ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών

Computerlinguistik

Lehreinheit 6 -7: Ressourcen

Dr. Christina Alexandris
Nationale Universität Athen
Deutsche Sprache und Literatur

Ressourcen

Ressourcen - Einführung

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.

Beispiel: Konstruktion eines Systems

- Zum Beispiel ist für die Konstruktion des Systems maschineller Übersetzung für Wetterberichte
- aber auch für die Konstruktion des Dialogsystems als Auskunftssystem für Wetterberichte

eine Sammlung von Texten (Korpus) aus Wetterberichten (geschriebener /gesprochener / multimodaler Form - Interaktion mit Videos) erforderlich,

um den **Inhalt** der Information, den **Wortschatz** und die **Art und Weise der Informationsvermittlung** der Wetterberichte zu bestimmen.

Beispiel: Evaluierung eines Systems

- In der letzten Phase der Konstruktion des Systems maschineller Übersetzung für Wetterberichte oder des Dialogsystems für Wetterberichte wird geprüft, ob das System **richtig funktioniert**
- (Phase der Evaluierung).

Dieser Vorgang

- findet mithilfe einer zweiten Sammlung von Texten (Korpus) aus Wetterberichten (in geschriebener, gesprochener oder multimodaler Form) statt.
- Bei der Konstruktion und der Evaluierung anderer Anwendungen der Verarbeitung natürlicher Sprache werden sowohl **Sammlungen von Texten** (**Korpora**) als auch **Sammlungen von Wörtern** (**Lexika**) verwendet.

Ressourcen als “Testsets”

- Ressourcen können somit für das Testen einzelner Komponenten oder für die Evaluierung eines gesamten Systems benutzt werden - als Datenbanken aufbereiteter und unaufbereiteter sprachlicher Daten.
- Diese Datenbanken, die für die Evaluierung eines gesamten Systems verwendet werden, werden "Testsets" genannt.

Quellen von Ressourcen

Quellen von Ressourcen sind:

(1) Sprachdatenbanken, wie Sammlungen von Texten (Korpora) oder auch Sammlungen von Wörtern (Lexika),

(2) das World-Wide-Web (WWW),

(3) Lexikalisch-semantische Wortnetze und

(4) sprachliche Daten in annotierter Form.

Lexikalisch-semantische Wortnetze

können generell als eine Art Lexika bezeichnet werden, die die semantischen und sonstigen linguistischen Beziehungen zwischen Wörtern oder Wendungen beschreiben (Anwendungsgebiet der theoretischen Linguistik).

Aufgabengebiete bezüglich der Ressourcen (1/2)

- Ein Aufgabengebiet der Computerlinguistik ist das Sammeln computerlinguistisch relevanter Informationen für die Erstellung von Ressourcen (z.B. Lexika und Korpora).
- Das Aufbereiten computerlinguistischer Daten konstituiert ein weiteres Aufgabengebiet bezüglich der Ressourcen.

Aufgabengebiete bezüglich der Ressourcen (2/2)

- Bei der Aufbereitung computerlinguistischer Daten wird die interne Darstellung der computerlinguistisch relevanten Informationen durch Datenmodelle, linguistische Repräsentationsformate und/oder Repräsentationsformate aus dem Bereich der Informatik (z.B. Markup-Sprachen) errreicht.
- Ein Aufgabengebiet in dem Computerlinguistischen Bereich der Ressourcen ist es auch computerlinguistische Daten zu verwalten und zur Verfügung zu stellen.

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.

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)**

Aufgabengebiete:

(1) Das **Sammeln computerlinguistisch relevanter Informationen,**

(2) das **Aufbereiten computerlinguistischer Daten und**

deren **interne Darstellung (Datenmodelle, Repräsentationsformate,** Markup-Sprachen)

(3) die Verwaltung und Zurverfügungstellung computerlinguistischer Daten.

Ressourcen

Lexika

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.

Beispiel aus einem Wirtschaftslexikon für Deutsch und Griechisch

- Rabattmarke = κουπόνι δώρων
- Rabatt = έκπτωση, προεξόφληση, υφαίρεση
- => (mit) Rabatt gewähren =προεξοφλώ (συναλλαγματική), υφαιρώ
- => (mit) Rabatt verkaufen =πουλώ με έκπτωση
- => (mit) Rabatt geben =δίδω/παρέχω έκπτωση

Schritte beim Aufbau eines Lexikons (1/4)

- Als erster Schritt beim Aufbau eines Lexikons kann der Prozess der systematischen Datensammlung bezeichnet werden.
- Mit der systematischen Datensammlung wird eine konsistente Version der Schriftformen und Transkriptionen der Wörter bzw. Texte erstellt.
- Um Konsistenz zu erzielen, werden Darstellungsstandards für Daten verwendet, die sogenannten "DTD" (Document Type Description). Heutzutage werden Darstellungsstandards für Daten in SGML oder XML benutzt (Gibbon, 2004).

