

# Korpora und Annotation

---

Sebastian Pado

20.01.2004

---

# Übersicht

- Zwei Arten von Linguistik
- Arten von Korpora
- Merkmale von Korpora
- Erstellung von Korpora (Annotation)
- Semantische Annotation in SALSA

---

# 1. 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.

---

# 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.

Korpus = Sammlung von von Menschen produziertem Text

---

---

# Charakterisierung des Unterschieds

- Theoretische Linguistik
  - Rationalismus
  - Modelle durch Nachdenken (Introspektion)
  - Beispiele:
    - Grammatiken schreiben
    - Welche Eigenschaften soll eine Grammatik im Allgemeinen haben?
- Korpuslinguistik
  - Empirismus
  - Modelle durch Sichtung von Beispielen
  - Beispiele:
    - Grammatiken automatisch aus Daten ableiten (lernen)
    - Neue Wörter (Neologismen) entdecken

---

# Wie verhalten sich die beiden?

- Armchair linguist: „Why are your results relevant?“
- Corpus linguist: „Why are your results true?“
  
- Computerlinguistische Praxis: Kombination von beidem
  - Korpusarbeit: Datensichtung
  - Introspektion: Entwicklung von Theorien über die Daten
  
- Aber: verschiedene Schwerpunkte ergeben (tendenziell) verschiedene Modelle
  - Korpuslinguistik: viel statistische Modellierung
  - TL: symbolische Repräsentationen (Grammatiken)

Vor/Nachteile?

---

---

# Ein bißchen Geschichte

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

---

# Beispiel: Grammatiken

- Frühe Korpuslinguistik
  - Handkodierte Regeln
  - Handkodierte Anwendung
- Theoretische Linguistik
  - Grammatiktheorien
  - Linguistisch motivierte Grammatiken
  - wenig Interesse an Anwendung
- Jetzt:
  - Komplett statistische Modelle einfacher Grammatiken
  - Anreicherung komplexer Grammatiken mit statistischer Information



---

## 2. Arten von Korpora

- Allgemeine sprachwissenschaftliche Definition:
  - Korpora sind „endliche Sammlungen von konkreten sprachlichen Äußerungen, die als Grundlage für sprachwissenschaftliche Untersuchungen dienen“ (Lexikon der Sprachwissenschaft)
- CoLi-Korpora enthalten typischerweise Annotation für bestimmte sprachliche Ebenen

# Einheiten der Annotation

## **Wortbasierte Korpora**

„Roher“ Text (nur Worte)

Wortarten (POS Tags)

Syntax (flach oder tief)

Semantik  
(Semantische Rollen)

Diskurs (Diskursrelationen)

## **Zeitbasierte Korpora**

Phonetische Korpora

## **Parallele Korpora**

---

# Anwendungen für Korpora

- Lexikographie
  - Was sind die Bedeutungen eines Wortes?
  - Erstellung und Erweiterung von Wörterbüchern (z.B. mit Neologismen, Idiomen, etc.)
- Sichten von Daten für alle linguistischen Zwecke
- Korpora dienen als Trainingsdaten für statistische Modelle („Gold-Standard“) für alle Bereiche des NLP
  - Statistische Modelle lernen Regelmäßigkeiten aus Korpus (Klassifikation von Instanzen)
  - Möglichst große Korpora nötig

---

# Phonetik-Korpora

- Training von Spracherkennungs-Systemen
    - Instanzen: Phone(me)
    - Klassen: Buchstaben
  
  - Training von Text-to-Speech-Systemen
    - Instanzen: Buchstaben
    - Klassen: Phone(me)
  
  - Standardkorpora: v.a. amerikanisches Englisch
    - Auskunftssysteme
      - ATIS: Air Travel Information Service
    - Telefonkonversation
      - Switchboard (>2000 Telefondialoge à 6 Min. = 1.5M Worte)
-

