Notwendiger Informationsverlust – Annotation von Korpusdaten

Anke Lüdeling
Humboldt-Universität zu Berlin
Methodenworkshop, April 2010

Daten

"Data is ontologically different from the world. The world is as it is; data is an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data – measurements of such things as air temperature are. A text corpus is not the linguist's data – measurements of such things as average sentence length are." (Moisl 2009, 876)

Korpusdaten

- „interpretation of the world"
  - Auswahl der Daten → Korpusdesign
  - Kategorisierung der Daten → Annotation
- notwendig
  - freier Zugang zu Texten/Korpora (multimodal)
  - gute Metadaten
  - Korpusarchitektur / Infrastruktur die es erlaubt, verschiedene Modelle auf unabhängigen Ebenen zu annotieren und zu durchsuchen

Korpusdaten: Hintergrund

- vorelektronische Korpora (Textsammlungen) für Grammatikschreibung, Lexikographie, Übersetzung, Dialektstudien, historische Linguistik, …
- Methoden
  - Beispielbank / ‚Belege' für grammatische Äußerungen
  - Konkordanzen
  - (selten) Zählungen
- siehe z.B. Köhler (2005), Meyer (2008)

- „Anm.: Gelegentlich erscheint auch sonst ein *s* in der Kompositionsfuge nach Femininum, ohne daß es in die Schriftsprache durchgedrungen ist, vgl. z.B. *Gemeindsversammlung* Hebel 452, 24, *Huldszeichen* Heine 2, 111, *über Naturs Größe* Le. 11, 209, 5, *Sprachsverbesserer*, Leibniz, Unvorgreifl. Ged. 67,3, *Vernunftswahrheiten* Le. 12, 434, 32. Belege für Anfügung eines s an einen weiblichen Genitiv sind noch: *Erdens-Götter* Lohenst., Cleop. 2291 ..."
  Hermann Paul (1959 (1920)), Band V, 13)

- „Anm.: Gelegentlich erscheint auch sonst ein *s* in der Kompositionsfuge nach Femininum, ohne daß es in die Schriftsprache durchgedrungen ist, vgl. z.B. *Gemeindsversammlung* Hebel 452, 24, *Huldszeichen* Heine 2, 111, *über Naturs Größe* Lessing 11, 209, 5, *Sprachsverbesserer*, Leibniz, Unvorgreifl. Ged. 67,3, *Vernunftswahrheiten* Lessing 12, 434, 32. Belege für Anfügung eines s an einen weiblichen Genitiv sind noch: *Erdens-Götter* Lohenstein, Cleop. 2291 ..."
  Hermann Paul (1959 (1920)), Band V, 13)

vorelektronische Korpora

- Design – nicht systematisch
- Suche/Auswertung – nicht systematisch
- Kategorisierung – nicht systematisch
- Datengrundlage oft nicht verfügbar
- Ist das heute wirklich anders?

Korpusdaten: Möglichkeiten

- Dokumentation
- Grammatikschreibung
- Sprachwandelforschung / historische Linguistik
- Dialektologie
- Lexikographie
- Psycholinguistik
- Soziolinguistik
- Spracherwerbsforschung (L1, L2)
- Computerlinguistik
- …

- Beispiele finden (systematisch)
- Konkordanzen
- Zählen
- Exploration eines Phänomens
- Experimentdesign (Hypothesen testen)
- Modellierung (Clustering, Simulation etc.)
- explizite Kategorisierung, Annotation

➢ Resultate werden (wenigstens im Prinzip) nachvollziehbar und reproduzierbar

Stichprobe (sampling)

- große Korpora (Milliarden Tokens),
- oft opportunistisch, oft unbekannte Inhalte
- Ziel: Computerlinguistik, Sprachtechnologie, Lexikographie, linguistische Forschung

- kleine Korpora (< 100 Million Tokens, oft viel kleiner)
- klares Design (für eine bestimmte Forschungsfrage)
- Ziel: linguistische Forschung

Stichprobe: Sprachvariation

- Sprachproduktion durch viele Faktoren beeinflusst (Modus, Situation, soziale Variablen, …)
- Variation: eine Variable und ihre Realisierungen
- Variation ist oft quantitativ und schwer zu erkennen
  - freie Variation
  - bedingte Variation

- Labov (2008), Moisl (2008), Biber (2009), …

offensichtliche Variation

- PRESIDENT: Well --
- CONNALLY: He, he could make, Mr. President, I suggest to you that somebody could make a little capital with the Speaker and with Wilbur. Now if you'll do this. Now somebody can do it. Now, they'll, they'll say, well, you, you know, they'll say, well, "You did it because
- PRESIDENT: Yeah.

http://nixon.archives.gov/forresearchers/find/tapes/watergate/trial/connally_exhibit_1.pdf; siehe z.B. Eklund (2004)

- We present a six-dimensional T2/Z2 orbifold model which arises as an intermediate step in the compactification of the heterotic string to the MSSM. The orbifold contains two pairs of inequivalent fixed points, with unbroken local gauge groups SU(5) and SU(2)×SU(4), respectively, the intersection of which gives the standard model gauge group. All bulk and brane anomalies are cancelled by the Green–Schwarz mechanism.

