Einführung in die Korpuslinguistik

Ines Rehbein

WS 09/10

Überblick

Überblick

Linguistische Daten

- Woher bekommen wir unsere linguistischen Daten?
  - Instrospektion (armchair linguistics)
  - Psycholinguistische Experimente
  - Datenerhebungen
  - Korpora

Linguistische Daten

- Woher bekommen wir unsere linguistischen Daten?
  - Instrospektion (armchair linguistics)
  - Psycholinguistische Experimente
  - Datenerhebungen
  - Korpora

Linguistische Daten

- Woher bekommen wir unsere linguistischen Daten?
  - Instrospektion (armchair linguistics)
  - Psycholinguistische Experimente
  - Datenerhebungen
  - Korpora

Linguistische Daten

- Woher bekommen wir unsere linguistischen Daten?
  - Instrospektion (armchair linguistics)
  - Psycholinguistische Experimente
  - Datenerhebungen
  - Korpora

Linguistische Daten

- Woher bekommen wir unsere linguistischen Daten?
  - Instrospektion (armchair linguistics)
  - Psycholinguistische Experimente
  - Datenerhebungen
  - Korpora

Linguistische Daten - Introspektion

- Generative Tradition, im Fokus steht die Sprachkompetenz
- Frage: Welche Äußerungen einer Sprache sind grammatikalisch?
- Vorteil:
  - kann jederzeit und überall praktiziert werden
- Nachteil:
  - keine Berücksichtigung von graduellen Grammatikalitätsurteilen
  - keine Berücksichtigung von quantitativen Aspekten
  - viele Phänomene werden bei Introspektion übersehen

Linguistische Daten - Introspektion

- Generative Tradition, im Fokus steht die Sprachkompetenz
- Frage: Welche Äußerungen einer Sprache sind grammatikalisch?
- Vorteil:
  - kann jederzeit und überall praktiziert werden
- Nachteil:
  - keine Berücksichtigung von graduellen Grammatikalitätsurteilen
  - keine Berücksichtigung von quantitativen Aspekten
  - viele Phänomene werden bei Introspektion übersehen

Linguistische Daten - Psycholinguistische Experimente

- Frage: Wie wird Sprache verarbeitet?
- Methoden: Reaktionszeitexperimente (lexical decision task), Produktionsexperimente, Bewertungsexperimente, eye tracking, ...
- Vorteil:
  - Erzeugung von Daten für spezifische Fragestellungen
- Nachteil:
  - künstlich erzeugte Äußerungen, evt. beeinflusst durch die Laborsituation
  - Repräsentativität?

Linguistische Daten - Psycholinguistische Experimente

- Frage: Wie wird Sprache verarbeitet?
- Methoden: Reaktionszeitexperimente (lexical decision task), Produktionsexperimente, Bewertungsexperimente, eye tracking, ...
- Vorteil:
  - Erzeugung von Daten für spezifische Fragestellungen
- Nachteil:
  - künstlich erzeugte Äußerungen, evt. beeinflusst durch die Laborsituation
  - Repräsentativität?

Linguistische Daten - Datenerhebungen

- Leute befragen (KollegInnen/Studierende/...)
- Methoden: Fragebogen, Interview, ...
- Repräsentativität?
  - wieviele Leute muss man befragen?
- Verlässlichkeit der Daten?

Linguistische Daten - Korpora

- Sammlung von
  - Texten (z.B. Zeitungstexte, historische Texte, Kochrezepte, transkribierte gesprochene Sprache, ...)
    $\Rightarrow$ Textkorpora
  - Audiodateien (Sprachaufnahmen, evt. mit Transkription und phonetischer Annotation)
    $\Rightarrow$ Sprachkorpora
  - Videos (z.B. Gebärdensprache, evt. mit Transkription)
    $\Rightarrow$ multimodale Korpora
  - ...
- meist mit linguistischen Annotationen versehen (Wortart, Syntax, Semantik, Diskurs, ...)
- Repräsentativität?

Beispiele Textkorpora

- **DDR-Korpus**
  - 1150 Texte von 1949 bis 1990, die in der DDR erschienen sind, bzw. von DDR-Schriftstellern geschrieben und in der BRD veröffentlicht wurden
  - 9 Millionen Textwörter (tokens) in 1150 Dokumenten
- **DWDS-Kerncorpus**
  - zeitlich und nach Textsorten ausgewogenes Korpus des gesamten 20. Jahrhunderts
  - 100 Millionen Textwörter (tokens) in 79.830 Dokumenten

Beispiele Textkorpora

- **DDR-Korpus**
  - 1150 Texte von 1949 bis 1990, die in der DDR erschienen sind, bzw. von DDR-Schriftstellern geschrieben und in der BRD veröffentlicht wurden
  - 9 Millionen Textwörter (tokens) in 1150 Dokumenten
- **DWDS-Kerncorpus**
  - zeitlich und nach Textsorten ausgewogenes Korpus des gesamten 20. Jahrhunderts
  - 100 Millionen Textwörter (tokens) in 79.830 Dokumenten

Beispiele Textkorpora

- **DDR-Korpus**
  - 1150 Texte von 1949 bis 1990, die in der DDR erschienen sind, bzw. von DDR-Schriftstellern geschrieben und in der BRD veröffentlicht wurden
  - 9 Millionen Textwörter (tokens) in 1150 Dokumenten
    unausgewogen, nur für bestimmte Fragestellungen geeignet

- **DWDS-Kerncorpus**
  - zeitlich und nach Textsorten ausgewogenes Korpus des gesamten 20. Jahrhunderts
  - 100 Millionen Textwörter (tokens) in 79.830 Dokumenten
    ausgewogen, balanciertes Korpus

Referenzkorpus vs. Monitorkorpus

- **Referenzkorpus**
  (reference corpus, fixed corpus)
  - feste Größe, Zusammensetzung bekannt
  - weit verfügbar, Standard, Ergebnisse können reproduziert werden
  - veraltet irgendwann
- **Monitorkorpus**
  (anwachsend)
  - Zusammensetzung und Größe evtl. nicht bekannt (manchmal gibt's aber bestimmte Herausgabedaten)
  - für lexikographische Zwecke gut geeignet, diachron
- **Datensammlung**
  (opportunistisch)
  - man nimmt was man kriegt
  - groß und kostengünstig
  - unausgewogen, nicht repräsentativ

