Korpuslinguistik

Anke Lüdeling
anke.luedeling@rz.hu-berlin.de

Doktorandenseminar Bochum
Oktober 2008

Organisatorisches: Kontakt

- email: anke.luedeling@rz.hu-berlin.de
- homepage: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/
- Telefon: 030-20939799

Korpuslinguistik

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

  von Korpora
- Korpora sind Sammlungen von linguistischen Daten (Texte, gesprochene Sprache, außersprachliche Daten wie Gestik etc.)

Organisatorisches: Plan

- heute:
  - Daten in der Linguistik → Korpusdaten/Überblick
  - Korpusdesign
  - Korpusvorverarbeitung
  - Experimentdesign
- morgen
  - Lernerkorpora

Daten in der Linguistik

Linguistische Daten

- Für welche Fragestellungen braucht man welche Art von Daten?
- Wie werden linguistische Daten erhoben?
- Welche Eigenschaften haben die unterschiedlichen Arten von Daten?
- Was kann man aus den vorhandenen Daten schließen?

Linguistische Daten

- Woher bekommen Linguisten/Linguistinnen eigentlich die Daten, die sie zur Überprüfung ihrer Theorien/Hypothesen brauchen?

- Introspektion
- psycholinguistische & neurologische Experimente
- Datenerhebungen
- Korpora

Introspektion

- Die Linguistin sitzt im Lehnstuhl ☺ und beurteilt Sprachdaten (arm-chair linguistics)
  - Kompetenzmodell, generative Tradition
  - Frage: welche Äußerungen kann die muttersprachliche Kompetenz eines Sprechers/einer Sprecherin hervorbringen?
  - Vorgehen: Grammatikalitätsurteile – eine Äußerung ist grammatisch oder nicht
  - (Vorsicht: 'graded grammaticality' (Keller 2001, Featherston 2003))

Psycholinguistische Experimente

- Frage: Wie wird menschliche Sprache verarbeitet?
- Zugriffszeiten, Speicher- und Verarbeitungsmodelle, Fehlerbehandlung, Interaktion mit anderen kognitiven Prozessen
- Vorgehen: Reaktionszeitexperimente (lexical decision), Produktionsexperimente, Bewertungsexperimente, eye tracking, fMRI + andere imaging-Techniken, ...

Datenerhebungen

- (verwandt mit psycholinguistischen Experimenten und Korpuslinguistik)
- man geht mal über'n Flur und fragt die Kollegen
  - was kann man mit den Auskünften anfangen?
- man entwickelt einen Fragebogen nach bestimmten Kriterien und fragt eine ausgewählte Zielgruppe (schriftlich oder mündlich)
- man sammelt Daten nach bestimmten Kriterien → Korpus

Korpora

- Sammlungen von
  - Texten (geschriebener Text, transkribierte gesprochene Sprache) → Textkorpora (text corpora)
  - Sprachdaten (Sprachsignal evtl. mit Transkription, phonetische Annotation) → Sprachkorpora (speech corpora)
  - Sprachdaten mit Transkription und weiterer Information wie Gestik, Mundbewegung etc. → multimodale Korpora

Korpora

- können nach bestimmten Kriterien gesammelt werden oder auch opportunistisch (man nimmt, was man bekommt)
- können riesig groß sein (>200 m Wörter) oder ganz klein
- können auf allen linguistischen Ebenen annotiert sein (Wortart, Semantik, Syntax, ...)

Linguistische Daten

- für psycholinguistische/neurolinguistische Experimente genauso wie für Fragebogenstudien gibt es methodologische Grundlagen
  (z.B. zu filler items, observer's paradox, ethischen Fragen, statistischen Methoden)
- ➢ genauso muss man für die Auswertung von Korpora methodische Grundlagen kennen (viele sind identisch oder eng verwandt mit anderen experimentellen Methoden)

Linguistische Daten

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
qualitativ (kategorial)		qualitativ + quantitativ (probabilistisch)

"Mantra"

- welche Daten geeignet sind, folgt aus einer möglichst präzise gestellten Forschungsfrage

Korpusdaten/Überblick

Linguistische Probleme

- Für welche linguistischen Probleme/Fragestellungen/Hypothesen sind Korpusdaten hilfreich?
- Wie sollten die Daten aussehen (in einer idealen Welt)?
  - Welche Daten?
  - Wie aufbereitet?
  - Wie zugänglich?
- Ihr linguistisches Lieblingsproblem?

Linguistische Bereiche,

- die schon lange quantitative Methoden/Korpora verwenden
  - Soziolinguistik/Sprachvariationsforschung
  - historische Linguistik/Sprachwandelforschung
  - Psycholinguistik
  - Lexikographie/Kollokationsforschung
  - Sprachbeschreibung (Feldforschung)

Soziolinguistik/Dialektforschung

- Wie unterscheiden sich Dialekte/Soziolekte voneinander und von der ‚Standardsprache'?
  - phonologisch
  - in der Lexik
  - in der Syntax
  - ...
- Wie verändern sich Dialekte/Soziolekte?
- Wie werden Dialekte voneinander abgegrenzt?
- Schreiben/sprechen Frauen anders als Männer?
- ...

Parameter

- Annahme:
  Jede Varietät hat ihre eigenen sprachlichen Mittel (qualitativ und quantitativ)
- Experimentchen 1:
  2 Textausschnitte, jeweils die oberste Meldung einer Online-Zeitung zum selben Zeitpunkt (08.10.2008, 10:40)
  - Welche Zeitungen
    (Börsen-Zeitung Online, Bild Online)?
  - Welche Evidenz für Ihre Entscheidung?

Parameter

Europa schnürt Rettungspaket

BZ - Die Mitgliedstaaten der EU wollen große systemrelevante Banken nicht fallen lassen. Die Finanzminister verständigten sich zudem darauf, als Signal der Stabilität den Einlagenschutz der Sparer in Europa zu erhöhen. Die amtierende EU-Ratsvorsitzende, die französische Finanzministerin Christine Lagarde, sagte nach Abschluss der Ministerberatungen in Luxemburg, die EU-Staaten wollten „alle nötigen Schritte unternehmen, um die Robustheit und die Stabilität des Bankensystems in Europa zu stärken". Die Finanzminister hätten sich darauf verständigt, über den Finanz- und Wirtschaftsausschuss in täglichem Kontakt zu stehen und Informationen auszutauschen. Die Ratsvorsitzende sprach von einem „Signal der Geschlossenheit", das Europa aussende. Hilfsaktionen müssten angesichts der prekären Lage der Finanzmärkte „schnell koordiniert und umgesetzt werden".

Blutiges Tierdrama bei den Connors
Sarahs Hunde zerfleischen Nachbars-Kaninchen

Von BIANCA WEINER und SVEN KUSCHEL

Ein aufgebrochener Gitterkäfig, graue Fellfetzen auf dem Boden. HIER haben zwei Promi-Hunde zwei Hauskaninchen zerfleischt!

Schock für die Nachbarn von Popstar Sarah Connor (28): Vergangenen Dienstagmorgen liefen Connors Neufundländer Bailey (4) und ihr neuer Hovawart-Welpe vom 3000-qm-Grundstück (10-Zimmer-Villa, Wert: 800 000 Euro) weg.

Die Tiere schlichen durch die Gärten der Villengegend im Landkreis Oldenburgs. Knapp einen Kilometer entfernt wohnt Manuela C. mit ihrem Mann, ihrer Tochter (3) und zwei Hauskaninchen. Gegen 8.30 Uhr beißen die Hunde zu. Sie drücken das Tor zum Stall auf und zerfleischen die Tiere. Als der Hausbesitzer die Bluttat bemerkt, verscheucht er die Hunde. Die Kaninchen sind tot.

Parameter

Europa schnürt Rettungspaket

BZ - Die Mitgliedstaaten der EU wollen große systemrelevante Banken nicht fallen lassen. Die Finanzminister verständigten sich zudem darauf, als Signal der Stabilität den Einlagenschutz der Sparer in Europa zu erhöhen. Die amtierende EU-Ratsvorsitzende, die französische Finanzministerin Christine Lagarde, sagte nach Abschluss der Ministerberatungen in Luxemburg, die EU-Staaten wollten „alle nötigen Schritte unternehmen, um die Robustheit und die Stabilität des Bankensystems in Europa zu stärken". Die Finanzminister hätten sich darauf verständigt, über den Finanz- und Wirtschaftsausschuss in täglichem Kontakt zu stehen und Informationen auszutauschen. Die Ratsvorsitzende sprach von einem „Signal der Geschlossenheit", das Europa aussende. Hilfsaktionen müssten angesichts der prekären Lage der Finanzmärkte „schnell koordiniert und umgesetzt werden". (Börsen-Zeitung Online)

Blutiges Tierdrama bei den Connors
Sarahs Hunde zerfleischen Nachbars-Kaninchen

Von BIANCA WEINER und SVEN KUSCHEL

Ein aufgebrochener Gitterkäfig, graue Fellfetzen auf dem Boden. HIER haben zwei Promi-Hunde zwei Hauskaninchen zerfleischt!

Schock für die Nachbarn von Popstar Sarah Connor (28): Vergangenen Dienstagmorgen liefen Connors Neufundländer Bailey (4) und ihr neuer Hovawart-Welpe vom 3000-qm-Grundstück (10-Zimmer-Villa, Wert: 800 000 Euro) weg.

Die Tiere schlichen durch die Gärten der Villengegend im Landkreis Oldenburgs. Knapp einen Kilometer entfernt wohnt Manuela C. mit ihrem Mann, ihrer Tochter (3) und zwei Hauskaninchen. Gegen 8.30 Uhr beißen die Hunde zu. Sie drücken das Tor zum Stall auf und zerfleischen die Tiere. Als der Hausbesitzer die Bluttat bemerkt, verscheucht er die Hunde. Die Kaninchen sind tot. (Bild-Online)

Parameter

- Evidenz: thematisch oder sprachlich?
- Experimentchen 2:
  Zeitungsausschnitte zum gleichen Thema, die zum gleichen Zeitpunkt gefunden wurden (08.10.2008, 10:20 Uhr)
  - Welche Zeitungen
    (Spiegel Online, FAZ Online, Bild Online)?
  - Welche Evidenz?
  - Was fällt auf?

Parameter

Dax fällt unter 5000 Punkte

Frankfurt am Main - Belastet von Konjunktursorgen ist der Dax am Mittwoch weiter abgerutscht. Der deutsche Leitindex verlor am Vormittag drastisch und notierte erstmals seit November 2005 wieder unter 5000 Punkten. Gegen 10 Uhr notierte er bei 4915 Zählern und damit mehr als sieben Prozent im Minus. "Es gibt große Angst, dass die Finanzkrise auf die Realwirtschaft durchschlägt und es zu einer Rezession kommt", sagte ein Börsianer. [...]

Ausverkauf an den Börsen

Dax unter 5000 Punkten - Panik in Tokio

Spezial Der Ausverkauf an den internationalen Aktienmärkten geht weiter. Nach Panikverkäufen in Tokio brechen die Aktienkurse in Deutschland ein. Zum ersten Mal seit November 2005 hat der Dax am Mittwoch unter 5000 Punkten notiert. „Die Märkte fangen an, eine globale Rezession einzupreisen", sagte ein Händler. [...]

Parameter

Deutsche Börse
Dax fällt unter 5000 Punkte

Belastet von Konjunktur-Sorgen ist der Dax weiter abgerutscht. Der deutsche Leitindex verlor am Morgen fast acht Prozent auf 4915 Punkte.
„Es gibt große Angst, dass die Finanzkrise auf die Realwirtschaft durchschlägt und es zu einer Rezession kommt“, sagte ein Börsianer.
Belastet wird der deutsche Markt zudem von hohen Kursverlusten an der Wall Street und in Tokio, wo der Nikkei-Index fast zehn Prozent tiefer schloss. [...]

Parameter

Dax fällt unter 5000 Punkte

Frankfurt am Main - Belastet von Konjunktursorgen ist der Dax am Mittwoch weiter abgerutscht. Der deutsche Leitindex verlor am Vormittag drastisch und notierte erstmals seit November 2005 wieder unter 5000 Punkten. Gegen 10 Uhr notierte er bei 4915 Zählern und damit mehr als sieben Prozent im Minus. "Es gibt große Angst, dass die Finanzkrise auf die Realwirtschaft durchschlägt und es zu einer Rezession kommt", sagte ein Börsianer. [...] (Spiegel Online)

Ausverkauf an den Börsen
Dax unter 5000 Punkten - Panik in Tokio

Spezial Der Ausverkauf an den internationalen Aktienmärkten geht weiter. Nach Panikverkäufen in Tokio brechen die Aktienkurse in Deutschland ein. Zum ersten Mal seit November 2005 hat der Dax am Mittwoch unter 5000 Punkten notiert. „Die Märkte fangen an, eine globale Rezession einzupreisen“, sagte ein Händler. [...] (FAZ Online)

Parameter

Deutsche Börse
Dax fällt unter 5000 Punkte

Belastet von Konjunktur-Sorgen ist der Dax weiter abgerutscht. Der deutsche Leitindex verlor am Morgen fast acht Prozent auf 4915 Punkte.
„Es gibt große Angst, dass die Finanzkrise auf die Realwirtschaft durchschlägt und es zu einer Rezession kommt“, sagte ein Börsianer.
Belastet wird der deutsche Markt zudem von hohen Kursverlusten an der Wall Street und in Tokio, wo der Nikkei-Index fast zehn Prozent tiefer schloss. [...] (Bild Online)

- Agenturmeldung wird verwendet (Gaizauskas & Clough, erscheint, über Textwiederverwendung, METER corpus)
- Was kann aus solchen Daten geschlossen werden?

Soziolinguistik: Wunsch

- Texte aus den verschiedenen Soziolekten/Dialekten/Varietäten, die in jeder anderen Hinsicht vergleichbar sind
- Annotation auf allen Ebenen: Phonologie, Syntax, Pragmatik etc.
- Methoden, Texte quantitativ zu vergleichen (mathematische Modelle)

Sprachwandel

- Wann hat man angefangen, *you* statt *thou* zu verwenden (wie verändert sich die Anrede im Deutschen)?
- Allgemeiner: Wann hört man auf, ein bestimmtes Wort/eine bestimmte Konstruktion/eine bestimmte phonetische Variante zu verwenden?
- Ist ein bestimmter Text mittelhochdeutsch oder eher schon neuhochdeutsch?
- Wie kann man Sprachwandel mathematisch modellieren?