C. Lüdeling (2007)

offensichtliche Variation

- PRESIDENT: Well --
- CONNALLY: He, he could make, Mr. President, I suggest to you that somebody could make a little capital with the Speaker and with Wilbur. Now if you'll do this. Now somebody can do it. Now, they'll, they'll say, well, you, you know, they'll say, well, "You did it because
- PRESIDENT: Yeah.

http://nixon.archives.gov/forresearchers/find/tapes/watergate/trial/connally_exhibit_1.pdf; see e.g. Eklund (2004) on disfluencies

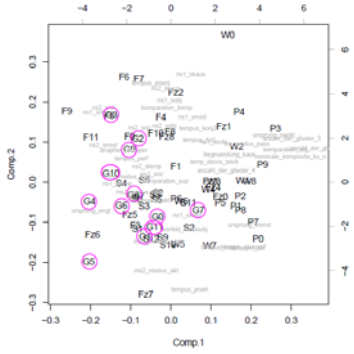
- We present a six-dimensional T2/Z2 orbifold model which arises as an intermediate step in the compactification of the heterotic string to the MSSM. The orbifold contains two pairs of inequivalent fixed points, with unbroken local gauge groups SU(5) and SU(2)×SU(4), respectively, the intersection of which gives the standard model gauge group. All bulk and brane anomalies are cancelled by the Green–Schwarz mechanism.

C. Lüdeling (2007)

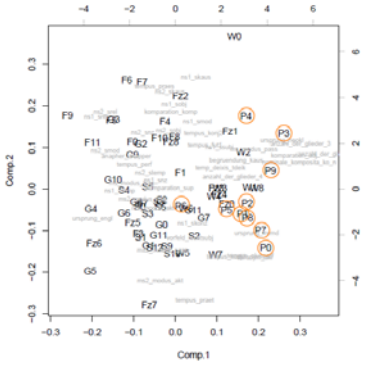
weniger offensichtliche Variation

- Einleitungen und Zusammenfassungen in wissenschaftlichen Artikeln (Biber 2009)
- Zeitungstexte vor und nach dem 11.09.2001 (Rehbein 2004)
- Lehrwerke für Friseure und Lehrwerke für Mechatroniker (Niederhaus 2009, in Vorb.)
- verschiedene Rubriken in der FAZ (Seminar Register WS 2009/2010)

Gesellschaft



Politik



Beispiel: kanonische und nichtkanonische Sätze

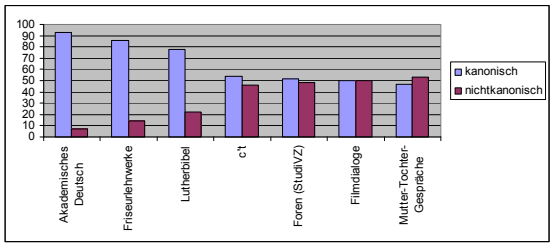
- kanonische Sätze: alle Äußerungen mit einem finiten Verb
  (können in gängigen Grammatikmodellen analysiert werden)
- nichtkanonische Sätze: Äußerungen ohne finites Verb
  (können in gängigen Grammatikmodellen nicht analysiert werden)

Beispiel: kanonische und nichtkanonische Sätze

- kanonische Sätze: alle Äußerungen mit einem finiten Verb
  (können in gängigen Grammatikmodellen analysiert werden)
- nichtkanonische Sätze: Äußerungen ohne finites Verb
  (können in gängigen Grammatikmodellen nicht analysiert werden):
  Überschriften, Einwortantworten, Bildunterschriften, Einwürfe etc.