Referenzkorpus vs. Monitorkorpus

- **Referenzkorpus**
  (reference corpus, fixed corpus)
  - feste Größe, Zusammensetzung bekannt
  - weit verfügbar, Standard, Ergebnisse können reproduziert werden
  - veraltet irgendwann
- **Monitorkorpus**
  (anwachsend)
  - Zusammensetzung und Größe evtl. nicht bekannt (manchmal gibt's aber bestimmte Herausgabedaten)
  - für lexikographische Zwecke gut geeignet, diachron
- **Datensammlung**
  (opportunistisch)
  - man nimmt was man kriegt
  - groß und kostengünstig
  - unausgewogen, nicht repräsentativ

Referenzkorpus vs. Monitorkorpus

- **Referenzkorpus**
  (reference corpus, fixed corpus)
  - feste Größe, Zusammensetzung bekannt
  - weit verfügbar, Standard, Ergebnisse können reproduziert werden
  - veraltet irgendwann
- **Monitorkorpus**
  (anwachsend)
  - Zusammensetzung und Größe evtl. nicht bekannt (manchmal gibt's aber bestimmte Herausgabedaten)
  - für lexikographische Zwecke gut geeignet, diachron
- **Datensammlung**
  (opportunistisch)
  - man nimmt was man kriegt
  - groß und kostengünstig
  - unausgewogen, nicht repräsentativ

Beispiel Sprachkorpus

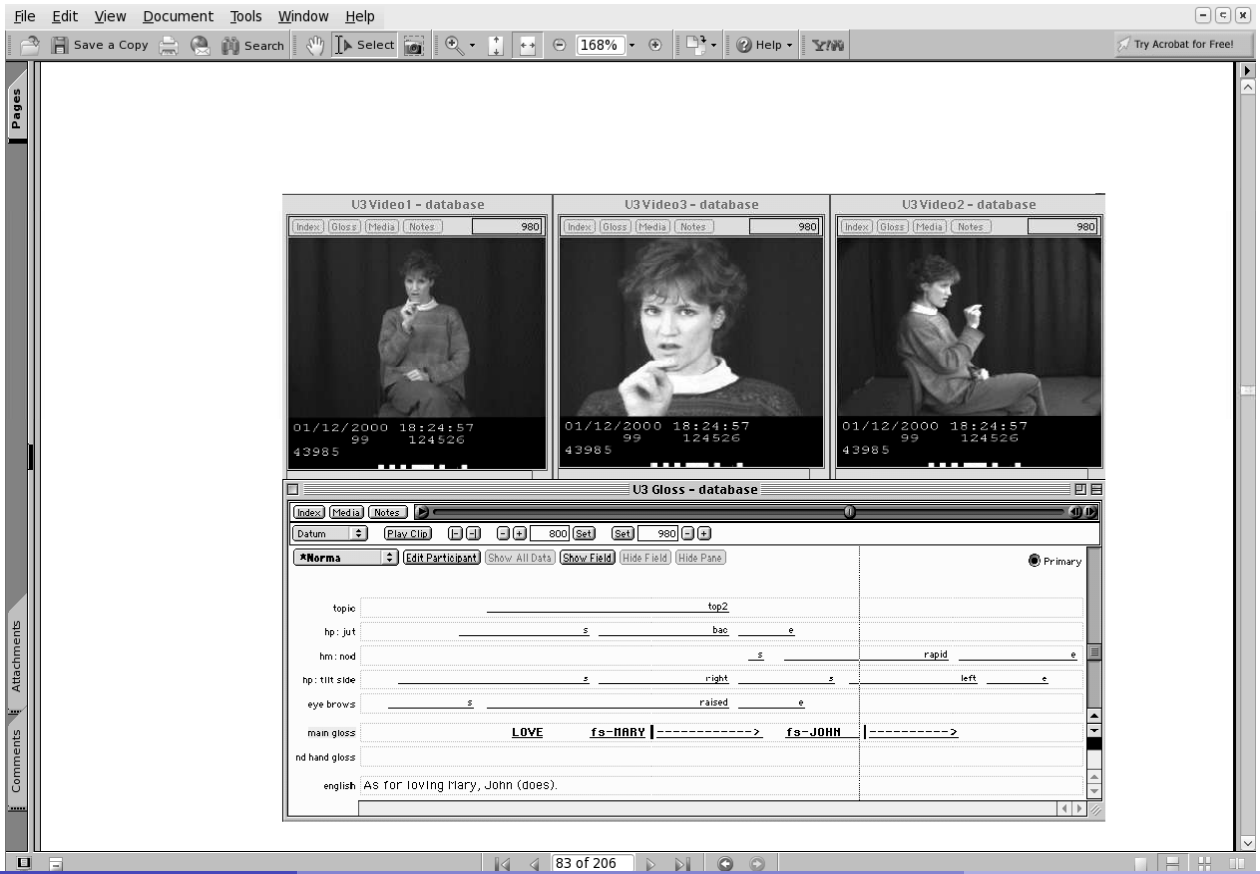
Alcohol Language Corpus (ALC)

"Ziel dieses Projektes ist die Schaffung eines umfangreichen Sprachkorpus mit Sprache unter Alkoholeinfluss. Anhand dieses Korpus soll es erstmals möglich werden, auf einer soliden statistischen Basis und auch für weibliche Sprecher Untersuchungen des Einflusses von Alkohol auf die Sprache zu untersuchen (200 Sprecher). Der Korpus enthält eine Vielzahl von Sprachstilen, von einfachen Zahlenketten, über gelesene Sprache, Zungenbrecher, Kommandos (situational prompting), Monologe bis hin zu echter Konversation. Der Grad der Alkoholisierung wird über Atem- und Blutalkohol gemessen. ALC wird in enger Kooperation mit dem Institut für Rechtsmedizin, LMU München, und dem Bund gegen Alcohol und Drogen im Straßenverkehr (B.A.D.S.) durchgeführt."

`http://phonetik.uni-muenchen.de/forschung/Bas/BasProjectsdeu.html#ALC`

Beispiel Gebärdensprachkorpus (multi-modal)

- **American Sign Language Linguistic Research Project Corpus (ASLLRP)**
  - künstlich erzeugte Sätze, die bestimmte Satzstrukturen in ASL abbilden
  - Sätze mit festem Vokabular für Computer Vision Research, Kurzgeschichten, Dialoge, verschiedene Sichten auf Handformen in ASL
  - geringe Größe, aufwendig in der Erstellung, nicht repräsentativ



Linguistische Daten - Fazit

Introspektion	**psycholinguistische Experimente**	**Korpusdaten**
Kompetenz: was ist grammatisch?	Verarbeitung: wie wird 'Sprache' verarbeitet	Performanz: was kommt vor?
Produktionssystem, das alle grammatischen Äußerungen einer Sprache hervorbringt	Modell, das die Organisation und den Zugriff auf verschiedene sprachliche Einheiten in Produktion und Rezeption im Gehirn beschreibt	Modell, das die Phänomene und Verteilungen innerhalb eines bestimmten Korpus beschreibt
nicht empirisch qualitativ (kategorial)	empirisch	empirisch qualitativ + quantitativ (probabilistisch)

Was ist am besten?