---

# Wortarten-Korpora

- Training von Wortartenbestimmern (POS-Taggern)
  - Instanzen: Wörter (im Kontext)
  - Klassen: Wortarten in Form eines Tagsets
    - Deutsch: STTS-Tagset (54 Tags)
    - Englisch: Penn Tagset (45 Tags), CLAWS2 tag set (132 Tags)
  
- Standardkorpora:
  - Englisch: British National Corpus (BNC), 100M Worte, gemischte Texte inkl. gesprochene Sprache
  - Alle Korpora mit syntaktischer Annotation sind auch mit Wortarten annotiert

---

# Syntax-Korpora

- Training von stochastischen Parsern:
  - Instanzen: Sätze (bzw. Satzteile)
  - Klassen: Parsebäume (bzw. Teilbäume)
- Es werden v.a. kontextfreie Grammatiken gelernt
- 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)

---

# Semantik-Korpora

- Training von semantischen Parseern
  - Instanzen: Satzteile
  - Klassen: semantische Rollen
- Korpora:
  - Englisch: PropBank, auf Grundlage der Penn Treebank: fast fertig
  - Deutsch: SALSA, auf Grundlage von TIGER: in Arbeit

---

# Diskurs-Korpora

- Training von „Diskurs-Parsern“
  - Instanzen: Paare von Sätzen oder Satzteilen
  - Klassen: Diskursrelationen
- Korpora:
  - DiscourseBank, auf Grundlage der Penn Treebank: kürzlich begonnen



---

# Parallele Korpora

- Paralleles Korpus: Gleiche (semantische) Information in zwei (oder mehr) Sprachen
  - Wenig verfügbar; noch weniger mit guter Annotation; viel Politik
    - Canadian Hansard (E/F)
    - Proceedings of the European Parliament (12 Sprachen)
    - UN-Material (E/F/S)
- Training von Systemen zu maschinellen Übersetzung
- (Fast) alle NLP-Anwendungen sind sprachspezifisch
  - Statistische Modelle können z.T. mithilfe eines parallelen Korpus auf eine andere Sprache übertragen werden
  - Voraussetzungen: Alinierung
    - Welcher/s Satz/Wort in einer Sprache entspricht welchem Satz/Wort in den anderen Sprache?

---

## 3. Merkmale von Korpora

- Größe des Korpus
- Rauschen (Noise)
- Tokenisierung
- Markup
- Repräsentativität
- Verfügbarkeit

---

# Wie groß ist ein großes Korpus?

- Faustregel 1: Ein Korpus kann nie groß genug sein
  - Zipf-Verteilung: Auf jeder sprachlichen Ebene gibt es wenige, sehr häufige, und sehr viele seltene Elemente
    - Wortebene: 1 Mio. Wörter des Deutschen enthalten weniger als 50% aller Lemmata des Deutschen
    - Syntax: NE V NP PP (Iris stört die Rentnerin mit Musik) ist häufig
      - ...daß NE NP PP V (daß Iris die Rentnerin mit Musik stört) kommt nicht in NEGRA vor
- Faustregel 2: Je „tiefer“ die Annotation, desto kleiner das Korpus
  - Rohes Text: mehrere G verfügbar
  - POS-Tags: BNC (100M), auch mehr
  - Syntax: 1-10M

Korpora mit automatischer Annotation können auch größer sein,  
sind aber weniger verlässlich!

---

---

# Rauschen (Noise)

- Manuelle Annotation ist relativ sicher, aber nie 100%
  - Es gibt immer Fehler in der Annotation
- Korpora mit automatischer Annotation sind i.A. deutlich unsauberer [dafür größer]
  - Manche Methoden leiden mehr unter Datenmangel, andere unter Fehlern in der Annotation
- Alle Ergebnisse aus Korpusstudien sollten verifiziert werden (am besten auf anderen Korpora)

---

# Tokenisierung

- Am Anfang ist jedes Korpus roher Text
  - Wort- und Satzgrenzen müssen (möglichst automatisch) erkannt werden
- 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

---

# Die Markup-Sprache XML

- XML beschreibt die Struktur von Texten
  - Verallgemeinerung von HTML
- `<p><s>This is a sentence.</s>`  
`<s> This is another sentence in the`  
`same paragraph.</s></p>`

`<p><s>And this is another paragraph.`  
`</s></p>`

- XML erlaubt beliebige Tags, solange alle in der richtigen Schachtelung angeordnet werden (Baumstruktur).

