Frankenpost

Das Deutsche Referenzkorpus (DeReKo) als Basis empirisch linguistischer Forschung

Marc Kupietz

19. März 2008
GLOBAL COE INTERNATIONAL WORKSHOP

DeReKo auch bekannt unter...

- IDS-Korpora
- Mannheimer Korpora
- COSMAS-Korpora
- Archiv der Korpora geschriebener Gegenwartssprache am IDS

- seit 2004: „Deutsches Referenzkorpus“
- kurz „DeReKo“

Verantwortliches Projekt:

- *Ausbau und Pflege der Korpora geschriebener Gegenwartssprache*
- 7 Mitarbeiter
- aber die meisten mit weniger als 25% ihrer Arbeitszeit

Institutionelle Einbettung

Institut für Deutsche Sprache (IDS)

- Abteilung Grammatik
- Abteilung Pragmatik
- Abteilung Lexik
  - Programmbereich Lexikologie und Lexikografie
  - Programmbereich Korpuslinguistik
    - Ausbau und Pflege der Korpora geschriebener Sprache
    - Methoden der Korpusanalyse und -erschließung

    Arbeitsgruppe mit: Cyril Belica, Marc Kupietz, Rainer Perkuhn

Allgemeiner Hintergrund

- Auftrag des IDS (zitiert aus der Satzung):
  „... die deutsche Sprache in ihrem gegenwärtigen Gebrauch zu erforschen und zu dokumentieren ...“