Eignung der verschiedenen Methoden hängt ab von der jeweiligen Forschungsfrage

Linguistische Daten - Fazit

Introspektion	**psycholinguistische Experimente**	**Korpusdaten**
Kompetenz: was ist grammatisch?	Verarbeitung: wie wird 'Sprache' verarbeitet	Performanz: was kommt vor?
Produktionssystem, das alle grammatischen Äußerungen einer Sprache hervorbringt	Modell, das die Organisation und den Zugriff auf verschiedene sprachliche Einheiten in Produktion und Rezeption im Gehirn beschreibt	Modell, das die Phänomene und Verteilungen innerhalb eines bestimmten Korpus beschreibt
nicht empirisch qualitativ (kategorial)	empirisch	empirisch qualitativ + quantitativ (probabilistisch)

Was ist am besten?

Eignung der verschiedenen Methoden hängt ab von der jeweiligen Forschungsfrage

Outline

Was macht die Korpuslinguistik?

- Korpuslinguistik beschäftigt sich mit:
  - dem Aufbau (⇒ Korpus-Erstellung)
  - der Auszeichnung (⇒ linguistische Annotation)
  - und der Auswertung von Korpora (⇒ linguistische Analyse)

Wozu braucht man Korpora?

Theoretische Linguistik

- Syntax
  - Ist eine bestimmte Konstruktion häufig / wahrscheinlich?
  - Ist eine bestimmte Konstruktion wirklich ungrammatisch?
- (Lexikalische) Semantik
  - Wie wird ein bestimmtes Wort verwendet?
  - Wie ist die Häufigkeitsverteilung der einzelnen Lesarten?
- Phonologie
  - Kann man anhand der Intonation Lesarten unterscheiden?
    IKEA leer gekauft Fischtank leer gekauft
- Historische Linguistik
  - Sprachwandelphänomene
- Soziolinguistik
  - Einfluss von Alter, Geschlecht, Herkunft, Klasse, ... auf die Sprache
- ...

Wozu braucht man Korpora?

Theoretische Linguistik

- Syntax
  - Ist eine bestimmte Konstruktion häufig / wahrscheinlich?
  - Ist eine bestimmte Konstruktion wirklich ungrammatisch?
- (Lexikalische) Semantik
  - Wie wird ein bestimmtes Wort verwendet?
  - Wie ist die Häufigkeitsverteilung der einzelnen Lesarten?
- Phonologie
  - Kann man anhand der Intonation Lesarten unterscheiden?
    IKEA leer gekauft Fischtank leer gekauft
- Historische Linguistik
  - Sprachwandelphänomene
- Soziolinguistik
  - Einfluss von Alter, Geschlecht, Herkunft, Klasse, ... auf die Sprache
- ...

Wozu braucht man Korpora?

Theoretische Linguistik

- Syntax
  - Ist eine bestimmte Konstruktion häufig / wahrscheinlich?
  - Ist eine bestimmte Konstruktion wirklich ungrammatisch?
- (Lexikalische) Semantik
  - Wie wird ein bestimmtes Wort verwendet?
  - Wie ist die Häufigkeitsverteilung der einzelnen Lesarten?
- Phonologie
  - Kann man anhand der Intonation Lesarten unterscheiden?
    IKEA leer gekauft Fischtank leer gekauft
- Historische Linguistik
  - Sprachwandelphänomene
- Soziolinguistik
  - Einfluss von Alter, Geschlecht, Herkunft, Klasse, ... auf die Sprache
- ...

Wozu braucht man Korpora?

Theoretische Linguistik

- Syntax
  - Ist eine bestimmte Konstruktion häufig / wahrscheinlich?
  - Ist eine bestimmte Konstruktion wirklich ungrammatisch?
- (Lexikalische) Semantik
  - Wie wird ein bestimmtes Wort verwendet?
  - Wie ist die Häufigkeitsverteilung der einzelnen Lesarten?
- Phonologie
  - Kann man anhand der Intonation Lesarten unterscheiden?
    IKEA leer gekauft Fischtank leer gekauft
- Historische Linguistik
  - Sprachwandelphänomene
- Soziolinguistik
  - Einfluss von Alter, Geschlecht, Herkunft, Klasse, ... auf die Sprache
- ...

Wozu braucht man Korpora?

Theoretische Linguistik

- Syntax
  - Ist eine bestimmte Konstruktion häufig / wahrscheinlich?
  - Ist eine bestimmte Konstruktion wirklich ungrammatisch?
- (Lexikalische) Semantik
  - Wie wird ein bestimmtes Wort verwendet?
  - Wie ist die Häufigkeitsverteilung der einzelnen Lesarten?
- Phonologie
  - Kann man anhand der Intonation Lesarten unterscheiden?
    IKEA leer gekauft Fischtank leer gekauft
- Historische Linguistik
  - Sprachwandelphänomene
- Soziolinguistik
  - Einfluss von Alter, Geschlecht, Herkunft, Klasse, ... auf die Sprache
- ...

Wozu braucht man Korpora?

Theoretische Linguistik

- Syntax
  - Ist eine bestimmte Konstruktion häufig / wahrscheinlich?
  - Ist eine bestimmte Konstruktion wirklich ungrammatisch?
- (Lexikalische) Semantik
  - Wie wird ein bestimmtes Wort verwendet?
  - Wie ist die Häufigkeitsverteilung der einzelnen Lesarten?
- Phonologie
  - Kann man anhand der Intonation Lesarten unterscheiden?
    IKEA leer gekauft Fischtank leer gekauft
- Historische Linguistik
  - Sprachwandelphänomene
- Soziolinguistik
  - Einfluss von Alter, Geschlecht, Herkunft, Klasse, ... auf die Sprache
- ...

