Korpuslinguistik

Sebastian Pado

15.02.2005

Ein Wort vorweg

- Bitte keine Word-Attachments von Übungsblättern!

Übersicht

- Linguistische Methodologie
  - Korpuslinguistik und theoretische Linguistik
- Korpora
  - Verwendung
  - Annotationsebenen
  - Technisches

Linguistische Methodologie

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

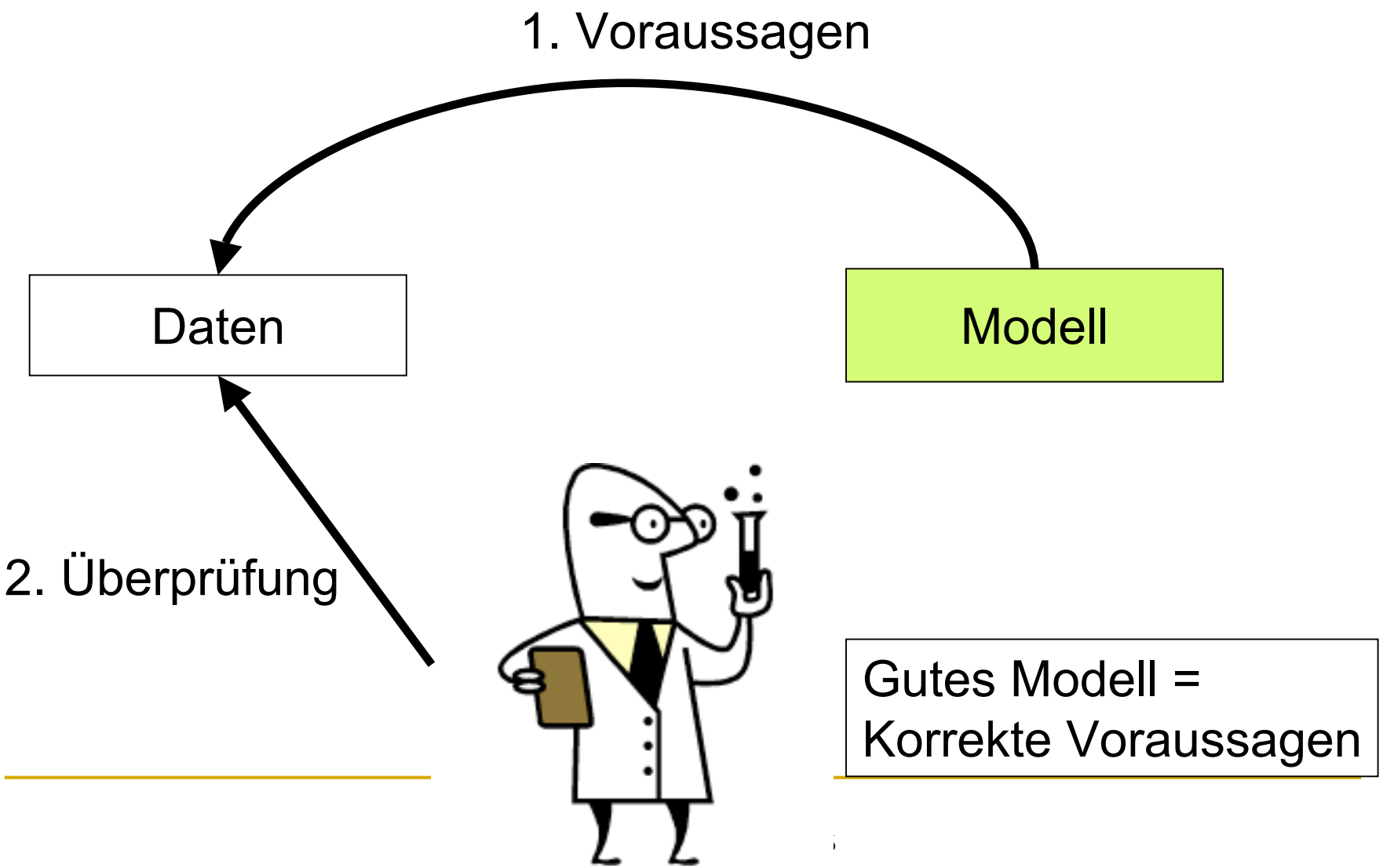
He has a corpus of approximately one zillion running words that contains all his primary facts. His work is deriving secondary facts from primary facts.
At the moment, he is busy determining the relative frequencies of the eleven parts of speech as the first words of a sentence versus as the second word of a sentence.