Sprachwandel: Wunsch

- Texte verschiedener Sprachstufen, die in anderer Hinsicht vergleichbar sind
  (besondere Schwierigkeiten: Verfügbarkeit, Edition, keine Sprachdaten vorhanden)
- Annotation auf allen linguistischen Ebenen
- mathematische Modelle und Methoden

Psycholinguistik

- Beeinflusst die Worthäufigkeit die Reaktionsgeschwindigkeit in der Erkennung (lexical decision)?
- Spielen die Häufigkeiten morphologisch verwandter Wörter vielleicht auch eine Rolle?
- bei ambigen Wörtern: welche Rolle spielen die Häufigkeiten der einzelnen Lesarten?
- Wie wird die sprachliche Performanz von anderen Faktoren beeinflusst? Wie kann man diese messen? Z.B.: Wie verändert sich sprachliches Verhalten in Stresssituationen?

Psycholinguistik: Wunsch

- repräsentative Textsammlungen, aus denen Worthäufigkeiten ermittelt werden können
- Annotation auf verschiedenen Ebenen, z.B. morphologisch und semantisch

Lexikographie

- Wie wird ein bestimmtes Wort verwendet? (Kontexte)
- Wie häufig sind die einzelnen Lesarten?
- Welche Wörter werden häufig zusammen verwendet (Kollokationen)?
- Welche Wörter werden nicht mehr verwendet?
- Welche Wörter werden in einer bestimmten Fachsprache verwendet?

auf

- *... ist es günstiger, die Babys auf den Bauch zu legen ....*
- *die Kinderwagen erhielten extra große Panoramafenster, um dem Säugling einen Blick auf seine Umgebung zu ermöglichen*
- *entsprechend fällt auch die Antwort des Kinderarztes auf die Frage aus, wie ...*
- *... die Krise des Wohlfahrtsstaates einzig und allein auf dem Feld der Ökonomie bewältigen zu können ...*
- *doch Damaskus drängt, den Wahltermin auf Juni vorzuziehen ...*
- *ein Resultat läßt allerdings auf sich warten*
- *damals, als 20jähriger verdingte er sich auf der Insel Hiddensee als Kellner*

Lexikographie: Wunsch

- ausgewogenes großes Korpus für allgemeines Lexikon; Fachkorpora für Fachlexika
- Annotation auf allen linguistischen Ebenen, insbesondere auch syntaktische Struktur
- gute Suchwerkzeuge, mit denen man z.B. Wörter einer bestimmten Wortart/in einer bestimmten Konstruktion suchen kann
- mathematische Verfahren, mit denen man ‚Kollokationsstärke' errechnen kann

Traditionell nicht quantitative Bereiche,

- die aber doch Korpusdaten verwenden können
  - (generative) Syntax
  - (lexikalische) Semantik
  - Phonologie
  - Morphologie

Syntax

- Ist eine bestimmte Konstruktion wirklich ungrammatisch?
- Ist eine bestimmte Konstruktion häufig/wahrscheinlich?
- Beispiel: Wortstellungsvarianten im deutschen Mittelfeld (Heylen 2004, Kempen & Harbusch 2004)
- Wunsch
  - möglichst große, syntaktisch annotierte Korpora

(Lexikalische) Semantik

- Wie wird ein bestimmtes Wort verwendet?
- Wie verteilen sich die Lesarten?
- Welche Verwendungen sind kollokativ?
- Kommt das Wort auch metaphorisch vor?
- Welches Wissen hilft bei der Auflösung von Anaphern/Epithetaketten?
- Wunsch:
  - großes Textkorpus mit Lesartenmarkierung

Marathon in Berlin
Gebrselassie läuft wieder Weltrekord
Von Michael Reinsch, Berlin

28. September 2008 „Jetzt ist alles egal: Es ist eine 2:03. Es ist erledigt." Haile Gebrselassie ist am Sonntag beim Berlin-Marathon seinen 26. Weltrekord gelaufen, als erster Mensch der Welt auf den 42,195 Kilometern eine Zeit unter 2:04 Stunden - exakt eine Sekunde darunter. Und er war erleichtert. In der Vorbereitung war der 35 Jahre alte Äthiopier in den Bergen bei Addis Abbeba zwanzig Kilometer in glatt 58 Minuten gelaufen - und hatte nach diesem Test Krämpfe. Sein Arzt erteilte ihm sechs Tage Laufverbot. Die vorletzte Woche, bevor er 2:03,59 Stunden lief, trainierte der erfolgreichste Läufer der Welt ausschließlich auf dem Fahrradergometer. „Dies ist der wichtigste Weltrekord von allen", sagte er nun in Berlin. „Für diesen Rekord muss alles stimmen: das Wetter, die Tempomacher, die Form." Allein der Veranstalter belohnte ihn mit 130.000 Euro Sieg- und Rekordprämie; zusätzlich zum Antrittsgeld und Prämien seiner Sponsoren. (FAZ Online)

Phonologie

- Wie werden Fremdwörter ausgesprochen?
- Kann man anhand der Prosodie Lesarten unterscheiden?
  dass er den Laden leer kauft
- Wie kann man Akzente klassifizieren?
- Wunsch:
  - möglichst natürlich aufgenommene Sprachdaten mit phonetischer und phonologischer Annotation

Ungeordnet

- Wieviele Wörter kommen in einer bestimmten Textsorte/Sprechergruppe/etc. vor?
- Wieviele Wörter gibt‘s eigentlich?
- Wie wahrscheinlich ist es, dass man ein noch nie gesehenes Wort (eines bestimmten Wortbildungsmusters) findet, wenn man zu 200 m Wörtern 100 Wörter hinzufügt?
- Wie unterscheiden sich übersetzte Texte von nicht übersetzten Texten?

Ungeordnet

- Was sind typische Tippfehler?
- Was sind typischer Lernerfehler?
- Welche Regularitäten gibt's im Spracherwerb?
- Wie lernen Kinder eigentlich orthographische Regeln?
- Von welchem Autor ist ein bestimmter Text? (Stylometry, auch forensische Linguistik)
- ...

Zum Erinnern: Korpuslinguistik

- beschäftigt sich damit, welche linguistischen Fragen anhand (großer) Textmengen behandelt werden können
- wie diese Texte ausgewählt und akquiriert werden
- wie sie aufbereitet (annotiert) werden und
- wie sie durchsucht/bearbeitet werden können
- (welche mathematischen Modelle für welche Probleme anwendbar sind)

Korpusdesign

Korpuserstellung - Themen

- feste Korpora vs. wachsende Korpora
- Repräsentativität (das ‚R-Wort')
- Ausgewogenheit
- Datenakquisition

- Biber (1993), Lauer (1995), Hundt (erscheint), Claridge (erscheint), Wichmann (erscheint), ... sehr viel ...

Was mach ich, wenn ich ein Korpus will?

Zuerst muss ich eine linguistische Fragestellung haben (Mantra!). Abhängig davon brauche ich

- ein Korpus mit Texten, die speziell für diese Fragestellung erzeugt wurden oder
- ein Korpus mit Texten, die zu anderen Zwecken erzeugt wurden
  - heterogen
    (in Bezug auf einen bestimmten Parameter)
  - homogen
    (in Bezug auf einen bestimmten Parameter),

Referenzkorpus

- ein Referenzkorpus
  (reference corpus, fixed corpus)
  - feste Größe, Zusammensetzung bekannt
  - weit verfügbar, Standard, Ergebnisse können dupliziert werden
  - veraltet irgendwann (für bestimmte Fragestellungen)

Monitorkorpus

- ein wachsendes Korpus
  (monitor corpus)
  Zusammensetzung und Größe evtl. nicht bekannt (manchmal gibt's aber bestimmte Herausgabedaten/Versionen)
  - für lexikographische Zwecke gut geeignet, diachron

Datensammlung: opportunistisch

- opportunistisch: ich nehm alles, was ich bekommen kann (Werbung, Romane, Märchen, die Bibel, Zeitungstexte, Foren, email, ...)
  - Vorteil: Verfügbarkeit, Kosten
  - Nachteile: naja ...
    unausgewogen, nicht repräsentativ (was wird eigentlich erforscht?), Parameter können nicht kontrolliert werden, evtl. ist keine einheitliche Annotation möglich

Italienische Reise 1

Den 3. September 1786.

Früh drei Uhr stahl ich mich aus Karlsbad, weil man mich sonst nicht fortgelassen hätte. Die Gesellschaft, die den achtundzwanzigsten August, meinen Geburtstag, auf eine sehr freundliche Weise feiern mochte, erwarb sich wohl dadurch ein Recht, mich festzuhalten; allein hier war nicht länger zu säumen. Ich warf mich ganz allein, nur einen Mantelsack und Dachsranzen aufpackend, in eine Postchaise und gelangte halb acht Uhr nach Zwota, an einem schönen stillen Nebelmorgen.

(Johann Wolfgang von Goethe)

Italienische Reise 1

Den 3. September 1786.

Früh drei Uhr stahl ich mich aus Karlsbad, weil man mich sonst nicht fortgelassen hätte. Die Gesellschaft, die den achtundzwanzigsten August, meinen Geburtstag, auf eine sehr freundliche Weise feiern mochte, erwarb sich wohl dadurch ein Recht, mich festzuhalten; allein hier war nicht länger zu säumen. Ich warf mich ganz allein, nur einen Mantelsack und Dachsranzen aufpackend, in eine Postchaise und gelangte halb acht Uhr nach Zwota, an einem schönen stillen Nebelmorgen.

(Johann Wolfgang von Goethe)

Re: Laufsocken
Antwort schreiben | Zurück zum Forum

von PeterS am 19.Apr.2002 13:59 (vorlesen)

>>Welche Marken bzw. Socken koennt Ihr denn empfehlen bzw. auf was muss ich beim Sockenkauf achten.

Habe selber einige Falke. Diese sind IMHO wirklich sehr gut. Bei "billigeren" Modellen wie Tchibo muß man u.U. Abstriche in der Qualität machen, obwohl cih selber auch einige Paar habe und sie sich bisher ganz gut halten. Teilweise gibt es die Falke als Ware mit kleinen Farbfehlern o.ä. erheblich billiger

Socken eher 'ne Idee zu klein als zu groß, sonst gibt's wirklich Blasen.

Gruß Peter

aus einem Chat-forum über Rennkleidung

Re: Laufsocken
Antwort schreiben | Zurück zum Forum

von PeterS am 19.Apr.2002 13:59 (vorlesen)

>>Welche Marken bzw. Socken koennt Ihr denn empfehlen bzw. auf was muss ich beim Sockenkauf achten.

Habe selber einige Falke. Diese sind IMHO wirklich sehr gut. Bei "billigeren" Modellen wie Tchibo muß man u.U. Abstriche in der Qualität machen, obwohl cih selber auch einige Paar habe und sie sich bisher ganz gut halten. Teilweise gibt es die Falke als Ware mit kleinen Farbfehlern o.ä. erheblich billiger

Socken eher 'ne Idee zu klein als zu groß, sonst gibt's wirklich Blasen.

Gruß Peter

aus einem Chat-forum über Rennkleidung

§ 27
Bestellung und Geschäftsführung des Vorstands

(1) Die Bestellung des Vorstands erfolgt durch Beschluss der Mitgliederversammlung.

(2) Die Bestellung ist jederzeit widerruflich, unbeschadet des Anspruchs auf die vertragsmäßige Vergütung. Die Widerruflichkeit kann durch die Satzung auf den Fall beschränkt werden, dass ein wichtiger Grund für den Widerruf vorliegt; ein solcher Grund ist insbesondere grobe Pflichtverletzung oder Unfähigkeit zur ordnungsmäßigen Geschäftsführung.

(3) Auf die Geschäftsführung des Vorstands finden die für den Auftrag geltenden Vorschriften der §§ 664 bis 670 entsprechende Anwendung.

§ 27
Bestellung und Geschäftsführung des Vorstands

(1) Die Bestellung des Vorstands erfolgt durch Beschluss der Mitgliederversammlung.

(2) Die Bestellung ist jederzeit widerruflich, unbeschadet des Anspruchs auf die vertragsmäßige Vergütung. Die Widerruflichkeit kann durch die Satzung auf den Fall beschränkt werden, dass ein wichtiger Grund für den Widerruf vorliegt; ein solcher Grund ist insbesondere grobe Pflichtverletzung oder Unfähigkeit zur ordnungsmäßigen Geschäftsführung.

(3) Auf die Geschäftsführung des Vorstands finden die für den Auftrag geltenden Vorschriften der §§ 664 bis 670 entsprechende Anwendung.

Die Spanferkel-Braten werden 1-2 Tage bevor die Ware in den Versand geht in einen Karton verpackt der mit 1 cm Styropur ausgelegt ist. Die Kartons werden in einen Tiefkühlraum der -18 Grad hat gelagert und das heisst das die Ware Tiefgekühlt auf die Reise geht. So kann es nicht passieren das die Kühlkette unterbrochen wird. Dennoch ist die Ware immernoch 6-8 Tage bei einer Kühlung von +2 grad haltbar. […] Falls Sie Fragen oder besondere Wünsche haben schreiben Sie uns eine Mail, wir werden jede Frage Beantworten. Hallo Feinschmecker, Gourmets, Wildliebhaber und alle die gerne und gut Essen wollen. Wenn man Heute die Medienberichte verfolgt und das sieht was da in Sachen Fleisch, Qualität und Frische sieht was da abgeht da kommt einem das Grausen. Darum können Sie ganz Sicher sein daß Sie bei uns nur Top Qualtät zu einem verbünftigen Preis bekommen. Denn unser Motto ist das was ich nicht Esse traue ich keim anderen zu.

[http://cgi.ebay.de/, Kategorie Feinschmecker]

- also gut also ja * mhm * äh * also wir ham jetzt vor * ein/ eineinhalb jahren ham wa uns seit eineinhalb jahren ham wa uns n hund angeschafft * und an für sich auch auf druck * mit unsrer kinder die so gerne n schönes großes tier wollen * und auch mit für MIch * und jetzt geht's immer um die pflege des hundes * weil die erziehung ham die eltern übernommen das war ja zu erwarten * äh aber was ich halt vermisse * äh is des engagement der kinder in bezug auf den hund * also freiwillig mit ihm spazierengehen […]

[aus den Mutter-Tochter-Dialogen, IDS Mannheim]

ich geb dir noch n andern tiprobier mal nach annahme der quest ODER OHNE se anzunehmen von draussen, also um dwen buckel rum, in die höhle zu gehn. erstens isses kürzer bis zur truhe und zweitens nich so nervig, dann kannst die warane – die dir schon mit 4 auf einem haufen grad am anfang – ziemlichen trouble machen können noch n wenig aufsparen bisde z b mit nem feuerball aufräumen kannst.