Wozu braucht man Korpora? (2)

Computerlinguistik

- Korpora als Trainingsdaten für statistische Systeme:
  - Wortarten-Tagger
  - Syntaktische Parser
  - Semantische Parser / Labelling von Semantischen Rollen
  - Systeme zur Lesarten-Disambiguierung
  - Anaphern-Auflösung
  - Maschinelles Übersetzen
  - Automatische Spracherkennung
  - ...
- Korpora als Benchmark zur Evaluation der oben genannten Systeme

Wozu braucht man Korpora? (2)

Computerlinguistik

- Korpora als Trainingsdaten für statistische Systeme:
  - Wortarten-Tagger
  - Syntaktische Parser
  - Semantische Parser / Labelling von Semantischen Rollen
  - Systeme zur Lesarten-Disambiguierung
  - Anaphern-Auflösung
  - Maschinelles Übersetzen
  - Automatische Spracherkennung
  - ...
- Korpora als Benchmark zur Evaluation der oben genannten Systeme

Outline

Erste Korpora

- Schon im 19. Jhdt. (und früher) Verwendung von Textsammlungen
  - zur Beschreibung von Sprachwandel
  - Wörterbucherstellung (z.B. Grimmsches Wörterbuch)
  - Dokumentation von Spracherwerb
  - Belege für grammatische Aussagen
- meist Belege aus der Literatur
- nicht repräsentativ

Erste digitale Korpora

- Anfang 60er:
  - Brown University Standard Corpus of Present-Day American English (Francis & Kucera)
    - ★ synchron, ausgewogen (balanced)
    - ★ ca. 1 Mio. Token (500 Samples mit je 2000 Token)
    - ★ geschriebene Sprache von 1961
    - ★ Korpus fertiggestellt in 1964
- Andere Korpora folgten:
  - Lancaster-Oslo/Bergen (LOB) Corpus (Leech)
    - ★ erstellt 1970-78
    - ★ englisches Gegenstück zum Brown Corpus (Größe, Design)
  - London-Lund Corpus (LLC, Swartvik)
    - ★ publiziert 1980
    - ★ gesprochenes Englisch, transkribiert
    - ★ ca. 50 000 Token
  - Kolhapur Corpus of Indian English (Shastri, 1988)
  - Australian Corpus of English (ACE)
  - Wellington Corpus of Written New Zealand English

Erste Reaktionen auf linguistische Korpora

- 1957: Noam Chomsky, *Syntactic Structures*
- Empirismus als herrschendes Paradigma in der Linguistik (und anderen Kognitionswissenschaften) wird vom Rationalismus abgelöst
- Fokus auf Sprachkompetenz, Sprachperformanz und quantitative Aspekte von Sprache gelten als uninteressant

> *"It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term."* (Chomsky, 1969)

- Wenig Interesse an empirischen, korpus-linguistischen Projekten
- Korpora als zufällige, nicht repräsentative Sammlungen von Texten, die keinen wirklichen Einblick in die Sprachkompetenz geben

Erste Reaktionen auf linguistische Korpora

- 1957: Noam Chomsky, *Syntactic Structures*
- Empirismus als herrschendes Paradigma in der Linguistik (und anderen Kognitionswissenschaften) wird vom Rationalismus abgelöst
- Fokus auf Sprachkompetenz, Sprachperformanz und quantitative Aspekte von Sprache gelten als uninteressant

  "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." (Chomsky, 1969)

- Wenig Interesse an empirischen, korpus-linguistischen Projekten
- Korpora als zufällige, nicht repräsentative Sammlungen von Texten, die keinen wirklichen Einblick in die Sprachkompetenz geben

Erste Reaktionen auf linguistische Korpora

- 1957: Noam Chomsky, *Syntactic Structures*
- Empirismus als herrschendes Paradigma in der Linguistik (und anderen Kognitionswissenschaften) wird vom Rationalismus abgelöst
- Fokus auf Sprachkompetenz, Sprachperformanz und quantitative Aspekte von Sprache gelten als uninteressant

  "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." (Chomsky, 1969)

- Wenig Interesse an empirischen, korpus-linguistischen Projekten
- Korpora als zufällige, nicht repräsentative Sammlungen von Texten, die keinen wirklichen Einblick in die Sprachkompetenz geben

Exkurs: Können Korpora Antworten auf linguistische Fragestellungen geben?

- 2 Beispiele:
  - **Beispiel I:** Partikelverben (Müller & Meurers, 2006)
  - **Beispiel II:** Idiome (Geyken, Sokirko, Rehbein & Fellbaum, 2004)

Beispiel I: Partikelverben

- *Theorie:* Verbpartikeln können nicht vorangestellt werden (Ausnahme: prädikative Partikeln wie *auf* in *aufmachen*)

- *Korpusevidenz:*
  Los$_{PART}$ **ging** es schon in dieser Woche. (taz, 11.10.1995)
  Vor$_{PART}$ **hat** er das jedenfalls. (taz, 15.07.1999)

Beispiel I: Partikelverben

- *Theorie:* Verbpartikeln können nicht vorangestellt werden
  (Ausnahme: prädikative Partikeln wie *auf* in *aufmachen*)

- *Korpusevidenz:*
  Los$_{PART}$ **ging** es schon in dieser Woche. (taz, 11.10.1995)
  Vor$_{PART}$ **hat** er das jedenfalls. (taz, 15.07.1999)

Beispiel II: Idiome

- *Theorie:* klassische Ansätze betonen die Invariabilität von Idiomen (Katz, 1973; Chomsky, 1980)
- *Korpusevidenz:* **ein Blatt vor den Mund nehmen**
  - Pluralisierung:
    - ohne Blätter vor den Mund zu nehmen
  - Quantifizierung:
    - Hier nahm er manches Blatt vor den Mund
    - der sich 100 Blätter vor den Mund nimmt
  - Adjektivische Modifikation eines oder beider Nomen:
    - mit einem postmodernen Blatt vor dem Munde
    - kein Blatt vor seinen republikfeindlichen Mund
  - Nomen-Modifikation:
    - ohne das geringste (Klee-)Blatt vor den vorlauten Mund zu nehmen

Beispiel II: Idiome

- *Theorie:* klassische Ansätze betonen die Invariabilität von Idiomen (Katz, 1973; Chomsky, 1980)
- *Korpusevidenz:* **ein Blatt vor den Mund nehmen**
  - Pluralisierung:
    - ohne Blätter vor den Mund zu nehmen
  - Quantifizierung:
    - Hier nahm er manches Blatt vor den Mund
    - der sich 100 Blätter vor den Mund nimmt
  - Adjektivische Modifikation eines oder beider Nomen:
    - mit einem postmodernen Blatt vor dem Munde
    - kein Blatt vor seinen republikfeindlichen Mund
  - Nomen-Modifikation:
    - ohne das geringste (Klee-)Blatt vor den vorlauten Mund zu nehmen

Können Korpora Antworten auf linguistische Fragestellungen geben?

