ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών

Computerlinguistik

Lehreinheit 9 : Computerlinguistik
Übersicht - Wiederholung

Dr. Christina Alexandris

Nationale Universität Athen

Deutsche Sprache und Literatur

Περιεχόμενα ενότητας

Επανάληψη – Κατευθυντήριες Γραμμές για Εκπόνηση Εργασίας

In der Computerlinguistik…

- werden die **Modelle der Theoretischen Sprachwissenschaft** und deren Disziplinen
- wie z.B die Modelle der Syntax, der Semantik, der Morphologie aber auch die Modelle der Pragmatik und der Phonetik und Phonologie
- in der Form von **Regeln** umgesetzt, die der Computer später als **Programm** erkennen und verarbeiten kann.
- Diese Modelle aus der Theoretischen Sprachwissenschaft versuchen das **Sprachliche Wissen** zu beschreiben.

Einer der Ziele der Computerlinguistik ist,

- dass der Computer dieses sprachliche Wissen **erwirbt**.
- Dieser Prozess wird nach Hanneforth (2001) auf folgende Weise beschrieben:
- “Das **sprachliche Wissen** wird von den Computerlinguisten in die **Praxis** umgesetzt, um die größte **Barriere** zwischen Mensch und Maschine zu überwinden.” (Hanneforth, 2001)

Kern aller Anwendungsbereiche der Computerlinguistik (1/2)

- ist ein System für die Verarbeitung der natürlichen Sprache
- (**Natural Language Processing system – NLP System**).
- So ein System für die Verarbeitung natürlicher Sprache kann die natürliche, bzw. „menschliche“ Sprache „verstehen“ und sie, anschließlich, auf verschiedenen Weisen, je nach der Anwendung des Systems, verarbeiten.
- Der Vorgang der Verarbeitung kann je nach Struktur und Anwendungsbereich des Systems variieren.

Kern aller Anwendungsbereiche der Computerlinguistik (2/2)

- Jedoch kann in groben Zügen dieser Vorgang **in drei Phasen** beschrieben werden:
- Ein System für die Verarbeitung natürlicher Sprache analysiert den (geschriebenen oder gesprochenen) Text, der vom System erkannt wird (**Analyse**),
- es verarbeitet den analysierten Text, je nach Art der Anwendung (**Verarbeitung**)
- und generiert den Text in der gewünschten Form, je nach Art der Anwendung (**Generierung**).

Die Programme mit den syntaktischen Regeln und den morphologischen Regeln,

- **sorgen dafür** daß jeder Satz und jedes individuelle Wort des Satzes von dem System **richtig verstanden**, bzw. analysiert wird (**Analyse**).
- In einigen Anwendungen der Computerlinguistik zum Beispiel, in der automatischen Textverfassung oder in der maschinellen Übersetzung sorgen **weitere syntaktische und morphologische Regeln** dafür, daß jeder Satz und jedes individuelle Wort des Satzes von dem System richtig **umgesetzt** (**Transfer**) und **erzeugt** (**Genereriung**) wird,

Programme mit syntaktischen Regeln und morphologischen Regeln-2

- oder in der Syntax und der Morphologie der **Zielsprache** richtig umgesetzt und erzeugt wird (je nach der Art der Anwendung und der Phase des Prozesses bzw. Modul des Systems, in der diese morphosyntaktischen Programme aktiviert werden).

Ein System künstlicher Intelligenz,

- also z.B. ein Computer, kann hierarchische Strukturen verstehen.
- Wenn man die natürliche Sprache als eine **hierarchische Struktur** beschreibt, dann kann der Computer die natürliche Sprache "verstehen" und "bearbeiten".

Nehmen wir an, dass unser Computer, z.B., den folgenden Satz "verstehen„ soll:

- "***Ein dicker Kater sitzt auf dem Stuhl***"
- Der Satz "*Ein dicker Kater sitzt auf dem Stuhl*" wird von dem Computer (dem System) nur als eine **Reihe von nicht-mathematischen Zeichen**, nämlich als "**alphanumerische Zeichen**" (**Strings**) und (leere Lücken) **Leerzeichen** "verstanden".
- Für den Computer ist das **Wort** eine Einheit aus **alphanumerischen Zeichen**, die rechts und links durch **Leerraumzeichen** (engl. "white space") oder durch **Interpunktion** begrenzt werden.
- Diese Reihe von Elementen bildet die "**Eingabedaten**" (**Input**) des Computers (des Systems).

Kontextfreie Grammatik

- Ohne diese hierarchische Struktur könnten die engeren Beziehungen, die manche Elemente (Strings) zueinander haben, nicht beschrieben werden, wie zum Beispiel die Beziehung "Verb – Verbalphrase".
- Mit diesen Regeln "weiß" der Computer an welchen Stellen er die Eingabe "Ein dicker Kater sitzt auf dem Stuhl" in weitere Stücke/Segmente teilen/segmentieren kann und sie in kleinere und noch kleinere Stücke segmentieren und analysieren kann.

