

Korpuslinguistik

Sebastian Pado
21.11.2006

Übersicht

- 1. Linguistische Methodologie
 - Theoretische Linguistik vs. [Korpuslinguistik](#)
- 2. Korpora
 - 2.1. Ebenen der Annotation
 - 2.2. Verwendung
 - 2.3. Aspekte der Annotation

1. Linguistische Methodologie

- “Wie betreibt man Linguistik?”

Einführung in die CL WS 06/07

The Armchair Linguist

He sits in a comfortable armchair, his eyes closed. Once in awhile he opens his eyes, shouting "Wow, what a neat fact", grabs his pencil, and writes something down. Then he struts around for a couple of hours, excited by his finding.

Einführung in die CL WS 06/07

The Corpus Linguist

He has a **corpus** of approximately one zillion running words that contains all his primary facts. His work is deriving secondary facts from primary facts.

At the moment, he is busy determining the relative frequencies of the eleven parts of speech as the first words of a sentence versus as the second word of a sentence.

Einführung in die CL WS 06/07

Gegenseitige Kritik

- Kritik an Korpuslinguistik:
 - „Why are your results **relevant**?“
- Kritik an theoretischer Linguistik:
 - „Why are your results **true**?“

Einführung in die CL WS 06/07

Wissenschaft = Modellbildung

Daten

Modell



Wissenschaft = Modellbildung

1. Voraussagen

Daten

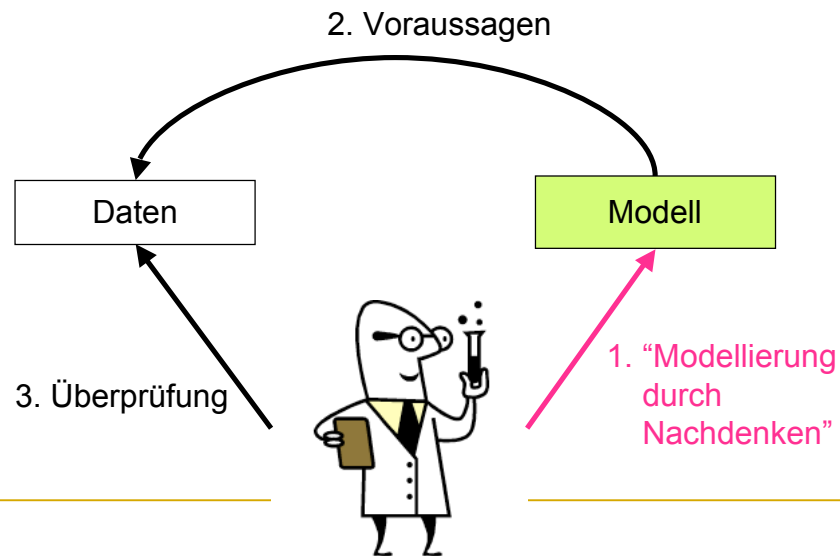
Modell

2. Überprüfung



Gutes Modell =
Korrekte Voraussagen

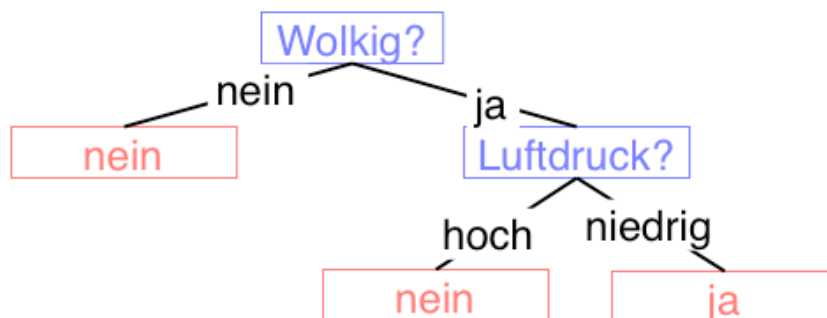
Rationalismus



Beispiel: Wettermodell

Frage: Wird es heute noch regnen?

Rationalistische Antwort: Was für **Faktoren** könnten dafür wichtig sein?