Beispiel: kanonische und nichtkanonische Sätze

(Seminar nichtstandardisierte Sprache WS 2007/2008)



Korpora für Variationsstudien

- Korpora, die sich *nur* in dem zu untersuchenden Parameter unterscheiden
- nutzbares Format
- nutzbare Metadaten
- Annotation auf verschiedenen Ebenen
- qualitative und quantitative Modelle

Biber (1998, 2006, 2009), Conrad (2001),

Kategorisierung der Korpusdaten

- implizite Kategorisierung schon bei der Aussage über Einzelfälle

- „Anm.: Gelegentlich erscheint auch sonst ein *s* in der Kompositionsfuge nach Femininum, ohne daß es in die Schriftsprache durchgedrungen ist, vgl. z.B. *Gemeindsversammlung* Hebel 452, 24, *Huldszeichen* Heine 2, 111, *über Naturs Größe* Le. 11, 209, 5, *Sprachsverbesserer*, Leibniz, Unvorgreifl. Ged. 67,3, *Vernunftswahrheiten* Le. 12, 434, 32. Belege für Anfügung eines s an einen weiblichen Genitiv sind noch: *Erdens-Götter* Lohenst., Cleop. 2291 ...“
Hermann Paul (1959 (1920)), Band V, 13)

Kategorisierung der Korpusdaten

- implizite Kategorisierung schon bei der Aussage über Einzelfälle
- klarer bei Abstraktion über Einzelfälle und
- ➢ Modellbildung
- daraus: Vorhersage über ungesehene Fälle

Kategorisierung - konzeptuell

- ideal: Angabe von (operationalisierbaren) hinreichenden und notwendigen Bedingungen
  (kein Interpretationsspielraum, im Prinzip verlustfrei automatisierbar)
- meistens: Angabe von hinreichenden und notwendigen Bedingungen nicht möglich

Beispiel Informationsstruktur – Tag: *Focus*

- (aus Dipper, Götze, Skopeteas 2007, http://www.sfb632.uni-potsdam.de/~d1/sfb632_guidelines/isis7.htm#_Toc160896392)
- Focus: That part of an expression which provides the most relevant information in a particular context as opposed to the (not so relevant) rest of information making up the *background* of the utterance. Typically, focus on a subexpression indicates that it is selected from possible alternatives that are either implicit or given explicitly, whereas the background can be derived from the context of the utterance.
- Unit: Focus can extend over different domains in the utterance (like affixes, words, clause constituents, whole clause) and can be discontinuous as well. One expression can contain more than one focus.

Beispiel: *Focus*

- Der jahrelang Kampf um den Erhalt der Kartoffelsorte "Linda" hat sich gelohnt: Das Bundessortenamt hat den Anbau als freies Saatgut ohne Lizenzregelung genehmigt. Vor vier Jahren hatte der Saatguterzeuger Europlant das Patent auslaufen lassen, ein Anbau war also nicht mehr erlaubt, auch wenn es viele Bauern nun erst recht getan haben. [http://www.slowfood.de/]

Beispiel: *Focus*

- Der jahrelang Kampf um den Erhalt der Kartoffelsorte "Linda" hat sich gelohnt: Das Bundessortenamt hat den Anbau als freies Saatgut ohne Lizenzregelung genehmigt. Vor vier Jahren hatte der Saatguterzeuger Europlant das Patent auslaufen lassen, ein Anbau war also nicht mehr erlaubt, auch wenn es viele Bauern nun erst recht getan haben. [http://www.slowfood.de/]

Beispiel: *Focus*

- Der jahrelang Kampf um den Erhalt der Kartoffelsorte "Linda" hat sich gelohnt: Das Bundessortenamt hat den Anbau als freies Saatgut ohne Lizenzregelung genehmigt. Vor vier Jahren hatte der Saatguterzeuger Europlant das Patent auslaufen lassen, ein Anbau war also nicht mehr erlaubt, auch wenn es viele Bauern nun erst recht getan haben. [http://www.slowfood.de/]

Beispiel: *Focus*

- die Richtlinien sind nicht klar operationalisierbar – unterschiedliche Annotatoren entscheiden sich unterschiedlich
  (das inter-annotator agreement ist niedrig)
- kann man darauf eine Theorie aufbauen?
- soll man deswegen gar nicht kategorisieren?

Beispiel: *Focus*

- soll man deswegen gar nicht kategorisieren?

➢ Das ist nicht möglich!
Jede Analyse beinhaltet eine Interpretation.

