



Korpuslinguistik

Sebastian Pado

1

This slide features a title box with a dark blue vertical bar on the left containing the text 'Korpuslinguistik'. Below it is a subtitle box with a light blue vertical bar on the left containing the text 'Sebastian Pado'. The slide number '1' is located in the bottom left corner.

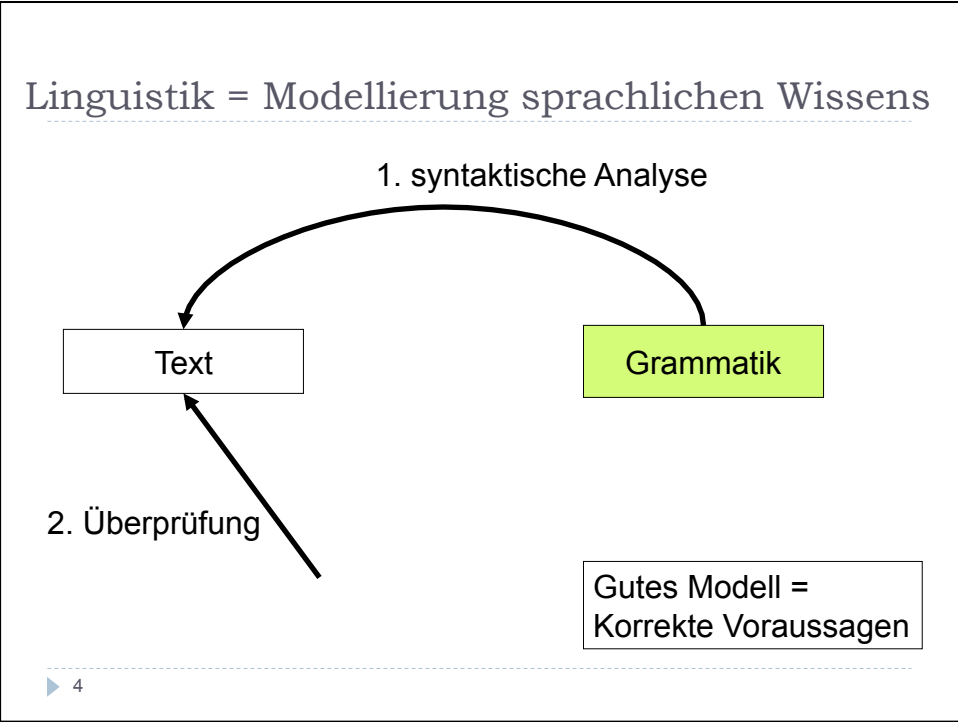
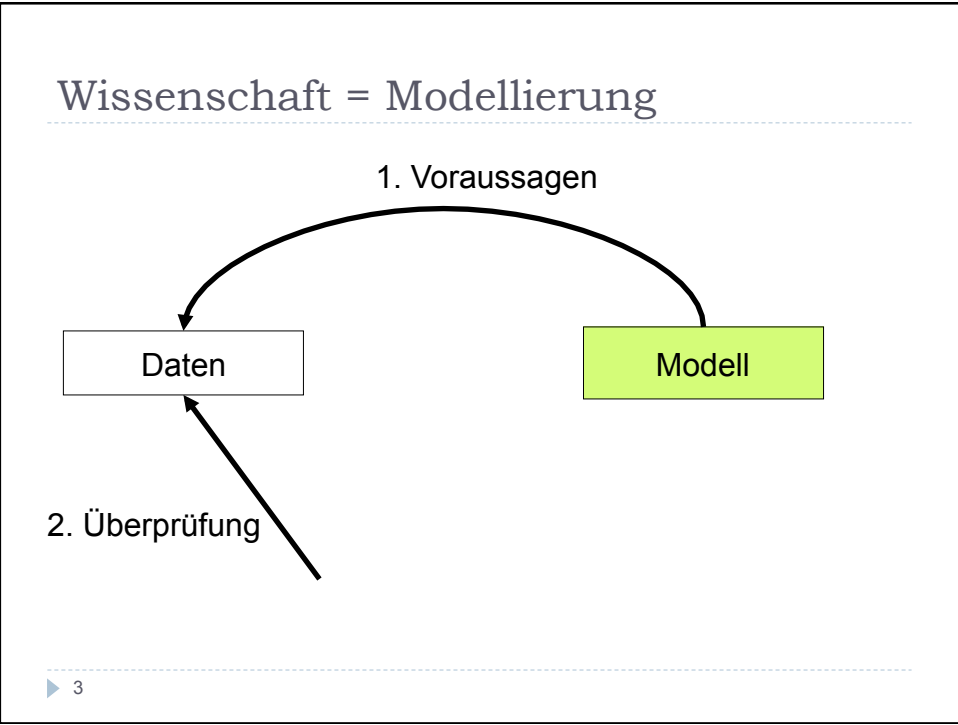
Wissenschaft = Modellierung