[Forumstext http://www.worldofgothic.de, 09.01.2007]

Repräsentativität …

- Begriff aus der Statistik:
  Man möchte bestimmte Eigenschaften einer Menge (von Personen, Wörtern, Bäumen etc.) untersuchen, die aber zu groß ist, um in ihrer Gesamtheit angeschaut werden zu können.
- Daher zieht man aus dieser sogenannten Grundgesamtheit (population) kleinere Stichproben
  - zufällig (random sample)
  - repräsentativ (representative sample)

Repräsentativität

- eine repräsentative Stichprobe muss für die Grundgesamtheit ‚typisch' sein, d.h. für einen bestimmten Parameter (z.B. Alter, Bildungsstand) die gleichen Anteile wie in der Grundgesamtheit enthalten
- Wichtig: Repräsentativität bezieht sich immer auf vorgegebene Parameter

repräsentative Korpora

- Idee: ein Korpus soll eine Sprache (oder einen Dialekt oder einen Soziolekt etc.) repräsentativ abbilden
  (das ist oft das Ziel der sog. ‚Nationalkorpora' oder ‚Referenzkorpora')
- aber man kennt meistens die Zusammensetzung der Grundgesamtheit nicht
  - welche Parameter sind wichtig? (gesprochen vs. geschrieben, Rezeption oder Produktion, Variation zwischen Sprechern etc.)
  - wie sind diese verteilt?

das ‚R-Wort'

- die meisten Korpora, die sich 'repräsentativ' nennen, können die Grundgesamtheit nicht angeben
- dazu sehr viel Diskussion/Aufsätze etc.

➢ Vorsicht beim Gebrauch des Begriffs ‚Repräsentativität'

kann man

- ein repräsentatives Korpus erstellen für
  - die Werke von Goethe?
  - alle schriftlichen Aufzeichnungen, die sich in diesem Moment im Raum befinden?
  - das Althochdeutsche?
  - die althochdeutsche Überlieferung?

Datensammlung: ausgewogen

- ein ausgewogenes (balanced) Korpus versucht, Texte nach gegebenen Parametern zusammenzustellen und in vorher festgelegten Mengen zu repräsentieren

➢ Beispiele: *Brown Corpus* (American English), *Lancaster-Oslo-Bergen Corpus* (LOB, gleiche Zusammenstellung, British English), *British National Corpus* (BNC), *Deutsches Referenzkorpus* (DeReKo), ... (alle synchron!)

Datensammlung: homogen, spezifisch

- man kontrolliert einen Parameter (homogen)
  - Texte eines Autors
  - Texte einer bestimmten Bevölkerungsgruppe
  - Texte eines Genres
  - Texte aus einer bestimmten Zeit
  - ...

➢ Beispiele: *Corpus of Early English Correspondence*, *Nibelungenlied*, Zeitungskorpora (*Wall Street Journal, Frankfurter Rundschau, Wendekorpus*, ...), Lernerkorpora, ...

Datensammlung: homogen, spezifisch

- Vorteile: geeignet für gezielte Forschung, man kann andere Parameter untersuchen, einheitliche Vorverarbeitung möglich
- Nachteile: evtl. nur wenig Material verfügbar, evtl. teuer, evtl. Copyrightprobleme
- und: man kann keine Fragen über ‚die Sprache' beantworten

Datensammlung: ausgewogen

- Welche Textsorten sollen aufgenommen werden?
- Wie groß sind die Anteile der jeweiligen Textsorten?

Beispiel für ein ausgewogenes Korpus: BNC

- Das British National Corpus ist eines der bekanntesten existierenden ausgewogenen Korpora
  (American National Corpus mit ähnlichem Design wird größer, bisher nicht so viel genutzt)
- Referenzkorpus (100 m Wörter)
- gute Vorverarbeitung, eigenes Suchtool
- weit erhältlich

Beispiel Englisch: BNC

- variiert über eine Anzahl Parameter
- 90% geschriebene Sprache,
  10% gesprochene Sprache
- gesprochene Sprache: nach Thema (educational, business, institutional, leasure, others), nach demographischen Parametern (Alter, soziale Gruppe, Geschlecht, Region)

Beispiel Englisch: BNC

- geschriebene Sprache:
  Zeit (1960 – 1974, 1975 – 1993),
  Medium (Buch, Zeitschrift, div veröffentlicht (Ephemera), div unveröffentlicht, ...),
  Thema (‚informativ', ‚imaginativ', ...),
  ‚Sprachebenen',
  Informationen über AutorIn,
  Informationen über Publikum, ...
- Chunks (samples) von nicht mehr als 40.000 Wörtern

Beispiel Englisch: ICE

- das BNC ist relativ homogen in Bezug auf Region
- Varietäten des Englischen aus anderen Regionen werden nicht berücksichtigt
- Das International Corpus of English sammelt Daten anderer Regionen (Westafrika, Australien, Canada, Singapur, Hong Kong, ...) – jedes Subkorpus (ca. 1 m Wörter) wird nach den gleichen Designprinzipien aufgebaut

Beispiel Englisch: Helsinki Corpus

- historisches Korpus: Old English bis Early Modern English, 1 m Wörter
- ausgewogen nach Region (Dialekt) und Genre
- Schwierigkeit: Verfügbarkeit in den älteren Sprachstufen

Beispiel Deutsch: Akademiekorpus

- erstellt als Textgrundlage für ein Lexikographieprojekt (Digitales Wörterbuch der Deutschen Sprache, http://www.dwds.de/textbasis)
- Kernkorpus: 100 m Wörter, in Dekaden von 1900 – 2000, annotiert, online verfügbar
- Erweiterungskorpus: 980 m Wörter, opportunistisch, nicht zugänglich

Beispiel Deutsch: DeReKo

- DeReKo (Deutsches Referenzkorpus), Projekt
  - deutsche Gegenwartssprache 1965-2000
  - entwickelt am IMS, Stuttgart, IDS Mannheim und SfS Tübingen
  - Textauswahlprinzipien nicht dokumentiert (?)
  - linguistische Annotation bis zu Chunks
- DeReKo (Deutsches Referenzkorpus), alle Korpora des IDS Mannheim
- 3,4 Milliarden Wörter, http://www.ids-mannheim.de/kl/projekte/korpora/ , Metadaten für eigenes ad hoc Korpusdesign, annotiert
- online verfügbar über Cosmas II

Beispiel Deutsch: Verbmobil

- Korpus wurde gesammelt als ein Beispiel und Trainingskorpus für ein großes computerlinguistisches Projekt zur maschinellen Übersetzung von gesprochener Sprache
- 'gestellte' Dialoge, Themen: Terminabsprache, Reiseplanung, Abendgestaltung

Zum Erinnern

- das Korpusdesign hängt von den Fragestellungen ab, die man mit dem Korpus beantworten möchte
- der Begriff ‚repräsentativ' ist problematisch, da man die Grundgesamtheit nicht kennt
- auch ausgewogene Korpora bilden nie ‚die Sprache' ab!

Vorverarbeitung: Themen

- Tokenisieren
- Wortartzuweisung (Tagging)
- Lemmatisieren

Einschub: Suchen

- in jedem Text kann man nach Zeichenketten suchen
- viele Such- und Auswertungsprogramme bieten zusätzlich zur genauen Zeichenkettensuche Wild Cards oder andere Suchoptionen an
- typische Ergebnis-Darstellung: keyword-in-context (kwic) Konkordanzen (ganz altes Konzept!)
- hier nur als Beispiele, Korpus: Parlamentsreden, Suchprogramm CQP (entwickelt an der Universität Stuttgart)

Einschub: Suchen

- PARLAMENT-schlafen.html
- PARLAMENT-dunkeln.html

Vorverarbeitung

- für viele linguistische Fragestellung muss ein Korpus vorverarbeitet werden
- Normalisierung auf Zeichenebene
- festlegen von kleinsten Einheiten (Tokenisierung)
- Anreichern mit linguistischer Information (Annotation)

Vorverarbeitung

- auf allen Vorverarbeitungsebenen werden (linguistische) Entscheidungen getroffen
- viele Vorverarbeitungstechniken sind fehleranfällig
- generelle Möglichkeiten:
  - statistisch vs. regelgeleitet vs. hybrid
  - viel oder wenig linguistisches Wissen

Tokenisierung

- Token
  - kleinste Einheit für Annotation und Suche
  - „eine von Leerzeichen (das umfasst Tabulatorzeichen und Zeilenumbrüche) oder Interpunktion begrenzte Folge von Buchstaben oder Ziffern“
    (Evert & Fitschen 2001, 371)
  - ≈ graphemisches Wort (?)
  - guter Überblick in Schmid (erscheint)

Tokenisierung

- alle Satzzeichen, Interpunktionszeichen etc. werden von den Wörtern abgetrennt (→ reguläre Ausdrücke, Listen, Heuristiken)

"Wir sind in der Defensive", sagt etwa Larry Williams

"	*Wir*	*sind*	*in der*	*Defensive*	*"*
,	*sagt*	*etwa*	*Larry*	*Williams*	

Aber

- sind solche rein graphemisch definierten Tokens die Einheit, mit der man weiterarbeiten will?

Tokenisierung

- größere Einheiten können danach wieder zusammengefügt werden
  - Zahlen:
    20 000, 030-2093 9799, BLZ 111 111 11
    → reguläre Ausdrücke, Heuristiken
  - Namen, feste Verbindungen:
    New York, Weil der Stadt, en passant, Vereinte Nationen, der Deutsche Bundestag
    → Listen
    enthalten schon Entscheidungen, sind endlich

Tokenisierung

- Partikelverben (Beispiele aus den Parlamentsreden)
  - *Dass man in der Landwirtschaft auf die Freigabe des Anbaus wartet und Hanf anbauen will ,*
  - *Wird Hanf auf stillgelegten Flächen angebaut , gibt es anstelle der Beihilfe*
  - *George Washington , liebe Kolleginnen und Kollegen , und Thomas Jefferson bauten auf ihren Plantagen Cannabis an .*
  - ➢ syntaktische Analyse nötig

Tokenisierung

- einige Tokens können noch aufgeteilt werden
  - *beim, zum, gibt's, siehste*
    → Listen, reguläre Ausdrücke, Heuristiken
    (will man Informationen über die ursprüngliche Form behalten?)
- bestimmte Sonderzeichen in Formeln u. ä.: Desambiguierung schwierig
  42,195 , 8:04:08
  Patienten der WOS-(West of Scotland)-Studie

Satzendeerkennung

- automatische Desambiguierung des • nicht trivial, aber wichtig für alle späteren Verarbeitungsschritte
  - Satzende
  - Abkürzung (Gehört der Punkt zum Token? Was macht man dann, wenn der Satz mit einer Abkürzung endet?)
    Ev. Elisabeth-Krankenhaus, usw., George W. Bush
  - in Zahlen
    5. Versuch, 3.100 Teilnehmer, 7.00 Uhr Ortszeit
  - in bestimmten chemischen/medizinischen Namen
    E.coli-Bakterien

Satzendeerkennung

- Liste von Abkürzungen, die einen Punkt enthalten; alle anderen Punkte sind dann Satzendepunkte
  Problem: Liste ist statisch, Abkürzungen können aber produktiv gebildet werden, Übertragbarkeit auf andere Domänen schwierig
- regelgeleitet (Heuristiken)
  (folgendes Wort wird groß geschrieben etc.)
  Problem: mehr linguistisches Wissen notwendig, fehleranfällig, komplexe Regeln
- statistische Verfahren
  mit Trainingskorpus, Assoziationsmaßen
  Problem: Fehleranfälligkeit, Übertragbarkeit auf andere Domänen nicht immer gewährleistet (man braucht ein neues Trainingskorpus)
  siehe auch: Kiss & Strunk (2006), Schmid (erscheint)

Tokenisierung

- auf dieser Ebene werden bereits Entscheidungen getroffen/Fehler gemacht, die sich auf alle späteren Vorverarbeitungsschritte auswirken
- PARLAMENT-beim.html
- PARLAMENT-new.html
- einige Entscheidungen lassen sich ohne weiteres linguistisches Wissen nicht sicher treffen

Taggen

- beim Taggen ordnet man
  (im Prinzip beliebige) linguistische Informationen
  (im Prinzip beliebigen) Texteinheiten zu
- verkürzt wird ‚Taggen' meist für Wortartzuweisung (part-of-speech, pos) verwendet

Wortarttaggen

- tokenbasiert
- als Eingabe für weitere computerlinguistische Anwendungen
  (Parser, sem. Verarbeitung, MT, ...)
- man möchte ambige Wortformen einschränken können
- PARLAMENT-meinen.html

Taggen

- man möchte alle Wörter in einer bestimmten Position in einer gegebenen Sequenz von Wortarten finden
  Bsp.: Sequenzen, in denen vor einem Nomen drei Adjektive stehen (Verbmobilkorpora, Suche mit CQP)
  Was halten Sie von <richtig schönen deutschen Universitätsstädten>?
  Aber bitte vergessen Sie auch nicht, ein schönes, <frisch gezapftes Radeberger Bier>.
  und das geht <ganz schlecht nächste Woche> bei mir die <einzigste wirklich freie Woche>, die ich habe ist im August
- hm ...

Taggen

- Ergebnis: jedes Token erhält ein pos-Tag
  was/PWS *halten*/VVFIN *Sie*/PPER *von*/APPR *richtig*/ADJD *schönen*/ADJA *deutschen*/ADJA *Universitätsstädten*/NN *?*/$
- Tagset: man muss sich eine Menge von pos-Tags definieren

NB: Wortarten

- Wortarten können schwer 'definiert' werden
- Anzahl der Wortarten wurde schon in der klassischen griechischen Grammatik diskutiert. Dionysius Thrax (100 v. Chr.): *Nomen, Verb, Pronomen, Präposition, Adverb, Konjunktion, Partizip, Artikel*

NB: Wortarten

- positionsbasierte Definitionen
  (Position im Satz/ relativ zu anderen Wörtern)
- merkmalsbasierte Definitionen
  (bestimmte Flexionsmerkmale, semantische Merkmale etc.)
- syntaktische Funktion, morphologische Eigenschaften, ...
  sprachübergreifende Definition?

Taggen: Tagset

- Tagset: Abwägen zwischen Genauigkeit (soll z.B. die grammatische Information kodiert werden?) und Handhabbarkeit
  Schlafmütze/NN
  Schlafmütze/NN NOM SG
  Schlafmütze/NN GEN SG
  Schlafmütze/NN DAT SG
  Schlafmütze/NN AKK SG

Taggen: Tagsets

Tagsets für deutsche Korpora (Rapp & Lezius 2001):

IBM Heidelberg	689	33
Uni Münster	143	54
STTS (Stuttgart/Tübingen Tag Set)		50
ISSCO (Genf)		56
Morphy (Paderborn)	500	52
...		