---

# Representativität

- Ein Korpus sollte im Idealfall repräsentativ (balanciert) sein:
  - Alle Genre
  - Alle Sprachebenen
  - Alle Gegenstandsbereiche (Domänen)
  
- Balancierte Korpora: BNC, Brown (altes Korpus)
  - Zeitungskorpora sind nicht balanciert (TIGER, Penn Treebank)

---

# Verfügbarkeit

- Benutzte Korpora sollten (zumindest prinzipiell) allgemein verfügbar sein
  - Replizierbarkeit der Ergebnisse
  - Oft Probleme wegen Copyright (z.B. Zeitungen)
    - Aktuelle Debatte über Urheberrecht



---

## 4. Herstellung von Korpora: Aufwand

- Annotation ist sehr aufwendiger Prozeß
  - Annotation eines Wortes: 30 Sekunden
  - Annotation von 1M Worten: 500 000 Minuten = 5 Personenarbeitsjahre
  
- Beschleunigung: Annotatoren unterstützen
  - (Semi)-Automatisierung und manuelle Überprüfung
    - Abhängig von der Schwierigkeit der Aufgabe
  - Grafische Oberfläche
    - Drag'n'drop viel schneller als Schreiben

---

# Herstellung von Korpora: Qualität

- 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 dasselbe annotieren lassen

---

# Detailliertheit der Annotation: Nominal-POS

- Penn Tagset (45 Tags)
  - NN – noun, singular
  - NNS – noun, plural
  - NNP – proper noun, singular
  - NNPS – proper noun, plural

---

# Detailliertheit der Annotation: Nominal-POS

- CLAWS2-Tagset (132 Tags)
  - 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 (ls. / 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)

---

# Herstellung von Korpora: Annotation

- Wie detailliert soll die Annotation sein?
  - Detaillierte Annotation
    - Viel Information
    - Viele Zweifelsfälle (schwer, Qualität zu halten)
  - Grobe Annotation
    - Wenig Information
    - Einfacher, Qualität zu halten
- Kontrollierter Ablauf der Annotation nötig
  - Annotationsrichtlinien: Wann annotiere ich was?
  - Problemfälle: Was passiert, wenn ich mir nicht sicher bin?