Daten

Modell

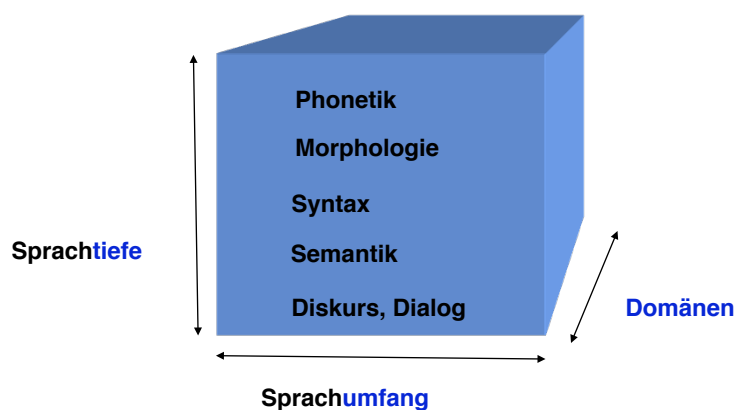
▶ 2

This slide has the title 'Wissenschaft = Modellierung' at the top, underlined. Below the title are two boxes: a white box on the left labeled 'Daten' and a light green box on the right labeled 'Modell'. At the bottom left, there is a small blue triangle followed by the number '2'.



## Linguistische Methodologie

- ▶ Woher kommt das **sprachliche Wissen**, das wir in der Sprachverarbeitung verwenden?



