

Korpora und Annotation

Sebastian Pado

15.02.2005

Übersicht

- Zwei Arten von Linguistik
 - Anwendungen für Korpora
 - Arten von Korpora
 - Mehr über Korpora
 - Erstellung von Korpora (Annotation)
-

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.

Korpuslinguistik

- Modelle durch Sichtung von Beispielen
 - Erlaubt schrittweise Annäherung an Phänomene
 - Erlaubt maschinelles Lernen
 - Empirismus
- Ergebnis: „Flache“ Modelle
 - Statistische Modelle, Mustererkennung
 - Robustheit zentraler als Korrektheit
- Funktioniert gut für:
 - Grammatiken automatisch aus Daten ableiten (lernen)
 - Neue Wörter (Neologismen) entdecken

Theoretische Linguistik

- Modelle durch Nachdenken (Introspektion)
 - Erlaubt Modellierung „verdeckter“ Phänomene
 - Analyse existierender Modelle
 - Rationalismus
- Ergebnis: komplexe, symbolische Modelle
 - Modelle sind für menschliche Analyse geeignet
 - Interaktion von Symbolen bestimmt Vorhersage
 - Korrektheit zentraler als Robustheit
- Funktioniert gut für:
 - Grammatiken schreiben
 - Meta-Analyse von Grammatiken (Grammatiktheorie)

Gegenseitige Kritik

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

Und die Computerlinguistik?