---

## 5. SALSA - Saarbrücken Lexical Semantics Annotation and Analysis

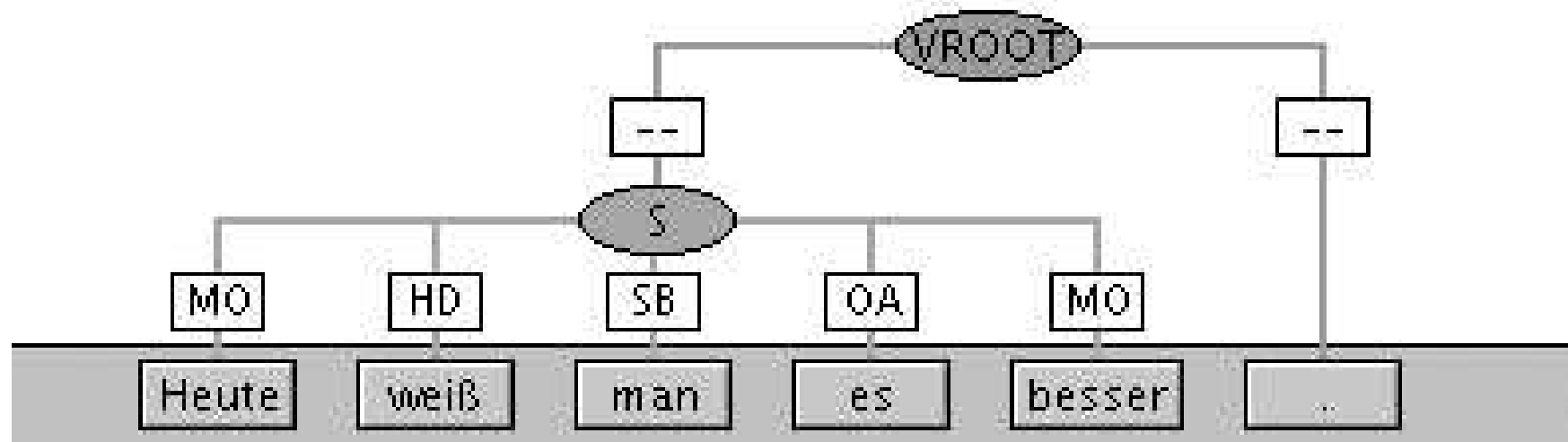
- Grundlage: TIGER-Korpus (80.000 Sätze) mit syntaktische Annotation
- Manuelle Annotation von TIGER mit semantischen Rollen
  - Grundlage: FrameNet
  - Repräsentation in XML
- Automatische Erweiterung auf größeres Korpus

---

# Semantische Rollen und FrameNet

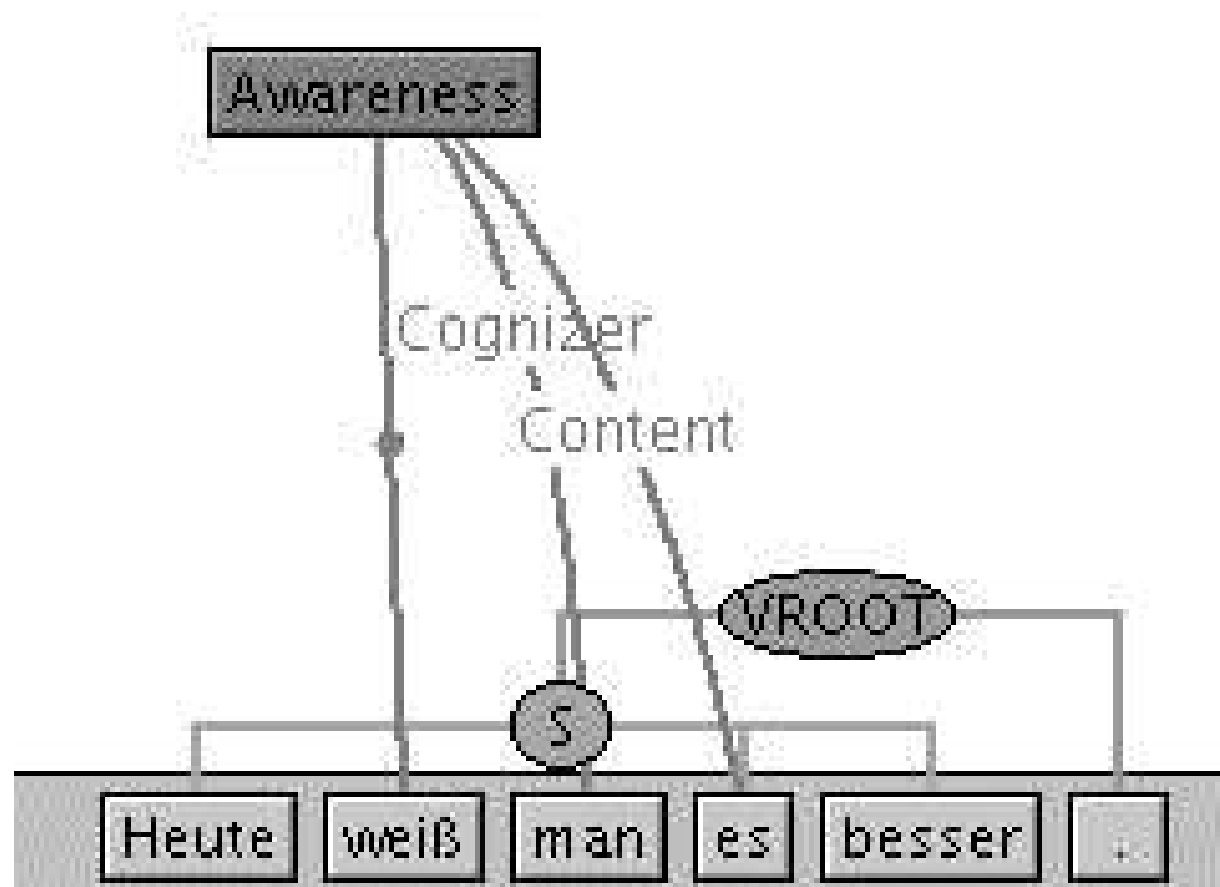
- Semantische Rollen: Who does what to whom
  - kaufen (v.): Käufer, Verkäufer, Waren, Geld
  - Wichtig zum Verständnis der Situation
    - Rollen kodieren Wissen über Vor- und Nachzustände
      - Käufer hat vorher Geld, hinterher Waren
      - Verkäufer hat vorher Waren, hinterher Geld
- FrameNet: Rollen sind spezifisch für prototypische Situationen (Frames)
  - Frame Commercial\_Transaction: The Seller gives Goods to the Buyer and receives Money in return.
  - Deutsche Lemmata von Commercial\_Transaction: kaufen, verkaufen, verscherben, abkaufen, Verkauf, Kauf, Verkäufer, versteigern (?)