Taggen

- automatisches Taggen meist hybrid (Lexikonlookup & Trigramme, siehe z.B. Manning & Schütze 1998, Schmid, erscheint)
- PARLAMENT-meinen-pposs.html
- PARLAMENT-meinen-VV.html
- beim Taggen gibt es Fehler, oft systematisch
- je unähnlicher der zu taggende Text dem Trainingskorpus (typisch: Zeitung) ist, desto mehr Fehler

Lemmatisierung/ morphologische Analyse

- bestimmte Entscheidungen kann man nur mit (flexions)morphologischer Information treffen, daher wollen wir die Token lemmatisieren (auf ein Lemma zurückführen) – im gleichen Schritt erhalten wir die morphologische Analyse der Wortform
  (Vorsicht: ‚morphologisch' bezieht sich hier immer auf Flexionsmorphologie!)

kleine Erinnerung ☺

- Lemma: abstrakte ‚Grundform': Name mit Flexionsklasse
- Wortform: bestimmte Form in einem Paradigma
  Problem: Synkretismus
- grammatisches Wort: Wortform mit eindeutiger morphologischer Zuweisung

Lemmatisierung

- zum Lemmatisieren braucht man also
  - ein Lexikon, in dem die Lemmata mit ihrer Flexionsklasse stehen
  - ein Regelapparat, der flektierte Formen auf die Lemmata zurückführen kann und dabei die Wortform analysiert
    (meistens Two-Level-Morphology)
- oder
  - ein Vollformenlexikon
- Qualität hängt von den verwendeten Ressourcen ab

Lemmatisierung

- PARLAMENT-schlafen-lemma.html

Lemmatisierung und Taggen

- mehrere Schritte
- einige Ambiguitäten bleiben erhalten
  Bank
  lying
- Problem: unbekannte Wörter
  - heuristische Kompositaanalyse (Morphy, dmor)
  - ‚echte' morphologische Analyse (DeKo, WordManager, GerTWOL, ...)

zum Erinnern ...

- Tagger & Lemmatisierer verwenden Lexika (& evtl Regeln)
  Qualität abhängig vom Lexikon
- Tagger verwendet oft statistische Analyse, trainiert auf einem handgetaggten Trainingskorpus
  Qualität abhängig vom Trainingskorpus
- Tokenisierer (und vielleicht die anderen Komponenten) verwendet Heuristiken
  Qualität abhängig von den Heuristiken

zum Erinnern ...

- in allen Komponenten werden linguistische Entscheidungen getroffen
- in allen Komponenten können Fehler gemacht werden
- weitere Annotationsebenen: Syntax, Phonologie, Gesten, Anaphern, Informationsstruktur, narrative Kategorien, ... (eigentlich alles, was man erforschen möchte)

Experimentdesign

Plan

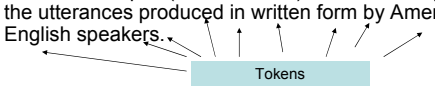
- Tokens, Types
- Datentypen und Kategorisierung
- Fragestellungen, für die man quantitative Daten braucht
- deskriptive Statistik
- inferentielle Statistik: Hypothesen testen

Tokens und Types

- Token: laufendes Wort (Zeichenkette zwischen Leerzeichen)
- Typ: Zuordnung von Tokens zu Kategorien
  - Wortformtypen
  - Lemmatypen
  - grammatisches-Wort-Typen
  - ...

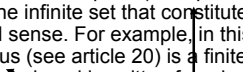
Linguists look for generalizations and explanations of various kinds for linguistic phenomena. While the interest is usually in an *intensional* view of these phenomena, to be explained in terms of the human language competence, such competence cannot be directly observed. Thus, evidence has to come from an external reflection of it, i. e., it has to be based on an *extensional* view of language. According to this extensional view, a language is defined as the set of all utterances produced by speakers of the language (with all the paradoxes that this view implies see, e. g., Chomsky 1986, chapter 2). Corpora are finite samples from the infinite set that constitutes a language in this extensional sense. For example, in this perspective, the Brown corpus (see article 20) is a finite sample of all the utterances produced in written form by American English speakers. [Baroni & Evert, erscheint]

Linguists look for generalizations and explanations of various kinds for linguistic phenomena. While the interest is usually in an *intensional* view of these phenomena, to be explained in terms of the human language competence, such competence cannot be directly observed. Thus, evidence has to come from an external reflection of it, i. e., it has to be based on an *extensional* view of language. According to this extensional view, a language is defined as the set of all utterances produced by speakers of the language (with all the paradoxes that this view implies see, e. g., Chomsky 1986, chapter 2). Corpora are finite samples from the infinite set that constitutes a language in this extensional sense. For example, in this perspective, the Brown corpus (see article 20) is a finite sample of all the utterances produced in written form by American English speakers.



Linguists look for generalizations and explanations of various kinds for linguistic phenomena. While the interest is usually in an *intensional* view of these phenomena, to be explained in terms of the human language competence, such competence cannot be directly observed. Thus, eviden Wortformtypen an external reflection of it, i. e., it has to be based on an *extensional* view of language. According to this extensional view, a language is defined as the set of all utterances produced by speakers of the language (with all the paradoxes that this view implies see, e. g., Chomsky 1986, chapter 2). Corpora are finite samples from the infinite set that constitutes a language in this extensional sense. For example, in this perspective, the Brown corpus (see article 20) is a finite sample of all the utterances produced in written form by American English speakers.

Linguists look for generalizations and explanations of various kinds for linguistic phenomena. While the interest is usually in an *intensional* view of these phenomena, to be explained in terms of the human language competence, such competence cannot be directly observed. Thus, evidence has to come from an external reflection of it, i. e., it has to be based on an *extensional* view of language. According to this extensional view, a language is defined as the set of all utterances produced by speakers of the language (with all the paradoxes that this view implies see, e. g., Chomsky 1986, chapter 2). Corpora are finite samples from the infinite set that constitutes a language in this extensional sense. For example, in this perspective, the Brown corpus (see article 20) is a finite sample of all the utterances produced in written form by American English speakers.



Variablen: Was wird gezählt/gemessen?

- kategorial (ja/nein)
  - nominal (keine Ordnung zwischen den Kategorien)
  - ordinal (numerische Ordnung zwischen den Kategorien)
- kontinuierlich

- *jede* Zählung, die auf Kategorien beruht, erfordert eine Interpretation der Daten

Einschub: Kategorisierung

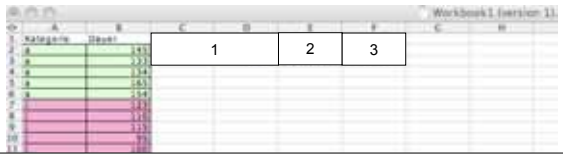
- Entscheidung über einen Untersuchungsgegenstand
- Welche Merkmale sind wesentlich?
- Können die Merkmale immer eindeutig zugewiesen werden?

Einschub: Kategorisierung

- Kategorisierung von eigentlich kontinuierlichen Daten

Wo ist ein Spatium?

was schreibt der Lerner?



Alle Vokale, die länger sind als 120ms werden der Kategorie LANG zugeschlagen, alle Vokale, die kürzer sind, werden der Kategorie KURZ zugeteilt:

	lang	kurz	total
/a/	5	0	5
/I/	1	4	5
Total	6	4	10

Einschub: Kategorisierung

- Kategorisierung von eigentlich kontinuierlichen Daten
- Interpretation der Daten schwierig
- Angabe von notwendigen und hinreichenden Bedingungen schwierig

Einschub: Kategorisierung

- Was ist ein Partikelverb?
- Genauer: Was sind die notwendigen und hinreichenden Bedingungen dafür, dass ein Verb als Partikelverb klassifiziert wird?
- Komplexität, Trennbarkeit?
  - *anfangen, aufgeben, einschlafen*
  - *kranklachen, totschlagen, wachküssen*
  - *autofahren, fußballspielen, fernsehengucken*

Einschub: Kategorisierung

- Kategorisierung oft schwierig
- das hat (erhebliche!) Auswirkungen auf die Ergebnisse der jeweiligen Studie
- daher (so viel Klarheit über die Daten und ihre Interpretation wie möglich):
  - Angabe von notwendigen und hinreichenden Bedingungen
  - Angabe von Kategorisierungsrichtlinien
  - Diskussion von Zweifelsfällen

Statistik

- deskriptive Statistik beschreibt eine gegebene Datenmenge quantitativ
- inferentielle Statistik generalisiert aus den gegebenen Daten auf ungesehen Daten

inferentielle Statistik

- für Aussagen, die über die gegebenen Daten hinausgehen
- das Korpus/die Daten werden als eine Stichprobe (sample) einer größeren Grundgesamtheit (population) gesehen
- Korpusdesign:
  - Zusammensetzung ("Repräsentativität")
    - anteilige Stichprobe (proportional sample) – bildet die Grundgesamtheit (nach den als relevant erschteten Kategorien) in ihrer Verteilung ab
    - ausgewogene Stichprobe (balanced/stratified sample) – enthält (ähnlich große) Anteile von jeder als relevant erachteten Kategorie
  - Größe
    - notwendige Größe von der Forschungsfrage und dem betrachteten Phänomen abhängig
    - ganze Texte/Textabschnitte

Fragestellungen

- nicht für jede Fragestellung braucht man quantitative Daten – genaue Formulierung der Fragestellung in einem gegebenen theoretischen Rahmen nötig
- mögliche Fragestellungen (unit of analysis) (Biber& Jones, erscheint)
  1. Untersuchung mehrerer Varianten eines linguistischen Merkmals in einem Korpus – wir interessieren uns für das Merkmal und seine Eigenschaften → Typ-A-Studien
  2. Untersuchungen einer bestimmten Varietät (Text, Korpus) – wir interessieren uns für die Varietät → Typ-B-Studien
     - bezogen auf eine Variable
     - bezogen auf viele Variablen

Typ A - Studien

- Merkmale
  - verschiedene kategorial unterschiedene Ausprägungen eines Merkmals
  - nominal
- Interesse: Wie sind die Ausprägungen bedingt?
- in der Korpuslinguistik sehr häufig, Beispiele
  - Aktiv vs. Passiv aus Ausprägungen des Merkmals Satzmodus
  - verschiedene Wortstellungsmuster im Mittelfeld ([NOM, AKK, DAT], [AKK, NOM, DAT], ...)
  - *you* oder *thou* als Ausprägung des Merkmals Anrede in der 2. P. S.
  - ...

Typ-A-Studien: Beispiel 1

- untergeordnete Sätze, die mit *that* eingeleitet werden vs. untergeordnete Sätze, die nicht overt eingeleitet werden (0)
  I do not think that the situation is slipping out of control.
  I don't think [ ] any of us would be willing to do that.
- Eigenschaften für jeden Satz: Register, Verb, Form des Subjekts, ...
- Tabelle mit allen wesentlichen Eigenschaften (jede Beobachtung in einer Zeile, alle Ausprägungen eines Merkmals in einer Spalte)

Typ-A-Studien: Beispiel 1

Tab. 61.1: Coded observations for the analysis of *that*-omission

Complementizer	Matrix verb	Subject	Register
that	indicate	noun	academic
that	suggest	noun	academic
that	imply	noun	academic
0	say	pro-he	newspaper
that	argue	noun	newspaper
that	say	pro-I	newspaper
0	think	pro-I	newspaper
that	report	pro-they	newspaper
0	think	pro-I	conversation
0	say	pro-I	conversation
that	feel	pro-I	conversation
0	think	pro-I	conversation
0	think	pro-he	conversation
0	know	pro-I	conversation

Typ-A-Studien: Beispiel 1

- mit eines solchen Tabelle können dann die Ausprägungen jedes Merkmals gezählt werden

Tab. 61.2: Frequencies of complementizer variants from Table 61.1.

Complementizer	Frequency
0	7
that	7
Total:	14

Typ-A-Studien: Beispiel 1

Tab. 61.3: Frequencies of matrix verb variants from Table 61.1.

Matrix verb	Frequency
think	4
say	3
indicate	1
suggest	1
imply	1
argue	1
report	1
feel	1
know	1
Total:	14

Typ-A-Studien: Beispiel 1

Tab. 61.4: Frequencies of matrix verb groups from Table 61.1.

Matrix verb	Frequency
think	4
say	3
other verbs	7
Total:	14

Typ-A-Studien: Beispiel 1

- dann kann man die Merkmale miteinander in Beziehung setzen (Kreutztabellen) und daraus evtl. Generalisiserungen erkennen

Typ-A-Studien: Beispiel 1
Generalisierung: der Komplementierer wird bei häufigen Verben öfter weggelassen

Tab. 61.5: Cross-tabulation frequencies of complementizer choice by matrix verb

Complementizer	Matrix verb			
	think	say	other	Total
0	4	2	1	7
that	0	1	6	7
total	4	3	7	14

Typ-A-Studien: Beispiel 1
Generalisierung: Register ist ein Einflussfaktor

Tab. 61.6: Cross-tabulation frequencies of complementizer choice by register.

Complementizer	Register			
	Academic	News	Conversation	Total
0	0	2	5	7
that	3	3	1	7
total	3	5	6	14

Typ-A-Studien: Beispiel 1

- bei mehr Daten kann man auch Kombinationen von Faktoren betrachten
- bisher nur deskriptive Statistik, Testen von Hypothesen wäre nötig

Typ-A-Studien: Beispiel 2

- Kollokationen
  (‚wie ein Wort mit anderen Worten auftritt')
  blaue Augen
  Zähne putzen
  unglaublich schön
- dazu sehr viel Literatur (Abgrenzung von Phraseologismen, Idiomen, Funktionsverbgefügen etc.) – wird hier nicht behandelt (siehe Evert, erscheint)

Typ-A-Studien: Beispiel 2

- Kollokationen

Tab. 61.7: Coded observations for an analysis of the collocates of *blue*.

Preceding word	Target word	Following word
your	blue	eyes
the	blue	sky
had	blue	and
her	blue	eyes
a	blue	napkin

Typ-A-Studien: Beispiel 2

- Kollokation ‚blue eyes'
  – Häufigkeitszählung (Frequenzzählung) von ‚blue' und rechtem Kontext

Tab. 61.8: Frequencies of the three most frequent right collocates of *blue*. (Note the corpus is a sub-sample from the *Longman Spoken and Written English Corpus*.)