Rationalismus

- Typischerweise **symbolische** Modelle
- Vorteile
 - Konzentriert sich auf das „Warum?“
 - Erlaubt Modellierung komplexer Phänomene:
 - Rückgriff auf menschliches Verständnis
 - (Meta-)Analyse existierender Modelle
- Nachteile
 - Modelle oft nicht allgemein genug
 - Repräsentation durch Regeln oft zu starr
 - **Abhängigkeit von vorgefaßten Meinungen**

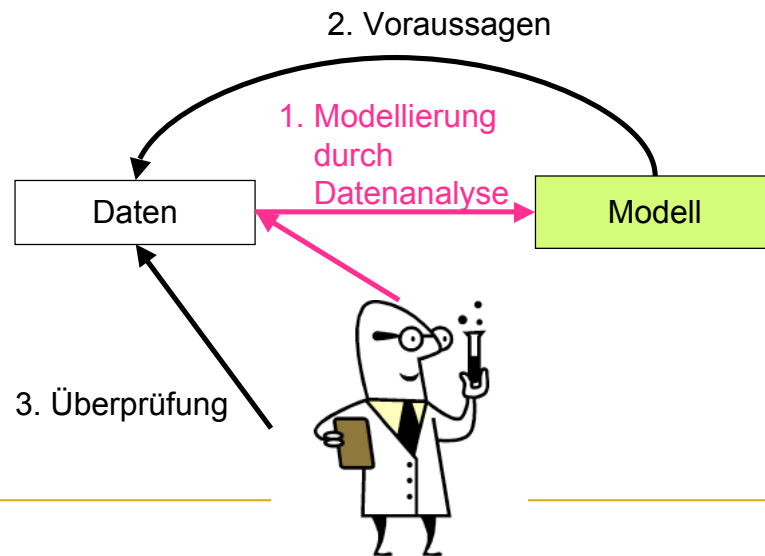
Einführung in die CL WS 06/07

Rationalismus in der Linguistik: Armchair Linguistics

- Oft Argumentation anhand einer kleiner Anzahl von Beispielsätzen:
„Satz X ist grammatisch, Satz Y ungrammatisch“
 - Gutes Modell: liefert **detaillierte (tiefe)** Analyse der Unterschiede zwischen X und Y
- Konzentration auf
 - Modellierung einzelner linguistischer Mechanismen
 - Meta-Analyse von Grammatiken (Grammatiktheorie)
- Herausforderung: die ganze **Breite** der Sprache abzudecken

Einführung in die CL WS 06/07

Empirismus



Beispiel: Wettermodell

Frage: Wird es heute noch regnen?

Empirische Antwort: Beobachte das Wetter ein paar Tage und berechne Häufigkeiten der Muster

Wolkig?	Luftdruck?	Wahrsch. fuer Regen
Nein	Hoch	5%
Ja	Niedrig	95%
Ja	Hoch	20%
Nein	Niedrig	50%

Empirismus

- Typischerweise **einfache, statistische** Modelle (Mustererkennung)
- Vorteile
 - Modelle bilden „beobachtete Wirklichkeit“ ab
 - Modelle robust
 - Maschinelle Lernverfahren verfügbar
- Nachteile
 - Modelle oft nur ungefähr richtig
 - **Abhängigkeit von den Daten**

Einführung in die CL WS 06/07

Empirismus in der Computerlinguistik

- Oft Argumentation anhand von grossen Textmengen
 - Gutes Modell: kann 84% anstatt 83% der Sätze einer gegebenen Textmenge analysieren
- Konzentration auf
 - Modellierung der häufigsten linguistische Mechanismen
 - Beobachtbare Eigenschaften der Sprache
- Herausforderung: die ganze **Tiefe** der Sprache abzudecken