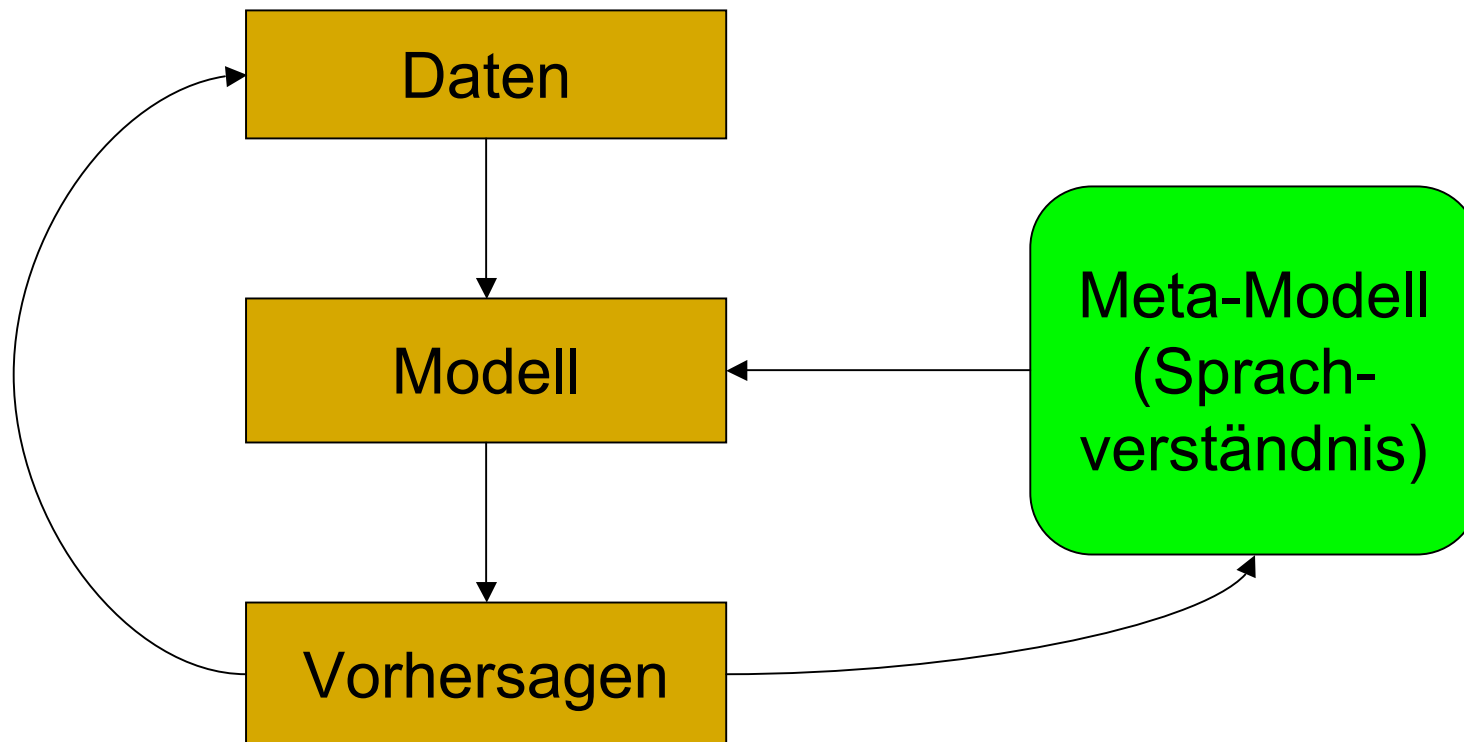
Geschichte der Computerlinguistik

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

Heutiger Methodologie: Kombination

Empirie

Theorie



Beispiel: Grammatiken

- 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

Was ist ein Korpus?

- Allgemeine sprachwissenschaftliche Definition:
 - Ein Korpus (n.!) ist eine „endliche Sammlung von konkreten sprachlichen Äußerungen, die als Grundlage für sprachwissenschaftliche Untersuchungen dienen“ (Lexikon der Sprachwissenschaft)
- Computerlinguistik:
 - Korpus muß maschinell verarbeitbar sein
 - Erfordert typischerweise Vorverarbeitung

Vorverarbeitung

- Wortbasierte Korpora
 - Grundlage: Text(datei)
 - Vorverarbeitung: Tokenisierung
- Zeitbasierte Korpora
 - Grundlage: phonetisches Signal (Aufnahme)
 - Vorverarbeitung: Segmentierung
- Dann: Annotation verschiedener Ebenen

Tokenisierung

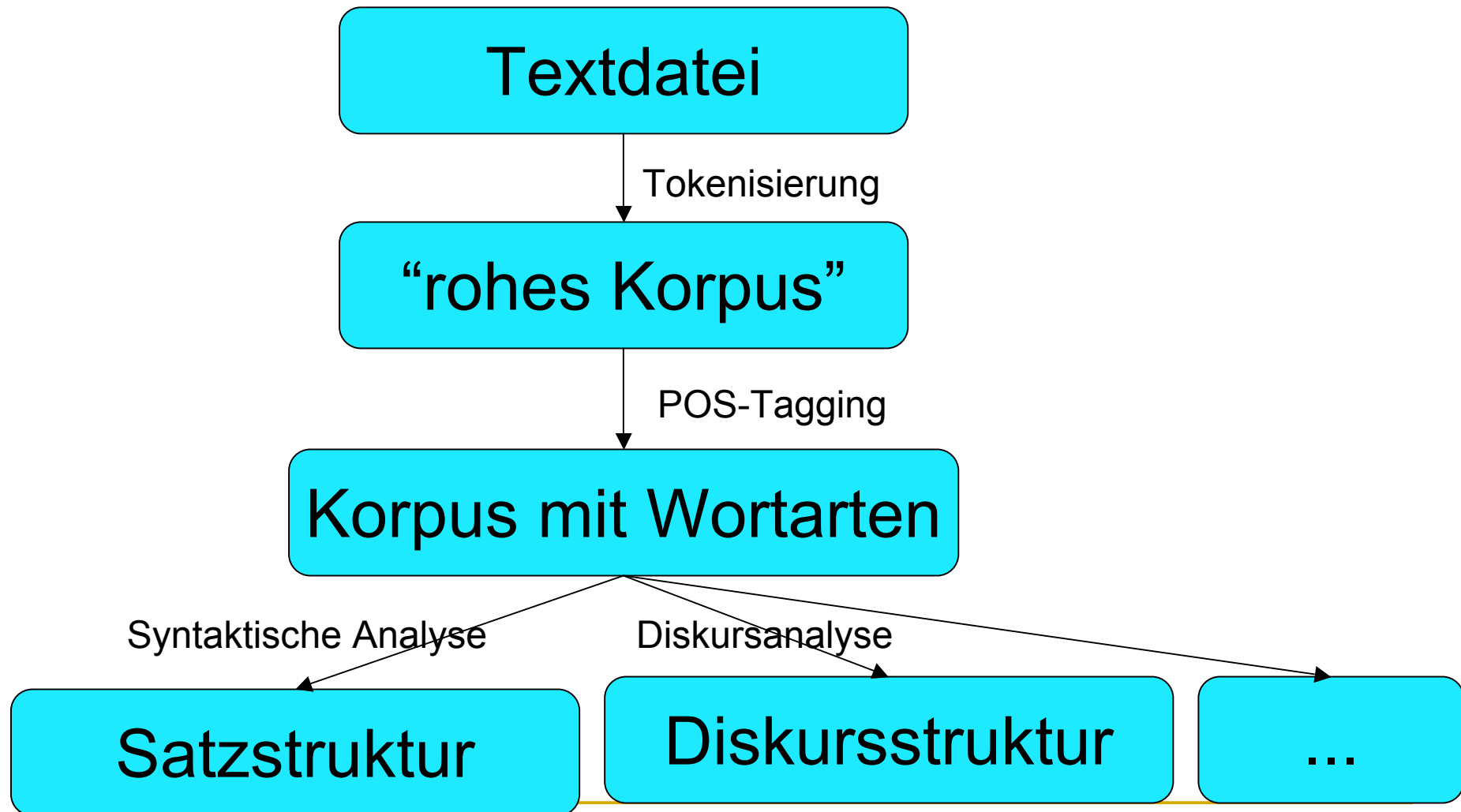
- Problem: Korpus ist Abfolge von Zeichen (Textdatei)
 - Wort- und Satzgrenzen müssen (möglichst automatisch) erkannt werden
- Was ist eine Satzgrenze?
 - Heuristik: Ein Satzzeichen (Punkt, ...)
 - 1., Mr., Std.
- Was ist eine Wortgrenze?
 - Heuristik: Alles, was kein Buchstabe ist
 - Tholey-Theley,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