➢ Viele Analysen (Modelle/Theorien) beruhen auf impliziten, d.h., für andere nicht nachvollziehbaren und nicht reproduzierbaren Dateninterpretationen

➢ In einem Korpus können Interpretationen sichtbar und nachvollziehbar angegeben werden
→ Annotation

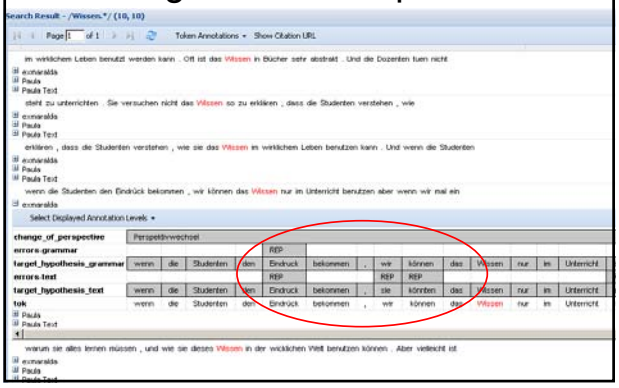
Beispiel 2: Lernerdaten

- *Und wenn die Studenten den Eindrück bekommen, wir können das Wissen nur im Unterricht benutzen aber wenn wir mal ein Job bekommen, dann sollen wir alles noch einmal lernen,*

konfligierende Interpretation

- Fehleranalyse nicht möglich ohne (implizite) Zielhypothese
- der (jeder) Satz kann verschiedene Zielhypothesen haben
- Zielhypothesen sind abhängig von theoretischen Annahmen
  – zeigt man, was man zeigen will?

konfligierende Interpretation



konfligierende Interpretation: relevant?

- Experiment: 17 Sätze, 5 Annotatoren (Lüdeling 2008)

➢ Interpretation (Annotation) der Daten *muss* verfügbar sein

➢ konfligierende Interpretationen derselben Daten sollten möglich sein!

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

Annotation in Korpora – technisch

- lange: inline Annotation
  (Annotation in derselben Datei wie der Text)
- seit einigen Jahren, aber noch selten:
  standoff Annotation
  (Annotation unabhängig von den Daten)
  tokenbasiert, spannenbasiert, Bäume (Graphen), …
- Multimodale Korpora
- Links zu externen Ressourcen
- mächtige Suchwerkzeuge

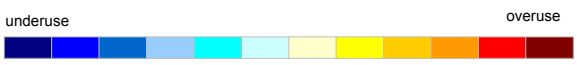
- Leech (1993), Carletta et al. (2003), Section IV of Lüdeling & Kytö (2008), Wittenburg (2008)

Annotation & Visualisierung

- möglicher Einwand: Annotation kostet Zeit & bringt nichts, da man nur rausbekommt, was man reinsteckt
- neben Nachhaltigkeit, Wiederverwendbarkeit, Transparenz und Reproduzierbarkeit
- Integration von verschiedenen Annotationsebenen und gute Visualisierung erlauben neue Erkenntnismöglichkeiten

Beispiel für Visualisierung: Mindergebrauch

- aus dem deutschen Lernerkorpus Falko
- Studien zu Übergebrauch/Mindergebrauch:
  Vergleich von L1-Daten und L2-Daten, normalisiert
- Forschungsfrage: finde Kandidaten für ‚schwierige' Strukturen



- Lüdeling et al. (2008), Zeldes, Lüdeling & Hirschmann (2008), WHIG-Projekt

Visualisierung im Korpus

Visualisierung des Experiments: lexikalische Elemente

lemma	tot_norm	de	da	en	fr	pl	ru
in	0.013188	0.012261	0.014041	0.014247	0.015272	0.012135	0.009534
es	0.010897	0.011945	0.010900	0.011379	0.013347	0.008163	0.012385
sie	0.010618	0.008193	0.010643	0.008835	0.010909	0.006067	0.005613
man	0.010164	0.007900	0.012438	0.008742	0.009754	0.006950	0.007306
dass	0.009522	0.007404	0.012823	0.008789	0.009625	0.008880	0.009890
von	0.007982	0.007122	0.007309	0.006846	0.007315	0.010259	0.007930
auch	0.007028	0.008362	0.008527	0.005828	0.005775	0.005461	0.004455
für	0.006683	0.007201	0.006091	0.007216	0.006802	0.005736	0.004188
sind	0.006465	0.004271	0.008976	0.007308	0.006930	0.004964	0.005346
sich	**0.006309**	**0.011687**	**0.006283**	**0.006291**	**0.006930**	**0.007170**	**0.005435**
ich	0.006262	0.003877	0.013272	0.005366	0.005465	0.001434	0.001426
aber	0.006048	0.003347	0.007309	0.006245	0.007315	0.003365	0.003831

sich ist mindergebraucht, unabhängig von der L1 der Lerner

Visualisierung des Experiments: Annotationen

bigram	tot_norm	de	da	en	fr	pl	ru
$.-PPER	0.042384	0.005297	0.009748	0.007963	0.006166	0.005801	0.007409
VVFIN-$,	0.042131	0.006457	0.00776	0.006343	0.006937	0.006243	0.008391
PPOSAT-NN	0.041739	0.008058	0.007247	0.007269	0.007066	0.006298	0.005802
ADV-ADV	**0.041604**	**0.012852**	**0.010518**	**0.006111**	**0.006166**	**0.003094**	**0.002856**
ADV-APPR	0.039742	0.009117	0.008016	0.005324	0.007837	0.004807	0.004642
PDAT-NN	0.03956	0.005409	0.004233	0.005509	0.007837	0.007735	0.008837
ADV-ART	0.037125	0.007629	0.006349	0.006898	0.005653	0.006133	0.004463