→sehr große Sprachkorpora

→systematischer Aufbau von Korpora seit 1969

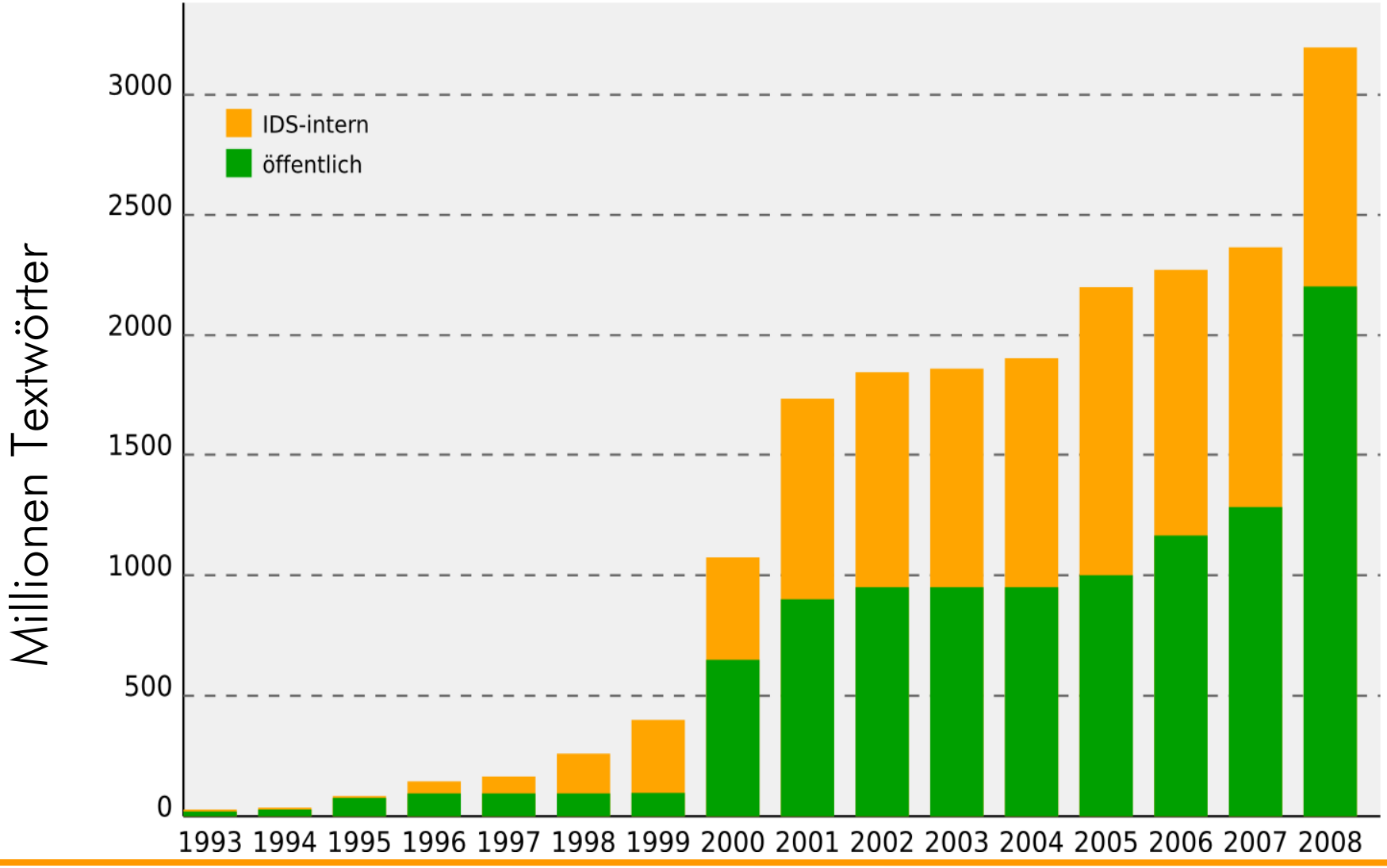
Korpora

- (Sprach-)Korpus:
  Sammlung mündlicher/schriftlicher Sprachproduktionen
  für linguistische Analysen
  → Stichprobe eines Ausschnitts des Sprachgebrauchs
- aussagekräftige quantitative Analysen;
  insbes. seltene Phänomene
  → große Korpora: *"more data is better data"*
  (Mercer, zitiert in Church 1993)
- extreme Größe & technische Analysemöglichkeiten
  → enormes Potenzial für empirische Forschung!!

DeReKo-Eigenschaften

- über 3,2 Milliarden Wörter
  (ca. 8 Millionen Buchseiten, Leseseit > 50 Jahre)
- weltweit größte linguistisch motivierte Sammlung deutschsprachiger Texte
- Texte ab 1956
- belletristische, wissenschaftliche und populärwissenschaftliche Texte, Zeitungstexte, weitere Textarten
- wird laufend erweitert
- urheberrechtlich abgesichert

Entwicklung des Gesamtumfangs



DeReKo-Nutzung

- weltweit ca. 13.000 Nutzer
- über die Korpus-Such- und Analysesoftware COSMAS II
  - separates Projekt
- über die das „gläserne Labor“, bzw. die Kookkurrenzdatenbank CCDB

Akquisition neuer Texte

- vorweg: das IDS kann keine Texte kaufen
- stattdessen: Spenden von Nutzungsgenehmigungen
  - das heißt auch: das IDS besitzt (fast) keine Texte
    - Ausnahmen: MK1, MK2, BZK
- Gegenleistungen, bescheidene...
  - Aufwandsentschädigungen
  - Werbung
- in Kontakt mit ca. 300 (potenziellen) Textspendern
- Laufend neue Texte von ca. 20 Textgebern

Textspender

Ablauf der Akquisition

- Kontaktaufnahme per Brief oder telefonisch
- Suche eines geeigneten Ansprechpartners
  - z.B. Chefredakteur, Verleger, Chef vom Dienst
- Aushandlung von Nutzungsvereinbarung und Gegenleistungen
- Probetexte zur Formatanalyse (Aufwandsabschätzung)
- Aushandeln des Datentransfers
- Unterzeichnung der Nutzungsvereinbarung

Rechtliche Rahmenbedingungen

- de jure:
  - Urheberrecht
    - in soweit beschränkt, dass kurze Passagen zitiert werden dürfen
    - aber: öffentliche Verfügbarmachung von Kopien illegal
  - Lizenzrecht
    - Nutzungsvereinbarungen mit Rechteinhabern und Urhebern für jeden Text!
- de facto:
  - Rechtsstreit mit Verlagen unmöglich
  - IDS ist abhängig von Verlagen
  - guter Ruf darf nicht gefährdet werden

Nutzungsvereinbarungen

- mit Verlagen / Zeitungen / Zeitschriften
- teilweise auch mit Autoren (insb. bei Monographien)
- nur einfache Nutzungsgenehmigung, Recht zur …
  - „datentechnischen“ Aufbereitung
  - Verfügbarmachug über spezielle Zugangssoftware
    - beschränkter (Wissenschaft…) authentifizierter Nutzerkreis
    - mittel- und unmittelbare kommerzielle Verwertung muss ausgeschlossen sein → (Endnutzervereinbarungen)
    - Volltexte dürfen nicht rekonstruierbar sein
    - Zugriffe müssen aufgezeichnet werden
    - Missbrauch muss möglichst technisch ausgeschlossen werden

Endnutzervereinbarungen (EULA)

- jeder Nutzer muss …
  - sich namentlich registrieren
  - sich mit einer ausschließlich wissenschaftlichen, nicht-kommerziellen Nutzung einverstanden erklären
  - elektronisch unterzeichnen

Hürden bei der Akquisition

- oft kein geeigneter Ansprechpartner
- Angst vor Missbrauch der gespendeten Texte (Download aus dem Netz)
- urheberrechtliche Probleme
- misstrauische Rechtsabteilungen
- Texte nicht elektronisch verfügbar oder nicht rekonstruierbar

Rohtextproblem

- Daten von Textgebern kommen in allen denkbaren Formaten:
  - z.B. Word, PDF, RTF, XML, HTML, Pseudo-XML, unstrukturierte Texte, Quark XPress, Spaltensatz, ...
- Fehler und Unsystematisches nicht direkt erkennbar
  - zu große Datenmengen!
- Metadaten (z.B. wer ist der Autor, was ist eine Überschrift) oft nur teilweise konstruierbar
- „manuelle“ Eingriffe in die Konvertierung unmöglich!