▶ 5

## 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.

▶ 6

## 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.

---

▶ 7

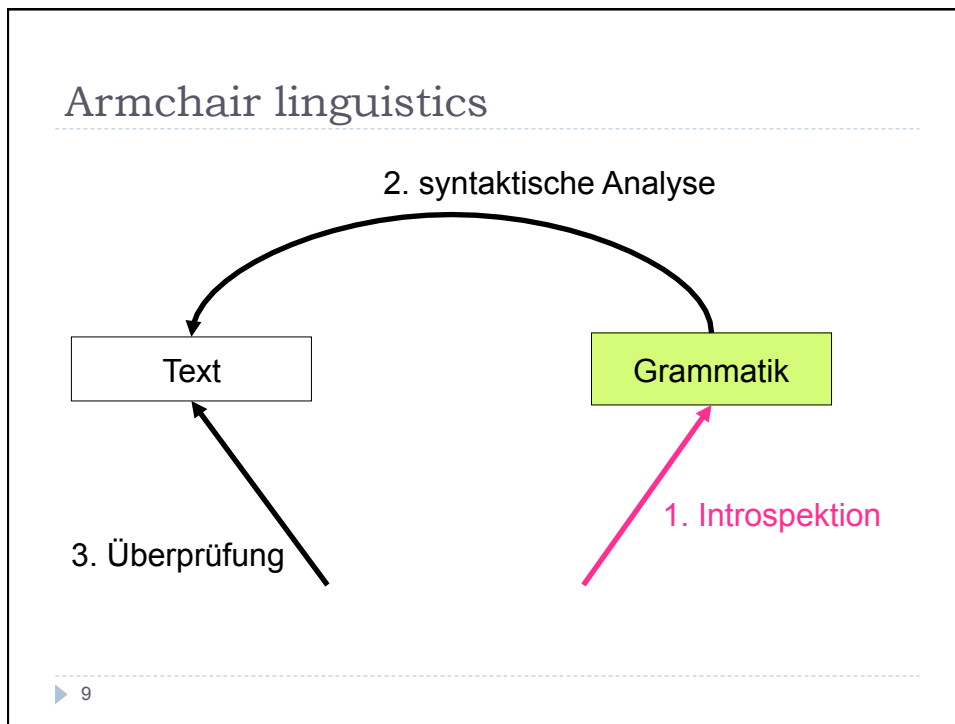
## Gegenseitige Kritik

---

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

---

▶ 8



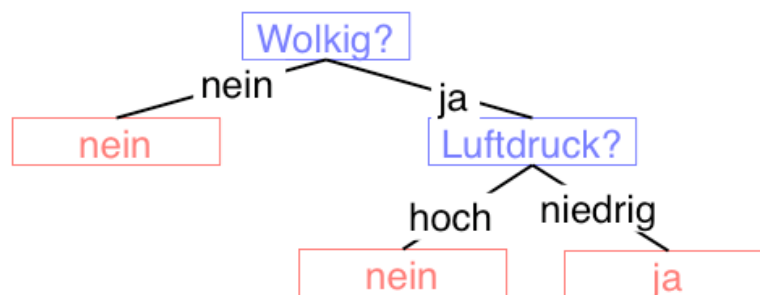
- ## Armchair linguistics
- **Modellierung durch Nachdenken**
    - Typischerweise komplexe, **regelbasierte** Modelle
  - **Vorteile**
    - Erlaubt Modellierung komplexer Phänomene
      - Rückgriff auf menschliches Verständnis
    - (Meta-)Analyse existierender Modelle
    - Ergebnis-Wissensrepräsentation für menschliche Analyse geeignet
  - **Nachteile**
    - Am besten geeignet für **diskrete Phänomene**
    - Erfasst alle möglichen Lesarten: **Ambiguität**
    - Präskriptivität vs. Deskriptivität
      - **Mangel an Robustheit**
    - Aufwand!!
- ▶ 10

## Eine kognitive Erkenntnis

- ▶ Menschen sind sehr gut darin, **Muster (Regeln)** zu erkennen
- ▶ Menschen sind ziemlich schlecht darin, **Zahlen** verlässlich abzuschätzen

▶ 11

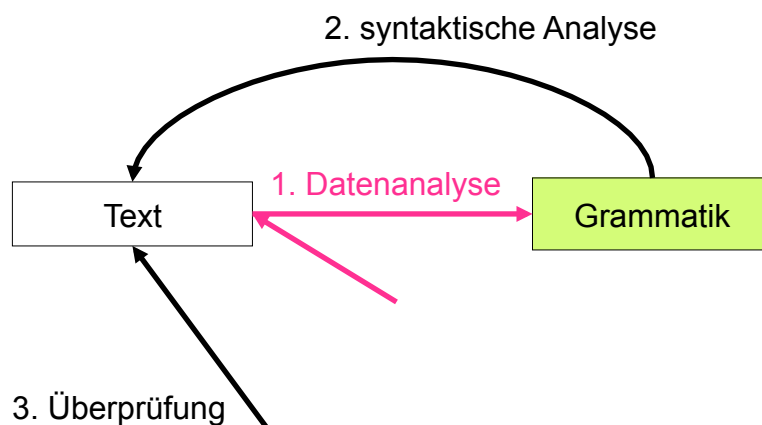
## Beispiel: symbolisches Modell



- In der Computerlinguistik besonders erfolgreich für
  - Grammatiken, Meta-Analyse von Grammatiken (Grammatiktheorie)
  - formale semantische Analyse (insbes. von strukturellen Aspekten, Diskursrepräsentation)

▶ 12

## Corpus linguistics



▶ 13

## Corpus linguistics

- ▶ Modelle durch Sichtung von Beispielen
  - ▶ Typischerweise **statistische Modelle** (Mustererkennung)
- ▶ Vorteile
  - ▶ Erlaubt Bestimmung von **Wahrscheinlichkeiten**
    - ▶ Welches ist die wahrscheinlichste aller möglichen Lesarten?
  - ▶ Einsatz maschineller Lernverfahren
    - ▶ Schnelle Modellierung neuer Domänen, Sprachen etc.
- ▶ Nachteile
  - ▶ Modelle oft nur approximativ richtig
  - ▶ Schwierige Probleme können oft nicht zuverlässig modelliert werden (Integration von externem Wissen schwierig)
  - ▶ Abhängigkeit von den Daten
  - ▶ Modelle schwierig von Menschen modifizierbar

▶ 14

## Beispiel: statistisches Wettermodell

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

- ▶ In der Computerlinguistik erfolgreich für:
  - ▶ Neologismen entdecken, Texte datieren
  - ▶ Wortartenanalyse, „grobe“ semantische Analyse
  - ▶ Automatische syntaktische Analyse

▶ 15

## Methodengeschichte der Computerlinguistik

- ▶ 1950er-1980er: Theoretische Linguistik
  - ▶ Linguistische Grundlagenarbeit (Grammatiktheorien)
  - ▶ Methode: Modellierung bestimmter Phänomene; Vergleich verschiedener Ansätze
    - ▶ Weniger Fokus auf praktische Anwendung
- ▶ Seit 1990: Verwaltung riesiger Datenmengen wird als zentrale Aufgabe (Internet!) der CL erkannt
  - ▶ Maschinelles Lernen zentrale Methode
    - ▶ „Jedes Problem läßt sich durch genügend Daten lösen“
    - ▶ „Every time I fire a linguist, my language processor improves“

▶ 16

## In der Semantik...

- ▶ “You shall know a word by the **company** it keeps” (Z. Harris)

- ▶ Ähnliche Kontexte → Ähnliche Bedeutung

\_\_\_\_\_ is sold in liter bottles.  
 \_\_\_\_\_ is a popular drink in Russia  
 \_\_\_\_\_ makes you drunk quickly.