Eine einfache kontextfreie Grammatik

für die Generierung des Satzes

"***Ein dicker Kater sitzt auf dem Stuhl***" (Analyse nach Jurafsky and Martin, 2008):

Regel:

+ S -> NP VP (S = Startysmbol)

+ NP -> D N’

+ VP -> V PP

+ PP -> P NP

+ N’ -> ADJ N

+ NP -> D N

Anhand einer kontextfreien Grammatik wird der Satz "Ein dicker Kater sitzt auf dem Stuhl" in verschiedenen Stufen allmählich geparst (Parsing) (1/2)

(i) 1. Ebene:

[ein dicker **Kater] [sitzt auf dem Stuhl]**

Griechisches Beispiel: [ένας χοντρός γάτος] [κάθεται πάνω στην καρέκλα]

Regel:

S -> NP VP (S = Startysmbol)

Parsing: "Ein dicker Kater sitzt auf dem Stuhl“ (2/2)

(ii) 2. Ebene:

[ein [dicker **Kater]] [sitzt [auf dem Stuhl]]**

Griechisches Beispiel:

[ένας [χοντρός **γάτος]] [κάθεται [πάνω στην καρέκλα]]**

Regel:

NP -> D N’

VP -> V PP

In jeder Stufe werden bestimmte Regeln der kontextfreien Grammatik verwendet bzw. aktiviert, die mit der Erzeugung der entsprechenden Ebenen der hierarchische syntaktischen (Baum-) Struktur korrespondieren.

(iii) 3. Ebene:

[ein [dicker [**Kater**]]] **[sitzt [auf [dem Stuhl]]]**

Griechisches Beispiel:

[ένας [χοντρός [**γάτος**]]] **[κάθεται [πάνω [στην καρέκλα]]]**

Regel:

PP -> P NP

N' -> ADJ N

(iv) 4. Ebene:

[ein [dicker [**Kater**]]] **[sitzt [auf [dem [Stuhl]]]]**

Griechisches Beispiel:

[ένας [χοντρός [**γάτος**]]] **[κάθεται [πάνω [στην [καρέκλα]]]]**

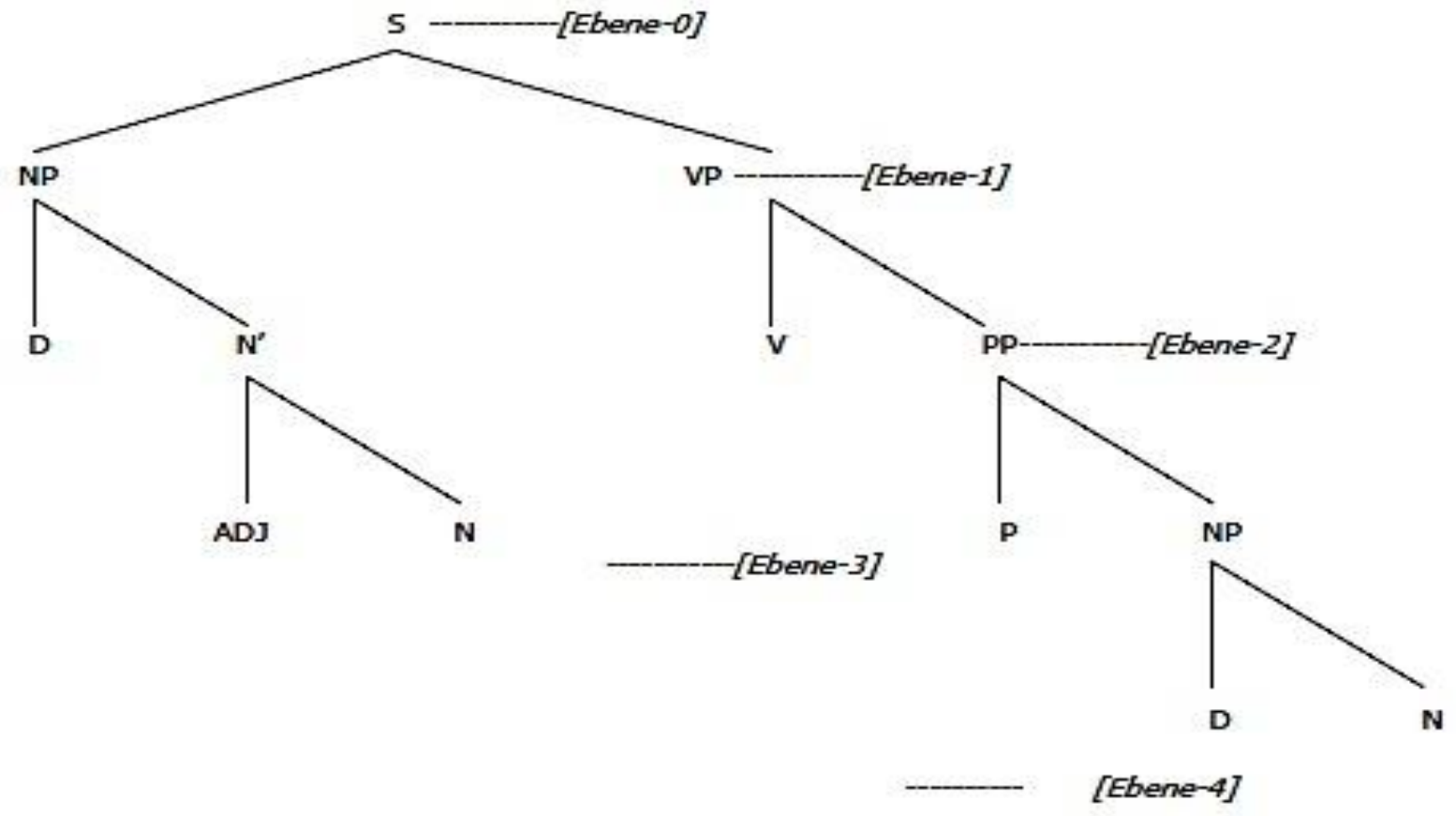
Regel:

NP -> D N

Syntaktische Struktur für die Eingaben:

[ein dicker **Kater sitzt** auf dem Stuhl]
[der große **Löwe springt** aus dem Käfig]

Griechisches Beispiel:
[ένας χοντρός **γάτος κάθεται** πάνω στην καρέκλα]



ANWENDUNGEN UND PROBLEME

Multilinguale Mensch-Maschine Kommunikation: Linguistische Aspekte und Anwendungen

Hier geht es um:

- **Anwendungen**:
  - Multilinguale Frage-Antwort-Systeme, Dialogsysteme (geschriebener und gesprochener Sprache)
  - Interaktive Maschinelle Übersetzung, bzw. Transkribierung
  - Maschinelle Übersetzung in Frage-Antwort-Systemen, Dialogsystemen
  - Multilinguale Hypermedia
- und **Daten** aus älteren und neueren Projekten der Europäischen Union (EU-Projekte und Nationale Projekte)

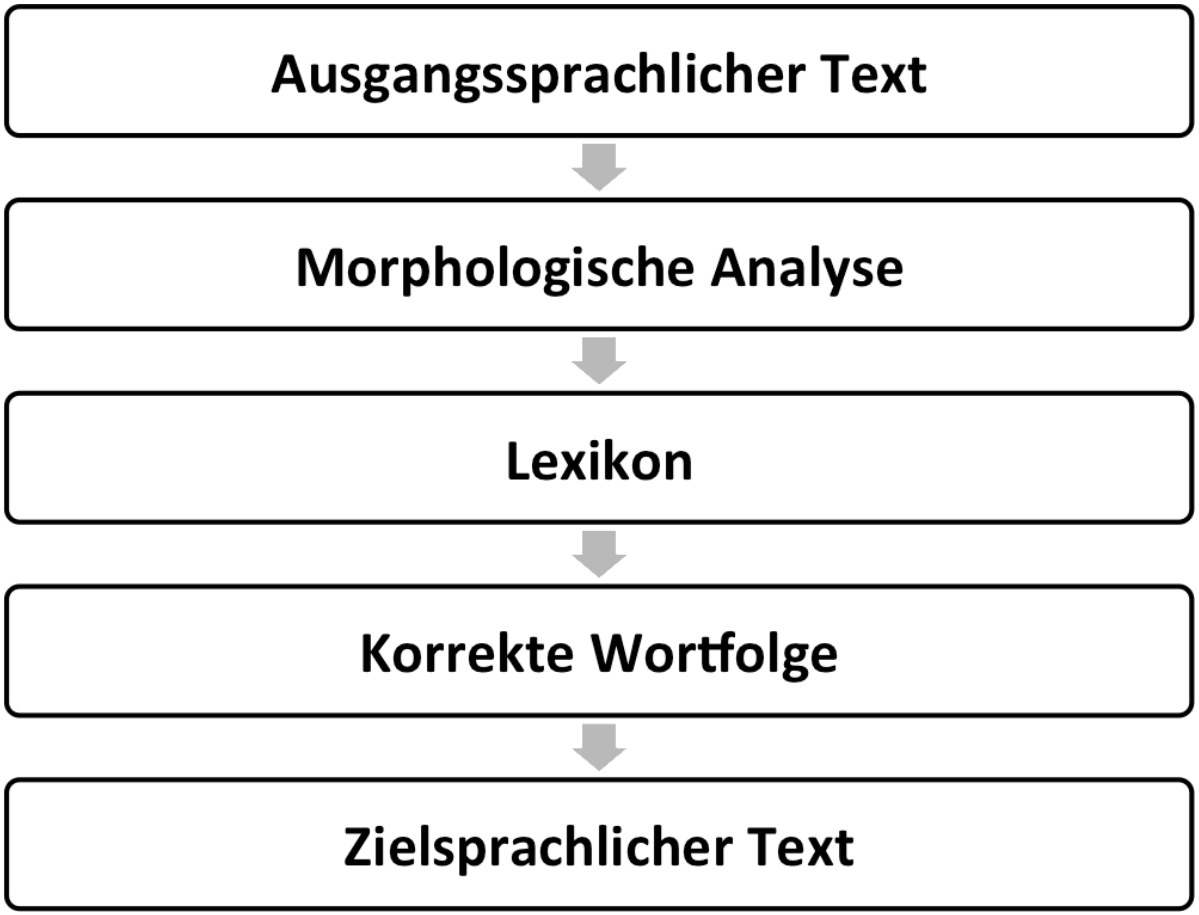
Direkte Systeme

- In der direkten Übersetzung wird in der Phase der Analyse der ausgangsprachliche Text (Quelltext) morphologisch analysiert und zwar auf einer relativ oberflächennahen Beschreibungsebene.
- In der Phase des Transfers werden die aus der Phase der Analyse erzeugten morphosyntaktischen Einheiten direkt in die Zielsprache mit Hilfe eines bilingualen Wörterbuchs übersetzt.
- Diese morphosyntaktischen Einheiten, die eine aus dem ausgangssprachlichen Text gewonnene abstrakte Repräsentation bilden, sind sehr nahe am orginären Quelltext der Ausgangsprache.
- Anschließend, werden in der Generierungsphase aus den in der Zielsprache übersetzten morphosyntaktischen Einheiten der Zielsprache Sätze bzw. Texte gebildet.
- In der Generierungsphase der direkten Systeme können einfache Operationen, wie die Veränderung der Wortreihenfolge, in der Zielsprache abgeschlossen werden (Dorna und Jekat, 2004).