Wissenschaft = Modellierung

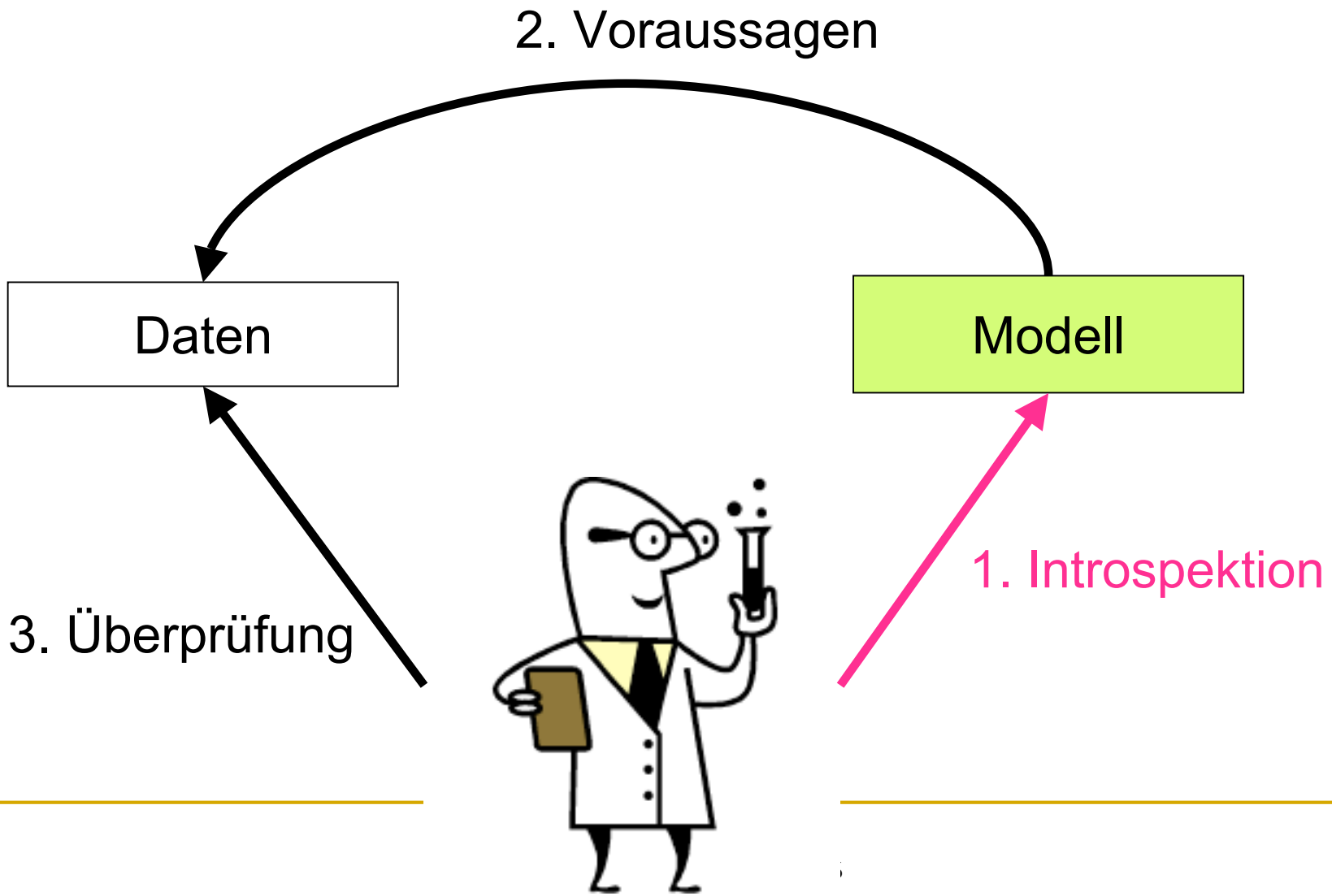
Daten

Modell

Wissenschaft = Modellierung



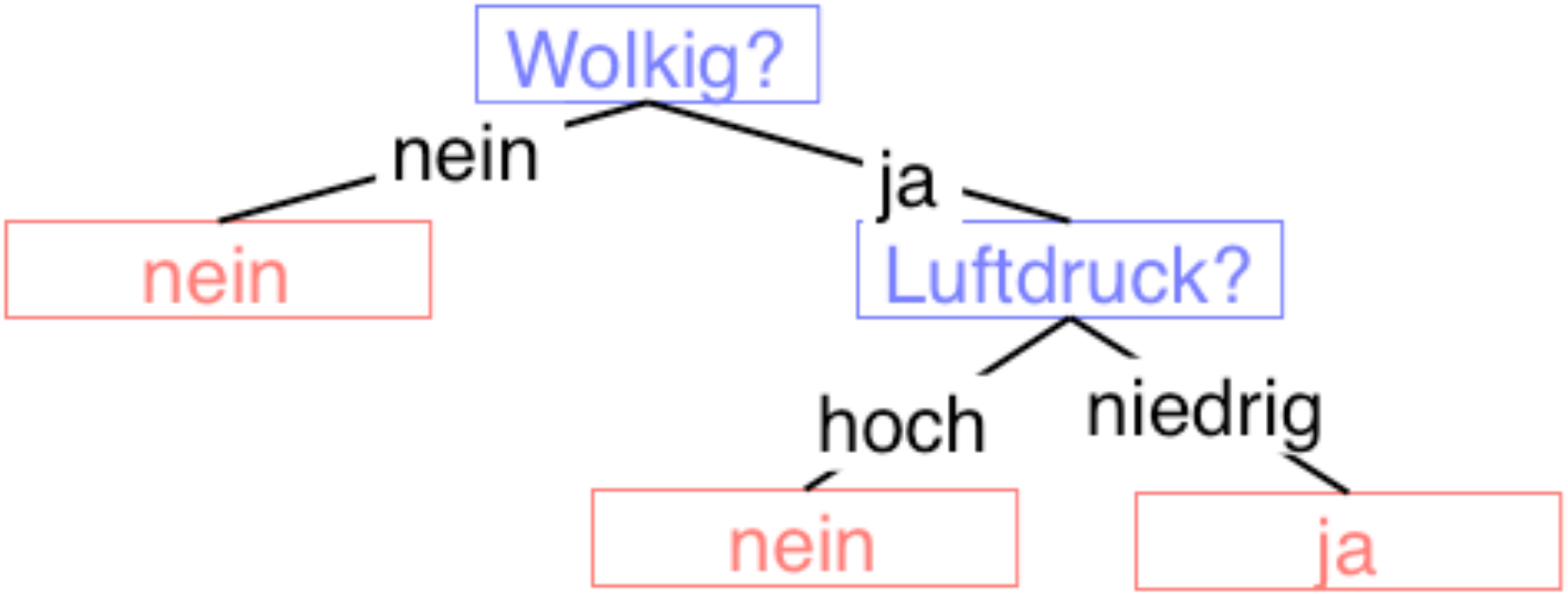
Modellierung durch Introspektion



Rationalismus

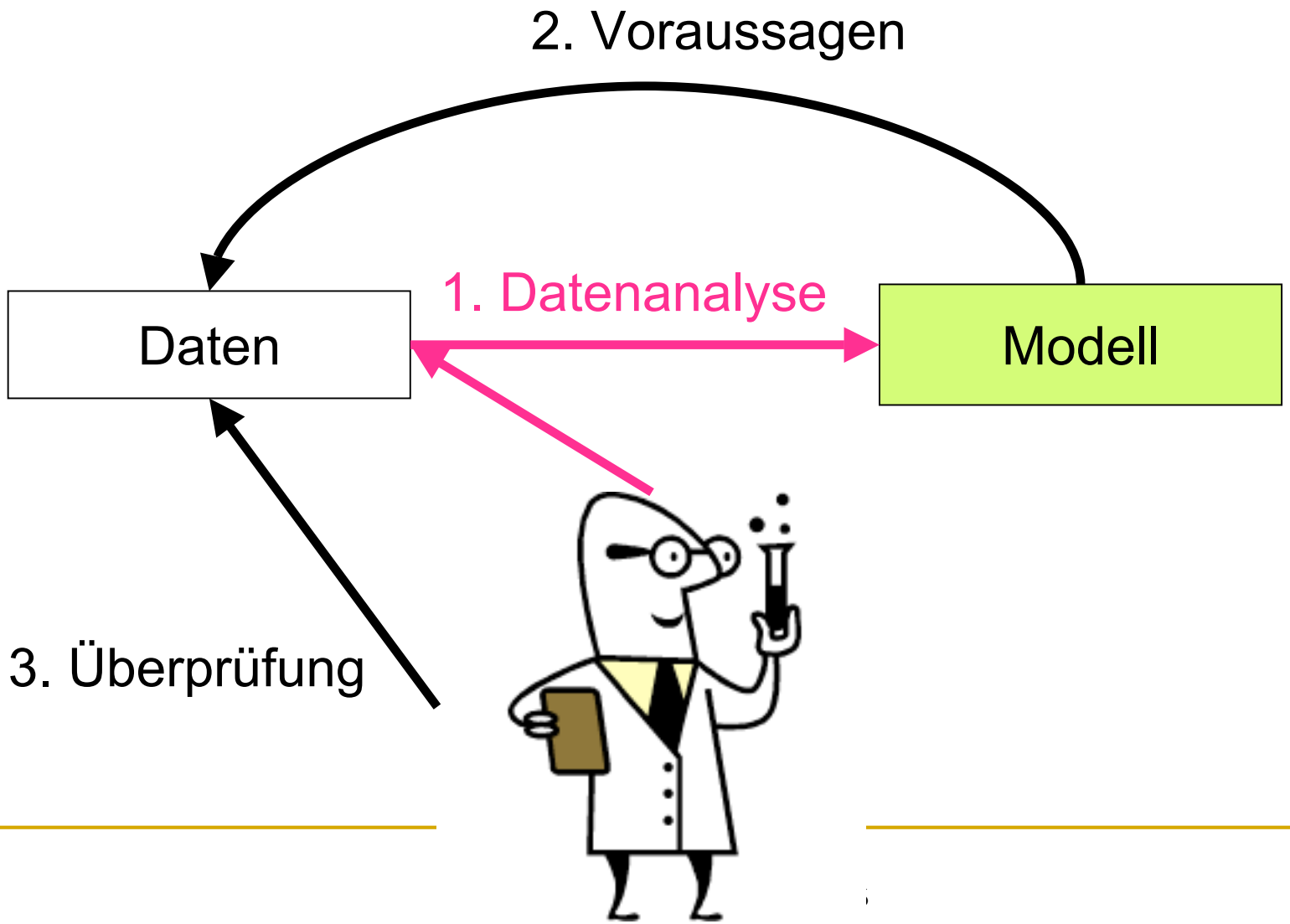
- Modelle durch Nachdenken
  - Typischerweise komplexe, regelbasierte Modelle
- Vorteile
  - Erlaubt Modellierung komplexer Phänomene
    - Rückgriff auf menschliches Verständnis
  - (Meta-)Analyse existierender Modelle
  - Modelle für menschliche Analyse geeignet
- Nachteile
  - Modellbau oft nicht robust (zu spezifisch)
  - Repräsentation durch Regeln oft zu starr
  - **Abhängigkeit von vorgefaßten Meinungen**