Target word	Following word	Frequency
blue	eyes	39
blue	and	25
blue	sky	11

Typ-A-Studien: Beispiel 2

- Kollokation ‚*blue* + rechter Kontext'
  - Zählungen: wie häufig kommen die Kombinationen ‚blue eyes' bzw. ‚blue napkin' vor?
- sind die häufigsten Kombinationen Kollokationen?
- dafür muss man berechnen, wie wahrscheinlich es ist, dass diese Kombinationen zufällig auftreten

Typ-A-Studien: Beispiel 2

- die Häufigkeit jedes Tokens in einem Korpus wird gezählt
- Annahme: alle Tokens sind in einem Sack (bag of words), und zwar genauso häufig, wie sie in dem Korpus vorkommen
- man zieht zufällig zwei Tokens hintereinander aus dem Sack und wiederholt dies oft
- dann kann man zählen, wie häufig zwei Tokens zufällig hintereinander vorkommen würden, wenn es keine syntaktischen, lexikalischen etc. Beziehungen gäbe
- Kollokationen sind dann Kombinationen von zwei Tokens die unwahrscheinlich häufig zusammen vorkommen

Typ-A-Studien: Beispiel 2

Expected frequency (fe) = (Target word frequency * Collocate word frequency) / Total corpus size

The observed frequencies of the individual words in the above example are:

Target word:
blue 371

Possible collocates:
eyes 1,649
and 49,598
sky 379

Target word	Following word	Frequency
blue	eyes	39
blue	and	25
blue	sky	11

Using the formula above, we can compute the expected frequency (f_e) for *blue eyes*:

f_e (*blue eyes*) = (371 * 1649) / 1,672,055 = .37

The expected frequencies for the other two combinations are:

f_e (*blue and*) = (371 * 49598) / 1,672,055 = 11.0

f_e (*blue sky*) = (371 * 379) / 1,672,055 = .08

Typ-A-Studien: Beispiel 2

Expected frequency (fe) = (Target word frequency * Collocate word frequency) / Total corpus size

The observed frequencies of the individual words in the above example are:

Target word:
blue 371

Possible collocates:
eyes 1,649
and 49,598
sky 379

f_e bezieht die absolute Frequenz der einzelnen Wörter in die Berechnung der erwarteten Frequenz mit ein!

and tritt alleine schon sehr häufig auf, daher ist anzunehmen, dass ***blue and*** durch Zufall eine so hohe erwartete Auftretenshäufigkeit (11.0) hat.

Using the formula above, we can compute the expected frequency (f_e) for *blue eyes*:

f_e (*blue eyes*) = (371 * 1649) / 1,672,055 = .37

The expected frequencies for the other two combinations are:

f_e (*blue and*) = (371 * 49598) / 1,672,055 = 11.0

f_e (*blue sky*) = (371 * 379) / 1,672,055 = .08

Typ-A-Studien: Beispiel 2

Expected frequency (fe) = (Target word frequency * Collocate word frequency) / Total corpus size

The observed frequencies of the individual words in the above example are:

Target word:
blue 371

Possible collocates:
eyes 1,649
and 49,598
sky 379

f_e bezieht die absolute Frequenz der einzelnen Wörter in die Berechnung der erwarteten Frequenz mit ein!

Weder ***sky*** noch ***blue*** kommen häufig vor, daher ist nicht anzunehmen, dass ***blue sky*** zufällig auftritt
f_e = (.08)

Using the formula above, we can compute the expected frequency (f_e) for *blue eyes*:

f_e (*blue eyes*) = (371 * 1649) / 1,672,055 = .37

The expected frequencies for the other two combinations are:

f_e (*blue and*) = (371 * 49598) / 1,672,055 = 11.0

f_e (*blue sky*) = (371 * 379) / 1,672,055 = .08

Typ-A-Studien: Beispiel 2

- man kann dann berechnen, wie stark die gezählte Frequenz einer Kombination von der erwarteten Frequenz der Kombination abweicht
  → Mutual Information Score, typischerweise noch logarithmiert (Church & Hanks 1990),
  f_o ist die gezählte Frequenz (observed frequency),
  f_e ist die erwartete Frequenz (expected frequency),
  Vorsicht: MI überschätzt bei geringen Frequenzen, siehe Evert (erscheint)

Mutual Information Score = f_o / f_e

Typ-A-Studien: Beispiel 2

- Mutual Information Scores für *blue eyes*, *blue and* und *blue sky*

Mutual info (*blue eyes*) = f_o (*blue eyes*) / f_e (*blue eyes*) = 39 / .37 = 105.4

Mutual info (*blue and*) = 25 / 11.0 = 2.3

Mutual info (*blue sky*) = 11 / .08 = 137.5

Typ-A-Studien

- sagen nichts über die Frequenz eines Merkmals in einem Text/Register/Korpus aus, sondern nur etwas über die Auswahl einer Variante eines Merkmals
- ➢ aus Tabelle 6.1 können wir nichts darüber lernen, wie häufig subordinierte Sätze überhaupt in jedem Register sind! Es kann sein, dass subordinierte Sätze mit *that* häufiger in conversation als in academic vorkommen!

Tab. 61.6: Cross-tabulation frequencies of complementizer choice by register.

Complementizer	Register			
	Academic	News	Conversation	Total
0	0	2	5	7
that	3	3	1	7
total	3	5	6	14

Typ-B-Studien

- in Typ-B-Studien werden Texte verglichen
- die Variablen sind Frequenzen von Merkmalen
- Vergleich von Modalverben in zwei Texten: in jedem Text sind 20 Modalverben. Kann man schließen, dass beide Texte sich bzgl. Modalverben gleich verhalten?
- nur wenn beide Texte gleich lang sind!
- bei unterschiedlichen Textlängen müssen die gezählten Frequenzen normalisiert werden
  gezählte Frequenz/Textlänge * Faktor
  = normalisierte Frequenz,
  wobei der Faktor die Textlängen approximieren sollte

Typ-B-Studien: Beispiel

- Vergleich von Modalverben in zwei Texten: in jedem Text sind 20 Modalverben. Kann man schließen, dass beide Texte sich bzgl. Modalverben gleich verhalten?
- Text A ist 750 Tokens lang, Text B ist 1200 Tokens lang, Normalisierung auf 1000 Tokens

Text A:
(20 modals / 750 words) × 1000 = 27.5 modals per 1,000 words

Text B:
(20 modals / 1200 words) × 1000 = 16.7 modals per 1,000 words

Typ-B-Studien: Beispiel

- Vergleich von 12 Texten bzgl. 3 Variablen, alle Freqenzen auf 1000 Tokens normalisiert
- wieder:
  - jede Zeile enthält eine Beobachtung (hier bezogen auf den ganzen Text)
  - jede Spalte enthält die Belegungen für eine Variable (TextID und Register sind nominal, alle anderen sind numerisch)

Typ-B-Studien: Beispiel

Tab. 61.9: Linguistic data for twelve texts

Text ID	Register	Word count	Past tense	Attrib adjs	1st person pronouns
n1.txt	news	2743	47.4	68.1	3.1
n2.txt	news	1932	49.2	63.0	9.2
n3.txt	news	2218	42.2	74.8	7.1
n4.txt	news	2383	45.3	72.1	2.2
n5.txt	news	1731	47.1	67.3	5.4
n6.txt	news	2119	51.2	70.0	5.2
c1.txt	conv	2197	32.2	43.1	62.6
c2.txt	conv	2542	37.4	36.3	59.1
c3.txt	conv	2017	36.8	39.7	58.7
c4.txt	conv	1896	29.2	35.2	65.5
c5.txt	conv	1945	31.3	34.0	58.2
c6.txt	conv	2072	23.8	38.3	60.4

Typ-B-Studien: Beispiel

- jetzt kann man das artithmetische Mittel ausrechnen, hier für die Variable past tense

(47.4 + 49.2 + 42.2 + 45.3 + 47.1 + 51.2 + 32.2 + 37.4 + 36.8 + 29.2 + 31.3 + 23.8) / 12 = 39.4

-

Typ-B-Studien: Beispiel

- man kann auch das artithmetische Mittel für jedes Register ausrechnen

Mean score of past tense verbs for newspapers:
(47.4 + 49.2 + 42.2 + 45.3 + 47.1 + 51.2) / 6 = 47.1

Mean score of past tense verbs for conversations:
(32.2 + 37.4 + 36.8 + 29.2 + 31.3 + 23.8) / 6 = 31.8

Typ-B-Studien: Beispiel

- (statistisch signifikante) Unterschiede in solchen Häufigkeiten müssen dann interpretiert werden
- dann kann man für jeden einzelnen Text die Abweichung vom arithmetischen Mittel ausrechnen
  (und dann natürlich wieder überlegen, worauf bestimmte (statistisch signifikante) Abweichungen zurückzuführen sind)

Einschub: arithmetisches Mittel, Median, Modalwert

- nicht immer ist das arithmetische Mittel aussagekräftig – Probleme bei starken Schwankungen
- Frage: Wie häufig kommen Wörter eigentlich so in einem Korpus vor?
- Beispiel aus Baroni (erscheint)
- Brown-Korpus (1 M Wörter), Typ-Token-Verteilung, die Wörter werden nach Häufigkeit sortiert

Einschub: arithmetisches Mittel, Median, Modalwert

Tab. 37.4: Top and bottom of the Brown frequency list

top frequencies			*bottom frequencies*		
rank	fq	word	rank range	fq	Randomly selected examples
1	62642	the	7967– 8522	10	recordings undergone privileges
2	35971	of	8523– 9236	9	Leonard indulge creativity
3	27831	and	9237–10042	8	unnatural Lolotte authenticity
4	25608	to	10043–11185	7	diffraction Augusta postpone
5	21883	a	11186–12510	6	uniformly throttle agglutinin
6	19474	in	12511–14369	5	Bud Councilman immoral
7	10292	that	14370–16938	4	verification gleamed groin
8	10026	is	16939–21076	3	Princes nonspecifically Arger
9	9887	was	21077–28701	2	blitz pertinence arson
10	8811	for	28702–53076	1	Salaries Evensen parentheses

Einschub: arithmetisches Mittel, Median, Modalwert

- arithmetisches Mittel (durchschnittliche Häufigkeit): 19
  stark beeinflusst durch die extrem häufigen Typen
- Modalwert (häufigster Wert): 1
- Median: 2

Typ-A-Studien vs. Typ-B-Studien

- Typ-A-Studien
- Vergleich von Varianten für ein Merkmal
- Variablen nominal

- Typ-B-Studien
- Vergleich von Texten (Korpora)
- Variablen numerisch (normalisierte Frequenzen)
- Mittelwert und Abweichung kann berechnet werden

inferentielle Statistik: statistische Tests

- Ist ein beobachteter Unterschied interessant oder ist er zufällig?
- dazu muss man zeigen, wie wahrscheinlich es ist, dass ein bestimmtes Ergebnis zufällig entsteht
- Annahme: die beobachteten Merkmale/Frequenzen gelten nicht nur für diesen einen beobachteten Text, sondern lassen sich generalisieren – der Text ist eine Stichprobe aus einer größeren Grundgesamtheit
- statistische Tests zeigen dann, wie ähnlich der betrachtete Text der angenommenen Grundgesamtheit ist (machen also Aussagen über nicht gesehenen Text)

inferentielle Statistik

- <Mantra> Forschungsfrage </Mantra>
- "The key to successful application of statistical techniques to linguistic problems lies in being able to frame interesting linguistic questions in operational terms that lead to meaningful significance testing." (Baroni & Evert, erscheint)

Hypothesen testen: Beispiel 1/binomiale Verteilung

- nach Baroni & Evert (erscheint)
- Nullhypothese: Der Anteil von passivischen Sätzen im Englischen beträgt 15%.
  H_0: $\pi = 15\%$
- Daten: schriftliche Texte des amerikanischen Englisch
- Stichprobe: 100 zufällig ausgewählte Sätze aus dem Brown Corpus
  (das Problem der Repräsentativität wird ausgeblendet)
  Stichprobengröße (sample size): n = 100
- alle Sätze werden in Passiv/nicht-Passiv eingeteilt
- erwartete Häufigkeit (expected frequency): *e* = 15

Hypothesen testen: Beispiel 1/binomiale Verteilung

- Ergebnis der Stichprobenauswertung:
  19 Sätze stehen im Passiv
  gefundene Häufigkeit (observed frequency): *f* = 19
- Kann man die Nullhypothese ablehnen?
- nein –
  wir haben nicht genügend Evidenz, denn es könnte – auch wenn die Nullhypothese gilt (15% aller englischen Sätze stehen im Passiv) – eine zufällige Stichprobe 19 Passive liefern würde
  → random variation
- Zufallsvariable X
- man muss testen, wie wahrscheinlich ein beobachtetes Ergebnis zufällig auftreten könnte
  Pr (X = k)

Hypothesen testen: Beispiel 1/binomiale Verteilung

- man könnte viele Stichproben von jeweils 100 Sätzen aus dem Brown Corpus ziehen und zählen, wie hoch jeweils der Anteil der passivischen Sätze ist (Urnenmodell)
- man kann das einfach ausrechnen, da hier eine binomiale Verteilung angenommen wird

$$\Pr(X = k) = \binom{n}{k} (\pi)^k (1 - \pi)^{n-k}$$

Hypothesen testen: Beispiel 1/binomiale Verteilung

- in 5,6% aller Fälle würde man, wenn H_0 gilt, auch bei einer zufälligen Stichprobe von jeweils 100 Sätzen 19 Passive finden
  Pr (X = 19) = 5,6%
- in 16,3% aller Fälle würde man 19 oder mehr Passive finden
  Pr (X ≥ 19) = 16,3%
- Pr nennt man auch Signifikanzniveau p. Man möchte ein möglichst kleines p. Akzeptiert sind
  p ≤ 0,05 (Wahrscheinlichkeit 5%)
  p ≤ 0,01 (Wahrscheinlichkeit 1%)
  p ≤ 0,001 (Wahrscheinlichkeit 0,1%)

Hypothesen testen: Beispiel 1/binomiale Verteilung

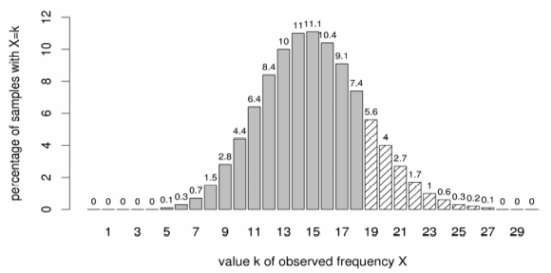


Fig. 36.1: Sampling distribution of X with $n = 100$ and $\pi = 15\%$

Hypothesen testen: Beispiel 1/binomiale Verteilung

- die Ablehnung der Nullhypothese bei 19 Passiven wäre riskant!
- Typ-1-Fehler: Ablehnung einer Nullhypothese ohne passende Evidenz
- eigentlich ist es noch schlimmer: wir haben einen einseitigen (one-tailed) Test verwendet. Die Nullhypothese kann auch nach unten verletzt werden. Daher nimmt man besser einen zweiseitigen (two-tailed) Test (z.B. Abstand 4 von 15%)
  Pr (X≥ 19 oder X ≤ 11) = 32,6%

Hypothesen testen: Beispiel 1/binomiale Verteilung

- Berechnung der Wahrscheinlichkeit, dass ein beobachtes Ergebnis zufällig auftreten kann durch verschiedene statisische Tests (χ2-tests, likelihood tests etc.)
- wichtig: alle Tests haben bestimmte Voraussetzungen bzgl. der zugrundeliegenden Verteilung, sample size etc. – man muss immer den passenden Test wählen!
- dabei gilt: je größer die Stichprobe, desto genauer das Ergebnis

Hypothesen testen: Beispiel 2/Abhängigkeiten

- nach Kessler (2001):
- Wie kann man statistische Tests verwenden, um Abhängigkeiten zwischen Variablen zu finden?
- Nullhypothese ist immer, dass die Variablen unabhängig sind. Dann wird getestet, wie wahrscheinlich die beobachteten Werte unter der Nullhypothese wären.
- Frage: Gibt es einen statistisch signifikanten Zusammenhang zwischen Kinder-haben und außer-Haus-arbeiten-gehen bei amerikanischen Frauen?
- Umfrage bei 1205 Frauen