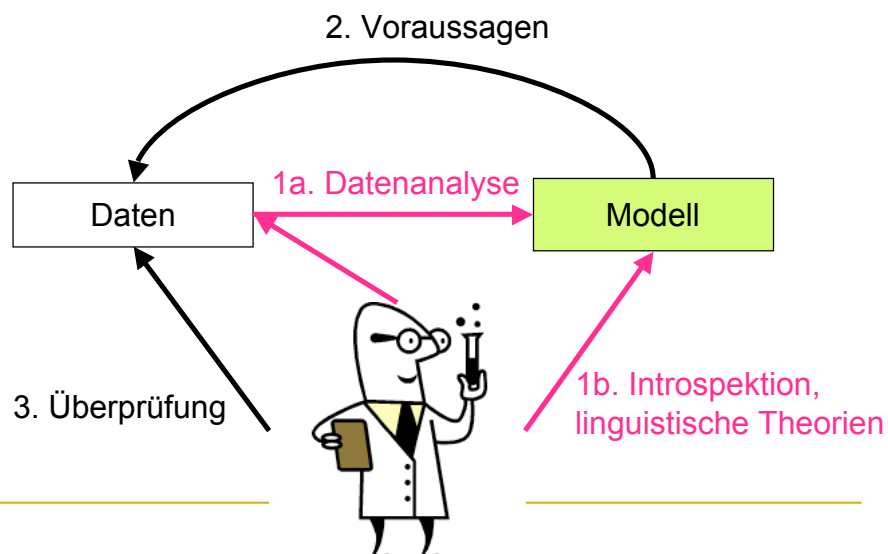
Einführung in die CL WS 06/07

Geschichte der Computerlinguistik

- Am Anfang (Ende 1940er): Reine Korpuslinguistik
 - Übersetzung Russisch – Englisch mit Mustererkennung
- Chomsky (1950er/60er): Theoretische Linguistik
 - Linguistische Grundlagenarbeit (Grammatiktheorien)
- Seit 1990: Grosse Erfolge der Korpuslinguistik
 - Große Datenmengen (Korpora, Internet)
 - Maschinelles Lernen zentrale Methode (Teilweise fundiert durch linguistische Theorien...)

Einführung in die CL WS 06/07

Heutige Methodik



Grammatiken im Wandel der Zeit

- Frühe Korpuslinguistik
 - Ziel: Modellierung einfacher Sätze im Korpus
 - Methode: einfache handkodierte Grammatik“muster”
 - wenig Interesse an abstrakten Erkenntnissen
- Theoretische Linguistik
 - Ziel: Entwicklung von Grammatikformalismen, Grammatiktheorien
 - Methode: Modellierung bestimmter Phänomene; Vergleich verschiedener Ansätze
 - wenig Interesse an Anwendung
- Moderne Computerlinguistik: Hybride Modelle
 - Expressive Grammatikformalismen, **plus** statistische Information

Einführung in die CL WS 06/07

2. Korpora

Einführung in die CL WS 06/07

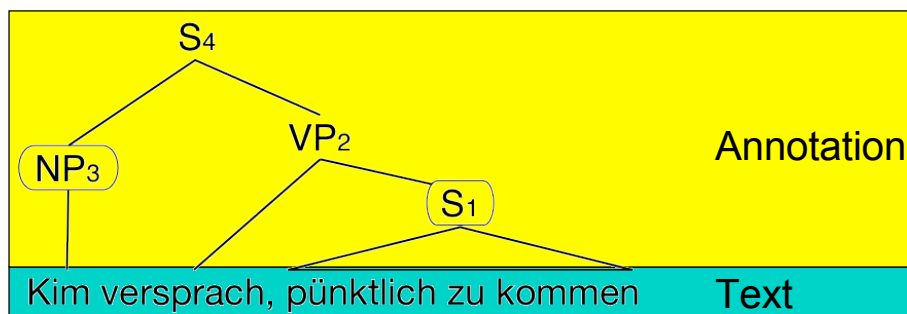
Was ist ein Korpus?

- Sprachwissenschaftliche Definition:
 - Ein Korpus (n.!) ist eine endliche Sammlung von konkreten sprachlichen Äußerungen, die als Grundlage für sprachwissenschaftliche Untersuchungen dienen (Lexikon der Sprachwissenschaft)