In den Transfer-Systemen

- findet in der Phase der Analyse des Quelltexts eine gründlichere Verarbeitung der ausgangsprachlichen Einheiten auf einer tieferen Beschreibungsebene als in dem Ansatz der direkten Systeme statt.
- In dem Transferansatz besteht die Phase der Analyse des Quelltexts aus zwei Verarbeitungsphasen: In der ersten Phase der Analyse (Phase 1a) wird die quellsprachliche Eingabe geparst und semantisch analysiert.
- In der zweiten Phase der Analyse (Phase 1b) wird während des Parsings und der semantischen Analyse eine abstrakte Repräsentation des Inhalts und der syntaktischen Struktur der Quellsprache (Ausgangssprache) erstellt, die (normalerweise) über morphosyntaktische Idiosynkrasien einzelner Sprachen, wie Tempusrealisierung oder Flexion, abstrahiert.
- Es muss bemerkt werden, dass diese abstrakte Repräsentation oft für jedes System verschieden ist (Dorna und Jekat, 2004).

Direkte Systeme
Direkte Übersetzung



[3]

Transfer Systeme

[4]

Maschinelle Übersetzung Divergenzen

- Beispiele aus Technischen Texten-Übersetzungsschwierigkeiten (Lehrndorfer, 1996)
  - ***Beispiel* :**
  - **Vorgangspassiv/Zustandspassiv**:
    - Die Datei wird geladen
    - =The file is loaded
    - Die Datei ist geladen
    - =The file is loaded
  - ***Beispiel :***
  - **Verbteil-Ellipse:**
    - ..be- und entladen
    - =to load and unload

Maschinelle Übersetzung
Lexikalische Lücken

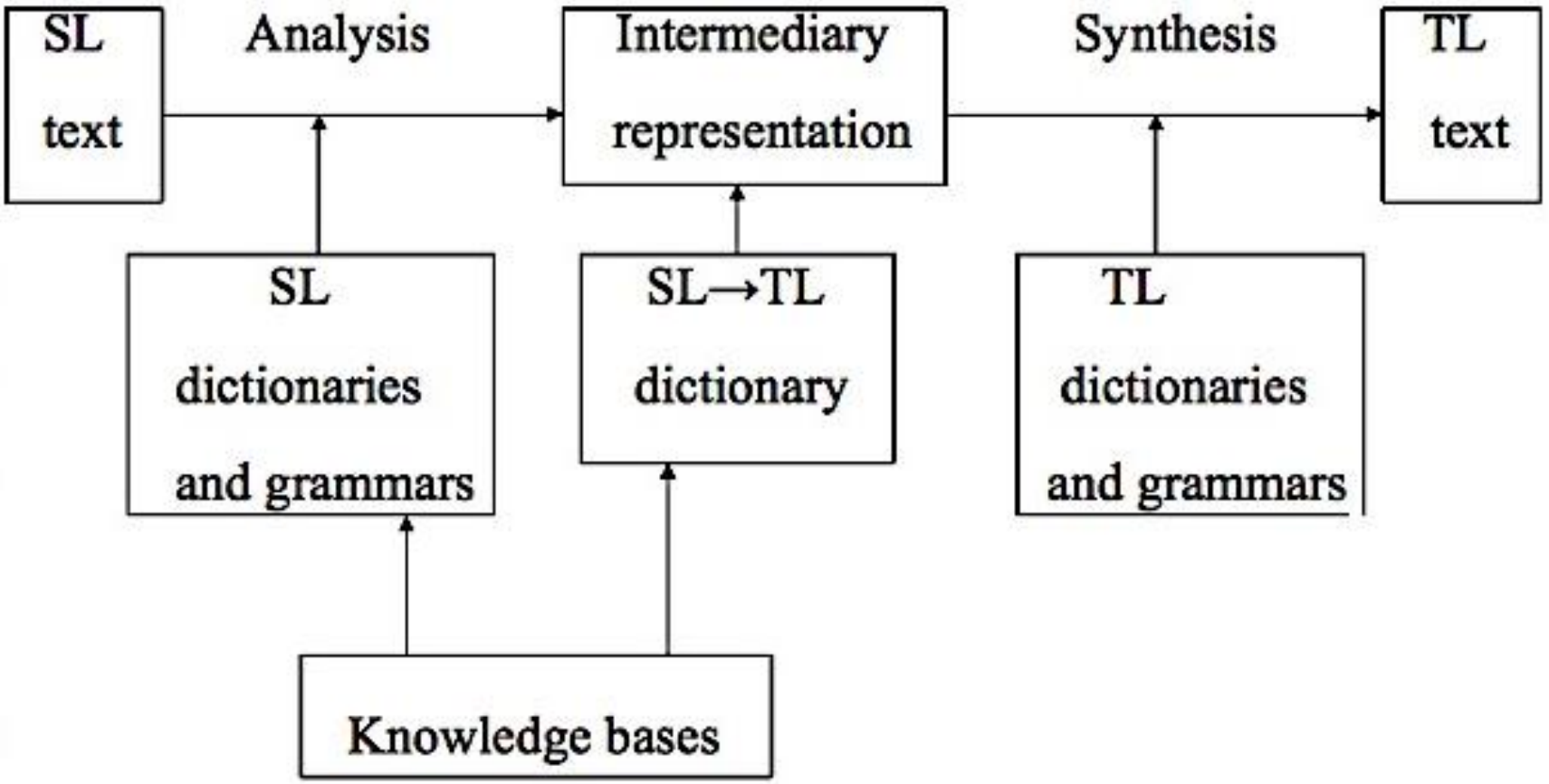
Typische Beispiele von Nichtenstprechungen oder „Lexikalische Lücken“ bezüglich des Sprachpaars Deutsch- Englisch:

- ***Beispiel 1*:**
  - (D): Gewalt (Macht), (Kraft), (Kontrolle), (Gewalttätigkeit)
  - (ENG): force, power, violence, threat, control (etc)
- ***Beispiel 2*:**
  - (D): Muster
  - (ENG): design, model, pattern, specimen, stich (beim Stricken), prototype, paragon (Vorbild)
- ***Beispiel 3*:**
  - (D): prüfen
  - (ENG): test, try, prove, check, survey, review, examine, audit, inspect, assay (etc)

Die Interlingua-Systeme,

- die auch als **Übersetzungstrategie 2. Generation** bezeichnet werden, haben den Vorteil, dass sie **Sprachunabhängig** sind und dass somit die Übersetzung des erzeugten zielsprachigen Textes wieder in der Quellsprache („**back-translation**“) möglich ist, auch für eine **Systemüberprüfung** (v. Hahn, 2001 - Vertan, 2002).
- Dass die Entwicklung bzw. die Darstellung der sprachunabhängigen Zwischenrepräsentation **schwer** ist, gehört zu den Nachteilen der Interlingua-Systeme (v. Hahn, 2001 - Vertan, 2002).
- Interlingua-Systeme werden oft für **multilinguale MÜ** Systeme benutzt, **weil die Interlingua unabhängig von der Quellsprache (Ausgangssprache) ist**.

Interlingua MT model



[7]

Beispiel

Gesprochene Äußerung: "Am Montag habe ich leider eine Konferenz"
(aus dem Babel-Verbmobil Korpus, Universität Bremen, 1998)
Vom System erzeugte entsprechende **Interlingua**:
(Alexandris, 1995) http://verbmobil.dfki.de/

```
((speech-act (*or* *state-constraint *reject))
  (sentence-type *state)
  (frame *booked)
  (who ((frame *i))
  (what ((frame *conference)
            (specifier indefinite))
  (when ((frame *simple-time)
             (day-of-week Monday))))
```

Satzinhalt:
Entsprechungen:
Multitasking Verbs: www.polias.gr und aus einem Griechischen Dialogsystem für Verbraucher-beschwerungen (Nottas et al., 2007).

- **Satzinhalt (“frame”):**
- ***Beispiel 1:***
- **Entsprechungen:**
- **(“Type-of-Product”, “Type-of-Complaint”)**
  - I **bought** this yoghurt
  - I **got** this yoghurt
  - This **is about** a yoghurt
  - I **was given** this yoghurt
  - I **have** this yoghurt here
- **Entsprechungen:**
- **(“Type-of-Complaint”)**
  - I **saw** some misguiding information on the label
  - There **is** misguiding information on the label
  - They **give** misguiding information on the label
  - Misguiding information **is shown** on the label
  - This **is about** misguiding information on the label

Wiederholung:

Dialogsysteme als Mensch-Maschine Kommunikation

Dialogsysteme

- Natürlichsprachliche Dialogsysteme erlauben es einem menschlichen Benutzer, mit einer Maschine mittels **sprachlicher Ein-und Ausgabe zu kommunizieren** (Kellner, 2004).
- Üblicherweise wird mit der Wendung „Dialogsystem“ ein System Verarbeitung gesprochener Sprache gemeint mit dem der Benützer **mündlich kommuniziert** und von dem System Antworten oder Reaktionen in der Form gesprochener Sprache bekommt.
- Die folgenden stammen aus einem Dialogsystem für Informationen (Auskunft) über **Fußballspiele** ((Moegle et al., 2006) (auf Englisch)

Dialogsystem (Moegle et al., 2006)

pro-010 Please think of the soccer WM. You want to get information about results and games of several teams.
First you want to know how the last game of Germany against Costa Rica ended.
4000
rec-010 *What was the result of the match Germany against Costa Rica?*
pro-010 Now you want to get information about who else plays in the group of England and the Netherlands.
4000
rec-020 *What other teams are in the group of Britain and Holland?*
pro-031 Next you are interested in the time of the next game of Mexico against Ukraine.
4000
rec-030 When is the match Mexico against Ukraine
...
Table 4: Example of SmartWeb recording dialogue using a standard prompt scheme. pro denotes the (female) instructor's voice; opr denotes the (male) operator's voice; rec indicates a recording of the users's elicited query.