Grosse Schwierigkeiten bei asiatischen Sprachen

Anwendung von „rohen“ Korpora

- Was sind die Bedeutungen eines Wortes?
 - Beispiele sichten und analysieren
- Erstellen und Erweitern von Wörterbüchern
 - Suche nach Neologismen (Neubildungen)
 - Suche nach Wörtern, deren Häufigkeit über die Zeit stark schwankt
 - Suche nach Kollokationen
 - Ins Gras beißen, sich einen schönen Tag machen, etc.
 - Suche nach Worten, die sehr häufig gemeinsam auftreten

Entstehung wortbasierter Korpora



Wofür überhaupt Annotation?

- Studien “tieferer” linguistischer Ebenen
 - Studium und Häufigkeitsbestimmung bestimmter Konstruktionen, Bedeutungen, etc.
- Bessere Modelle durch mehr Information
 - Ausschliessen falscher Kollokationen (“an das”)
- Wichtig: **Statistische Modellierung**
 - Erinnerung: Grammatik lernen setzt **syntaktisch annotiertes Korpus** voraus!
 - Frequenz von **Regeln im Korpus**

Annotationsebenen

Wortbasierte Korpora

„Rohes“ Korpus (Worte)

Wortarten (POS-Tags)

Syntax (flach, tief)

Semantik (z.B. Rollen)

Diskurs (Diskursrelationen)

Zeitbasierte Korpora

Segmentierung

Phone / Phoneme

Prosodie

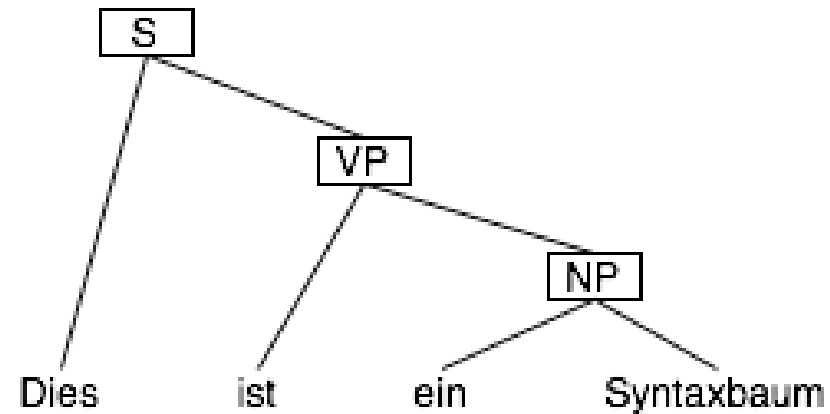
...