Einführung in die CL WS 06/07

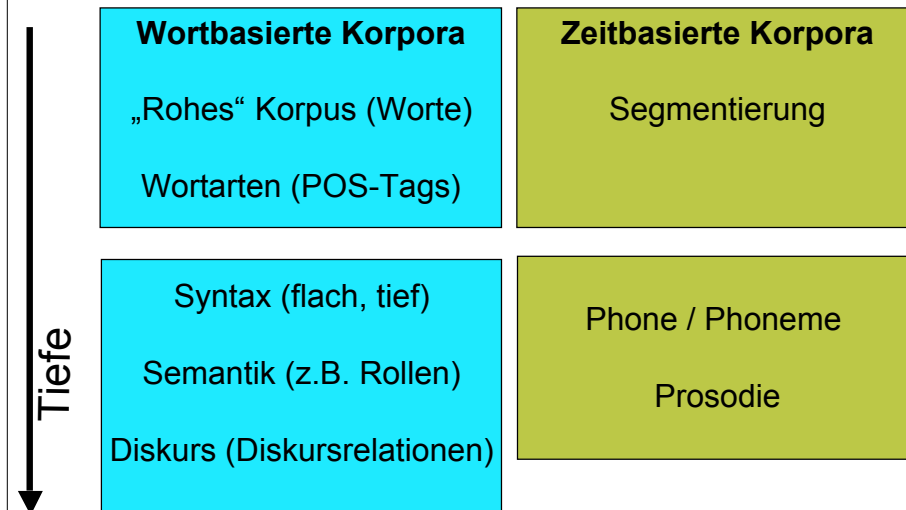
Annotation

- Fast alle in der Computerlinguistik verwendeten Korpora sind **annotiert**, d.h. sie sind linguistisch analysiert



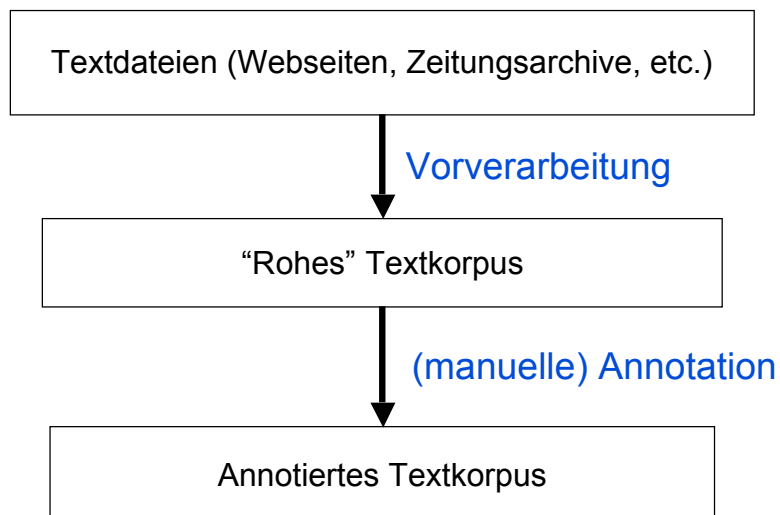
Einführung in die CL WS 06/07

2.1. Annotationsebenen



Einführung in die CL WS 06/07

Korpusverarbeitung



Einführung in die CL WS 06/07

Vorverarbeitung

- Vorverarbeitung ist alles, was notwendig ist, um das Korpus annotieren zu können
- Wortbasierte Korpora
 - Grundlage: Text(datei)
 - Vorverarbeitung: Tokenisierung
- Zeitbasierte Korpora
 - Grundlage: phonetisches Signal (Aufnahme)
 - Vorverarbeitung: Segmentierung

Einführung in die CL WS 06/07

Tokenisierung

- Aufgabe: Erkennung von Wort- und Satzgrenzen
 - Problem: Textdatei ist Abfolge von Zeichen
- Was ist eine Satzgrenze?
 - Heuristik: Ein Satzzeichen (Punkt, ...)
 - 1., Mr., Std.
- Was ist eine Wortgrenze?
 - Heuristik: Alles, was kein Buchstabe ist
 - Tholey-Theley, i18n, it's

Sehr schwierig bei asiatischen Sprachen

Einführung in die CL WS 06/07

Annotationsebenen in Korpora

Einführung in die CL WS 06/07

„Rohe“ Korpora

Dies ist ein Korpus ohne Annotation.