Beispiel: symbolisches Modell



- In der Computerlinguistik:
  - Grammatiken
  - Meta-Analyse von Grammatiken (Grammatiktheorie)

Modellierung durch Datenanalyse



Empirismus

- Modelle durch Sichtung von Beispielen
  - Typischerweise statistische, einfache Modelle (Mustererkennung)
- Vorteile
  - Erlaubt Modellierung unbekannter Phänomene
  - Erlaubt maschinelles Lernen
- Nachteile
  - Modelle oft nur approximativ richtig
  - Schwierige Probleme können oft nicht modelliert werden (Integration von externem Wissen schwierig)
  - **Abhängigkeit von den Daten**

Schwierige Probleme

- Konversationeller Kontext
  - Machen wir es so, wie du vorgeschlagen hast.
- Weltwissen
  - Ich bin zu spät, weil mein Tank leer war.

Beispiel: statistisches Wettermodell

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

- In der Computerlinguistik
  - Neologismen entdecken, Texte datieren
  - Automatische syntaktische Analyse

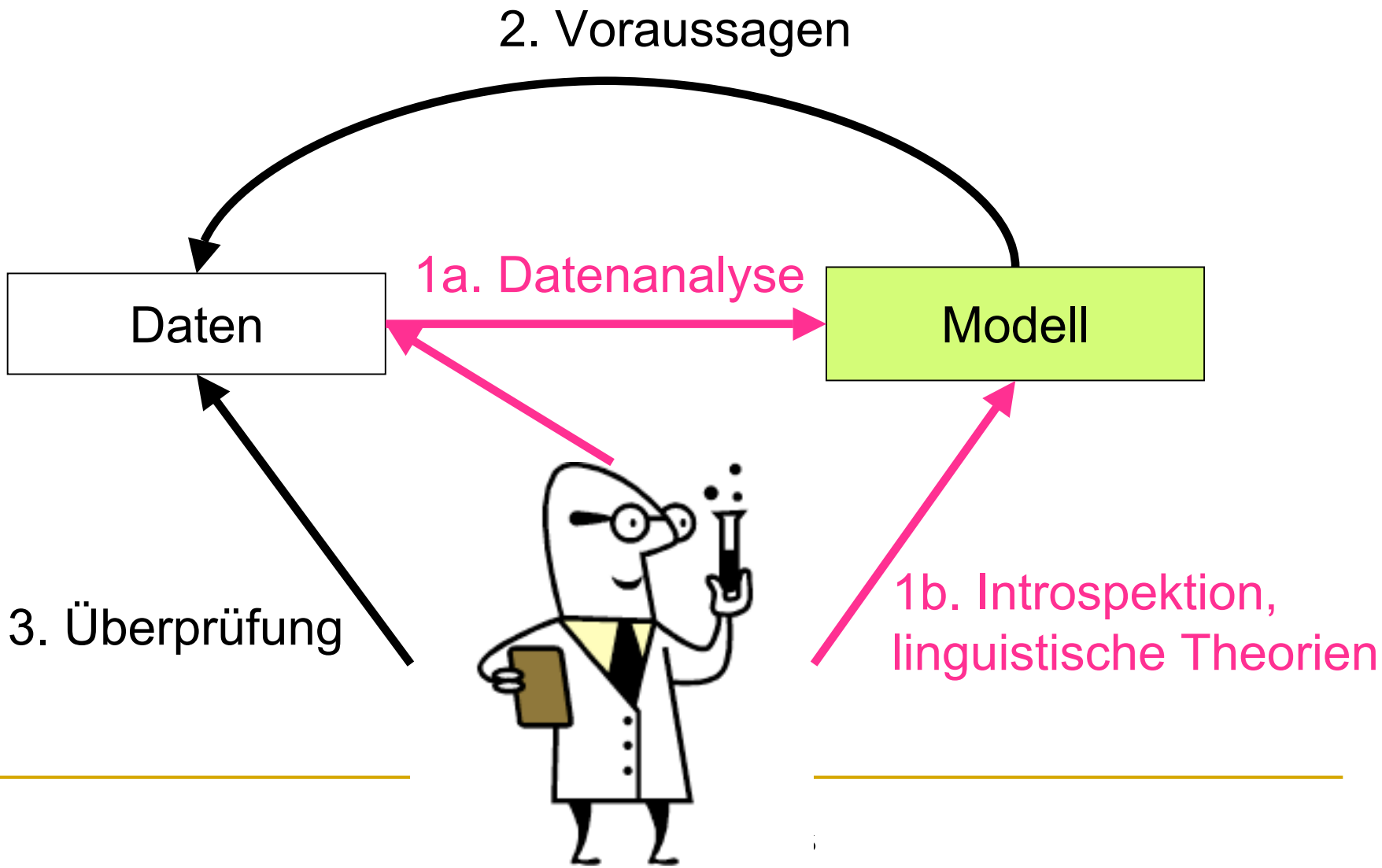
Gegenseitige Kritik

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

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…)