# Ein Satz mit syntaktischer Annotation

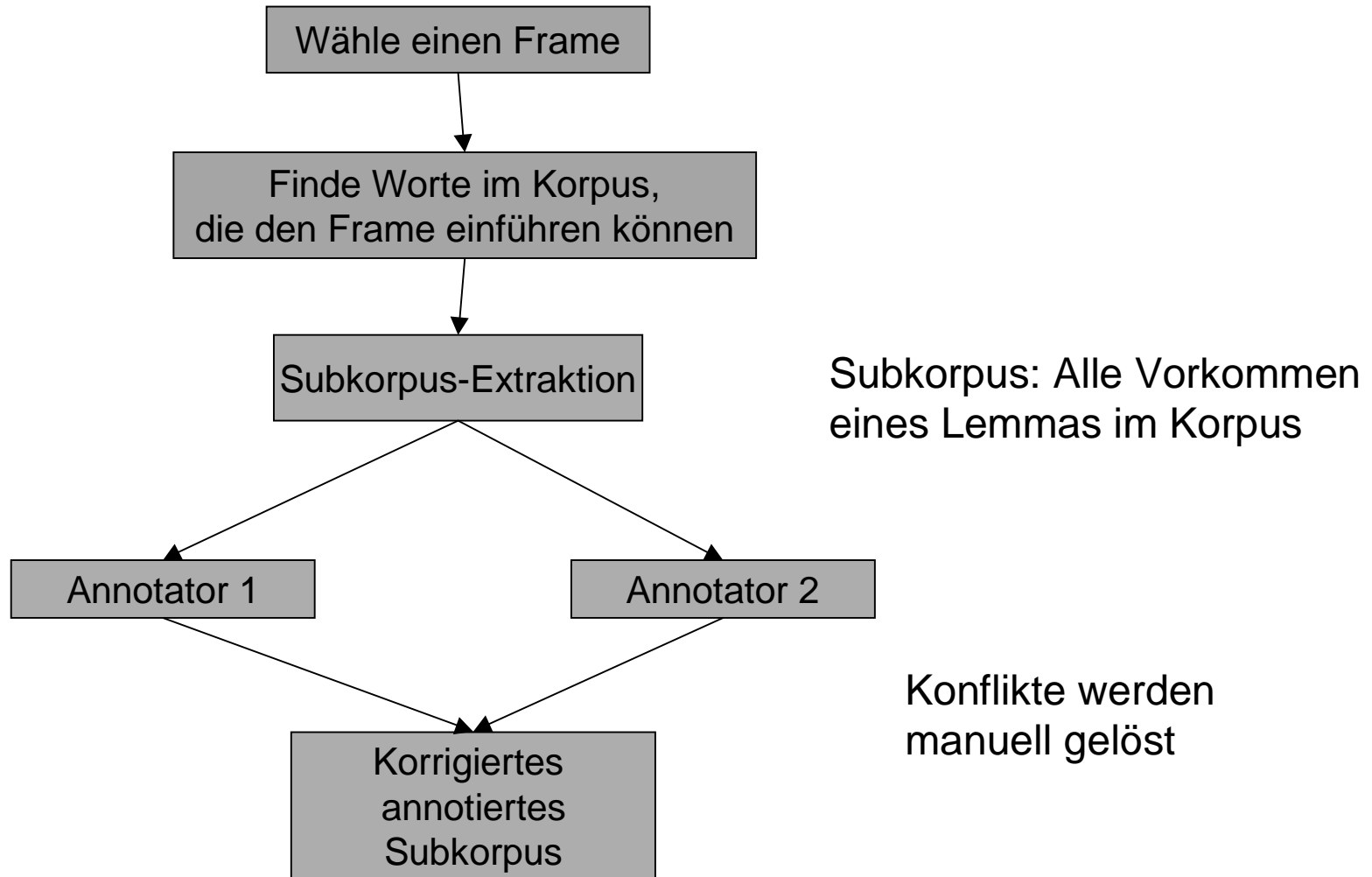




# Ein Satz mit syntaktischer und semantischer Annotation



# Ablauf der Annotation in SALSA



---

# Probleme bei der SALSА-Annotation

- Vagheit / Ambiguität
  - Der Journalist bemerkte, daß die Preise gestiegen waren
  - BECOMING\_AWARE (merken) oder COMMUNICATION (sagen)?
  
- Metonymie
  - Klassische Sprachfigur: „Ersetze Konzept durch verwandtes Konzept“
    - „Ich stehe da hinten“ -> Person für Auto
  - REQUEST (Forderung): Speaker (Sprecher) vs. Medium (Übermittlung)
    - Der Sprecher fordert                      - Speaker
    - Der Parteitag fordert                      - Speaker (?)
    - Die Bildzeitung fordert                      - Speaker oder Medium
    - Der Antrag fordert                      - Speaker oder Medium
    - Im Radio wird gefordert                      - Medium