Das sehr beschränkte Weltwissen eines Dialogsystems

- erlaubt es nicht überall und in jeder Gelegenheit wie ein Mensch zu kommunizieren.
- Deswegen werden Dialogsysteme bis heute vornehmlich **in klar abgeschlossenen Anwendungsdomänen** eingesetzt.
- In einem Dialogsystem soll das System die gesprochene Sprache erkennen und erzeugen und zugleich auch auf jede Äußerung des Benutzers die **angemessene** Antwort oder Reaktion produzieren.
- Der Prozess der Erzeugung angemessener Reaktionen des Systems wird mit Hilfe
- **pragmatischer Regeln durchgeführt die die Form eines Programs haben.**

Besonders wichtig in der Multilingualen Mensch-Maschine Kommunikation: Faktoren

I. **Ziel** (nach der Anwendung des Systems)

II. **Benutzer** (z.B. Experten oder das breite Publikum)

III. Inhalt - Ausdruck

I. Ziel – Kriterien für ein erfolgreiches System

- Ist die Ausgabe des Systems angemessen? Verständlich?
- Ist der Inhalt der Ausgabe richtig?
- Werden die Aufgaben vom System richtig durchgeführt?
- Können mögliche Probleme (z.B. falsche Eingaben) vom System erfolgreich behandelt werden?
- Ist die Interaktion fehlerfrei und vergleichbar mit einem menschlichen Sprecher?

II. Benutzer und Benutzermodelierung

- Die Benutzermodelierung („user-modelling“) kann als ein unentbehrlicher Teil in dem Prozess der Konstruktion von Dialogsystemen bezeichnet werden und konstituiert zugleich ein Anwendungsgebiet der Pragmatik in die Computerlinguistik.
- Die Benutzermodelierung ist nicht nur für die Konstruktion von Dialogsystemen unentbehrlich, sondern auch für die Konstruktion und Entwicklung fast jeder Form interaktiver Software Programme.

Faktor III der Multilingualen Mensch-Maschine Kommunikation: Aspekte

A. Pragmatische Aspekte

B. Semantische Aspekte

C. Phonetisch-Phonologische Aspekte – Paralinguistische Merkmale

Deutsch-Japanisch: Gestik und Bewegung Virtuelle Lehrer (Lipi et al., 2008)

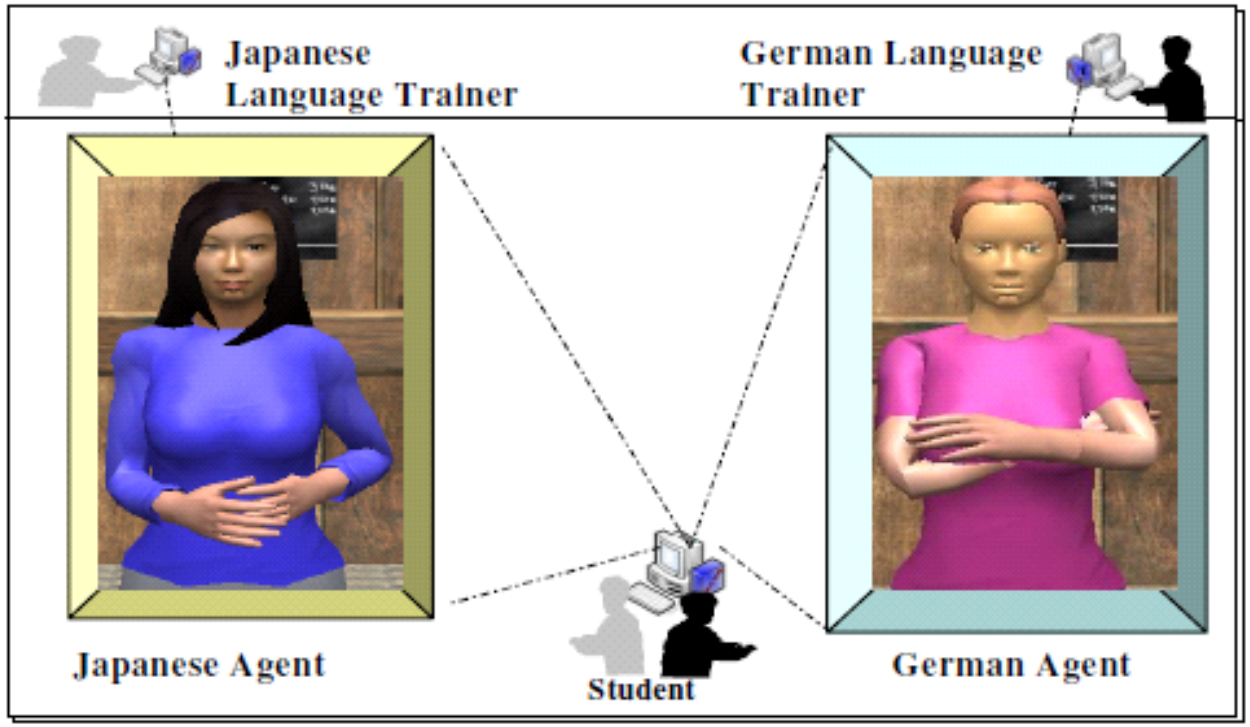


Fig. 1. Language- trainer Agents

Funktion der Ressourcen

Grundvoraussetzung für computerlinguistische Tätigkeiten ist die Verfügbarkeit großer Datenmengen wie Textkorpora und Sprachdatenbanken. Diese Datenmengen können entweder

(1)sprachliche Datenmengen (zum Beispiel Lexika und Korpora) oder

(2) nicht-sprachliche Datenmengen zur Repräsentation nicht-sprachlichen Wissens (zum Beispiel Videos) sein (Carstensen, 2004).

•Diese Datenmengen werden auch als "Ressourcen" bezeichnet (Engl. "resource").

•Ressourcen dienen als Quellen linguistischer Information.