Heutige Methodologie



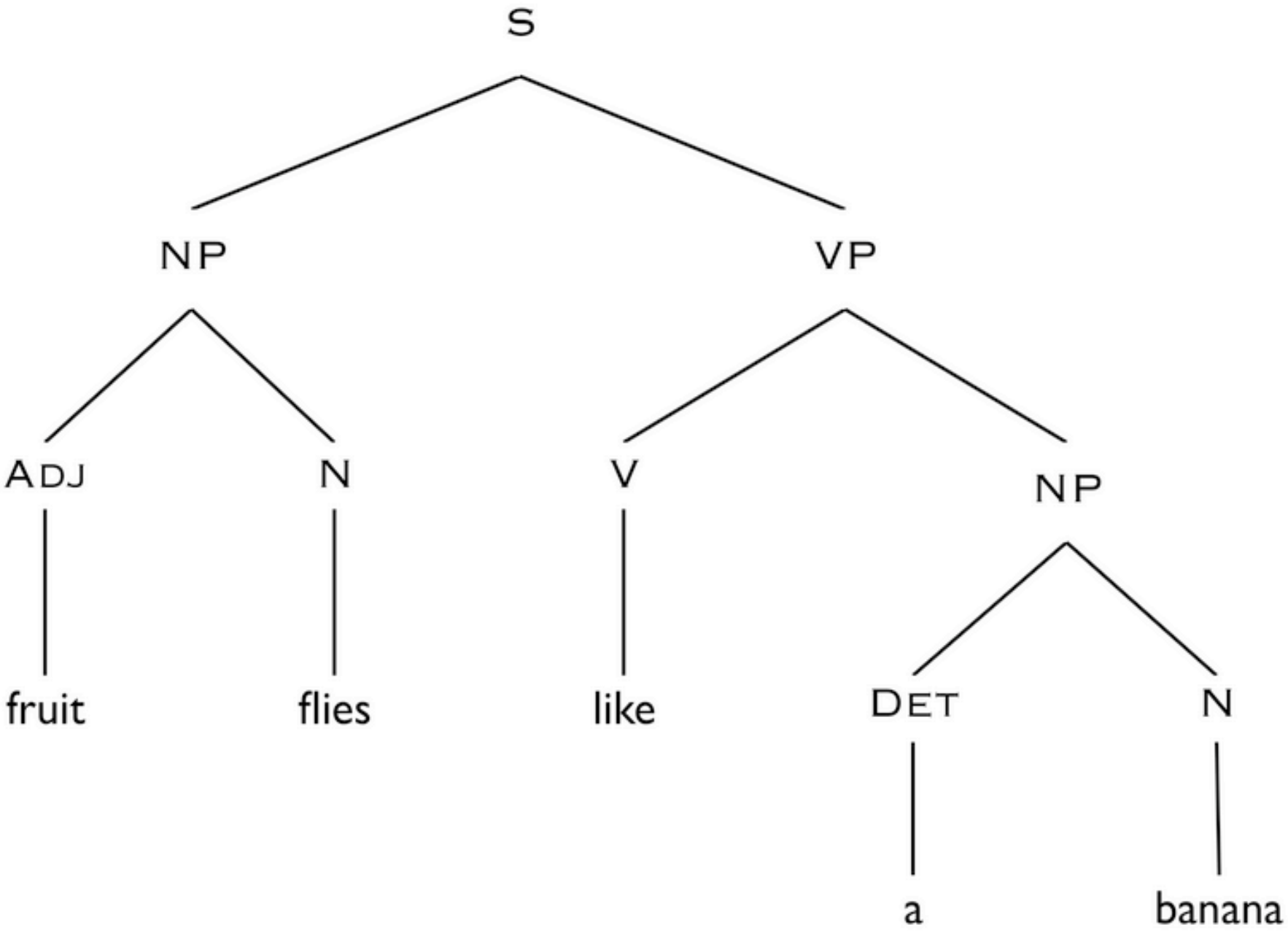
Grammatiken im Wandel der Zeit

- 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

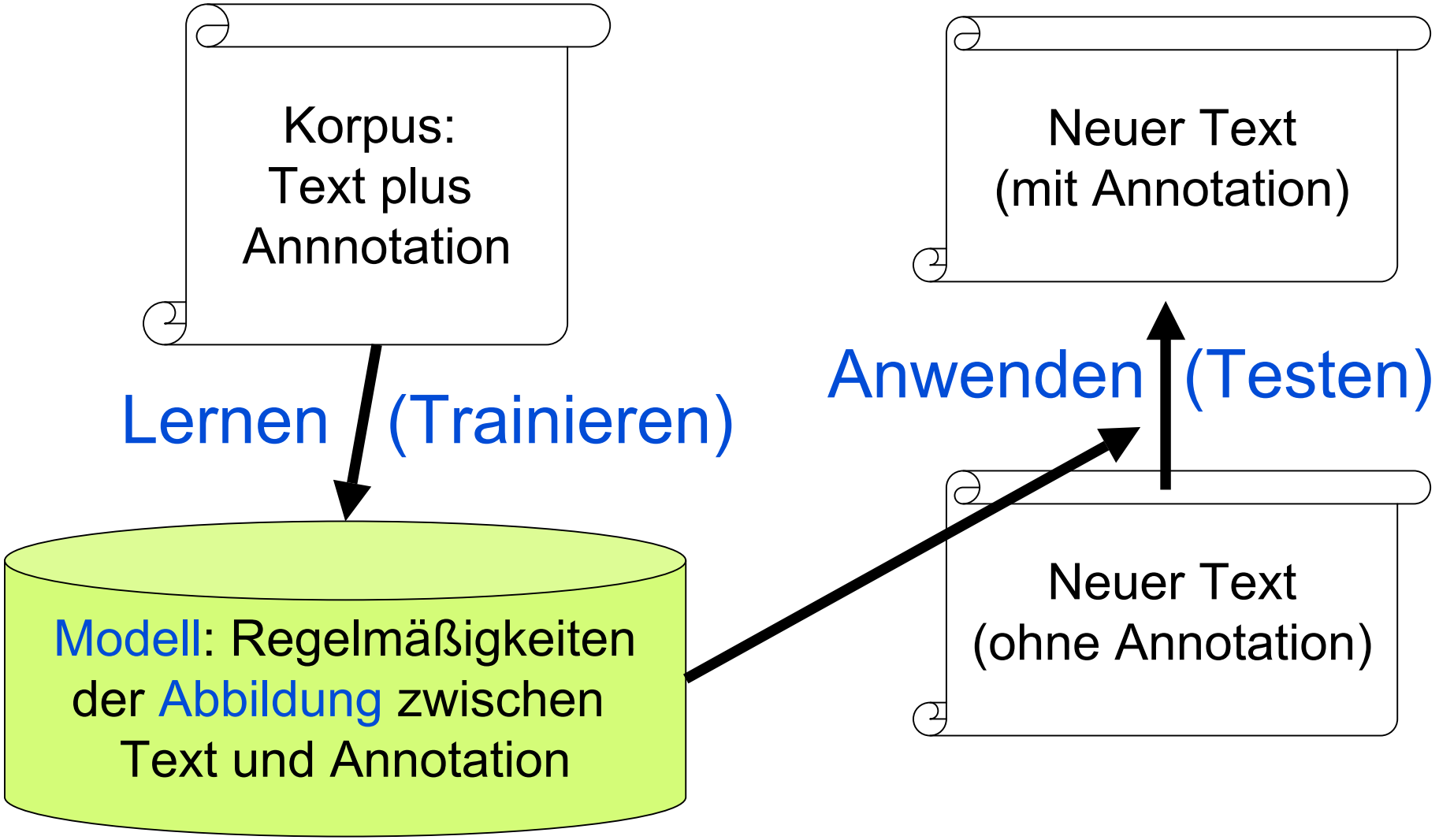
Korpora

Was ist ein Korpus?