- ▶ Robuste Bedeutungsbeschreibung
- ▶ Aber: Klappt nur für grobe Bedeutungsunterschiede

- ▶ Gegensätze und Ober/Unterbegriffe kommen auch in ähnlichen Kontexten vor...

a \_\_\_\_\_ dress  
 \_\_\_\_\_ ice  
 \_\_\_\_\_ matter

- ▶ Nötig: Genauere Konzentration auf bestimmte Kontexte
- ▶ ...erfordert **zusätzliches linguistisches Wissen!**

▶ 17

## In der Syntax

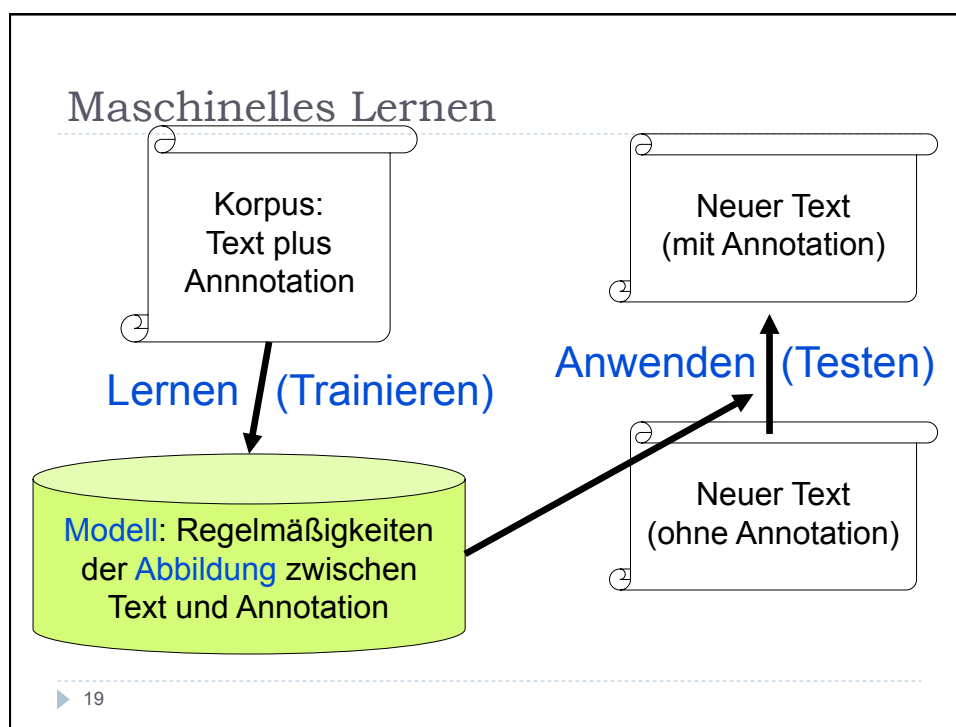
- ▶ Ziel: Automatische syntaktische Analyse von Sätzen wie „Peter sieht den Mann mit dem Fernrohr“

- ▶ Schritt I: PP „sieht X mit dem Fernrohr“ hat zwei Lesarten
- ▶ Schritt II: VP-Anbindung (Instrumentenlesart) wahrscheinlicher als NP-Anbindung

- ▶ Was ist dazu nötig?

- ▶ Viel **syntaktisch annotierter** Text
- ▶ Lernen von Beziehungen zwischen Wörtern (sehen, mit, Fernrohr) und syntaktischer Struktur (VP, PP, ...)