- Korpora erweisen sich als fruchtbare Hilfsmittel für linguistische Forschung:
  - ermöglichen die Überprüfung linguistischer Theorien
  - sinnvolle Ergänzung der Introspektion
- Daher steigender Bedarf nach
  - mehr Daten
  - mehr Annotation (Syntax, Semantik, Prosodie, Metadaten, ...)
  - mehr Sprachen

Und was sind Metadaten?

Metadaten aus dem British National Corpus (BNC)

```
<person
    age="Ag0"
    dialect="XLO"
    xml:id="PS5A1"
    role="self"
    sex="m"
    soc="C2">
    <name>Terry</name>
    <age>14</age>
    <occupation>student</occupation>
    <dialect>London</dialect>
</person>
```

Outline

Kriterien für die Korpuserstellung

- Fragestellung
- Welche Texte/Textsorten? Wieviele Daten? Balanciert vs. spezialisiert vs. opportunistisch?
- Was wird annotiert? Annotationsschema?
  - Feinkörnigkeit der Annotation - oft Kompromiss zwischen Detailgenauigkeit und Konsistenz
  - Konsistenz der Annotation (Inter-Annotator Agreement)
  - Dokumentation: Was wurde wie annotiert?
  - Originaltext muss wieder reproduzierbar sein
- Welche Meta-Daten? (AutorIn, Herkunft, Erstellungsdatum, Geschlecht, Alter, soziale Klasse, ...)

- Welche Fragen kann ich mit meinem Korpus beantworten?
- Wie kann ich mein Korpus durchsuchen?

Kriterien für die Korpuserstellung

- Fragestellung
- Welche Texte/Textsorten? Wieviele Daten? Balanciert vs. spezialisiert vs. opportunistisch?
- Was wird annotiert? Annotationsschema?
  - Feinkörnigkeit der Annotation - oft Kompromiss zwischen Detailgenauigkeit und Konsistenz
  - Konsistenz der Annotation (Inter-Annotator Agreement)
  - Dokumentation: Was wurde wie annotiert?
  - Originaltext muss wieder reproduzierbar sein
- Welche Meta-Daten? (AutorIn, Herkunft, Erstellungsdatum, Geschlecht, Alter, soziale Klasse, ...)

- Welche Fragen kann ich mit meinem Korpus beantworten?
- Wie kann ich mein Korpus durchsuchen?

Exkurs: Repräsentativität

- Wir wollen *repräsentative* Korpora
- Was bedeutet *repräsentativ*?

- Korpusdaten sollen typisch sein für die Grundgesamtheit, die der Forschungsfrage zugrunde liegt
- Was wäre eine repräsentative Stichprobe für Studien zum Thema:
  - Sprachgebrauch der Deutschen
  - Verwendung von technischen Begriffen in der DDR
  - Unterschiede im Sprachgebrauch in der DDR/BRD
  - Sprache von Jugendlichen in Webblogs
  - Sprachwandelprozesse der letzten 100 Jahre

Ein Korpus kann repräsentativ sein im Hinblick auf eine bestimmte Fragestellung, und gleichzeitig nicht repräsentativ für eine andere Fragestellung (z.B. Goethes Gesamtausgabe $\Rightarrow$ erlaubt Aussagen über Goethe, nicht über gesamte deutsche Literatur)

Exkurs: Repräsentativität

- Wir wollen *repräsentative* Korpora
- Was bedeutet *repräsentativ*?

- Korpusdaten sollen typisch sein für die Grundgesamtheit, die der Forschungsfrage zugrunde liegt
- Was wäre eine repräsentative Stichprobe für Studien zum Thema:
  - Sprachgebrauch der Deutschen
  - Verwendung von technischen Begriffen in der DDR
  - Unterschiede im Sprachgebrauch in der DDR/BRD
  - Sprache von Jugendlichen in Webblogs
  - Sprachwandelprozesse der letzten 100 Jahre

Ein Korpus kann repräsentativ sein im Hinblick auf eine bestimmte Fragestellung, und gleichzeitig nicht repräsentativ für eine andere Fragestellung (z.B. Goethes Gesamtausgabe ⇒ erlaubt Aussagen über Goethe, nicht über gesamte deutsche Literatur)

Exkurs: Repräsentativität

- Wir wollen *repräsentative* Korpora
- Was bedeutet *repräsentativ*?

- Korpusdaten sollen typisch sein für die Grundgesamtheit, die der Forschungsfrage zugrunde liegt
- Was wäre eine repräsentative Stichprobe für Studien zum Thema:
  - Sprachgebrauch der Deutschen
  - Verwendung von technischen Begriffen in der DDR
  - Unterschiede im Sprachgebrauch in der DDR/BRD
  - Sprache von Jugendlichen in Webblogs
  - Sprachwandelprozesse der letzten 100 Jahre

Ein Korpus kann repräsentativ sein im Hinblick auf eine bestimmte Fragestellung, und gleichzeitig nicht repräsentativ für eine andere Fragestellung (z.B. Goethes Gesamtausgabe ⇒ erlaubt Aussagen über Goethe, nicht über gesamte deutsche Literatur)

Exkurs: Repräsentativität

- Wir wollen *repräsentative* Korpora
- Was bedeutet *repräsentativ*?

- Korpusdaten sollen typisch sein für die Grundgesamtheit, die der Forschungsfrage zugrunde liegt
- Was wäre eine repräsentative Stichprobe für Studien zum Thema:
  - Sprachgebrauch der Deutschen
  - Verwendung von technischen Begriffen in der DDR
  - Unterschiede im Sprachgebrauch in der DDR/BRD
  - Sprache von Jugendlichen in Webblogs
  - Sprachwandelprozesse der letzten 100 Jahre

Ein Korpus kann repräsentativ sein im Hinblick auf eine bestimmte Fragestellung, und gleichzeitig nicht repräsentativ für eine andere Fragestellung (z.B. Goethes Gesamtausgabe ⇒ erlaubt Aussagen über Goethe, nicht über gesamte deutsche Literatur)

Korpuserstellung - Vorverarbeitung

- "Roher" Text
  Was gibt's in New York zu sehen?
- **Satzendeerkennung**
  Probleme mit Datumsangaben, Uhrzeit (7.00 Uhr), Abkürzungen, URLs, ...
- **Tokenisierung** (Zerteilung in kleinste Einheiten, Abtrennung von Satzzeichen)
  Was gibt 's in New York zu sehen ?
  Fragen: Wie soll *gibt's* getrennt werden? *New York* ein oder zwei Token? Und Komposita? (z.B. E.coli-Bakterien)
- **Lemmatisierung**
  was geben es in New York zu sehen ?
- **Part-Of-Speech (POS) Tagging** (Stuttgart-Tübingen-Tag-Set)
  Was/PWS gibt/VVFIN 's/PPER in/APPR New/NE York/NE zu/PTKZU sehen/VVINF ?