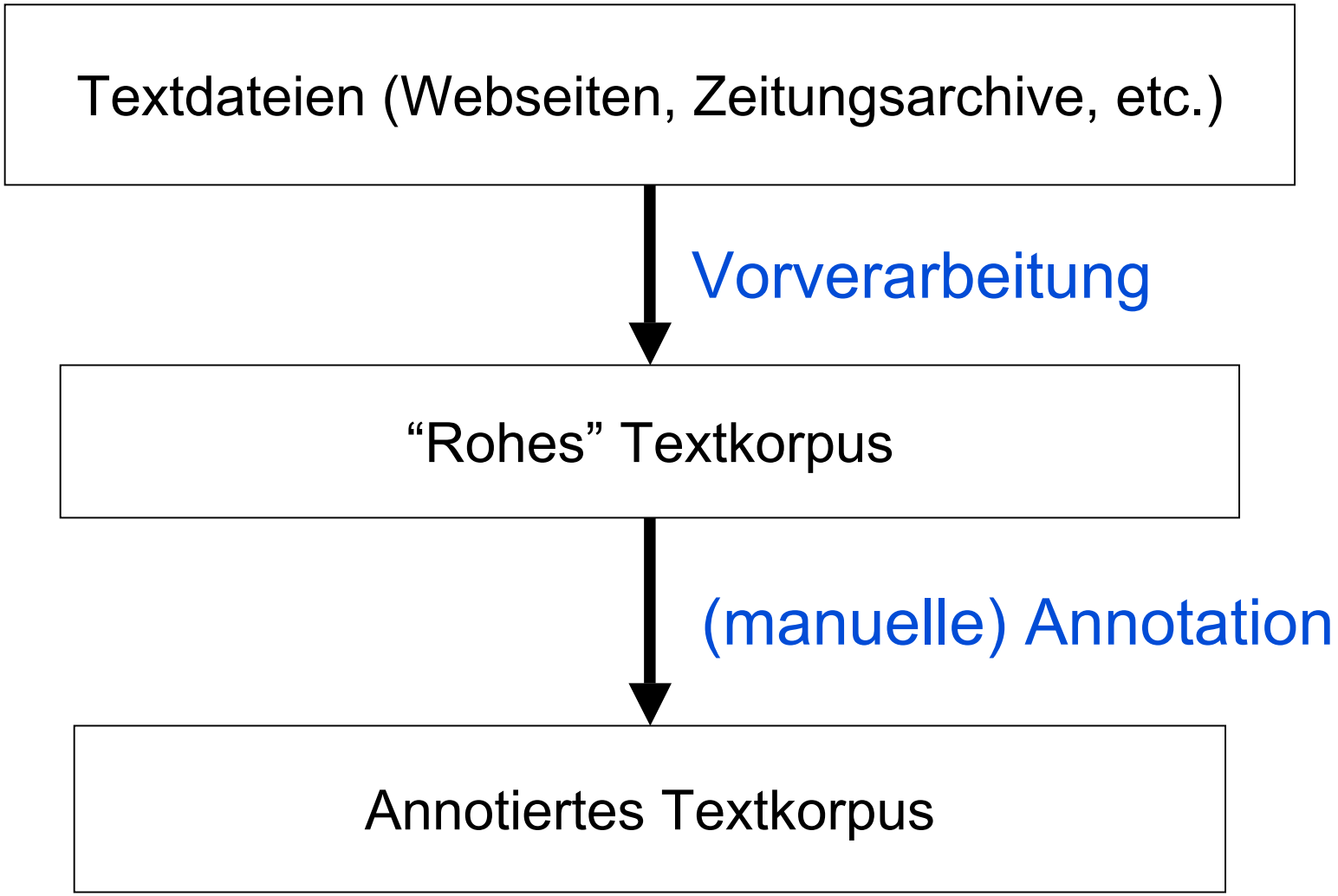
- Sprachwissenschaftliche Definition:
  - Ein Korpus (n.!) ist eine endliche Sammlung von konkreten sprachlichen Äußerungen, die als Grundlage für sprachwissen-schaftliche Untersuchungen dienen (Lexikon der Sprachwissenschaft)
- In der Computerlinguistik:
  - Typischerweise mit **Annotation**
    - Linguistische Analyse (z.B. synt. Struktur, etc.)