▶ 18

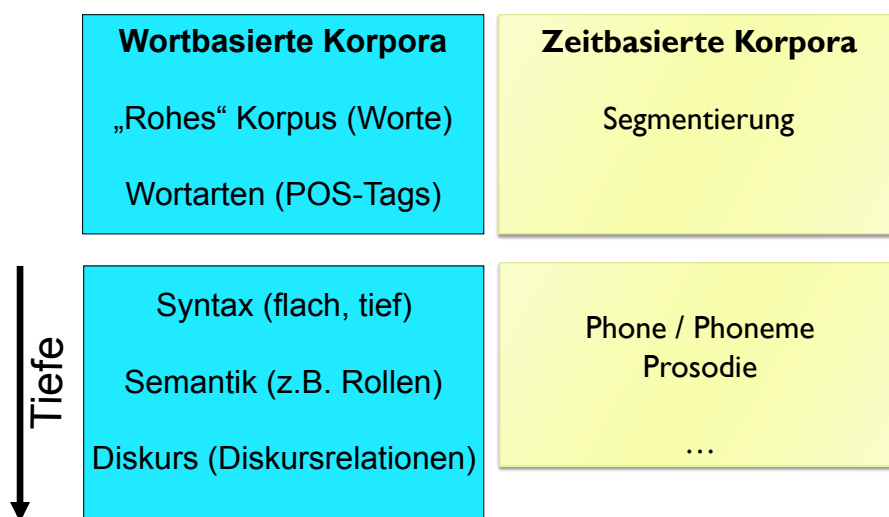


## Korpuslinguistik und CL

- ▶ Ein Korpus (n.!) ist eine endliche Sammlung von konkreten sprachlichen Äußerungen, die als Grundlage für sprachwissenschaftliche Untersuchungen dienen (Lexikon der Sprachwissenschaft)
- ▶ Computerlinguistische Korpora haben quasi immer zusätzliche **linguistische Annotation**
  - ▶ Zentrale Grundlage für datengetriebene Modellierung
    - ▶ Und damit eines Großteils der aktuellen CL-Forschung und Anwendung
  - ▶ Es gibt inzwischen Korpora für (fast) jede linguistische Ebene – fürs Englische

▶ 20

## Annotationsebenen



▶ 21

## „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

▶ 22

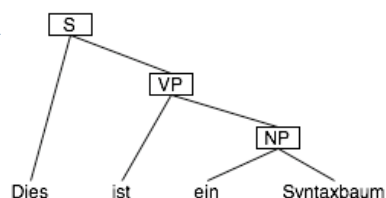
## Korpora mit Wortarten

Dieser Satz ist mit Wortarten annotiert.  
 ART NN VAFIN PRP NN VVPP / ADJ

- ▶ **Training von Wortartenbestimmern (POS-Taggern)**
  - ▶ Aufgabe: Ordne jedem Wort eine Wortart zu
- ▶ **Standardkorpora:**
  - ▶ British National Corpus (BNC), 100M Worte
  - ▶ American National Corpus (ANC), 10M Worte
  - ▶ Huge German Corpus (HGC), 200M Worte

▶ 23

## Syntax-Korpora



- ▶ **Training von stochastischen Parsern:**
  - ▶ Aufgabe: ordne jedem Satz eine **wahrscheinlichste** syntaktische Analyse zu
- ▶ **Standardkorpora („Baumbanken“): Zeitungstexte**
  - ▶ Englisch: Penn Treebank (1M Worte Wall Street Journal)
  - ▶ Deutsch:
    - ▶ NEGRA (20.000 Sätze Frankfurter Rundschau = 400K Worte)
    - ▶ TIGER (80.000 Sätze Frankfurter Rundschau = 1.5M Worte)

▶ 24

## Semantik-Korpora

---

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

- ▶ Training von semantischen Parsern
  - ▶ Aufgabe: ordne Satzteilen „semantische Rollen“ zu
- ▶ Korpora:
  - ▶ Englisch: PropBank, auf Grundlage der Penn Treebank
  - ▶ Deutsch: SALSA, auf Grundlage von TIGER

---

▶ 25

## Diskurs-Korpora

---

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

- ▶ Training von „Diskurs-Parsern“
  - ▶ Ordne Paaren von Sätzen Diskursrelationen zu
    - ▶ z.B. Begründung (weil), Zweck (damit), ...
  - ▶ Herausforderung: „Leere“ Diskurspartikel
    - ▶ Peter kam nicht mit. Er schlief.
- ▶ Korpora:
  - ▶ DiscourseBank, auf Grundlage der Penn Treebank

---

▶ 26

## Phonetik-Korpora

---

- ▶ **Training von Spracherkennungs-Systemen**
  - ▶ Ordne einer Schwingung / Schwingungsfolge eine orthographische Einheit zu
  
- ▶ **Standardkorpora: v.a. amerikanisches Englisch**
  - ▶ Auskunftssysteme
    - ▶ ATIS: Air Travel Information Service
  - ▶ Telefonkonversation
    - ▶ Switchboard (>2000 Telefondialoge à 6 Min. = 1.5M Worte)