Korpuserstellung - Vorverarbeitung

- "Roher" Text
  Was gibt's in New York zu sehen?
- **Satzendeerkennung**
  Probleme mit Datumsangaben, Uhrzeit (7.00 Uhr), Abkürzungen, URLs, ...
- **Tokenisierung** (Zerteilung in kleinste Einheiten, Abtrennung von Satzzeichen)
  Was gibt 's in New York zu sehen ?
  Fragen: Wie soll *gibt's* getrennt werden? *New York* ein oder zwei Token? Und Komposita? (z.B. E.coli-Bakterien)
- **Lemmatisierung**
  was geben es in New York zu sehen ?
- **Part-Of-Speech (POS) Tagging** (Stuttgart-Tübingen-Tag-Set)
  Was/PWS gibt/VVFIN 's/PPER in/APPR New/NE York/NE zu/PTKZU sehen/VVINF ?

Korpuserstellung - Vorverarbeitung

- "Roher" Text
  Was gibt's in New York zu sehen?
- **Satzendeerkennung**
  Probleme mit Datumsangaben, Uhrzeit (7.00 Uhr), Abkürzungen, URLs, ...
- **Tokenisierung** (Zerteilung in kleinste Einheiten, Abtrennung von Satzzeichen)
  Was gibt 's in New York zu sehen ?
  Fragen: Wie soll *gibt's* getrennt werden? *New York* ein oder zwei Token? Und Komposita? (z.B. E.coli-Bakterien)
- **Lemmatisierung**
  was geben es in New York zu sehen ?
- **Part-Of-Speech (POS) Tagging** (Stuttgart-Tübingen-Tag-Set)
  Was/PWS gibt/VVFIN 's/PPER in/APPR New/NE York/NE zu/PTKZU sehen/VVINF ?

Korpuserstellung - Vorverarbeitung

- "Roher" Text
  Was gibt's in New York zu sehen?
- **Satzendeerkennung**
  Probleme mit Datumsangaben, Uhrzeit (7.00 Uhr), Abkürzungen, URLs, ...
- **Tokenisierung** (Zerteilung in kleinste Einheiten, Abtrennung von Satzzeichen)
  Was gibt 's in New York zu sehen ?
  Fragen: Wie soll *gibt's* getrennt werden? *New York* ein oder zwei Token? Und Komposita? (z.B. E.coli-Bakterien)
- **Lemmatisierung**
  was geben es in New York zu sehen ?
- **Part-Of-Speech (POS) Tagging** (Stuttgart-Tübingen-Tag-Set)
  Was/PWS gibt/VVFIN 's/PPER in/APPR New/NE York/NE zu/PTKZU sehen/VVINF ?

Korpuserstellung - Vorverarbeitung

- "Roher" Text
  Was gibt's in New York zu sehen?
- **Satzendeerkennung**
  Probleme mit Datumsangaben, Uhrzeit (7.00 Uhr), Abkürzungen, URLs, ...
- **Tokenisierung** (Zerteilung in kleinste Einheiten, Abtrennung von Satzzeichen)
  Was gibt 's in New York zu sehen ?
  Fragen: Wie soll *gibt's* getrennt werden? *New York* ein oder zwei Token? Und Komposita? (z.B. E.coli-Bakterien)
- **Lemmatisierung**
  was geben es in New York zu sehen ?
- **Part-Of-Speech (POS) Tagging** (Stuttgart-Tübingen-Tag-Set)
  Was/PWS gibt/VVFIN 's/PPER in/APPR New/NE York/NE zu/PTKZU sehen/VVINF ?

Outline

Was sind und wofür braucht man Baumbanken?

- Baumbanken sind
  - Korpora mit syntaktischen Annotationen (über Part-of-Speech Ebene hinausgehend)
  - Syntax-Bäume a la Chomsky (Konstituenten) oder Dependenzen
  - manuell erstellt *oder*
  - automatisch erstellt und manuell korrigiert
- Baumbanken werden gebraucht zur
  - Untersuchung linguistischer Phänomene
  - Überprüfung linguistischer Theorien
  - Ressourcen zum Training von Methoden des Maschinellen Lernens/ für die Entwicklung von Sprachtechnologien:
    - ★ Training und Evaluation von Parsern
    - ★ Ressourcen für Maschinelle Übersetzung (Parallele Baumbanken)
    - ★ Extraktion von Subkategorisierungsrahmen für die Erstellung von Lexika
    - ★ ...

Baumbanken: Die Penn Treebank

- Penn Treebank (Englisch, 1989-1995)
- Phase I (1989-1992)
  - Wall Street Journal (50 000 Sätze, 1 Mio. Worte)
  - Zusätzlich: geparste Version des Brown Korpus (1 Mio. Worte),
  - Automatisch getagged (POS)
  - Manuell annotiert mit Phrasen-Struktur (skeletal parse)

```
(SBARQ (WHNP Who)
       (SQ (NP T)
           will
           (VP come
               (PP to
                   (NP the party))))
       ?)
```

- Phase II (1993-1995)
  - Anreicherung mit "tiefen" linguistischen Informationen