Korpora mit Wortarten

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

- Training von Wortartenbestimmern (POS-Taggern)
 - Aufgabe: Ordne jedem Wort eine Wortart zu
 - Wortarten-Inventar: **Tagset**
 - Deutsch: STTS-Tagset (54 Tags)
 - Englisch: Penn Tagset (45 Tags), CLAWS2 tag set (132 Tags)
 - Features: Wort, Worte und Wortarten im Kontext
- Standardkorpora:
 - Englisch: British National Corpus (BNC), 100M Worte
 - Alle Korpora mit syntaktischer Annotation

Syntax-Korpora



- Training von stochastischen Parsern:
 - Aufgabe: ordne jedem Satz eine syntaktische Analyse zu
 - Syntaxbaum-Inventar: abhängig von syntaktischer Theorie
 - Features: Worte, Wortarten

- 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

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

- Training von semantischen Parsern
 - Aufgabe: ordne Satzteilen semantische Rollen zu
 - Rollen-Inventar: z.B. Agent, Patient
 - Features: grammatische Funktion (Subjekt, Objekt, ..), Prädikat, Worte, Wortarten...
- Korpora:
 - Englisch: PropBank, auf Grundlage der Penn Treebank
 - Deutsch: SALSA, auf Grundlage von TIGER (in Arbeit)

Diskurs-Korpora

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

- Training von „Diskurs-Parsern“
 - Ordne Paaren von Sätzen Diskursrelationen zu
 - Diskursrelationen-Inventar: z.B. Begründung (weil), Zweck (damit), Ausführung (und dann)
 - Features: Konjunktionen, Satzbau, etc.

- Korpora:
 - DiscourseBank, auf Grundlage der Penn Treebank

Phonetik-Korpora

- Training von Spracherkennungs-Systemen
 - Ordne einer Schwingung einen Buchstaben zu
 - Features: Formanten
- Training von Text-to-Speech-Systemen
 - Ordne einem Buchstaben eine Schwingung zu
- Standardkorpora: v.a. amerikanisches Englisch
 - Auskunftssysteme
 - ATIS: Air Travel Information Service
 - Telefonkonversation
 - Switchboard (>2000 Telefondialoge à 6 Min. = 1.5M Worte)

Parallele Korpora

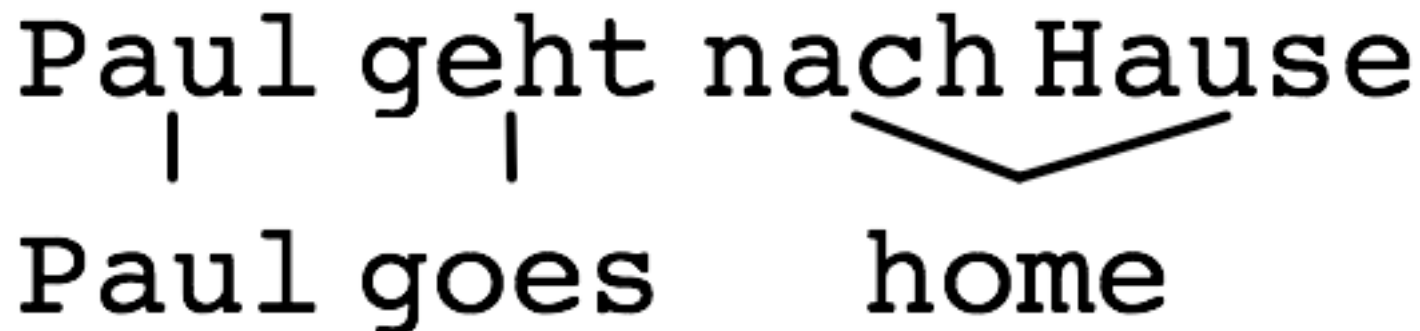
Peter liest nun ein Buch über Umweltverschmutzung.
Peter is reading a book about pollution.