Hypothesen testen: Beispiel 2/Abhängigkeiten

beobachtete Frequenzen		Kinder		
		ja	nein	Summe
Arbeit	ja	349	325	674
	nein	169	362	531
	Summe	518	687	1205

Hypothesen testen: Beispiel 2/Abhängigkeiten

- Annahme (Nullhypothese): Es gibt keinen Zusammenhang zwischen Kinder-haben und außer-Haus-arbeiten gehen.
- In den Marginalfrequenzen sehen wir, dass 674 von 1205 Frauen (55,9%) arbeiten gehen
- Wenn es keine Abhängigkeit zwischen den Variablen gibt, dann sollten 55,9% der Frauen mit Kindern arbeiten und 55,9% der Frauen ohne Kinder arbeiten

Hypothesen testen: Beispiel 2/Abhängigkeiten

erwartete Frequenzen unter H_0		Variable 2		
		ja	nein	Summe
Variable 1	ja			R1
	nein			R2
	Summe	C1	C2	N

Hypothesen testen: Beispiel 2/Abhängigkeiten

erwartete Frequenzen unter H_0		Variable 2		
		ja	nein	Summe
Variable 1	ja	R1*C1/N	R1*C2/N	R1
	nein	R2*C1/N	R2*C2/N	R2
	Summe	C1	C2	N

Hypothesen testen: Beispiel 2/Abhängigkeiten

beobachtete Frequenzen und erwartete Frequenzen		Kinder		
		ja	nein	Summe
Arbeit	ja	349 290	325 384	674
	nein	169 228	362 303	531
	Summe	518	687	1205

Hypothesen testen: Beispiel 2/Abhängigkeiten

- Wie wahrscheinlich ist es, dass die beobachteten Frequenzen zufällig sind?
- Oder: Wie unerwartet ist die Abweichung?
- eine Testmöglichkeit (unter vielen) ist der χ^2-Test (E für expected, O für observerd)
  $(E-O)^2/E$
  $(290-349)^2/290=12$
- das muss für alle Zellen wiederholt werden, die Summe hier ist 48
- diese Zahl kann dann mit Zahlen für andere Tabellen verglichen werden
- außerdem kann man daraus die Signifikanzniveaus bestimmen (dafür gibt es Tabellen)
  $p < 0{,}001$ (hochsignifikant!)
- die Variablen sind nicht unabhängig!
- Vorsicht bei der Interpretation!

Typ-A-Studien vs. Typ-B-Studien

- Typ-A-Studien
- Vergleich von Varianten für ein Merkmal
- Variablen nominal
- Tests: χ2 (Chi-Quadrat, chi sqare), Fishers exakter Test, z-score, ...

- Typ-B-Studien
- Vergleich von Texten (Korpora)
- Variablen numerisch (normalisierte Frequenzen)
- Mittelwert und Abweichung kann berechnet werden
- Tests: t-test, ...

Zusammenfassung

- wichtig für quantitative Studien
  - Welche Auswertungsmöglichkeiten fordert/erlaubt die Forschungsfrage? (Typ-A-Studien vs. Typ-B-Studien, deskriptive oder inferentielle Statistik etc.)
  - Datentypen: haben Sie kategorische (nominale, ordinale) oder kontinuierliche (numerische) Daten?
  - Korpusdesign: ist das verwendete Korpus repräsentativ für eine größere Grundgesamtheit oder kann ich nur Aussagen über das Korpus selbst machen? (Die Statistik hilft nicht dabei zu entscheiden, wie eine Grundgesamtheit aussehen kann und welche Parameter wichtig sind!)

Zusammenfassung: Hypothesen testen

- formuliere eine Nullhypothese so präzise, dass sie quantitativ anhand der vorhandenen Daten überprüfbar ist
- nimm eine zufällige Stichprobe und werte diese aus
- errechne eine Verteilung
- daraus ergibt sich die Wahrscheinlichkeit für jede Belegung der Zufallsvariable X
- teste, wie wahrscheinlich das beobachtete Ergebnis auch zufällig erzeugt sein kann

Plan

- Korpora in der Spracherwerbsforschung und im Fremdsprachunterricht
- Lernerkorpora
- Falko
- Studie: komplexe Verben im Spracherwerb

Motivation:

Korpora in der Fremdsprach-erwerbsforschung und im Fremdsprachunterricht

Korpora im Sprachunterricht

- L1-Korpora
- Übersetzungskorpora
- Lernerkorpora

L1-Korpora

- zur Verbesserung von Lehrmaterial
  - 'authentische/natürliche' Beispiele in Lehrmaterial und Lernerwörterbüchern
  - Frequenzinformationen, um Wortschatz festzulegen
  - Grammatiküberprüfung
  - Lehrstrategien / Lehrmaterialien ('entdeckendes Lernen')

➢ viele Studien, z.B. Sinclair et al. 1991, Wichmann et al. 1997, Granger et al. 2002, Nesselhauf 2005, Römer 2006

L1-Korpora

- nach Römer (2006):
  direkter Ansatz vs. indirekter Ansatz

L1-Korpora

- direkter Ansatz
  data-driven learning (DDL), entdeckendes Lernern: Korpora im Unterricht
  "confront the learner as directly as possible with the data, and to make the learner a linguistic researcher" (Johns 2002, 108), siehe auch Bernardini (2002)

☞Lehrende und Studierende

- ‚freie' Konkordanzen
- kontrollierte Übungen

Beispiel *the* vs. Nullartikel

- **Beispiel aus:** http://www.eisu.bham.ac.uk/johnstf/def_art.htm
- In Gwynedd, a bedrock of **the Welsh language**, there are 25 film-making companies.
- We must accept that the salvation of **the French language** involves learning one or more of the languages in neighbouring countries.
- The research also showed increases in the frequency of **bad language** and sex on television.
- Inspectors said behaviour was generally good, but features "such as free use of **colloquial language** and non-attendance at lessons are tolerated much more than in conventional schools".
- 1. proud of their command of _____ English language and engage in quite of lot of patting them
- 2. but it does not mean that _____ everyday language is bad: it is simply the way of things tha
- 3. cluded that cerebral dominance for _____ language is established before the age of five. Dur
- 4. abulary is one thing and _____ technical language is another, Vocabulary is words, lists of
- 5. avic-speakers. Orthodoxy and _____ Greek language remain the two markers of modern Greek ide

L1-Korpora

- indirekter Ansatz: Korpusuntersuchungen, um Materialien zu verbessern (COBUILD),
  Einzelthemen (Granger 1999, Nesselhauf 2005, Römer 2005 und viele andere)

☞Forscher, Lehrwerkschreiber

L1-Korpora

- ein Untersuchungstyp: Vergleich zwischen 'authentischer' L1 und 'Lehrwerks-L1'
- Beispiel:
  progressive im Englischen (Römer 2005, 2006), Vergleich zwischen den gesprochenen Teilen von BNC und BoE und der 'gesprochenen' Sprache in zwei Lehrbüchern
- progressive ist für deutsche Lerner schwierig
  - weil Deutsch keinen Progressiv hat
  - weil die Lehrwerke den Progressiv nicht so einführen, wie er wirklich verwendet wird ?

Beispiel: *looking* (Römer 2006, 236)

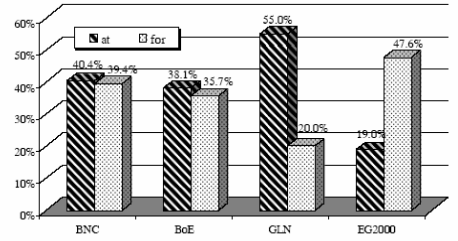


Figure 2: Looking and prepositions: looking at vs. looking for in BNC_spoken, BoE_brspok, GLN, and EG2000

Beispiel: *looking* (Römer 2006)

- der Progressiv kann bei einmaligen Handlungen und bei wiederholten Handlungen verwendet werden
- Well we*'re* really *looking* for a vegetarian one aren't we now (BoE_brspok)
- Yes. I remember that from when we *were looking* at houses # down there (BoE_brspok)

Beispiel: Römer 2006, 238

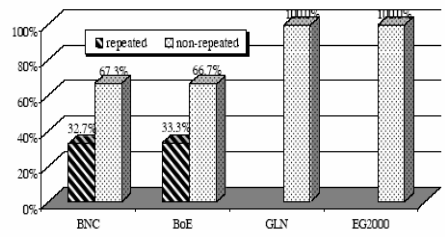


Figure 3: Looking and 'repeatedness': repeated vs. non-repeated in BNC_spoken, BoE_brspok, GLN, and EG2000

L1-Korpora

- Konsequenzen?
  noch nicht klar – weitere Studien nötig
  - vielleicht keine: es könnte sein, dass die Lehrwerke mit gutem Grund von der authentischen Sprache abweichende Verteilungen verwenden (mit zunehmender Fortgeschrittenheit unwahrscheinlicher)
  - vielleicht große: authentische Beispiele

L1-Korpora / Übersetzungskorpora

- „However, despite the progress that has unquestionably been made in the past two or three decades, I would still be hesitant to say that corpora have after all fully 'arrived' on the pedagogical landscape." (Römer 2006a, 121)
- aber: TALC, Workshops auf vielen Konferenzen (CL 2005, DGfS 2006, …)
- die besten Ressourcen/Studien/Lehrwerke für Englisch
- die anderen Sprachen hängen hinterher

Erforschung von Erwerbsverläufen: Lernerkorpora

Daten

- Intuition
- Sammlungen von authentischen Lernerdaten
  - unsystematisch, episodisch
  - Fehlersammlungen
  - Lernerkorpora
- experimentelle Daten
  - Elizitationsdaten
  - psycholinguistische Experimente
  - …

Daten

- Intuition – nicht vorhanden
- Sammlungen von authentischen Lernerdaten
  - unsystematisch, episodisch – problematisch
  - Fehlersammlungen – problematisch
  - Lernerkorpora
- experimentelle Daten
  - Elizitationsdaten
  - psycholinguistische Experimente
  - …

Daten

- Intuition – nicht vorhanden
- Sammlungen von authentischen Lernerdaten
  - unsystematisch, episodisch – problematisch
  - Fehlersammlungen – problematisch
  - Lernerkorpora
- experimentelle Daten
  - Elizitationsdaten
  - psycholinguistische Experimente
  - …

Lernerkorpora

- "Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance."
  (Granger 2002: 7)

Lernerkorpora für DaF/DaZ

- wahrscheinlich viele private Sammlungen
- aber kaum frei zugängliche fehlerannotierte Lernerkorpora
  - Belz (2004) – geschrieben, nicht fehlerannotiert (?), nicht zugänglich
  - LeaP – gesprochen, zugänglich (Milde & Gut 2002)
  - Weinberger (2002) – geschrieben, fehlerannotiert, bisher nicht zugänglich
  - ESF-Korpora (MPI Nijmegen) – gesprochen, zugänglich, DaZ
  - ???

Lernerkorpora: Das Ideal

- Design (Zusammensetzung)
  - auf die jeweilige Forschungsfrage abgestimmt (idealerweise vergleichbare Korpora)
  - gut dokumentiert
  - Größe
- Architektur
  - Metadaten
  - Zielhypothese
  - mehrere Ebenen, konfligierende Hypothesen
  - flexible Fehlerexponenten
- frei verfügbar, verteiltes Arbeiten möglich

Lernerkorpora: Die Wirklichkeit

- die meisten Lernerkorpora (es gibt Ausnahmen)
  - nicht gut designed und dokumentiert
  - nur eine implizite Zielhypothese
  - konfligierende Analysen können nicht dargestellt werden
  - verteiltes Arbeiten nicht möglich

an Falko arbeiten zur Zeit:

Anke Lüdeling, Maik Walter, Karin Schmid, Hagen Hirschmann, Seanna Doolittle, Vicky Oketch

zu Falko beigetragen haben auch:

Emil Kroymann, Karsten Hütter, Marc Reznicek, Max Möller, Heidi Byrnes, Castle Sinicrope

Falko ist zugänglich unter: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

Die Analyse von Lernerdaten

- Datenerhebung
- Datenaufbereitung
- Datenauswertung

Die Analyse von Lernerdaten

- Datenerhebung
- Datenaufbereitung
- Datenauswertung

Datenerhebung

Am Anfang jeder Datenerhebung…
…steht eine Frage!

Wie gelingt es *fortgeschrittenen* Lernern, komplexe Texte in der Fremdsprache Deutsch zu produzieren?

Datenerhebung

- Betrachtung von ausgewählten (zB Konnektoren, Struktur der Vorfeldbesetzung, Definitheit…)
- zwei Textsorten
  - Zusammenfassungen
  - freie Essays

Falko

Texttyp	Zusammenfassungen	Essays
Modus	geschrieben, z.T. handschriftlich, z.T. elektronisch	
Lernstand	fortgeschritten (DSH)	fortgeschritten (C-Test)
Größe (Tokens)	ca. 54.000	ca. 67.000, wächst

Daten: Zusammenfassungen

- Textzusammenfassungen eines literaturwissenschaftlichen bzw. linguistischen Fachtextes (ca. 1-2 Seiten) im Rahmen der Sprachstandsbestimmung von ausländischen Germanistikstudierenden an der Freien Universität
- Erhebungsdauer: 90 Minuten
- Formales Kriterium: abgelegte DSH-Prüfung
- 6 Erhebungszeiträume, 107 Lerner, 40592 Tokens (abgeschlossen) und
- Vergleichskorpus, 41 L1-Sprecher, 12757 (wächst)

Daten: Zusammenfassungen

12.3 Pragmatische Erwerbsprinzipien

Eine wichtige Phase des Spracherwerbs ist erreicht, wenn Kinder in den Wortschatzspurt eintreten. Nach manchen Schätzungen lernen Kinder ab zwei Jahren durchschnittlich zehn neue Wörter am Tag und verfügen mit etwa sechs Jahren schon über einen Wortschatz von ungefähr 14.000 Wörtern (Clark 1993: 13). Die Erwerbsaufgabe besteht für die Kinder darin, neue Wörter aus dem Input zu isolieren und sie in ihr mentales Lexikon zu übernehmen. Nach und nach werden die Wörter dort mit einer Angabe über ihre Eigenschaften versehen, wobei phonologische, morphologische, syntaktische, semantische und pragmatische Eigenschaften zu unterscheiden sind. Wie diese enorme Aufgabe bewältigt wird, ist eine spannende Frage der Spracherwerbsforschung (vgl. Rothweiler/Meibauer 1998).