Ressourcen helfen

- sowohl bei der **Konstruktion** eines Systems bei der Verarbeitung natürlicher Sprache
- als auch beim **Testen** (Evaluierung) des Systems.

Quellen von Ressourcen:

(1) **Lexikalisch-semantische Wortnetze**
(Anwendungsgebiet der theoretischen Linguistik)

(2) **Sprachdatenbanken**

(3) **Sprachliche Daten in annotierter Form** (unter Anwendung computerlinguistischer Analysemethoden), z.B: Baumbanken.

(4) das **World-Wide-Web (WWW)**

Wiederverwertbarkeit der Ressourcen

- Ein entscheidender Faktor bei der Aufbereitung computerlinguistischer Daten ist ihre **Wiederverwertbarkeit** (Re-usability).
- Aus Gründen der Wiederverwertbarkeit (Re-usability) wird versucht zu erreichen, dass die computerlinguistischen Daten in den Ressourcen auch außerhalb eines Systems (d.h. auch in anderen Systemen) brauchbar und nicht nur system-spezifi sch sind.
- Es wird versucht, sich an **allgemein akzeptierten Standards** sowohl im Bereich der Linguistik als auch im Bereich der Informatik (z.B. XML) zu orientieren.

Lexikographie- Grad der Komplexität eines Lexikonaufbaus

- Ein Lexikon kann als ein Inventar von Wörtern, Wortbestandteilen oder Wortgruppen definiert werden, in dem verarbeitungsrelevante Informationen gegeben werden (Heid, 2004).
- Der Grad der Komplexität reicht dabei von eher einfachen Aussprachewortlisten bis hin zu komplexen, terminologischen Handbüchern (Gibbon, 2004).
- Ein Übersetzungslexikon kann als ein sehr komplexes Lexikon bezeichnet werden, in dem subtile Unterschiede zwischen Sprachen definiert werden.

Schritte beim Aufbau eines Lexikons

- In der linguistischen Analyse werden die **Wurzeln** oder die **Grundform** der Wörter zu **Identifikationszwecken** ausgewählt (Infinitiv bei Verben, Nominativ Singular bei Substantiven), alle anderen Flexionsformen werden in der Mesostruktur des Lexikons verarbeitet.
- Es sollen z.B. nicht unbedingt alle Flexionsformen eines Verbs getrennt in die Ressource aufgenommen werden: füllen, füllst, füllt, füllte, fülltest, füllten, füllend, gefüllt. Es genügt die Wurzel "füll" zu erfassen, eine Normalform zu Identifikationszwecken auszuwählen und alles andere der Mesostruktur des Lexions zu überlassen, in der die Eigenschaften von Hunderten von ähnlichen Verben zusammengefasst werden können

Computergestützte Lexikographie und Terminologie

- In der computergestützten Lexikographie und Terminologie (oder "computational terminology") werden computerlinguistische Verfahren verwendet, um Fachwortschatz zu beschreiben (Heid, 2004).
- Beispiele für Aufgabengebiete der computergestützten Lexikographie und Terminologie sind die Extraktion von Fachtermini (Termen) aus Texten, die lexikographische Erfassung und Beschreibung dieser Fachtermini und die Strukturierung des Fachwortschatzes.

Die Strukturierung des Fachwortschatzes

- kann nach ontologischen Prinzipien durchgeführt werden, wie das in der Standardsprache in "**Wortnetzen**" geschieht.
- Die Extraktion von Fachtermini wird im Bereich der Informatik und der Künstlichen Intelligenz angewendet.
- Für die Suche und Extraktion von Informationen geschiet das in den Gebieten (1) der Information Retrieval, (2) der Text Mining und (3) der Informationsextraktion.
- Auch in rein linguistischen Anwendungen findet es (4) als lexikalisches Hilfsmittel für die computerunterstützte Übersetzung und (5) für den Aufbau mehrsprachiger **Terminologiesammlungen** und Translation Memories Gebrauch.

Anwendungen der Fachtermini:

(1) Information Retrieval

(2) Text Mining

(1) Informationsextraktion

(2) lexikalisches Hilfsmittel (computerunterstützte Übersetzung)

(3) Aufbau mehrsprachiger Terminologiesammlungen und Translation Memories

Aufgabe der Akquisition lexikalischer Informationen

- Die Akquisition **linguistischer Informationen aus Textkorpora** zielt auf die Bereitstellung von **Hilfsmitteln** für die Lexikographen.
  - Zum Beispiel kann mit der Akquisition linguistischer Informationen aus Textkorpora ein Lexikograph die **Konkordanz** zu einem Wort, zu einem Ausdruck oder zu einem **Lemma** und all seinen Flexionsformen finden. Die Konkordanz ist ein Kontext, der links und rechts von dem gesuchten Wort, Ausdruck oder Lemma angegeben wird.

Wortnetz GERMANET-Beispiele (1/2)

- In dem Wortnetz werden Beziehungen zwischen drei Verben beschrieben, nämlich "laufen", "starten" und "kommen".
- Der Begriff "laufen" wird nach der Art und Weise (des Laufens) und nach der Richtung des Laufens analysiert.
- Die Art und Weise des Laufens wird im vorliegenden Wortnetz mit "?Art_laufen" symbolisiert, während die Richtung des Laufens als "? Pfad_laufen" dargestellt wird.
- Die Verbindung zwischen den Verben "laufen" und "starten" wird mit den Verben "loslaufen" und "losschlurfen" realisiert, die sowohl als Unterbegriffe des Oberbegriffs "laufen" als auch als Unterbegriffe des Oberbegriffs "starten" kategorisiert werden können.