- Parallele Sätze in zwei (oder mehr) Sprachen
 - Wenig verfügbar; viel Politik
 - Canadian Hansard (E/F)
 - Proceedings of the European Parliament (12 Sprachen)
 - UN-Material (E/F/S)
 - Sehr interessant
 - Test, ob linguistische Theorien sprachübergreifend gelten
 - Training von Systemen zur maschinellen Übersetzung
 - Möglichkeit für **sprachübergreifende Modelle**
-

Projektion in parallelen Korpora

- Statistische Modelle können mithilfe eines parallelen Korpus auf eine andere Sprache **projiziert** werden
 - Konzeptuelle Voraussetzung: **Parallelismus** der Strukturen in beiden Sprachen
 - Technische Voraussetzung: **Alinierung**

Paul geht nach Hause
Paul goes home



3. Mehr über Korpora

- Repräsentation
- Repräsentativität
- Rauschen (Noise)
- Größe des Korpus

Repräsentation von Korpora

- Korpora enthalten Annotation verschiedener Ebenen
 - Wie repräsentieren?
- **XML** (Extensible Markup Language)
 - XML besteht aus Elementen und Attributen
 - XML ist eine Verallgemeinerung von HTML
 - Elemente können selbst bestimmt werden
 - XML kann **Bäume** beschreiben
 - Verbreitete Repräsentationssprache für Korpora

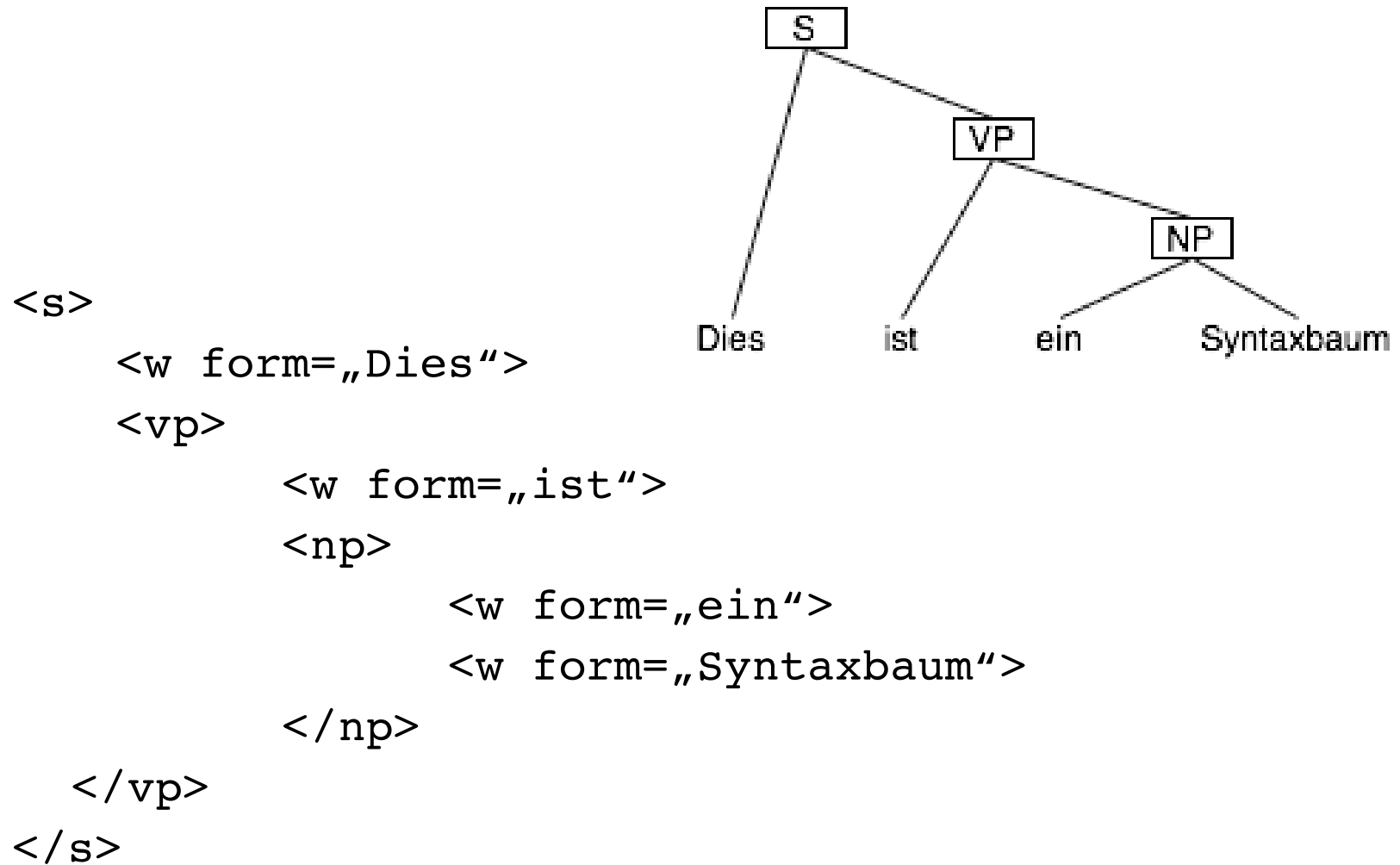
XML-Repräsentation (I)