---

▶ 27

## Keine Korpora verfügbar

---

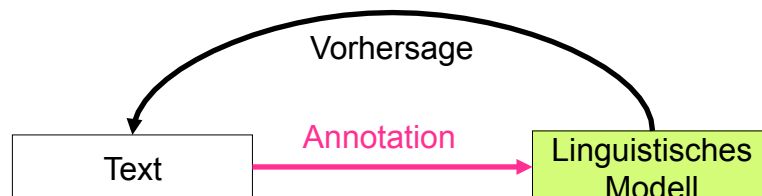
- ▶ **Pragmatik**
  - ▶ Intentionen der Sprecher
  - ▶ “was wirklich gemeint ist”
  
- ▶ **Viele andere Sprachen!**

---

▶ 28

## Eigenschaften der Annotation

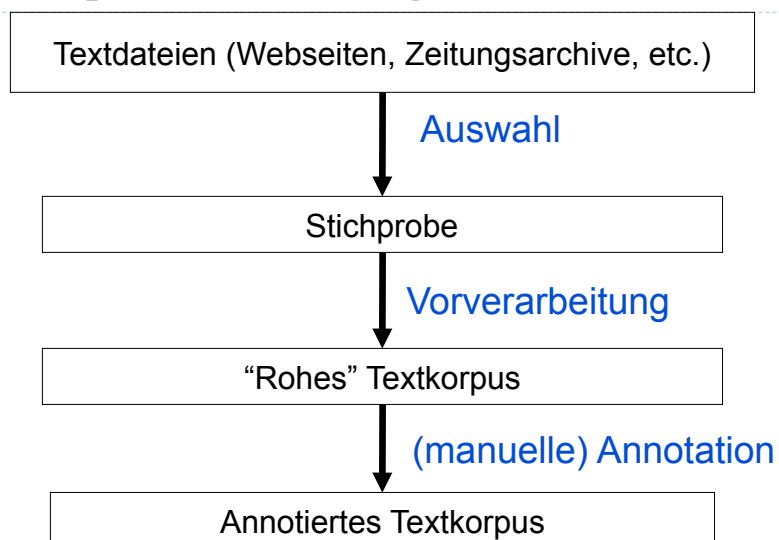
### ▶ Erinnerung:



- ▶ Annotation bildet die Grundlage der Modellierung
  - ▶ Annotation entspricht der Entwicklung linguistischer Theorien
- ▶ Was muß man bei der Annotation beachten, um ein verlässliches Korpus zu konstruieren?
  - ▶ TINSTAFL

▶ 29

## Korpusverarbeitung



▶ 30

## Representativität

- ▶ 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
  - ▶ Zeitungskorpora unbalanciert (TIGER, Penn Treebank)
  - ▶ Politische Korpora auch nicht

▶ 31

## Größe von Korpora

- ▶ **Wie groß sollten Korpora sein?**
  - ▶ Groß für flachere sprachliche Ebenen
  - ▶ Größer für tiefere sprachliche Ebenen
    - ▶ Abbildung Text -- Annotation ist schwieriger
- ▶ **Aber: tiefe Annotation sehr aufwendig**
  - ▶ Roher Text: mehrere G Wörter verfügbar
  - ▶ POS-Tags: BNC (100M Wörter)
  - ▶ Syntax/Semantik: 1-10M Wörter

**Es gibt nie genug Daten, um alles zu lernen**

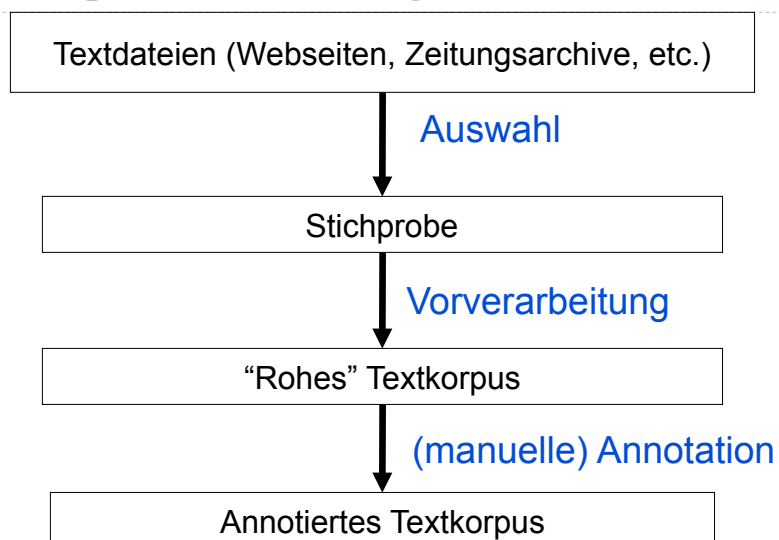
▶ 32

## „The Web as Corpus“

- ▶ **Vorschlag: Nutzen von Internet-Daten**
  - ▶ Problem 1: Repräsentativität
    - ▶ Bedeutung von „amazon“
  - ▶ Problem 2: Automatische Annotation nötig
    - ▶ Korrektheit der Daten zweifelhaft
- ▶ **Empirisches Ergebnis: Nutzen hängt von linguistischer Ebene ab**
  - ▶ Flache Analyse: Zusatzdaten vorteilhaft trotz Fehlern
  - ▶ Tiefe Analyse: Fehler überwiegen Vorteil