S
NP
VP
ADJ
N
V
NP
fruit
flies
like
DET
N
a
banana

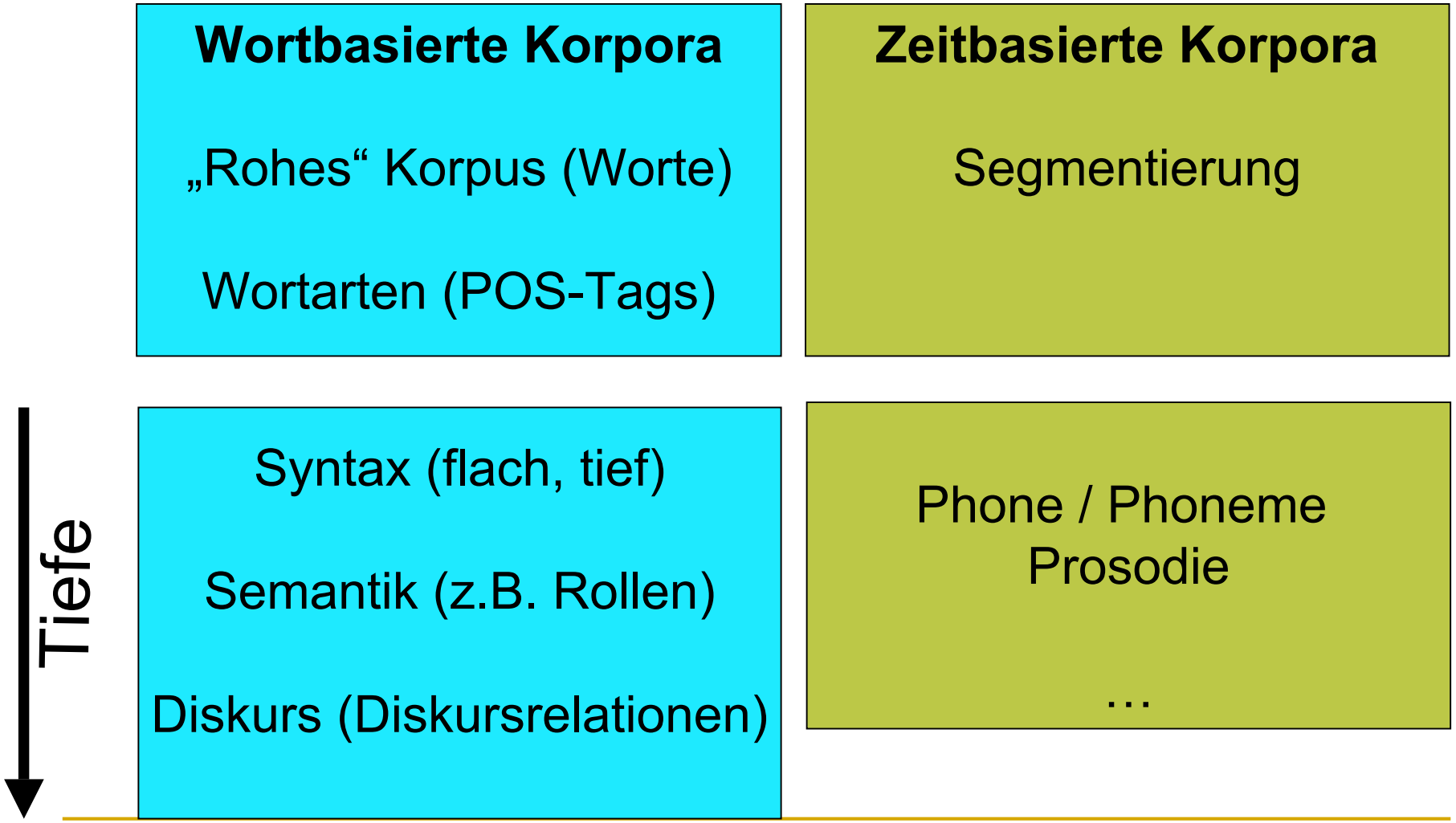
Maschinelles Lernen



Korpusverarbeitung



Annotationsebenen



Vorverarbeitung

- Vorverarbeitung ist das, was automatisch durchgeführt werden kann

- Wortbasierte Korpora
  - Grundlage: Text(datei)
  - Vorverarbeitung: Tokenisierung, Lemmatisierung, (Wortartenbestimmung)

- Zeitbasierte Korpora
  - Grundlage: phonetisches Signal (Aufnahme)
  - Vorverarbeitung: Segmentierung

Vorverarbeitung: 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

Vorverarbeitung: Lemmatisierung

- Aufgabe: Grundformen von Worten finden
  - Problem: Korpus enthält Wort**formen**
    - 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 “sich”, von Artikeln?

Korpora mit Annotation verschiedener Ebenen

„Rohe“ Korpora

```
Dies ist ein Korpus ohne Annotation.
```

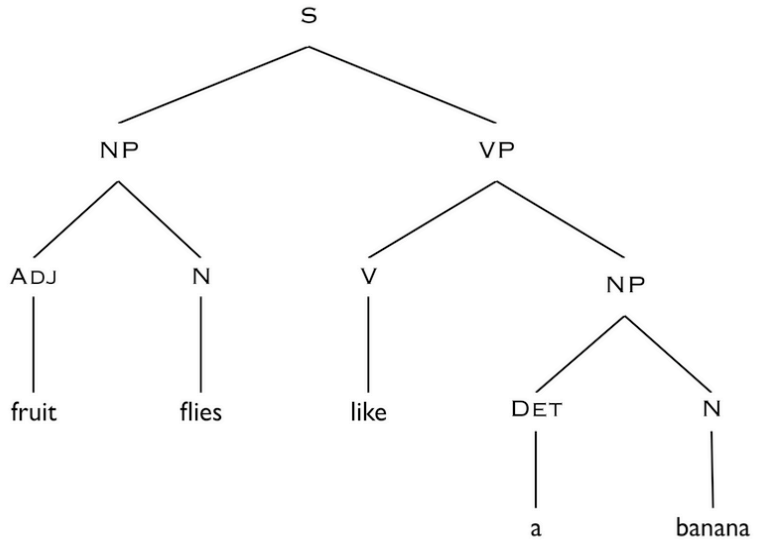
- Lexikographie: Manuelle Sichtung der Beispiele
  - Bestimmung von Wortbedeutungen
- Erstellen und Erweitern von Wörterbüchern
  - Suche nach Neologismen (Neubildungen)
    - Konkret: Suche nach Wörtern mit schwankender Häufigkeit
  - Suche nach Kollokationen
    - Ins Gras beissen, sich einen schönen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

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:
  - 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

- 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
- 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