Dieser Satz ist mit Wortarten annotiert.

ART NN VAFIN PRP NN ADJ

```
<s><w form=„Dieser“ pos=„ART“>
  <w form=„Satz“ pos=„NN“>
    <w form=„ist“ pos=„VAFIN“>
      <w form=„mit“ pos=„PRP“>
        <w form=„Wortarten“ pos=„NN“>
          <w form=„annotiert“ pos=„ADJ“>
</s>
```

XML-Repräsentation (II)



Repräsentativität

- Ein Korpus sollte im Idealfall **repräsentativ (balanciert)** sein:
 - Alle Genres, Sprachebenen, Gegenstandsbereiche (Domänen)
- Zweck: **Generelle** Forschungsergebnisse
- Balancierte Korpora: BNC, Brown-Korpus
 - Zeitungskorpora sind nicht balanciert (TIGER, Penn Treebank)
 - Politische Korpora auch nicht...

Korrektheit in Korpora

- Die Annotation in Korpora sollte **richtig sein**
 - **Manuelle Annotation**
 - Zweck: **korrekte** Forschungsergebnisse
 - Selbst manuelle Annotation ist nie fehlerfrei
 - Grund 1: Probleme der Annotatoren
 - Grund 2: Inhärente Vagheit
 - Beispiel: Wortbedeutungen
 - Zwiebel (1): Zwiebelpflanze
 - Zwiebel (2): Frucht der Zwiebelpflanze
 - Was ist „Ich habe eine Zwiebel gepflanzt“?
 - Sehr aufwendig! (siehe später)
 - Automatische Annotation erst recht nicht korrekt
 - Henne-Ei-Problem
-

Größe von Korpora

- Wie groß **sollte** ein Korpus sein?
 - Lexikographische Arbeit: groß genug, um die interessanten Phänomene beobachten zu können
 - Maschinelles Lernen: **Ein Mehrfaches der Größe des Feature-raumes**
 - Komplexere Modelle benötigen mehr Trainingsdaten
- Wie groß **sind** verfügbare Korpora (Englisch)?
 - Roher Text: mehrere G Wörter verfügbar
 - POS-Tags: BNC (100M Wörter)
 - Syntax/Semantik: 1-10M Wörter

Wenig Daten für tiefere sprachliche Ebenen

„The Web as Corpus“

- **Vorschlag: Nutzen von Internet-Daten**
 - Problem 1: Repräsentativität
 - Bedeutung von „amazon“
 - Problem 2: Korrektheit der Daten
 - Automatische Annotation nötig

- **Empirisches Ergebnis: Nutzen hängt von linguistischer Ebene ab**
 - Flache Analyse: zusätzliche Daten bringen Vorteil trotz Fehlern
 - Tiefe Analyse: Fehler überwiegen Vorteil der zusätzlichen Daten

Annotation: 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
 - Aber: Kann zu **systematischen Fehlern** führen

Annotation: 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 annotieren dasselbe

Echter Aufwand ist noch deutlich höher

Annotation: Annotationsschemata

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

Annotationsschema: Nominal-Wortarten

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

Annotationsschemata: Nominal-Wortarten

- 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)