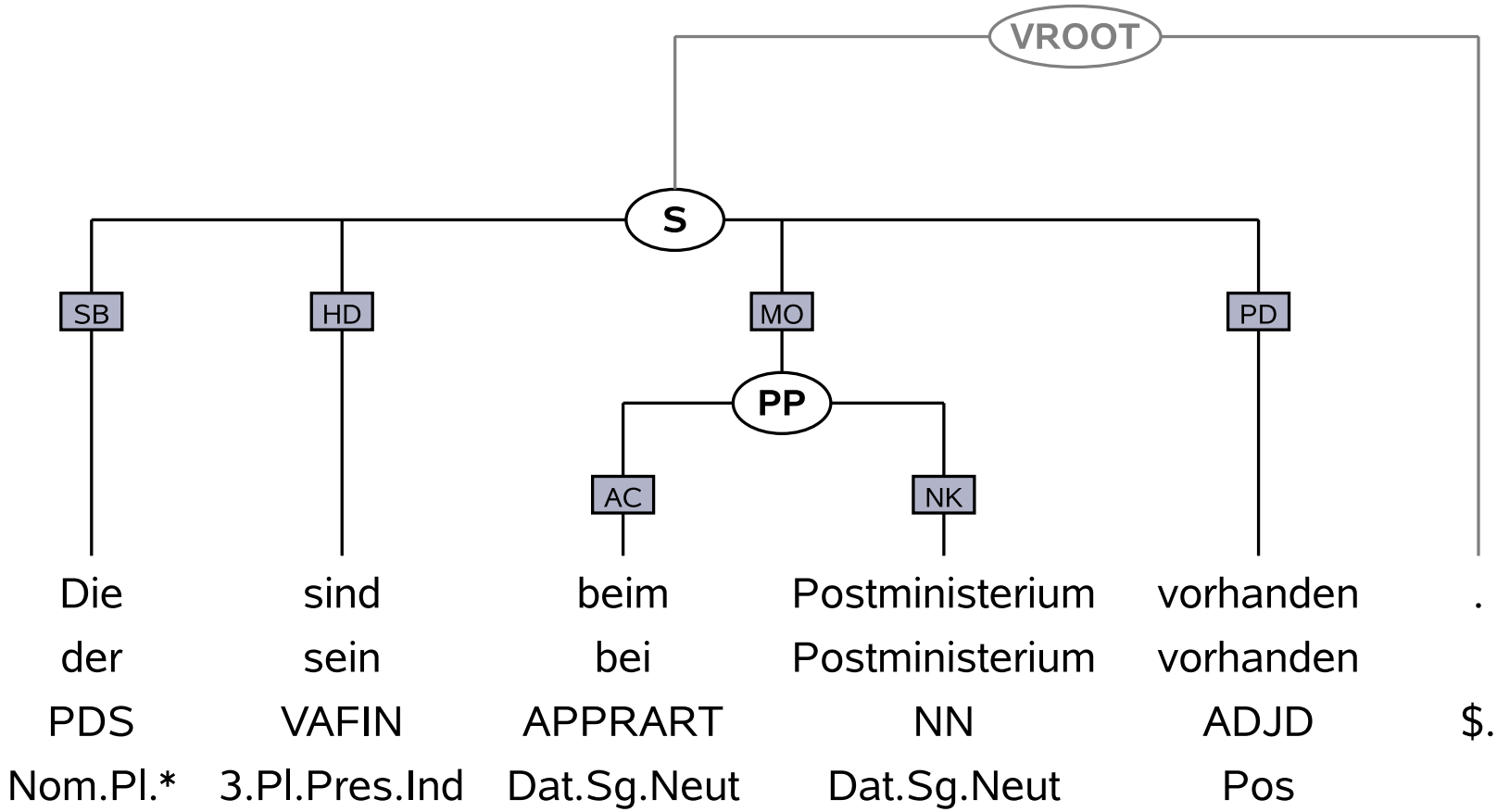
Digitale Korpora / Baumbanken - Zwischenfazit

- Erste digitale Korpora seit Mitte 60er, erste syntaktisch annotierte digitale Korpora seit Anfang 80er
- Wichtige Hilfsmittel für linguistische Forschung:
  - Überprüfung linguistischer Theorien
  - "Echte" Daten als Ergänzung für Introspektion
- Penn Treebank als erstes großes, syntaktisch annotiertes Korpus ermöglicht neue Herangehensweisen in NLP, probabilistische Methoden gewinnen an Bedeutung
- "Tiefe" linguistische Annotation der Penn-II Treebank erhöht die Nützlichkeit der Baumbank
- Bedeutung von linguistisch annotierten Korpora wächst, Erstellung von Korpora für andere Sprachen, Ausweitung der Annotation (Syntax, Semantik, Named Entities, Diskursstruktur, ...)

Baumbanken - Zwischenfazit

- Baumbanken sind syntaktisch annotierte Korpora
- Konstituenten versus Dependenzen
  - Penn Treebank (Wall Street Journal, Konstituenten)
  - Prague Dependency Bank (Dependenzen)
- hybride Baumbanken (z.B. die deutsche TiGer Baumbank)

Beispielbaum - TIGER Treebank



General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
         )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...
- Terminale Knoten: Die, sind, beim, ...

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...
- Terminale Knoten: Die, sind, beim, ...
- Part-of-Speech (POS) Tags: PDS, VAFIN, APPRART, NN, ...

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...
- Terminale Knoten: Die, sind, beim, ...
- Part-of-Speech (POS) Tags: PDS, VAFIN, APPRART, NN, ...
- Grammatikalische Funktionen: SB, HD, OA, DA, AG, ...

TiGerSearch

- Graphisches User-Interface zur Suche in syntaktisch annotierten Korpora
- User Manual:
  `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/manual_html.html`
- Kurze Einführung in TiGerSearch:
  `http://www.uni-potsdam.de/u/germanistik/ls_dgs/tiger1-intro.pdf`

TiGerSearch

Wort-Suche	[word="Fledermaus"]
Lemma-Suche	[lemma="Politiker"]
Morphologie	[morph="1.Dat.Sg.Fem"]
Kombinationen	[morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
	[lemma="Politiker" & morph=("Gen.Sg.Fem" |"Gen.Sg.Masc")]
ein syntaktischer Knoten mit der Kategorie PP (Präpositionalphrase)	#pp:[cat="PP"]
ein Artikel, direkt gefolgt von einem Nomen	[pos="ART"] . [pos="NN"]

TiGerSearch

Wort-Suche	[word="Fledermaus"]
Lemma-Suche	[lemma="Politiker"]
Morphologie	[morph="1.Dat.Sg.Fem"]
Kombinationen	[morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
	[lemma="Politiker" & morph=("Gen.Sg.Fem" |"Gen.Sg.Masc")]
ein syntaktischer Knoten mit der Kategorie PP (Präpositionalphrase)	#pp:[cat="PP"]
ein Artikel, direkt gefolgt von einem Nomen	[pos="ART"] . [pos="NN"]

TiGerSearch

Wort-Suche	[word="Fledermaus"]
Lemma-Suche	[lemma="Politiker"]
Morphologie	[morph="1.Dat.Sg.Fem"]
Kombinationen	[morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
	[lemma="Politiker"
	& morph=("Gen.Sg.Fem" |"Gen.Sg.Masc")]
ein syntaktischer Knoten	
mit der Kategorie PP	
(Präpositionalphrase)	#pp:[cat="PP"]
ein Artikel, direkt	
gefolgt von einem Nomen	[pos="ART"] . [pos="NN"]