Schritte beim Aufbau eines Lexikons (2/4)

- Als zweiter Schritt beim Aufbau eines Lexikons findet sowohl eine linguistische als auch eine statistische Analyse der Daten statt.
- Die relevanten Einheiten für die Lexikonerstellung werden identifiziert und linguistisch analysiert.
- In der statistischen Analyse (Manning und Schütze,1999), die parallel zu der linguistischen Analyse stattfindet, werden u.a. auch statistische Modelle für das Verhalten lexikalischer Einheiten im Kontext erstellt.

Schritte beim Aufbau eines Lexikons (3/4)-I

- 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

Schritte beim Aufbau eines Lexikons (3/4)-II

- 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

(**Operationalisierung** einer **Funktion**, die zunächst Korpuswörter auf **voll flektierte Lexikonwörter** überträgt, und diese dann auf **nichtflektierte Stämme** (morphologische Grundformen abbildet).

Schritte beim Aufbau eines Lexikons (4/4)

- Im dritten Schritt werden die gewonnenen lexikalischen Informationen in geeigneter Weise in verschiedene mediale Formen transformiert.
- Die Art der medialen Formen, in die lexikalischen Informationen transformiert werden, hängt von dem Anwendungskontext ab.
- Schließlich werden die transformierten lexikalischen Informationen in Datenbanken archiviert.

Schritte beim Lexikonaufbau (Gibbon, 2004):

- **Erster Schritt: Systematische Datensammlung (Gibbon et al, 1997)**, Erstellung einer konsistenten Version der Schriftformen und Transkriptionen.
- **Zweiter Schritt (a): Identifi zierung und linguistische Analyse** relevanter Einheiten für die Lexikonerstellung.
- **Zweiter Schritt (b): Parallel zur linguistischen Analyse: statistische Analyse** (Manning und Schütze, 1999) und Erstellung statistischer Modelle für das Verhalten lexikalischer Einheiten im Kontext.
- **Dritter Schritt: Transformation** (in verschiedene mediale Formen, je nach Anwendungskontext) der gewonnenen lexikalischen Informationen in geeigneter Weise und Archivierung in Datenbanken.

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.

Für die Suche und Extraktion von Informationen aus einem Text

- zeigt ein Lexikon die wichtigsten Wortgruppen an, aus denen der Inhalt des Textes ermittelt werden kann.
- Somit kann aus den Wortgruppen eines Textes der Schluss gezogen werden, dass ein Text z.B. sowohl Informationen aus dem Bereich des Fernmeldewesens als auch Informationen aus dem Bereich der Wirtschaft enthält.
- Die Beziehungen zwischen den Wortgruppen im Text können mit einer hierarchischen Struktur beschrieben werden.

Anwendungen der Fachtermini:

(1) Information Retrieval

(2) Text Mining

(1) Informationsextraktion

(2) lexikalisches Hilfsmittel (computerunterstützte Übersetzung)

(3) Aufbau mehrsprachiger Terminologiesammlungen und Translation Memories

Aufgaben und Aspekte der computergestützten Lexikographie und Terminologie

Die Aufgaben der computergestützten Lexikographie und Teminologie, können in vier Kategorien aufgeteilt werden:

(1) die Akquisition lexikalischer Informationen,

(2) die Akquisition linguistischer Informationen aus Textkorpora,

(3) die Akquisition linguistischer Informationen aus traditionellen Wörterbücher und

(4) die Repräsentation lexikalischer Informationen (Heid, 2004).

Aufgabe der Akquisition lexikalischer Informationen (1/2)-I

- ist die Beschaffung von **Daten**, auf deren Grundlage die **lexikographische Beschreibungsarbeit** stattfinden kann.
- Die Akquisition **linguistischer Informationen aus Textkorpora** zielt auf die Bereitstellung von **Hilfsmitteln** für die Lexikographen.

Aufgabe der Akquisition lexikalischer Informationen (1/2)-II

- 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.