Wortnetz GERMANET-Beispiele (2/2)

- Auf ähnliche Weise wird die Verbindung zwischen den Verben "laufen" und "kommen" mit den Verben "herlaufen" und "herschlurfen" realisiert, die sowohl als Unterbegriffe des Oberbegriffs "laufen" als auch als Unterbegriffe des Oberbegriffs "kommen" kategorisiert werden können.
- Wortnetze können auch für die kontrastive Darstellung der semantischen Verknüpfungen von Wörtern als Konzeptknoten zwischen zwei Sprachen verwendet werden. So kann z.B. die Information aus einem Wirtschaftslexikon für Deutsch und Griechisch in Form eines Wortnetzes beschrieben werden.

Assoziative Semantische Wortnetze

- Wortnetze (als Bezeichnung der Verwandtschaften zwischen Wörtern und Wortgruppen) können nicht nur semantisch sondern auch assoziativ sein.
- **Assoziative Beziehungen** sind wichtig für Anwendungen in Sprachlehr- und -lernsystemen, wie zum Beispiel, in Kinderlexika aber auch in Anwendungen mit Bezug auf Textsorten, wie journalistische Texte.
- In den semantischen Wortnetzen sind Unterschiede zwischen der griechischen und der deutschen Sprache zu erkennen.
- Es existieren Gemeinsamkeiten und Unterschiede. Letztere sind auch auf kulturelle Unterschiede zurückzuführen.

Textkorpora und Korpustypen (1/2)

- In den **reinen Textkorpora** ist die Grundeinheit das **Token**. Eine spezielle Form von Textkorpora mit der Grundeinheit des Satzes sind die **Baumbank**en, die aus syntaktisch analysierten Sätzen bestehen.
- In den **Baumbanken** werden die Analysen meist durch Syntaxbäume oder DAGs (Graphen) repräsentiert.
- **Linguistische Annotationen** (z.B. orthographische Transkription) werden mit phonetischen Annotationen (Phonemgrenzen, Grundfrequenz) in Sprachkorpora repräsentiert.

Textkorpora und Korpustypen (2/2)

- Außer in den Sprachkorpora sind Sprachsignale in den multimodalen Korpora zu finden, allerdings enthalten die Sprachsignale der multimodalen Korpora eine **zusätzliche Annotation** von **Prosodien, Mimik, Gestik** und **Mauszeigerbewegungen**.
- Diese Annotation kann mit einer entsprechenden **Videoaufnahme** verbunden werden.

Textkorpora und Abfragesysteme

- Für Textkorpora können Abfragesysteme benutzt werden, um alle Vorkommen einer gewählten Wortform (oder die Grundform) aufzufinden (Konkordanzsuche) , eine gewählte Wortartenfolge zu spezifizieren (musterbasierte Suche), die Häufigkeit der Erscheinung einer Wortart (Kolligationen) oder die Häufigkeit einer wiederholt auftauchenden Wortform (Konkurrenzen, Kollokationen) zu finden (statistische Analysen).

Erstellung von Korpora-Schritte in der Erstellung von Korpora (1/2)

- Die Aufbereitung von Textkorpora ist keine vollautomatische Prozedur.
- Der Prozess der Aufbereitung von Textkorpora wird in zwei Schritten beschrieben.
- Im ersten Schritt wird der Text in definierte Einheiten zerlegt. Diese Einheiten werden "Tokens" genannt. Dabei wird beachtet, dass der ganze (in Tokens) zerlegte Text in einem einheitlichen Repräsentationsformat vorgelegt wird. Der Prozess der Zerlegung des Textes in defi nierte Einheiten (Tokens) ist als "Tokenisierung" ("tokenization") bekannt.

Erstellung von Korpora-Schritte in der Erstellung von Korpora (2/2)

- Im zweiten Schritt wird jedem Token seine Wortart (engl. "part of speech") zugeordnet. Die Wortart wird von dem Tagger bestimmt . Der Prozess, in dem jedem Token seine Wortart zugeordnet wird, wird "Tagging" (tagging) genannt.
- Beim Tagging wird ein Lexikon konsultiert, das Taggerlexikon.

“Sauberkeit” des Ausgangsmaterials

- Eine weitere Aufgabe bei der Aufbereitung von Textkorpora ist die Sicherstellung der Textqualität, die sogenannte "Sauberkeit" des Ausgangsmaterials.
- Übliche Probleme im Ausgangsmaterial sind
- (1) Tippfehler,
- (2) Häufigkeit von ungrammatischen Sätzen (besonders in der gesprochenen Sprache) und
- (3) Fehler bei der linguistischen Vorverarbeitung und Annotation.

Τέλος Ενότητας

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αθηνών**» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Έργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

Έχουν προηγηθεί οι κάτωθι εκδόσεις:

- Έκδοση διαθέσιμη εδώ. http://eclass.uoa.gr/courses/GS158/

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.

Σημείωμα Χρήσης Έργων Τρίτων (1/2)

Το Έργο αυτό κάνει χρήση των ακόλουθων έργων:

Εικόνες/Σχήματα/Διαγράμματα/Φωτογραφίες

Σημείωμα Χρήσης Έργων Τρίτων (2/2)

Το Έργο αυτό κάνει χρήση των ακόλουθων έργων:

Πίνακες