TiGerSearch

Wort-Suche	[word="Fledermaus"]
Lemma-Suche	[lemma="Politiker"]
Morphologie	[morph="1.Dat.Sg.Fem"]
Kombinationen	[morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
	[lemma="Politiker" & morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
ein syntaktischer Knoten mit der Kategorie PP (Präpositionalphrase)	#pp:[cat="PP"]
ein Artikel, direkt gefolgt von einem Nomen	[pos="ART"] . [pos="NN"]

TiGerSearch

Wort-Suche	[word="Fledermaus"]
Lemma-Suche	[lemma="Politiker"]
Morphologie	[morph="1.Dat.Sg.Fem"]
Kombinationen	[morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
	[lemma="Politiker"
	& morph=("Gen.Sg.Fem" |"Gen.Sg.Masc")]
ein syntaktischer Knoten mit der Kategorie PP (Präpositionalphrase)	#pp:[cat="PP"]
ein Artikel, direkt gefolgt von einem Nomen	[pos="ART"] . [pos="NN"]

TiGerSearch

Wort-Suche	[word="Fledermaus"]
Lemma-Suche	[lemma="Politiker"]
Morphologie	[morph="1.Dat.Sg.Fem"]
Kombinationen	[morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
	[lemma="Politiker"
	& morph=("Gen.Sg.Fem" |"Gen.Sg.Masc")]
ein syntaktischer Knoten mit der Kategorie PP (Präpositionalphrase)	#pp:[cat="PP"]
ein Artikel, direkt gefolgt von einem Nomen	[pos="ART"] . [pos="NN"]

TiGerSearch

Wort-Suche	[word="Fledermaus"]
Lemma-Suche	[lemma="Politiker"]
Morphologie	[morph="1.Dat.Sg.Fem"]
Kombinationen	[morph=("Gen.Sg.Fem"|"Gen.Sg.Masc")]
	[lemma="Politiker"
	& morph=("Gen.Sg.Fem" |"Gen.Sg.Masc")]
ein syntaktischer Knoten mit der Kategorie PP (Präpositionalphrase)	#pp:[cat="PP"]
ein Artikel, direkt gefolgt von einem Nomen	[pos="ART"] . [pos="NN"]

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

TiGerSearch (2)

Boolsche Operatoren	& und | oder ! nicht
Reguläre Ausdrücke	[word=/.*tier/]
.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele Folgen von ab, abab, ababab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab

Zusammenfassung - Korpuslinguistik

- Korpuslinguistik beschäftigt sich mit
  - dem Aufbau,
  - der Auszeichnung und
  - der Auswertung von Korpora

- Korpora
  - sind Sammlungen von linguistischen Daten (geschriebene/gesprochene Sprache, multimodal)
  - meist mit Metadaten und linguistischer Annotation

- Korpora werden benutzt
  - zur Beantwortung linguistischer Fragestellungen (Phonologie, Morphologie, Syntax, historische Linguistik, Soziolinguistik, ...)
  - als Trainingsdaten für statistische Systeme
  - zur Evaluation von statistischen Systemen

Zusammenfassung - Korpuslinguistik

- Korpuslinguistik beschäftigt sich mit
  - dem Aufbau,
  - der Auszeichnung und
  - der Auswertung von Korpora

- Korpora
  - sind Sammlungen von linguistischen Daten (geschriebene/gesprochene Sprache, multimodal)
  - meist mit Metadaten und linguistischer Annotation

- Korpora werden benutzt
  - zur Beantwortung linguistischer Fragestellungen (Phonologie, Morphologie, Syntax, historische Linguistik, Soziolinguistik, ...)
  - als Trainingsdaten für statistische Systeme
  - zur Evaluation von statistischen Systemen

Zusammenfassung - Korpuslinguistik

- Korpuslinguistik beschäftigt sich mit
  - dem Aufbau,
  - der Auszeichnung und
  - der Auswertung von Korpora
- Korpora
  - sind Sammlungen von linguistischen Daten (geschriebene/gesprochene Sprache, multimodal)
  - meist mit Metadaten und linguistischer Annotation
- Korpora werden benutzt
  - zur Beantwortung linguistischer Fragestellungen (Phonologie, Morphologie, Syntax, historische Linguistik, Soziolinguistik, ...)
  - als Trainingsdaten für statistische Systeme
  - zur Evaluation von statistischen Systemen

Referenzen I

- Korpuslinguistik
  - Viele der hier gezeigten Folien basieren auf Lehrmaterial von Anke Lüdeling: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/BochumFolien.pdf
- Korpora
  - Graeme Kennedy. 1998. An Introduction to Corpus Linguistics. Longman.
  - Corpus Linguistics. Tony McEnery and Andrew Wilson. Edinburgh Textbooks in Empirical Linguistics.
  - Stefan Müller. 2004. *Complex NPs, Subjacency, and Extraposition*. Snippets 8, pages 10-11. `http://www.cl.uni-bremen.de/~stefan/Pub/subjacency.html`
  - Stefan Müller and Walt Detmar Meurers. 2006. Corpus Evidence for Syntactic Structures and Requirements for Annotations of Tree Banks. *Proceedings of the Int. Conference on Linguistic Evidence*. Tübingen, Germany. `http://purl.org/net/dm/papers/mueller-meurers-06.html`
  - Alexander Geyken, Alexej Sokirko, Ines Rehbein and Christiane Fellbaum. 2004. *What is the Optimal Corpus Size for the Study of Idioms?* Paper delivered at the Annual Meeting of the German Linguistic Society, Mainz, Germany.

Referenzen II

- Baumbanken
  - Penn Treebank: `http://www.cis.upenn.edu/~treebank`
  - Susanne: `http://www.grsampson.net/RSue.html`
  - NEGRA: Skut, Wojciech, Brigitte Krann, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. *In Proceedings of ANLP 1997*, Washington, D.C.
  - TIGER:
    - Brants, Sabine, and Silvia Hansen. 2002. Developments in the TIGER Annotation Scheme and their Realization in the Corpus. *In Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)* pp. 1643-1649 Las Palmas.
    - Dipper, S., T. Brants, W. Lezius, O. Plaehn, and G. Smith. 2001. The TIGER Treebank. *In Third Workshop on Linguistically Interpreted Corpora LINC-2001*, Leuven, Belgium.
  - TüBa-D/Z: Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister. 2005. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).* Seminar für Sprachwissenschaft, Universität Tübingen, Germany.
  - POS-Tagging
    - Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Technical Report, IMS-CL, University Stuttgart, 1995.