Aufgabe der Akquisition lexikalischer Informationen (2/2)

- Für die Akquisition linguistischer Informationen aus **traditionellen Wörterbüchern** werden die Definition aber die morphologischen und syntaktischen Angaben eines Wortes, eines Ausdrucks oder eines Lemmas extrahiert.
- Aus den Textmustern der Definitionen werden **Oberbegriffe** extrahiert und daraus **Begriffshierarchien** aufgebaut

Aspekte der computergestützten Lexikographie und Terminologie

können in inhaltliche und technische Aspekte unterteilt werden

Die inhaltlichen Aspekte hängen mit der angestrebten Anwendung und der linguistischen Theorie bzw. mit dem **Beschreibungsansatz** eng zusammen.

- Die Anwendung kann zum Beispiel ein Lexikon für Wortklassentagging oder ein Lexikon für eine maschinelle Übersetzung sein.

Aspekte der computergestützten Lexikographie und Terminologie-II

Die inhaltlichen Aspekte der computergestützten Lexikographie und Terminologie betreffen

- (1) die Auswahl der in einem elektronischen Wörterbuch abzudeckenden linguistischen **Information**,
- (2) die **Klassifikation** und
- (3) **Kodierung** der Information sowie die Form der Definition einer lexikalischen Spezifikation, die den Grad seiner Zugreifbarkeit bestimmt.

Die technischen Aspekte der computergestützten Lexikographie und Terminologie-I

- befassen sich mit der Wahl des Repräsentationsformats und/ oder mit dem **Repräsentationsformalismus** für lexikalische Daten und mit der **Standardisierung** der **lexikalischen Daten**.
- Mit der Standardisierung werden die **Richtlinien** für Morphologie, Morphosyntax, Inhalt, Form (zum Beispiel für zwei- oder mehrsprachige **Wörterbücher**) bei der Sprachverarbeitung bestimmt.

Die technischen Aspekte der computergestützten Lexikographie und Terminologie -II

- Mit der Standardisierung werden die **Richtlinien** für Morphologie, Morphosyntax, Inhalt, Form (zum Beispiel für zwei- oder mehrsprachige **Wörterbücher**) bei der Sprachverarbeitung bestimmt.
- **Beispiele** von Repräsentationsformaten und/oder Repräsentationsformalismen für lexikalische Daten sind **Text-dateien** oder **Datenstrukturen** von computerlinguistischen Formalismen wie zum Beispiel **Merkmalsstrukturen**.

Aspekte der computergestützten Lexikographie und Terminologie

- Inhaltliche Aspekte (Was?):
  - Auswahl der linguistischen Information
  - Klassifikation der Information
  - Kodierung der Information
  - Zugreifbarkeit (Definition einer lexikalischen Spezifikation)
- Technische Aspekte (Wie?):
  - Wahl des Repräsentationsformats und/oder Repräsentationsformalismus für lexikalische Daten
  - Standardisierung

Aspekte der computergestützten Lexikographie und Terminologie

- Lexikalisch-semantische Wortnetze bilden die häufigsten und wichtigsten Wörter einer Sprache und ihre bedeutungstragenden Beziehungen zu anderen Wörtern der Sprache ab (siehe auch Portz, 2005).
- In einem Wortnetz wird ein Wort als Konzeptknoten mit seinen semantischen Verknüpfungen repräsentiert.
  - So wird das Wort "Stuhl" zum Beispiel mit dem Oberbegriff Sitzmöbel und den Unterbegriffen Drehstuhl, Klappstuhl, Kinderstuhl u.a verknüpft. Der Oberbegriff ist darüberhinaus mit den Konzepten Lehne, Sitzfläche und Bein verbunden, die Teile eines Sitzmöbels repräsentieren (Beispiel aus Kunze, 2004).

GERMANET (1/3)

- In den Beispielen aus GermaNET wird der Begriff "froh" mit den Begriffen "fröhlich" und "freudig" verbunden.
- Im Weiteren kann der Begriff "fröhlich" sowohl mit den Begriffen "heiter" und "lustig„ als auch mit dem Begriff "humorvoll" semantisch verknüpft werden.
- Der Begriff "lustig" kann wiederum auch mit den Begriffen "komisch", "drollig„ und "spaβig" interpretiert werden.
- Schlieβlich kann der Begriff "komisch„ mit dem Begriff "lächerlich" verbunden sein.

GERMANET (2/3)

- 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.

GERMANET (3/3)

- 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.