▶ 33

## Korpusverarbeitung



▶ 34

## Vorverarbeitung

---

- ▶ Vorverarbeitungsschritte werden automatisch erledigt
  
- ▶ **Wortbasierte Korpora**
  - ▶ Grundlage: Text(datei)
  - ▶ Vorverarbeitung: Säuberung, Tokenisierung, Lemmatisierung, (Wortartenbestimmung)
  
- ▶ **Zeitbasierte Korpora**
  - ▶ Grundlage: phonetisches Signal (Aufnahme)
  - ▶ Vorverarbeitung: Segmentierung

---

▶ 35

## Vorverarbeitung: Säuberung

---

- ▶ Im Idealfall enthält die Eingabe einen grammatischen, kohärenten Text
  
- ▶ **Häufige Probleme:**
  - ▶ Eingescannte Texte: OCR-Fehler
  - ▶ Webseiten: HTML-Markup, Navigationselemente
  - ▶ Zeitungsartikel: Insets etc.
  - ▶ Blogs/Forums: Unvollständige und ungrammatische Sätze

---

▶ 36

## Vorverarbeitung: Tokenisierung

- ▶ **Aufgabe: Erkennung von Wort- und Satzgrenzen**
  - ▶ Problem: Textdatei ist Abfolge von Zeichen
- ▶ **Was ist eine Satzgrenze?**
  - ▶ Heuristik: Ein Satzzeichen (Punkt, ...)
  - ▶ I., Mr., Std.
- ▶ **Was ist eine Wortgrenze?**
  - ▶ Heuristik: Alles, was kein Buchstabe ist
  - ▶ Villingen-Schwenningen, l33t, it's

Sehr schwierig bei asiatischen Sprachen

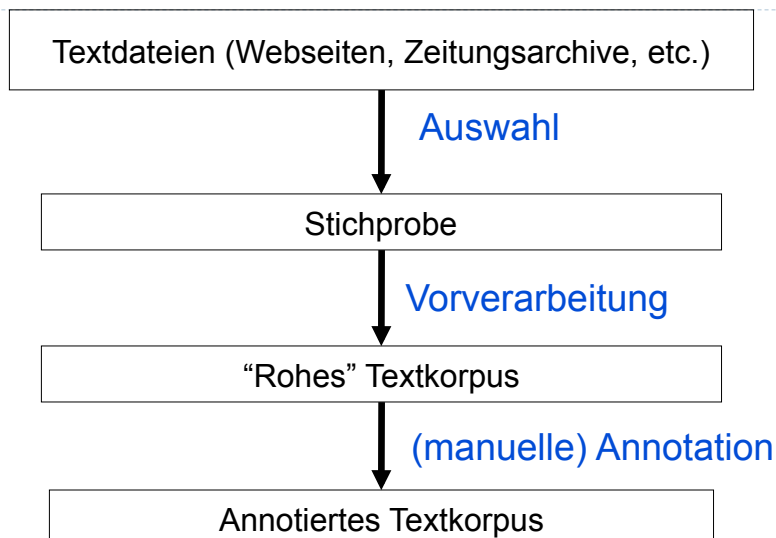
▶ 37

## Vorverarbeitung: Lemmatisierung

- ▶ **Aufgabe: Grundformen von Worten finden**
  - ▶ Problem: Korpus enthält Wortformen
    - ▶ Flexion, Konjugation, (Derivation)
  - ▶ Grundformen oft informativer
    - ▶ Wie oft kommt "sich sicher sein" im Korpus vor?
- ▶ **Problem 1: Mehrdeutigkeit**
  - ▶ "Stand": Präteritum des Verbs, oder Nomen?
- ▶ **Problem 2: Was genau ist eine Grundform?**
  - ▶ Grundform von Reflexivpronomen, von Artikeln?