→ teilweise opportunistische Auswahl von Texten

Konvertierung von Rohtexten

- nicht für jedes Format einen kompletten Konvertierer
- stattdessen: einige wenige Zwischenformate, von denen aus immer die selben Konvertierer verwendet werden können
- d.h. zunächst in ein XML-Format
- von da aus weiter mit einer Hierarchie von Transformationswerkzeugen
- XML-Transformationssprache XSLT
- möglichst wenig neu entwickeln…

IDS-Textmodell

- möglichst originalgetreue Abbildung von Inhalt und Struktur der Quelltexte
- breites Spektrum der erfassbaren Textarten
- hierarchische Gliederung des Datenbestandes auf den drei Ebenen Korpus, Dokument und Text
- Auszeichnung von bibliografischen, textstrukturellen und weiteren Informationen, die für die Recherche und Auswertung sinnvoll/nützlich erscheinen

Konkretes Format: IDS-XCES

- XCES: Corpus Encoding Standard for XML
- leider kein ISO-Standard und sehr unvollständig
- de-facto-Standard zur Enkodierung von Korpusdaten
- deshalb IDS-XCES mit eigenen Erweiterungen
- Mitarbeit im zuständigen ISO-Gremium
  (ISO TC 37 SC4 *Language Resources*)

XCES-Illustration

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE idsCorpus PUBLIC "-//IDS//DTD IDS-XCES 1.0//EN"
   "/usr/local/kl/textmodell/IDSXCES/ids.xcesdoc.dtd">
<idsCorpus version="1.0" TEIform="teiCorpus.2">
  <idsHeader type="corpus" pattern="Ztg/Zschr" status="new"
   version="1.0" TEIform="teiHeader">
    <fileDesc>
      <titleStmt>
        <korpusSigle>T06</korpusSigle>
        <c.title>die tageszeitung 2006</c.title>
      </titleStmt>
      <editionStmt version="1.0"/>
      <publicationStmt>
        <distributor>     Institut für Deutsche Sprache
    </distributor>
        <pubAddress>      Postfach 10 16 21, D-68016 Mannheim
    </pubAddress>
        <telephone>       +49 (0)621 1581 0         </telephone>
```

„Repräsentativität“

- häufiger Wunsch: „ein repräsentatives Korpus für die deutsche Sprache“
- aber: Was ist denn die deutsche Sprache?
- Antwort:
  - jeder beantwortet die Frage anders
  - in jedem Kontext anders
- Repräsentativität nur in Abhängigkeit der avisierten Grundgesamtheit definierbar!
  - von jedem Anwender
  - spezifisch für jede Fragestellungen

Maximen der Akquisition

- Menge
- Stratifikation, Streuung entlang Dimensionen wie:
  - Zeit
  - Ort
  - Thema
  - Register
  - Genre
  - Textsorte
  - ....
- Zusammensetzung der Stichprobe muss in der Nutzungsphase festgelegt werden!

Lösung: Virtuelle Korpora

- … Name der Lösung
- z. Zt. in COSMAS2:
  - dokumtentweise „von Hand“
  - vorgefertigte virtuelle Korpora
  - auf Anfrage …
- in Zukunft anhand einer Text-Metadaten-Datenbank:
  - z.B. Zeitungstexte von 1985-1990 zum Thema „Wirtschaft“
  - z.B. 20% Belletristik, 80% Zeitungstexte, nur wenn Überschriften als solche ausgezeichnet sind
  - z.B. keine Texte aus Österreich und Schweiz
  - bestimmtes Vokabular (→ Lernerkorpora?)

Woher Metadaten?

- bzw.: wie bestimmt man die Zusammensetzung eines (virtuellen) Korpus?
- bzw.: woher weiß man, für welche Grundgesamtheit eine Stichprobe repräsentativ sein könnte?
- Antworten:
  - Angaben der Textgeber, z.B.
    - Zeitungsressort