[Klausurvorlage: Anfang des Pragmatiktextes]

Daten: Zusammenfassungen

Pragmatische Erwerbsprinzipien

In diesem Text befaßt sich der Autor mit zwei pragmatischen Prinzipien, nämlich das Prinzip der Konventionalität und das Prinzip des Kontrasts. Im ersten Abschnitt wird eine spannende Frage der Spracherwerbsforschung gestellt, wie eine enorme Erwerbs-Aufgabe von Kindern bewältigt wird. Der Autor geht davon aus und interpretien die Argumentation von Clark. Der Autor betont, dass das Prinzip der Konventionalität und das Prinzip des Kontrasts in dieser Erwerbsaufgabe eine wichtige Rolle spielen. Obwohl diese Prinzipien gleichermaßen für Kinder und Erwachsene gültig sind, haben sie aber für Kinder andere Einflüsse, weil ihr Wortschatz ja noch ansteigen muss. Beispiel aus einem Lernertext…

[Falko-Text 38]

Daten: Essays

- vier kontroverse Themen (in Anlehnung an ICLE)
- 90 Minuten, keine Hilfsmittel, zT handschriftlich, zT digital
- Erhebungen in Berlin (HU, Sommerunversität), Kopenhagen, Taschkent, Mombasa, Nairobi, Nyeri, Adana, Stellenbosch
- 132 Texte, 65389 Tokens, wächst
- Vergleichskorpus an Berliner Gymnasien erhoben

Metadaten

- für die Summaries und die Essays wurden von allen Teilnehmern Metadaten erhoben
  - Alter
  - Geschlecht
  - Sprachbiographie (Muttersprache, andere Sprachen, wie gelernt etc.)
- man kann daraus entsprechende Subkorpora erstellen

Daten: Longitudinalkorpus

- Georgetown University, Washington
- 4 Sprachlevel
- unterschiedliche Aufgabenstellungen

Die Analyse von Lernerdaten

- Datenerhebung
- Datenaufbereitung
- Datenauswertung

Datenaufbereitung

- Digitalisierung der handschriftlichen Daten
- korpuslinguistische Aufbereitung der Daten
- Fehlerannotierung
  (Entwurf eines Schemas und „Umsetzung“)

Datenaufbereitung

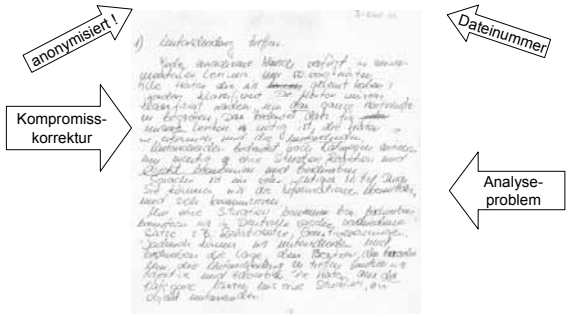
handschriftliche Daten

- Ausschluss von „Tippfehlern“ des Lerners, also Fehlern, die durch das Medium Computer evoziert werden
- kein Einfluss von Rechtschreibkontrolle
- Handschriftentzifferung als Analyseproblem

digitale Daten

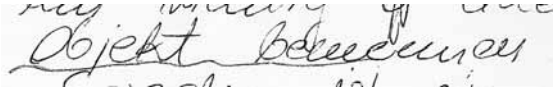
- eventuell Fehler, die durch das Medium Computer evoziert sind
- eventuell (bei uns aber kontrolliert) automatische Rechtschreibkontrolle
- einfach weiterzuverarbeiten

Datenaufbereitung

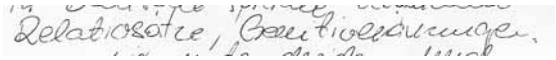


Analyseproblem „Handschrift“

Analyseproblem „Handschrift“



Analyseproblem „Handschrift“



Datenaufbereitung

- es ist ganz schwierig, falsche Texte nicht zu korrigieren

➢ Fazit: Der Digitalisierer hat Entscheidungen getroffen!

Datenaufbereitung

Die nächsten Schritte im Schnelldurchlauf:

- Automatische Wortartenzuweisung (mit dem TreeTagger) mit anschließender manueller Korrektur (unterschiedliche Ebenen)
- Aufstellung einer Zielhypothese
- Fehlerannotierung auf mehreren Ebenen
- Aufnahme und Einbindung der Metadaten

Erforschung von Erwerbsverläufen
Fehleranalyse
Kontrastive Analyse

Lernerkorpora

- “Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.” (Granger 2002: 7)

Lernerkorpora

- Forschung
  - Fehleranalyse (Error Analysis, EA)
  - kontrastive Analyse (Contrastive Interlanguage Analysis, CIA)
- Lernerkorpora
  - Design (welche Texte?)
  - Vorverarbeitung
  - Studien

Fehleranalyse – Ziele

- eine Lerneräußerung wird auf ihre Fehler/Abweichungen hin untersucht
- in einem Korpus können Lerneräußerungen mit standardisierten Fehlermarkierungen (Tags) versehen werden (→ Fehlerannotation)
  - standardisierte Suche (alle Kasusfehler)
  - quantitative Auswertung

Fehleranalyse – was ist ein Fehler?

- Bruch einer Regel
  - ungrammatisch
  - Woher bekommt man die Regeln – explizite Regeln vs. implizite Regeln
  - strukturelle Fehler, müssen algorithmisch zu finden sein (?)
- Abweichung von einer Norm, "breaches of code" (Corder 1973)
  - unakzeptabel, 'inappropriate'
  - Norm erfordert Setzung
  - verschiedene Normen (sprachliche, soziale, ...)
  - strukturelle und nichtstrukturelle Fehler, nicht unbedingt algorithmisch zu finden

Fehleranalyse – was ist ein Fehler?

- "A linguistic form, ... which, in the same context would in all likelihood not be produced by the learner's native speaker counterparts."
  (Lennon 1991, 182)

- brauchbare Definition ?
- kontrollierte Erhebungen notwendig

Fehleranalyse – was ist ein Fehler?

- Unterschied zwischen *error* (≈ Kompetenzfehler) und *mistake/lapse* (≈ Performanzfehler)

Fehleranalyse – Beispiel

- *In dem Text wird es über Eigenheiten einer - bis jetzt - uneigenständigen Gattung gesprochen. Die Benennung "Kunstmärchen" gibt uns einerseits eine Erklärung über den Gegenstand, aber andererseits bleibt es eher im Dunkelen was genau damit gemeint ist. Doch die Ungenaue Schätzungen können auch auf eine andere Schiene bringen, und eben dieses Variieren mit Ideen macht der Gegenstand interessant.* (Falko-Text 48)

Fehleranalyse – Beispiel

- *In dem Text wird es über Eigenheiten einer - bis jetzt - uneigenständigen Gattung gesprochen. Die Benennung "Kunstmärchen" gibt uns einerseits eine Erklärung über den Gegenstand, aber andererseits bleibt es eher im Dunkelen was genau damit gemeint ist. Doch die Ungenaue Schätzungen können auch auf eine andere Schiene bringen, und eben dieses Variieren mit Ideen macht der Gegenstand interessant.* (Falko-Text 48)

kontrastive Analyse (CIA)

- quantitativer Vergleich von Lernerdaten mit
  - Muttersprachlerdaten
  - Lernerdaten anderer Art (Muttersprache, Lernstand, Genre etc.)
- in Bezug auf
  - Lexik
  - syntaktische Unterschiede
  - Länge
  - Fehler
  - …

kontrastive Analyse

- Ziel:
  Übergebrauch (overuse, Lerner benutzen ein Wort/eine Konstruktion etc. im Vgl. mit Muttersprachlern zu häufig) und Mindergebrauch (underuse, Lerner benutzen ein Wort/eine Konstruktion etc. im Vgl. mit Muttersprachlern zu selten) finden

qualitative und quantitative Analyse

- jede quantitative Analyse setzt eine Kategorisierung (qualitative Analyse) voraus

Lernerkorpora: Zielhypothese

- es ist nicht möglich, einen Fehler zu annotieren, ohne eine implizite Zielhypothese im Kopf zu haben
  → Fehler werden also relativ zu einer Zielhypothese markiert

Lernerkorpora: Zielhypothese

- *die Erklärung für* ***<MoArInGn> diese*** *Phänomen ist*
  Mo – morphology, Ar – article, In – Inflection, Gn – gender

(Weinberger 2002)

➢ Zielhypothese implizit

➢ keine konkurrierenden Annotationen möglich
die Erklärung für ***diese****s Phänomen ist*
die Erklärung für diese ***Phänomen****e ist*

Lernerkorpora: Zielhypothese

- *Der Realismus ist eine im 19. Jahrhunder, als Gegenbewegung zu klassisch-romantischen Kunstauffassung literarische Richtung, die sich bis zum Ende des Jahrhunderts international weit erstreckte.*
- *Der Realismus ist eine literarische Richtung, die im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründet wurde und sich …*
- *Der Realismus ist eine im 19. Jahrhundert, als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründete literarische Richtung, die sich …*

Lernerkorpora: Zielhypothese

- *Der Realismus ist eine im 19. Jahrhunder, als Gegenbewegung zu klassisch-romantischen Kunstauffassung literarische Richtung, die sich bis zum Ende des Jahrhunderts international weit erstreckte. (Falko L2)*
- *Der Realismus ist eine literarische Richtung, die im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründet wurde und sich …*
  - *ORTH: Jahrhundert*
  - *WORTST: Relativsatz*
  - *AUSLASSUNG: gegründet wurde*

Lernerkorpora: Zielhypothese

- *Der Realismus ist eine im 19. Jahrhunder, als Gegenbewegung zu klassisch-romantischen Kunstauffassung literarische Richtung, die sich bis zum Ende des Jahrhunderts international weit erstreckte. (Falko L2)*
- *Der Realismus ist eine im 19. Jahrhundert, als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründete literarische Richtung, die sich …*
- *ORTH: Jahrhundert*
- *ORTH: Komma*
- *AUSLASSUNG: gegründete*

Fehleranalyse – Zielhypothese

- "reconstruction of those utterances in the target language" (Ellis 1994: 54)
- oft sind mehrere Zielhypothesen möglich

Zielhypothese: Experiment

- 5 Annotationen für 17 Sätze (fortlaufender Text) (Lüdeling, 2008)

Inhaltswörter	Funktionswörter
15	13
24	26
17	25
16	12
14	22

Textübernahme

- *Sprecher nehmen an, daß Unterschiede hinsichtlich der Form auch Unterschiede in der Wortbedeutung signalisieren. (PE)*
- *Sprecher nehmen an, dass Unterschiede hinsichtlich der Form auch Unterschiede in der Wortbedeutung signalisieren.*
- *Kontrast bedeutet also: Sprecher nehmen an, dass Unterschiede hinsichtlich der Form auch Unterschiede in der Wortbedeutung signalisieren.*

Textübernahme (Text 033)

Der Realismus ist eine überragende geistige und künstlerische Tendenz des 19. Jahrhunderts. Er erstreckt sich als international weit ausgreifende Epochenströmung bis gegen Ende des Jahrhunderts. In der Literatur tritt er unter verschiedenen Namen auf. Die Literatur- und Kunstgeschichte einigte sich erst im nachhinein über die Grenzen des Realismus. Anfangs haben sich die Naturalisten hauptsächlich als Realisten verstanden. Hieraus erkennt man, dass die genauere Bezeichnung des Realismus nach dem damaligen Sprachgebrauch noch nicht vorhanden war. In der heutigen Bezeichnung ist sie eine literatur- und kunstgeschichtliche Richtung, welche in Poetischer und Bürgerlicher Realismus eingeteilt wurde, um sie von der Naturalismus Variante unterscheiden zu können. Diese beiden Bezeichnungen werden vielfach synonym gebraucht. Bei einer stärkeren Unterscheidung würde man den Bürgerlichen Realismus mit der Gründungsphase der realistischen Bewegung um die Jahrhundertmitte identifizieren, welche auch als Programmatischer Realismus bezeichnet wird. In den fünfziger Jahren, hat es in Deutschland eine lebhafte literaturtheoretische Debatte um die Zielsetzung und das Wesen ein realistischer Literatur gegeben. Eines ihrer Zentren hat es, in der von Gustav Freytag und Julian Schmidt herausgegebenen Zeitschrift "Die Grenzboten". Der populäre Roman "Soll und Haben" von Gustav Freytag gilt als ein Exempel für die gehegten nationalpädagogischen Erwartungen an einen dezidiert bürgerlichen Realismus. Das bedeutendste poetolgische Manifest des deutschen Realismus ist Fontanes Aufsatz "Unsere lyrische und epische Poesie seit 1848" (1853). Hier wird das Ideal einer umfassenden Wirklichkeits-Repräsentanz formuliert. Die deutschen Realisten wollten auf keinen Fall als Kopisten der Wirklichkeit verstanden werden. Durch die Erfindung der Daguerrotypie (Photographie) rückte die Mimesis der Realität sehr nahe. Somit konnte die herrschende Ästhetik mit Hilfe der Photographie, sich von der herkömmlichen Kunst distanzieren. Rudolf Gottschall kritisiert an Beispielen aus dem Romanschaffen von Daudet und Zola die direkte Beziehung der Literatur auf die zeitgenössische Wirklichkeit. Er sieht es als unzulässige Überschreitung der Grenze zwischen Kunst und Realität. Gleichzeitig kritisiert er auch Fontane, Zolas und Alexander Kielands "Reportertum". Er erklärt, dass er an dem exakten Bericht einen ungeheuren Literaturfortschritt erkennt, welches uns auf einen Schlag vom öden Geschwätz zurückliegender Jahrzente befreit. Weiter stellt er fest, dass sich "Meisterstücke und Berichterstattung" erst dann zur Höhe des Kunstwerks erheben <>. Damit greift er auf eine Norm der klassischen Ästhetik zurück, dem "Idealrealismus". Das idealistische Ressentiment ist eines der Gründe für das Zurückbleiben des deutschsprachigen Realismus gegenüber der Schonungslosigkeit der Gesellschaftskritik eines Dickens oder Balzac und der unbestechlichen Psychologie Flauberts. Der deutsche Kritiker Emil Homberger kritisiert Flaubert. Für ihn, zergliedert Flaubert in seinen Erzählungen das Seelenleben seiner Figuren leidenschaftslos. So wie Fontane vermisst auch er die belebende Seele. Er versucht sogar, den französischen Autor einen Irrtum in der Auffassung von "Objektivität" nachweisen zu können. Denn wer objektiv sein wolle, dürfe sich nicht einseitig an der Dokumentation einzelner Faktoren berauschen.