▶ 38

## Korpusverarbeitung



▶ 39

## Annotation: Korrektheit

- ▶ Wichtigstes Kriterium: Korrektheit
  - ▶ Falsche Annotation führt zu falschen Modellen
  - ▶ Manuelle Annotation
- ▶ Selbst manuelle Annotation ist nie fehlerfrei
  - ▶ Grund 1: Unaufmerksamkeit der Annotatoren
  - ▶ Grund 2: Schwierigkeit der Aufgabe
- ▶ Nötig:
  - ▶ Überprüfung der Korrektheit
  - ▶ Entscheidung über Granularität

▶ 40

## Annotation: Qualitätssicherung

---

- ▶ Annotation muß über die Zeit gleich bleiben (hohes **Intra-Annotator Agreement**)
  - ▶ Denselben Annotator **mehrmals** annotieren lassen (in zeitlichem Abstand)
- ▶ Mehrere Annotatoren müssen gleich annotieren (hohes **Inter-Annotator Agreement**)
  - ▶ Mehrere **unabhängige** Annotatoren annotieren dasselbe

---

▶ 41

## Annotationsschema

---

- ▶ Definiert die Kategorien, die annotiert werden
- ▶ Definiert die Bedingungen, unter denen jede Kategorie annotiert wird
  - ▶ Richtlinien

---

▶ 42

## Annotationsschema: Nomen-Wortarten

### ▶ Penn Tagset (45 Kategorien)

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

▶ 43

## Annotationsschema: Nomen-Wortarten

### ▶ CLAWS2-Tagset (132 Kategorien)

- ▶ NDI – singular noun of direction (north, southeast)
- ▶ NN / NNI / NN2 – common noun, neutral / sg / pl (cod / book / books)
- ▶ NNI\$ -- genitive singular common noun (domini)
- ▶ NNJ / NNJ1 / NNJ2 – organization noun (department / assembly)
- ▶ NNL / NNLI />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)

▶ 44

## Annotationsschemata

---

### ▶ Wie detailliert soll die Annotation sein?

- ▶ Detaillierte Annotation
  - ▶ Viele Kategorien, viel Information
  - ▶ Viele Zweifelsfälle (schwer, Qualität zu halten)
  
- ▶ Grobe Annotation
  - ▶ Wenige Kategorien, wenig Information
  - ▶ Einfacher, Qualität zu halten

---

▶ 45

## Zweifelsfälle

---

- ▶ Annotationsschema muß Richtlinien für Zweifelsfälle beinhalten
  - ▶ Aber: oft bleiben systematische Zweifelsfälle
- ▶ Produktive Verwendung von Sprache kann Theorien sprengen
- ▶ Problem für tiefere Ebenen (Semantik!): häufige **Vagheit/ Ambiguität**

Zwiebel (1): Zwiebelpflanze  
 Zwiebel (2): Frucht der Zwiebelpflanze

- ▶ Was ist „Ich habe eine Zwiebel gepflanzt“?

---

▶ 46

## Annotation: Aufwand

---

- ▶ Annotationsaufwand für ein Wort: 30 Sekunden
- ▶ 1M Worte: 500 000 Minuten = 5 Jahre
- ▶ plus Aufwand fuer Qualitaetssicherung
  
- ▶ Beschleunigung: Annotatoren unterstützen
  - ▶ (Semi)-Automatisierung und manuelle Überprüfung
  - ▶ Zweischneidiges Schwert: Kann zu **systematischen Fehlern** führen

---

▶ 47

## Parallele Korpora

---

Peter liest nun ein Buch über Umweltverschmutzung.  
 Peter is reading a book about pollution.

- ▶ Parallele Sätze in zwei (oder mehr) Sprachen
  - ▶ Wenig verfügbar; viel Politik
    - ▶ Canadian Hansard (E/F)
    - ▶ Proceedings of the European Parliament (9 Sprachen)
    - ▶ UN-Material (E/F/S)
  
- ▶ Sehr interessant
  - ▶ Test, ob linguistische Theorien sprachübergreifend gelten
  - ▶ Training von Systemen zur maschinellen Übersetzung
  - ▶ Möglichkeit für **sprachübergreifende Modelle**

---

▶ 48

## “Stand der Kunst” im Lernen aus Korpora

---

- ▶ **Vorverarbeitung:**
  - ▶ für sehr viele Sprachen verfügbar
  - ▶ sehr hohe Korrektheit (>95%)
- ▶ **Syntaktische Analyse**
  - ▶ für viele Sprachen verfügbar
  - ▶ ausreichend gut (75-85%)
- ▶ **Semantische Analyse**
  - ▶ Teilweise lernbar, für einzelne Sprachen verfügbar
  - ▶ Noch nicht sehr gut (60-75%)