- Lexikographie: Manuelle Sichtung der Beispiele
 - Bestimmung von Wortbedeutungen
- Erstellen und Erweitern von Wörterbüchern
 - Suche nach Neologismen (Neubildungen)
 - Konkret: Suche nach Wörtern mit schwankender Häufigkeit
 - Suche nach Kollokationen
 - Ins Gras beißen, sich einen schönen Tag machen, etc.
 - Konkret: Suche nach Worten, die häufig gemeinsam auftreten
- Sehr grosse Korpora (aus dem Internet), mehrere G Wörter

Einführung in die CL WS 06/07

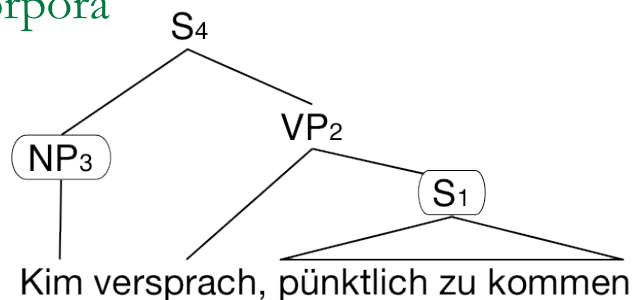
Korpora mit Wortarten

Dieser Satz ist	mit	Wortarten	annotiert.
ART NN VAFIN	PRP NN	VVPP / ADJ	
dies Satz sein	mit	Wortart	annotieren

- Jedem Wort wird eine Wortart zugeordnet
- Standardkorpora:
 - Englisch: British National Corpus (BNC), 100M Worte
- Oft gekoppelt mit Lemmatisierung (Grundform von Wörtern bestimmen)

Einführung in die CL WS 06/07

Syntax-Korpora



- Standardkorpora („Baumbanken“): Fast ausschliesslich Zeitungstexte
 - Englisch: Penn Treebank (1M Worte Wall Street J.)
 - Deutsch:
 - NEGRA (20.000 Sätze Frankfurter Rundschau = 400K Worte)
 - TIGER (80.000 Sätze Frankfurter Rundschau = 1.5M Worte)

Einführung in die CL WS 06/07

Semantik-Korpora

[Peter]	gibt	[Maria]	[ein Buch]
[Von Peter]	kriegt	[Maria]	[ein Buch]
Agent		Recipient	Theme

- Annotation von Satzteilen mit „semantischen Rollen“
 - „Who does what to whom“?
- Korpora:
 - Englisch: PropBank, auf Grundlage der Penn Treebank
 - Deutsch: SALSA, auf Grundlage von TIGER (in Arbeit)

Einführung in die CL WS 06/07

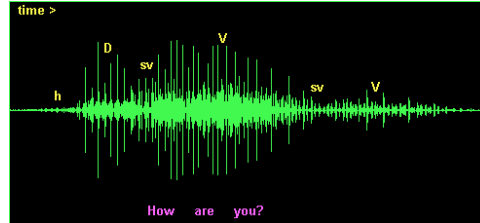
Diskurs-Korpora

[Peter ist müde].	Deshalb	[schläft er].
Grund	DPART	Folge

- Annotation von Paaren von Sätzen mit Diskursrelationen
 - z.B. Begründung (weil), Zweck (damit), ...
- Korpora:
 - DiscourseBank, auf Grundlage der Penn Treebank

Einführung in die CL WS 06/07

Phonetik-Korpora



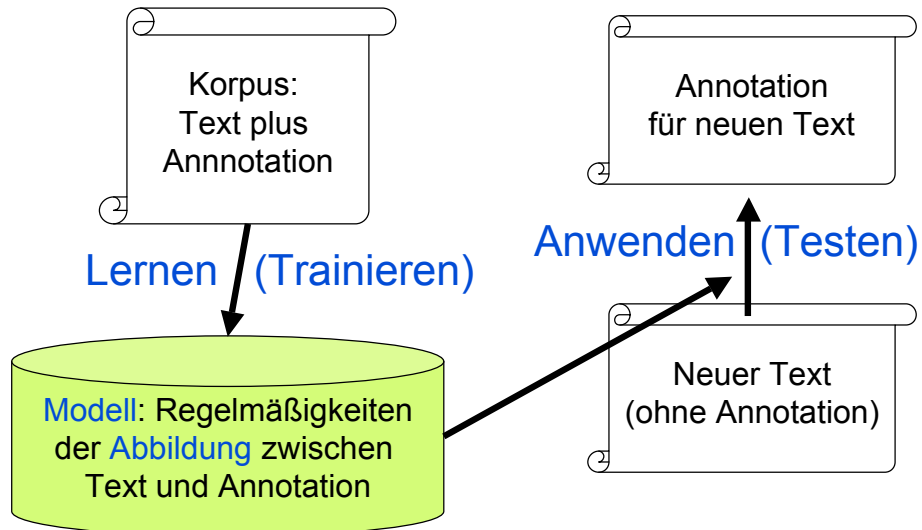
- Annotation von Aufnahmen (Oszillogrammen) mit Phon(em)en (Transkription)
- Standardkorpora: v.a. amerikanisches Englisch
 - Auskunftssysteme
 - ATIS: Air Travel Information Service
 - Telefonkonversation
 - Switchboard (>2000 Dialoge à 6 Min. = 1.5M Worte)