Studie: komplexe Verben in Lerneräußerungen

(noch 'in progress')

komplexe Verben in Lerneräußerungen

- *Ich habe **angefängt** viele Firmen (Production Companies) **zu anrufen** für Arbeit während der Dezember Ferien, aber keine Firmen **sind beeindruck** durch mein Grad.*

kurze Wiederholung: komplexe Verben im Deutschen

Präfixverben

- *ver•kaufen*
- *[…] dass Peter Schokolade verkauft.*
- *Peter verkauft Schokolade.*
- Infinitiv: *zu verkaufen*
- Partizip II: *verkauft*

Partikelverben

- *auf•essen*
- *[…] dass Peter die Schokolade aufisst.*
- *Peter isst die Schokolade auf.*
- Infinitiv: *aufzuessen*
- Partizip II: *aufgegessen*

Fragestellungen

- Grundlage: Theorie über komplexe Verben
- Grundlage: Erwerbstheorie
- theoretisch
  - (Wie) kann man aus den Erwerbsmustern auf eine Theorie des Erwerbs schließen?
- empirisch
  - Wie werden komplexe Verben im Deutschen als Fremdsprache erworben? Gibt es typische Muster? Gibt es typische Fehler?
- methodisch
  - Welche Studien sind möglich? Welche Daten braucht man? Welche Schlüsse kann man ziehen?
  - Fokussieren Lerngrammatiken wirklich auf das, was unterrichtet werden muss? Wie kann man Lerngrammatiken vielleicht verbessern?

Fragestellungen

- Grundlage: Theorie über komplexe Verben
- Grundlage: Erwerbstheorie
- theoretisch
  - (Wie) kann man aus den Erwerbsmustern auf eine Theorie des Erwerbs schließen?
- empirisch
  - Wie werden komplexe Verben im Deutschen als Fremdsprache erworben? Gibt es typische Muster? Gibt es typische Fehler?
- methodisch
  - Welche Studien sind möglich? Welche Daten braucht man? Welche Schlüsse kann man ziehen?
  - Fokussieren Lerngrammatiken wirklich auf das, was unterrichtet werden muss? Wie kann man Lerngrammatiken vielleicht verbessern?

Fehler?

- *Ich habe **angefängt** viele Firmen (Production Companies) **zu anrufen** für Arbeit während der Dezember Ferien, aber keine Firmen **sind beeindruck** durch mein Grad.*
- *Wenn man z.B. an der Universität von Stellenbosch Engineering studiert, hat man alles diese Theorie im Kopf, aber man kann es nicht **anwendet**.*
- *Ich **stimme** deshalb **mit** dass Universitätabschlüsse nicht man wirklich **vorbereitet** auf die wirkliche Welt.*

Fehler?

- *Ich habe **angefängt** viele Firmen (Production Companies) **zu anrufen** für Arbeit während der Dezember Ferien, aber keine Firmen **sind beeindruck** durch mein Grad.*
- falsche Flexionsform:
  angefängt statt *angefangen*
  beeindruck statt *beeindruckt* (Tippfehler?)
- keine morphologische Trennung beim Infinitiv:
  zu anrufen statt *anzurufen*

Fehler?

- *Wenn man z.B. an der Universität von Stellenbosch Engineering studiert, hat man alles diese Theorie im Kopf, aber man kann es nicht **anwendet**.*
- falsche Flexionsform:
  anwendet statt *anwenden*

Fehler?

- *Ich **stimme** deshalb **mit** dass Universitätabschlüsse nicht man wirklich **vorbereitet** auf die wirkliche Welt.*
- falsche Flexionsform:
  vorbereitet statt *vorbereiten*
- falsches Lexem:
  mitstimmen statt
  zustimmen, übereinstimmen

Lernaufgaben – Lerngrammatiken

- meist Mittelstufe
- meist gemeinsam in einem Kapitel (außer Griesbach – Präfixverben in Wortbildung)
- Grammatiken: Griesbach (1991), Hall & Scheiner (2001), Helbig & Buscha (1998, 2000), Rug & Tomaszewski (2001)

Lernaufgaben – Lerngrammatiken

- strukturell
  - Trennbarkeit von Partikeln (‚trennbare Vorsilben' / ‚Präverbien'), Betonung
  - Nichttrennbarkeit von Präfixen (‚untrennbare Vorsilben' / ‚Präverbien'), kein ‚*ge-*', Betonung
- semantisch
  - einige oberflächliche Regularitäten
  - Argumentstrukturveränderungen für einige Beispiele, keine allgemeinen Regeln
  - Rug/Tomaszewski & Griesbach: produktive Wortbildung, keine Regeln
- besonderes Problem: ‚doppelförmige' Verben

Lernaufgaben – strukturell

- Partikelverben
  - Trennbarkeit
  - Betonung
- Präfixverben
  - Betonung
    (generelle Regel: *infor'mieren - informiert*)

Lernaufgaben – semantisch

- lexikalisierte Fälle (*enthalten, anfangen*)
- produktive Reihen
  entkernen, entgräten, entehren, entvölkern, entgiften, entschlacken, entwaffnen, …
  ent- + denominales V → V
  immer transitiv
  anschalten, anknipsen, andrehen, anmachen, …
  an + V → V
  immer transitiv, resultativ
- Lüdeling (2001), Olsen (1998), Stiebels (1996)

Lernaufgaben

- Lexikon: Wörter und deren grammatische Eigenschaften (Flexionsklasse, Argumentstruktur etc.)
- morphologische und syntaktische Trennbarkeit von Partikelverben
- semantische Reihen, Unterscheidung von lexikalisierten Fällen und produktiven Mustern

Auswertung von Daten

1. Entscheidung über das Phänomen
2. Formulieren von Hypothesen innerhalb eines theoretischen Rahmens
3. Überprüfen und ggf. Anpassung der Hypothesen

Auswertung von Lernerkorpora

- Error Analysis (Fehleranalyse, EA)
- Contrastive Interlanguage Analysis (kontrastive Analyse, CIA)

komplexe Verben in Falko

- manuelle Analyse nötig

Exkurs: Vorverarbeitung

- getrennt stehende Partikeln sind generell schwer zu taggen (hier TreeTagger, Schmid 1994)
- *Für die Kinder sollten nicht die gleichbedeutenden Wörter verwendet werden, sondern die schon in ihren Lexikonen existierenden konventionellen Wörter, sonst würden die Kinder nicht verstehen. Bspw. „Lehrer“ statt/PTKVZ „Unterrichtender“, „Polizist“ statt/PTKVZ „Polizeimann“*
- *für Wegener war der Dativ am Anfang ein/PTKVZ im wesentlichen semantisch bestimmter Kasus*

Exkurs: Vorverarbeitung

- Tagger trainieren auf Zeitungstexten – Performanceabfall für abweichende Texte
- *Er verbringt das ganze Leben vor dem Tor, unternimmt Versuche von dem Türhüter eigelassen/ADV zu werden und ist völlig auf den Türhüter fixiert.*
- van Rooy & Schäfer (2003), siehe auch Rayson, Archer & Smith (2005)

komplexe Verben in Falko

- manuelle Analyse nötig
- an vielen Stellen Entscheidungen nötig
  - Partikelverben und Präfixverben in klar verbaler Funktion (keine adjektivischen Partizipien)
  - ‚Textnähe': *daraufhinweisen* in *Einleitend weist Meibauer daraufhin, dass […].*
  - Partizipbildung: Wörter, die synchron nicht als komplex wahrgenommen werden *(empfinden, gewinnen)?*

Fehleranalyse für komplexe Verben

- ‚lexikalische Fehler' (Ausdruck, Subkategorisierung)
  - für lexikalisierte Verben (nicht charakteristisch für komplexe Verben)
  - aber evtl. Indikation dafür, dass ein produktives Wortbildungsmuster nicht gelernt wurde
- strukturelle Fehler
  - morphologische/syntaktische Trennung
  - Wortstellungsfehler

Fehleranalyse – Auswertung

- AUSDRUCK
  *Der Verfasser **klärt** seinen Lesern den Aufsatz von Fontane **auf**.*
- SUBKAT
  *Der Autor betont aber, dass die Forschung über die Spracherwerbprinzipien noch **erweitern** und durch die Bereicherung der pragmatischen Wissen von Erwachsenensprache entwickeln wird.*

Fehleranalyse – Auswertung

- ORTH
  *Eine nicht informations übermittelnde Kommunikation mit nicht ernsthaften Menschen kann nur dann **stadt finden**, wenn sie entweder sich über das Thema der Diskusion nich geeignet haben […]*
- WORTSTELLUNG
  *Es liegt daran, das ein Gesprächspartner **nimmt ernst**, das was der nicht ernste Gesprächspartner redet .*

Fehleranalyse - Auswertung

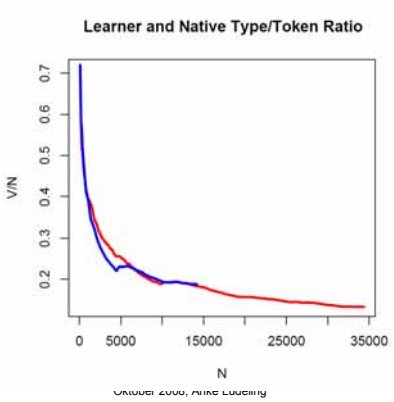
- Annahme: Wenn man alle Fehler annotiert hat, dann zählt man sie und kann die Forschungshypothesen bestätigen oder ablehnen
- viele Probleme ...
  - Zielhypothese
  - Textübernahme

Zusammenfassung: Erwerb von komplexen Verben

- kaum Probleme mit komplexen Verben – kaum Trennbarkeitsfehler
- viele generelle Verbprobleme mit Flexion und Argumentstrukur
- schwierig: semantische Reihen, produktive Wortbildung

Lernaufgabe – Produktivität

- ‚neue' Wörter (?)
- erstes Indiz:
  Maß für Produktivität – Hapaxe
- Voraussetzung: Lernervokabular generell nicht geringer als Muttersprachlervokabular
- Baayen (2001), Lüdeling & Evert (2005)



Daten

- zwei Vorlagentexte (Pragmatische Erwerbsprinzipien PE, Stile S)
- alle Daten zu einer Vorlage sind zusammengefasst
  (Vorsicht: hier wird über große Unterschiede gemittelt)
- L1
  Stile: 5637 tokens
  PE: 4022 tokens
- L2
  Stile: 3470 tokens
  PE: 5029 tokens

L2, PE

	Partikelverben	Präfixverben
Tokens	54	188
Typen	29	53
Hapaxe	20	27

L2, Stile

	Partikelverben	Präfixverben
Tokens	42	95
Typen	25	51
Hapaxe	16	31

Lernaufgabe – Produktivität

- scheint auch gelöst –
  Median der Hapaxe liegt jeweils bei 1 –
  hohes Maß an Produktivität
- Lernaufgabe Produktivität – auch gelöst?

Typen, Hapaxe

- mathematisches Problem: zu wenig Daten (data sparseness, Daten immer in der LNRE-Zone)
- Datenprobleme
  - Vorlage – Textsorte
  - Fehler
- viele der Typen (Hapaxe & häufige Typen) sind direkt aus den Vorlagen übernommen

L2, PE

	Partikelverben/ davon in Vorlage	Präfixverben/ davon in Vorlage
Tokens	54	188
Typen	29 / 11	53 / 17
Hapaxe	20 / 5	27 / 4

Fehler bei hapax legomena

- zusätzlich sind einige der Typen fehlerhaft – hier werden nur die Typen gezählt, die nicht in der Vorlage vorkamen
- in PE
  Partikelverben: 2 Fehler
  Präfixverben: 6 Fehler

Textsorte

- Interpretation der Zahlen:
  Sprachstand oder Textsorte?
- Vergleich mit L1-Texten

L1 & L2, PE

	Partikelverben		Präfixverben	
	L1	L2	L1	L2
Tokens	62	54	132	188
Typen	34 / 7	29 / 11	55 / 16	53 / 17
Hapaxe	20 / 2	20 / 5	24 / 2	27 / 4

L1 & L2, Stile

	Partikelverben		Präfixverben	
	L1	L2	L1	L2
Tokens	96	42	114	95
Typen	62 / 7	25 / 7	58 / 14	51 / 18
Hapaxe	44 / 1	16 / 4	40 / 6	31 / 6

Lernaufgabe – Produktivität

- produktive Verwendung von komplexen Verben stellt eine Schwierigkeit für Lerner dar

Zusammenfassung: Erwerb von komplexen Verben

- kaum Probleme mit komplexen Verben – kaum Trennbarkeitsfehler
- viele generelle Verbprobleme mit Flexion und Argumentstrukur
- schwierig: semantische Reihen, produktive Wortbildung

Zusammenfassung

- Voraussetzung: Auswahl & theoretische Analyse eines sprachlichen Phänomen, Erwerbstheorie
- Datenauswahl
- Lernerkorpora
- Fehleranalyse
  - theoretische Probleme: Zielhypothese
  - empirische Probleme: Textübernahme

Referenzen

- Baayen, R. Harald (2008) *Analyzing Linguistic Data. A Practical Guide to Statistics*. Cambridge University Press, Cambridge.
- Baroni, Marco & Evert, Stefan (erscheint) Statistical Methods for Corpus Exploitation. In: Lüdeling, Anke & Kytö, Merja (Hrsg.) *Corpus Linguistics. An International Handbook*. Walter de Gruyter, Berlin.
- Biber, Douglas 1993. Representativeness in corpus design. In: *Literary and Linguistic Computing 8*: 243-257.
- Biber, Douglas & Jones, James K. (erscheint) Quantitative Methods in Corpus Linguistics. In: Lüdeling, Anke & Kytö, Merja (Hrsg.) *Corpus Linguistics. An International Handbook*. Walter de Gruyter, Berlin.
- Church, Kenneth W. & Hanks, Patrick (1990) Word Association Norms, Mutual Information and Lexicography. In: *Computational Linguistics 16*, 22-29.
- Evert, Stefan & Fitschen, Arne (2001) Textkorpora. In: Carstensen et al. (Hrsg) *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Spektrum Akademischer Verlag, Heidelberg, 369 – 376
- Kessler, Brett (2001) *The Significance of Word Lists*. CSLI Publications, Stanford.
- Kiss, Tibor & Strunk, Jan (2006) Unsupervised Multilingual Sentence Boundary Detection. In: *Computational Linguistics* 32 (4): 485-525
- Lauer, M. 1995. "How much is enough? Data requirements for Statistical NLP". In 2nd. Conference of the Pacific Association for Computational Linguistics. Brisbane, Australia.
- Leech, Geoffrey (1993) Corpus Annotation Schemes. In: *Literary and Linguistic Computing* 8(4), 275 - 281
- Lüdeling, Anke (2008) Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: Walter, Maik & Grommes, Patrick (Hrsg.) *Fortgeschrittene Lernervarietäten*. Niemeyer, Tübingen, 119-140.
- Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin & Walter, Maik (erscheint) Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache 2*(2008), 67-73.
- Manning, Christopher D. & Schütze, Hinrich (1999) *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge,

- Software/Websites
  - R: http://cran.r-project.org/
  - http://faculty.vassar.edu/lowry/VassarStats.html (Statistik-Pakete)
  - http://sigil.collocations.de/wizard.html
- .