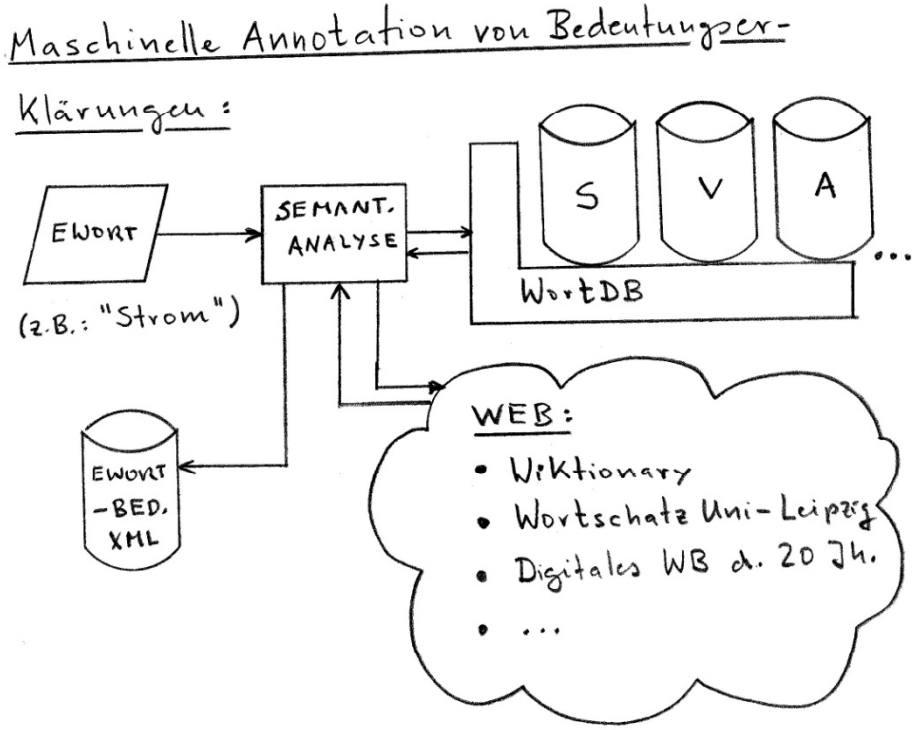
```
Wort              Wortart                       Phrase     Phrasenart /***/
|----------------------+-----------------------+--------+------------------|
 Hohe                  Adjektiv                 1
 Ströme                Substantiv               1         Nominalphrase
 erfordern             Verb im Infinitiv        3         Verbalphrase /***/
 starke                Adjektiv                 2                     /***/
 Isolatoren            Substantiv               2         Nominalphrase
 .                     Satzzeichen              0
```

5.2 Syntaktische Analyse (Forts.) (10)

Das Ergebnisprotokoll der syntaktischen Analyse (E_PRT.XML) im Auszug:

```
<Protokoll>
 <Wort>Hohe
   <Wortart>Adjektiv            </Wortart>
   <Phrase>1</Phrase>
 </Wort>
 <Wort>Ströme
   <Wortart  GF="Strom">Substantiv          </Wortart>
   <Phrase>1</Phrase>
   <Umstandsbestimmung>Nominalphrase</Umstandsbestimmung>
 </Wort>
 <Wort>erfordern
   <Wortart>Verb im Infinitiv     </Wortart>
   <Phrase>3</Phrase>
   <Umstandsbestimmung>Verbalphrase</Umstandsbestimmung>
 </Wort>
 <Wort>starke
   <Wortart>Adjektiv              </Wortart>
   <Phrase>0</Phrase>
   <Umstandsbestimmung></Umstandsbestimmung>
 </Wort>
 <Wort>Isolatoren
   <Wortart  GF="Isolator">Substantiv          </Wortart>
   <Phrase>2</Phrase>
   <Umstandsbestimmung>Nominalphrase</Umstandsbestimmung>
 </Wort>
....
</Protokoll>
```

5.2 Semantische Analyse (11)



5.2 Semantische Analyse (12)

Erster Schritt ist eine Annotation der bedeutungstragenden Wörter mit Bedeutungserklärungen (Quelle: Wiktionary) GF=“Strom“:

```
<Wort>Ströme
    <Wortart  GF=“Strom“>Substantiv            </Wortart>
    <Phrase>1</Phrase>
<BED Nr=“1“>allgemein eine Menge (Teilchen, Wasser, Menschen,
Informationen), die sich im Fluss befindet (fließt)
</BED>
<BED Nr=“2“>großes, fließendes Gewässer in Form eines Flusses
</BED>
<BED Nr=“3“>kurz für: elektrischer Strom - bewegte
Ladungsträger
</BED>
</Wort>
```

5.2 Semantische Analyse (13)

Eine maschinelle Annotation der bedeutungstragenden Wörter mit Bedeutungserklärungen aus der Quelle Wiktionary kann auf **Bedeutungserklärungen** und **Sachgebietszuordnungen/Synonyme** zugreifen**.** BSP: GF="**Strom**" (bearbeitete Auszüge):

„Aus Wiktionary, dem freien Wörterbuch: ... Strom, Plural: Strö·me…

Bedeutungen:

[1] allgemein eine Menge (Teilchen, Wasser, Menschen, Informationen), die sich im Fluss befindet (fließt)

[2] großes, fließendes Gewässer in Form eines Flusses

[3] *kurz für:* elektrischer Strom - bewegte Ladungsträger

Abkürzungen:

[3] *Technik, Naturwissenschaft:* I, i

Synonyme:

[1] Fluss

[2] breiter Fluss

[3] elektrischer Strom, Elektronenstrom

Oberbegriffe:

[2] Gewässer

[3] physikalische Größe

Unterbegriffe:

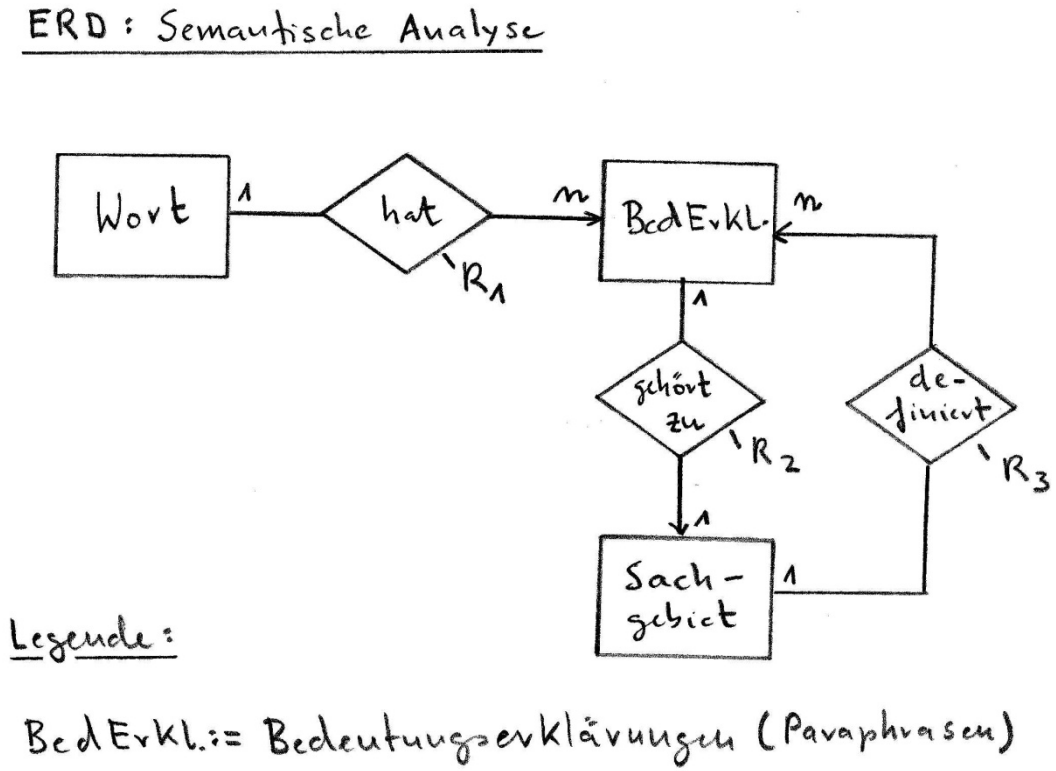
[1] Flüchtlingsstrom Teilchenstrom

[3] Gleichstrom, Laststrom, Wechselstrom

…"

5.2 Semantische Analyse (ERD) (14)

Mit einem ERD kann ein allgemeiner Zusammenhang zwischen einem Wort, seinen Bedeutungserklärungen und der Zuordnung von Bedeutungserklärungen zu Sachgebieten beschrieben werden :



5.2 Semantische Analyse (Forts.) (15)

BSP: GF="**Isolator**" (bearbeitete Auszüge):

"Aus Wiktionary, dem freien Wörterbuch: … Isolator, Plural: Isolatoren

Bedeutungen:

[1] *Physik:* isolierendes Material

[2] *Elektrotechnik:* elektrisch isolierendes Bauteil

Synonyme:

[1, 2] Nichtleiter

Gegenwörter:

[1] elektr. Leiter

Oberbegriffe:

[1]

Beispiele:

[1] Porzellan ist ein guter *Isolator*.

[2] Freileitungen sind an *Isolatoren* aufgehängt.

Abgeleitete Begriffe:

Wärmeisolator, Porzellanisolator

…"

5.2 Semantische Analyse (Forts.) (16)

Annotation von Bedeutungserklärungen mit Fachgebietsangabe (Quelle: Wiktionary) GF="Isolator":

```
<Wort>Isolatoren
    <Wortart  GF="Isolator">Substantiv          </Wortart>
    <Phrase>2</Phrase>
<BED Nr="1">Physik:  isolierendes Material
</BED>
<BED Nr="2">Elektrotechnik: elektrisch isolierendes Bauteil
</BED>
</Wort>
```

Desiderat der maschinellen Bedeutungsannotation: Kontextsensitive Auswahl der zutreffenden Bedeutung (hier Kontext = "Elektrotechnik" : bei "Strom" BED-Nr.3, bei "Isolator" BED-Nr.2). Der Kontext kann mit dem Sachgebiet identifiziert werden (s.o.: ERD: Semantische Analyse).