Semantische Rollen sind schwer abzugrenzen

---

# Probleme bei der SALSA-Annotation (II)

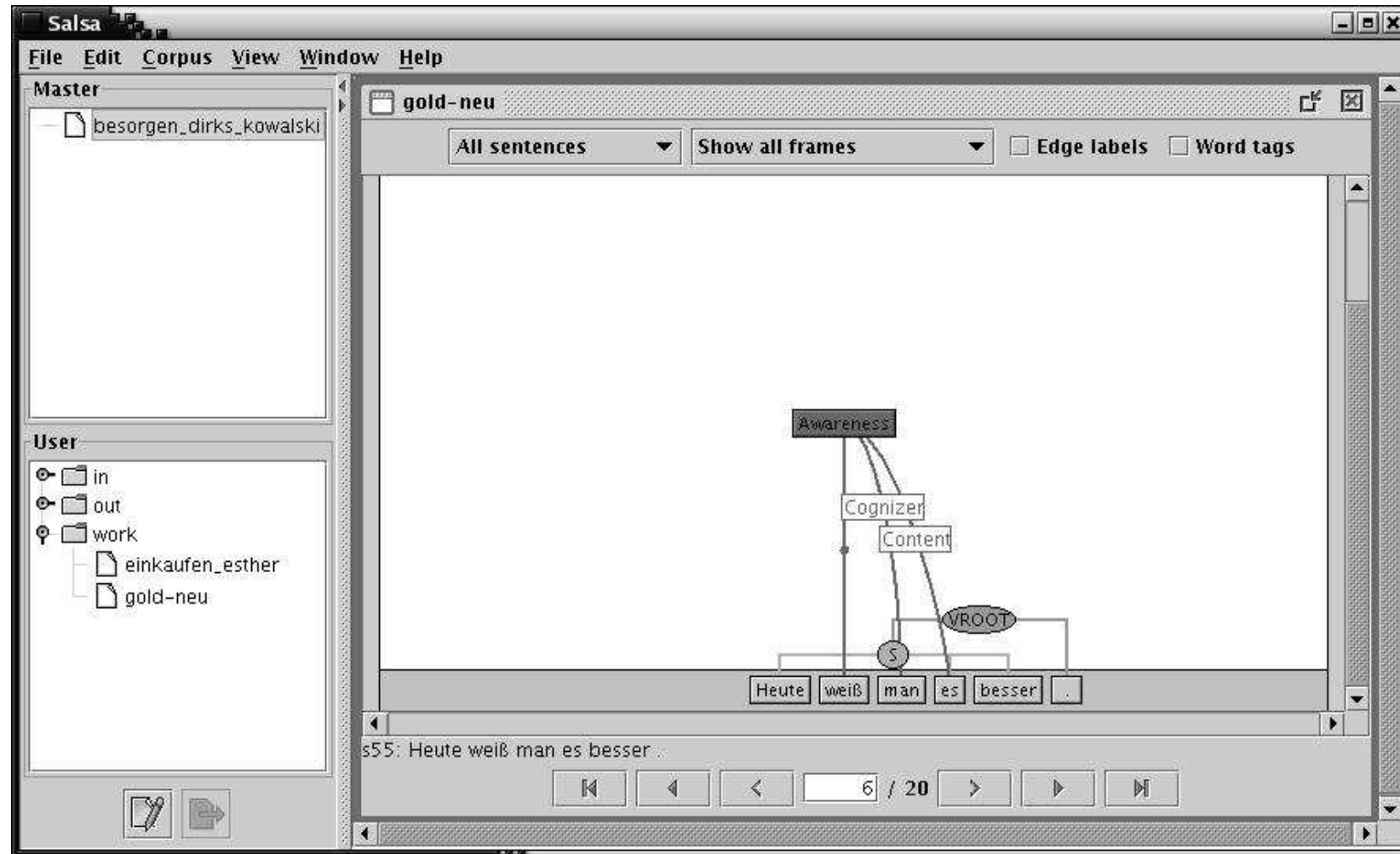
- Metaphern:  
„Rao erklomm den Premierministersessel“
- Wörtliche Bedeutung: MOTION
  - Rao ist Mover
  - Premierministersessel ist Place
    - „Rao hat sich auf eine bestimmte Art Stuhl gesetzt“
- Verstandene Bedeutung: LEADERSHIP
  - Rao ist Leader
  - Premierminister(sessel) ist Role
    - „Rao hat eine Führungsposition (als Premierminister) angetreten“

---

# Aufwand der Annotation

- Anfang: Annotation mit XML-Editor
  - Frames und Frame-Elemente von Hand einfügen
  - 5 Min pro Frame
  - Hochrechnung Aufwand: Siehe Folie von vorhin...
- Beschleunigung in zwei Schritten:
  - Grafische Annotationsoberfläche: Faktor 5
  - Semi-automatische Vorannotation: wird erforscht

# Grafische Annotationsoberfläche



---

# Semi-automatische Annotation

- Problem: Systematischer Fehler
  - Automatische Vorschläge enthalten immer Fehler
  - Vermutlich systematisch fehlerhaft
    - Verfälschung der Annotation
    - Inter-Annotator Agreement kann trotzdem hoch sein...