Beispiel

- Gemeinsamkeiten gibt es in Feldern wie zum Beispiel "Ferien" oder "Bauernhof", während in anderen thematischen Gebieten, wie "Feiertage", mehrere kulturelle Unterschiede erkennbar sind.
- Bei der Anwendung eines Lexikons in einem Sprachlehr- und –lernsystem werden z.B. beim Thema "Weihnachten" der Adventskranz und der Adventskalender der traditionellen deutschen (Vor-)Weihnachtszeit zugeordnet , während das Schiff, die Triangel und die Granatäpfel zur traditionellen griechischen Weihnachtsfeier gehören.
- Mit Hilfe eines wortnetzbasierten Computerlexikons können auch Beziehungen und Assoziationen gelernt werden (z.B. Räucherkerzen, Adventskranz, die in in der Griechischen Tradition nicht existieren).

Ressourcen

Textkorpora

Textkorpora und Korpustypen (1/2)-I

- Textkorpora bilden eine große Kategorie von Ressourcen.
- **Reine Textkorpora** bestehen aus geschriebenen oder transkribierten gesprochenen Texten, während **Sprachkorpora** aus Texten gesprochener Sprache in Form von Sprachsignalen (Audioaufnahmen) bestehen.

Textkorpora und Korpustypen (1/2)-II

- 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.

Korpustypen: (1/2)

(1) **Reine Textkorpora: Bestehen aus geschriebenen oder transkribierten** gesprochenen Texten, Grundeinheit ist das Token;

(2) **Sprachkorpora: Bestehen aus Sprachsignalen (Audioaufnahmen)** mit phonetischen und linguistischen Annotationen (Phonemgrenzen, Grundfrequenz, orthographische Transkription etc.).

Korpustypen: (2/2)

(3) **Multimodale Korpora: Bestehen aus Sprachsignalen mit zusätzlicher** Annotation von Prosodien, Mimik, Gestik, Mauszeiger- Bewegungen (oft in Verbindung mit einer Videoaufnahme)

(4) **Baumbanken (eine spezielle Form von Textkorpora): Bestehen** aus geschriebenen oder transkribierten gesprochenen Texten (aus syntaktisch analysierten Sätzen), Grundeinheit ist der Satz. Die Analysen werden meist durch Syntaxbäume oder DAGs (Graphen) repräsentiert.

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.

Lemmatisierung

- 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.

“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.

Literaturhinweise (1/2)

- Carstensen, K.U. (2004): "Ressourcen". In: *Computerlinguistik und*
- *Sprachtechnologie, Eine Einführung. Carstensen, K.U., Ebert, C., Endriss,* C., Jekat, S., Klabunde, R., Langer, H. (Hrsg.), 2te überarbeitete und erweiterte Aufl age, München: Spektrum Akademischer Verlag, 405-460.
- Evert, S., Fitschen, A. (2004): "Textkorpora". In: *Computerlinguistik und Sprachtechnologie, Eine Einführung. Carstensen, K.U., Ebert, C., Endriss,* C., Jekat, S., Klabunde, R., Langer, H. (Hrsg.), 2te überarbeitete und erweiterte Aufl age, München: Spektrum Akademischer Verlag, 406-413.
- Gibbon, D. (2004): "Lexika für multimodale Systeme". In: *Computerlinguistik und Sprachtechnologie, Eine Einführung. Carstensen, K.U., Ebert,* C., Endriss, C., Jekat, S., Klabunde, R., Langer, H. (Hrsg.), 2te überarbeitete und erweiterte Aufl age, München: Spektrum Akademischer Verlag, 432-439.

Literaturhinweise (2/2)

- Heid, U. (2004): "Computergestützte Lexikographie und Terminologie". In: *Computerlinguistik und Sprachtechnologie, Eine Einführung. Carstensen,* K.U., Ebert, C., Endriss, C., Jekat, S., Klabunde, R., Langer, H. (Hrsg.), 2te überarbeitete und erweiterte Aufl age, München: Spektrum Akademischer Verlag, 471-478.
- Kunze, C. (2004): "Lexikalisch-semantische Wortnetze". In: *Computerlinguistik und Sprachtechnologie, Eine Einführung. Carstensen, K.U.,* Ebert, C., Endriss, C., Jekat, S., Klabunde, R., Langer, H. (Hrsg.), 2te überarbeitete und erweiterte Aufl age, München: Spektrum Akademischer Verlag, 423-431.
- Portz, R. (2005): *Wort und Wortschatz, Lexikologische Betrachtungen zum Deutschen. Athen: Praxis.*

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