Einführung in die CL WS 06/07

2.2. Verwendung von Korpora

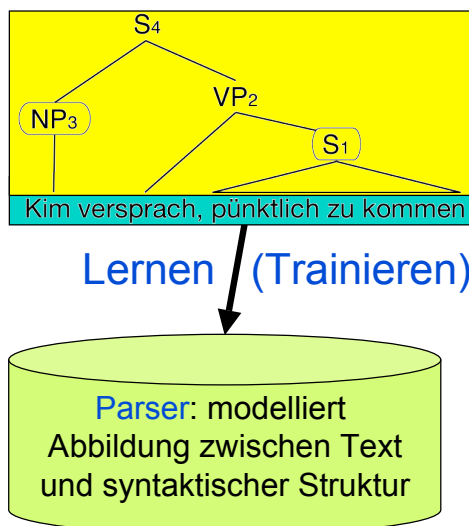
Einführung in die CL WS 06/07

Maschinelles Lernen



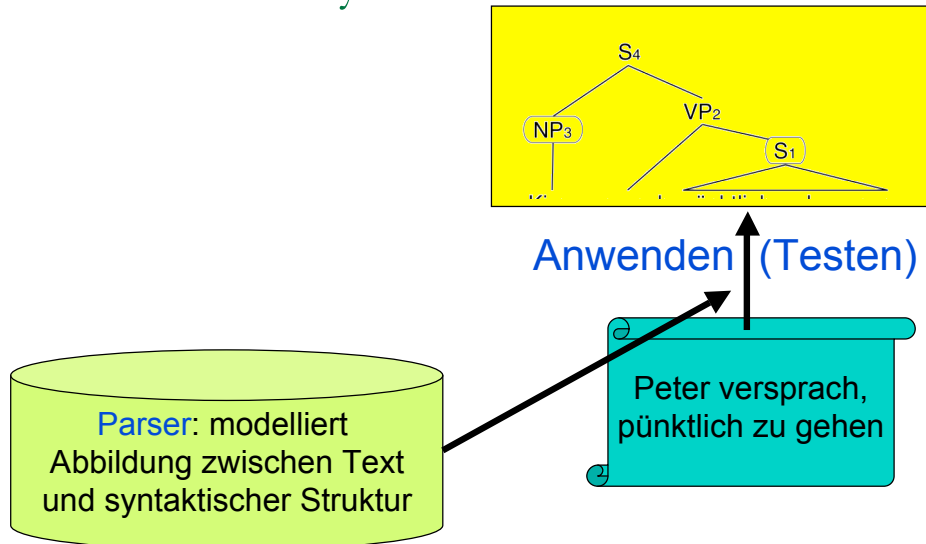
Einführung in die CL WS 06/07

Lernen von Syntax



Einführung in die CL WS 06/07

Lernen von Syntax



Einführung in die CL WS 06/07

Anwendung

- Maschinelles Lernen erlaubt die korpusbasierte Konstruktion von Modellen, die **automatisch neuen Text** analysieren können
 - Z.B. mit Wortarten ("part-of-speech tagger")
 - Z.B. mit syntaktische Struktur ("Parser")
- Das Gebiet, das sich mit der Anwendung solcher Werkzeuge befaßt, heißt Natural Language Processing (NLP) bzw. Sprachtechnologie

Einführung in die CL WS 06/07

2.3. Aspekte von Annotation

- Erinnerung: Die Qualität solcher Werkzeuge hängt von der Qualität der Korpora ab
 - Frage 1: Ist das Korpus repräsentativ?
 - Frage 2: Ist das Korpus korrekt annotiert?