    - z.B. Begründung (weil), Zweck (damit), ...
- Korpora:
  - DiscourseBank, auf Grundlage der Penn Treebank

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)

Keine Korpora verfügbar

- Pragmatik
  - Intentionen der Sprecher
  - “was wirklich gemeint ist”

Annotation

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

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

Annotationsschema: Nominal-Wortarten

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

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 (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 (Koreas)
  - NPD1 / NPD2 – weekday noun (Sunday / Sundays)
  - NPM1 / NPM2 – month noun (October / Octobers)

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
- Gute Annotationsschemata nötig
  - Richtlinien: Wann annotiere ich was?
  - Problemfälle: Was passiert, wenn ich mir nicht sicher bin?

Vagheit

- Problem für tiefere Ebenen (Semantik!): häufige **Vagheit**
  - Schwierig, **konsistent** zwischen Annotationskategorien zu unterscheiden

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

  - Was ist „Ich habe eine Zwiebel gepflanzt“?

Annotation: Aufwand

- Annotation ist sehr aufwendiger Prozess
  - Ein Wort: 30 Sekunden
  - 1M Worte: 500 000 Minuten = 5 Jahre
  - plus Aufwand fuer Qualitaetssicherung
- Beschleunigung: Annotatoren unterstützen
  - (Semi)-Automatisierung und manuelle Überprüfung
    - Kann zu **systematischen Fehlern** führen
  - Sehr problematisch für „tiefe“ Annotation