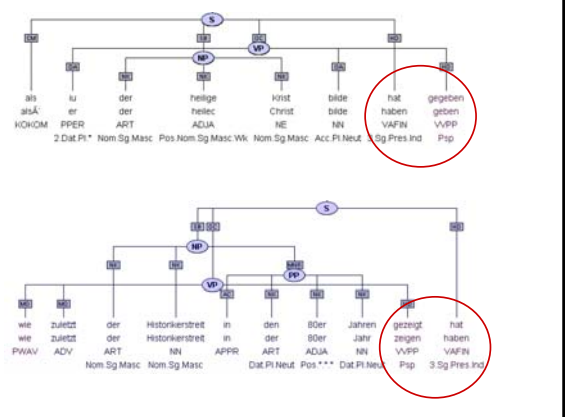
ADV-ADV ist mindergebraucht, unabhängig von der L1 der Lerner, Hirschmann 2010

Beispiel: Sprachwandel

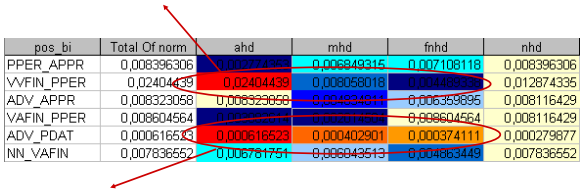
- Wandel von syntaktischen Konstruktionen
- SPEC Baumbank, (winzige) parallele Baumbank von AHD, MHD, FNHD, NHD, Tiger-Format
  (Hirschmann & Linde, in Vorb.)

	pos_bi	Total Of norm	ahd	mhd	fnhd	nhd
2	ART_NN	0,055600322		0,055600322	0,053497942	0,047579065
3	ADJA_NN	0,035052377	0,014796547	0,035052377	0,025813692	0,022949902
4	APPR_ART	0,024174053		0,024174053	0,022072578	0,021270641
5	VVPP_VAFIN	0,012874335	0,012874335	0,010475423	0,009726899	0,012874335
6	PPOSAT_NN	0,024352651	0,024352651	0,022965551	0,015336571	0,017912119
7	NN_VVFIN	0,029284834	0,029284834	0,010475423	0,010849233	0,015673104
8	NN_KON	0,040382244	0,040382244	0,019742143	0,013468013	0,013154212

- Mindergebrauch nimmt ab – Wandel?



- keine einheitliche Tendenz
  – andere Faktoren?



pos_bi	Total Of norm	ahd	mhd	fnhd	nhd
PPER_APPR	0,008396306	0,008396306	0,006849315	0,007108118	0,008396306
VVFIN_PPER	0,02404439	0,02404439	0,008058018	0,014489362	0,012874335
ADV_APPR	0,008323058	0,008323058	0,004634931	0,006359895	0,008116429
VAFIN_PPER	0,008604564	0,008604564	0,008604564	0,008604564	0,008116429
ADV_PDAT	0,000616523	0,000616523	0,000402901	0,000374111	0,000279877
NN_VAFIN	0,007836552	0,006781751	0,006943513	0,004863449	0,007836552

- Übergebrauch nimmt ab – Wandel?

Zusammenfassung: Korpusdaten und Annotation

- Ziele: wissenschaftliche Standards – Transparenz & Reproduzierbarkeit
- Korpusdaten sind eine wichtige Datenquelle für linguistische Studien
- *jede* Datenarbeit ist Interpretation
  - Auswahl der Daten
  - Kategorisierung der Daten
- viele Faktoren, die den Sprachgebrauch qualitativ und quantitativ beeinflussen, sind noch nicht wirklich gut erforscht
  → Metadaten, ad hoc-Korpora
- konfligierende Interpretation müssen sichtbar gemacht werden
- Daten und Interpretation *müssen* verfügbar sein

Vielen Dank!

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

Annis: http://www.sfb632.uni-potsdam.de/~d1/annis/

Kontakt: anke.luedeling@rz.hu-berlin.de