Einführung in die CL WS 06/07

Repräsentativität

- Aus Korpora gelernte Modelle modellieren Eigenschaften **des Korpus**
- Korpora sind im Idealfall **repräsentativ (balanciert)**:
 - Alle Genres, Sprachebenen, Gegenstandsbereiche (Domänen)
- Die meisten sind es nicht!
 - Balancierte Korpora: BNC, Brown-Korpus (beide englisch)
 - Zeitungskorpora sind unbalanciert (TIGER, Penn Treebank)

Einführung in die CL WS 06/07

Korrektheit

- Fast alle Korpora sind manuell annotiert
- Selbst manuelle Annotation ist nie fehlerfrei
 - Grund 1: Unaufmerksamkeit der Annotatoren
 - Grund 2: Schwierigkeit der Aufgabe
- Nötig:
 - Qualitätssicherung
 - Geeignetes Annotationsschema

Einführung in die CL WS 06/07

Qualitätssicherung

- Annotation soll über die Zeit gleich bleiben (hohes **Intra-Annotator Agreement**)
 - Idee: Jeder Annotator **wiederholt** seine Annotation (in zeitlichem Abstand)
- Mehrere Annotatoren sollen gleich annotieren (hohes **Inter-Annotator Agreement**)
 - Idee: Mehrere **unabhängige** Annotatoren annotieren dasselbe

Einführung in die CL WS 06/07

Aufwand: Der begrenzende Faktor

- Wenn die Annotation eines Wortes im Durchschnitt 30 Sekunden dauert
- ...dann dauert die Annotation von 1 Million Wörtern mehrere Jahre (Qualitätssicherung nicht eingerechnet!)
- Oft wird das Agreement nur stichprobenartig überprüft

Einführung in die CL WS 06/07

Annotationsschema

- Ein Annotationsschema spezifiziert:
 - Kategorien: Was wird annotiert?
 - Richtlinien: Wann annotiere ich was?
 - Problemfälle: Was passiert, wenn ich mir nicht sicher bin?
- Das Annotationsschema kann über Erfolg oder Mißerfolg eines Annotationsprojekts entscheiden

Einführung in die CL WS 06/07

Annotationsschema: Nominal-Wortarten

■ Penn Tagset (45 Kategorien)

- NN – noun, singular
- NNS – noun, plural
- NNP – proper noun, singular
- NNPS – proper noun, plural

Einführung in die CL WS 06/07

Annotationsschemata: Nominal-Wortarten

■ CLAWS2-Tagset (132 Kategorien)

- ND1 – singular noun of direction (north, southeast)
- NN / NN1 / NN2 – common noun, neutral / sg / pl (cod / book / books)
- NN1\$ -- genitive singular common noun (domini)
- NNJ / NNJ1 / NNJ2 – organization noun (department / assembly / governments)
- NNL / NNL1 / NNL2 – locative noun (Is. / street / roads)
- NNO / NNO1 / NNO2 – numeral noun (dozen / ? / hundreds)
- NNS / NNS1 / NNS2 – noun of style (? / president / viscounts)
- NNSA1 / NNSA2 – following noun of style abbreviation (M.A.)
- NNSB / NNSB1 / NNSB2 – preceding noun of style abbreviation (Prof.)
- NNT / NNT1 / NNT2 – temporal noun (? / day / days)
- NNU – unit of measurement (in., inch / inches)
- NP / NP1 / NP2 – proper noun (Andes / London (Korea)
- NPD1 / NPD2 – weekday noun (Sunday / Sundays)
- NPM1 / NPM2 – month noun (October / Octobers)

Einführung in die CL WS 06/07

Annotationsschemata

- Wie detailliert soll die Annotation sein?
 - Detaillierte Annotation
 - Viele Kategorien, somit viel Information
 - Aus der Annotation gelernte Modelle werden detailliert
 - Viele Zweifelsfälle (schwer, Qualität zu halten)
 - Grobe Annotation
 - Wenige Kategorien, somit wenig Information
 - Einfacher, Qualität zu halten