    - Datum der Veröffentlichung, ...
  - Vergleiche

Vergleiche womit?

- Texte eines Korpus untereinander → z.B.
  - Wiederholungen
  - Dubletten, teilweise Duplikate
- mit einem anderen Korpora → z.B.
  - Eigenschaften auf Lexemebene (auffällige Wörter)
  - n-Gramme, Wortverbindungen
  - Andere Eigenschaften: z.B. Satzlängen
- mit manuell klassifizierten Vergleichstexten → z.B.
  - Thema
  - Register

Thematischer Vergleich mit dewac

DIFF	dewac	dereko	Thema
10.32	12.71	2.39	Staat_Gesellschaft:Biographien_Interviews
3.15	4.96	1.81	Kultur:Literatur
2.51	4.66	2.15	Staat_Gesellschaft:Recht
2.47	4.71	2.24	Staat_Gesellschaft:Familie_Geschlecht
1.75	5.63	3.88	Freizeit_Unterhaltung:Reisen
1.69	4.09	2.40	Kultur:Musik
1.46	2.59	1.13	Kultur:Film
0.99	2.49	1.49	Staat_Gesellschaft:Kirche
0.72	1.55	0.82	Technik_Industrie:EDV_Elektronik
0.69	0.75	0.06	Staat_Gesellschaft:Tod
			⋮
-0.98	3.87	4.85	Freizeit_Unterhaltung:Vereine_Veranstaltungen
-1.04	2.15	3.19	Wirtschaft_Finanzen:Sozialprodukt
-1.17	0.95	2.12	Sport:Vermischtes
-1.84	0.98	2.83	Staat_Gesellschaft:Verbrechen
-2.01	0.82	2.83	Technik_Industrie:Unfaelle
-2.43	7.13	9.56	Politik:Kommunalpolitik
-2.72	1.76	4.48	Wirtschaft_Finanzen:Oeffentliche_Finanzen
-3.42	5.54	8.96	Politik:Ausland
-3.65	3.94	7.59	Sport:Fussball
-4.95	5.75	10.70	Politik:Inland

Autom. Erkennung von Duplikaten

T03/JUL.36384 die tageszeitung, 25.07.2003, S. 28, Ressort: tazplan-Programm; Diese Woche frisch

Neu im Kino:

Diese Woche frisch

Brandzeichen – Momente der Rebellion: Doku über den Kampf gegen die Neoliberalisierung in Argentinien **Das verordnete Geschlecht:** Interviews mit Hermaphroditen **Die Blume des Bösen:** Claude Chabrol seziert wieder die französische Provinzbourgeoisie und deren Kellerleichen **Früchte der Liebe:** ein schwuler Pianistengott, sein jugendlicher Liebhaber und dessen Mutter bilden ein Dreieck **Natürlich blond 2:** Lustig gemeinter Blondinenfilm **Planet der Kannibalen:** Düsterer Sciene Fiction, schwarzweiß mit einem kleinen Lichtstreif am Horizont **Raumpatrouille Orion – Rücksturz ins Kino:** Das Weltraumabenteuer unserer Eltern jetzt endlich im Kino **Sindbad – Herr der 7 Meere:** Der Held aus 1001 Nacht als cooler Slacker **The Gathering:** Horror mit Christina Ricci **Vampire Hunter D:** Zeichentrickfassung des beliebten japanischen Vampircomics, ganz ohne Bisse

NEU KINO WOCHE FRISCH BRANDZEICHEN MOMENTE REBELLION DOKU KAMPF NEOLIBERALISIERUNG ARGENTINIEN VERORDNETE GESCHLECHT INTERVIEWS HERMAPHRODITEN BLUME BSEN CLAUDE CHABROL SEZIERT FRANZSISCHE PROVINZBOURGEOISIE DEREN KELLERLEICHEN CHTE LIEBE SCHWULER PIANISTENGOTT SEIN JUGENDLICHER LIEBHABER DESSEN MUTTER BILDEN DREIECK NATRLICH BLOND _ LUSTIG GEMEINTER BLONDINENFILM PLANET KANNIBALEN DSTERER SCIENE **FICTION** SCHWARZWEI KLEINEN LICHTSTREIF HORIZONT RAUMPATROUILLE ORION RCKSTURZ INS KINO WELTRAUMABENTEUER UNSERER ELTERN JETZT ENDLICH KINO SINDBAD HERR _ MEERE HELD _ NACHT COOLER SLACKER THE GATHERING HORROR CHRISTINA RICCI VAMPIRE HUNTER D ZEICHENTRICKFASSUNG BELIEBTEN JAPANISCHEN VAMPIRCOMICS GANZ OHNE BISSE

T03/JUL.37208 die tageszeitung, 30.07.2003, S. 28, Ressort: tazplan-Programm; Diese Woche frisch

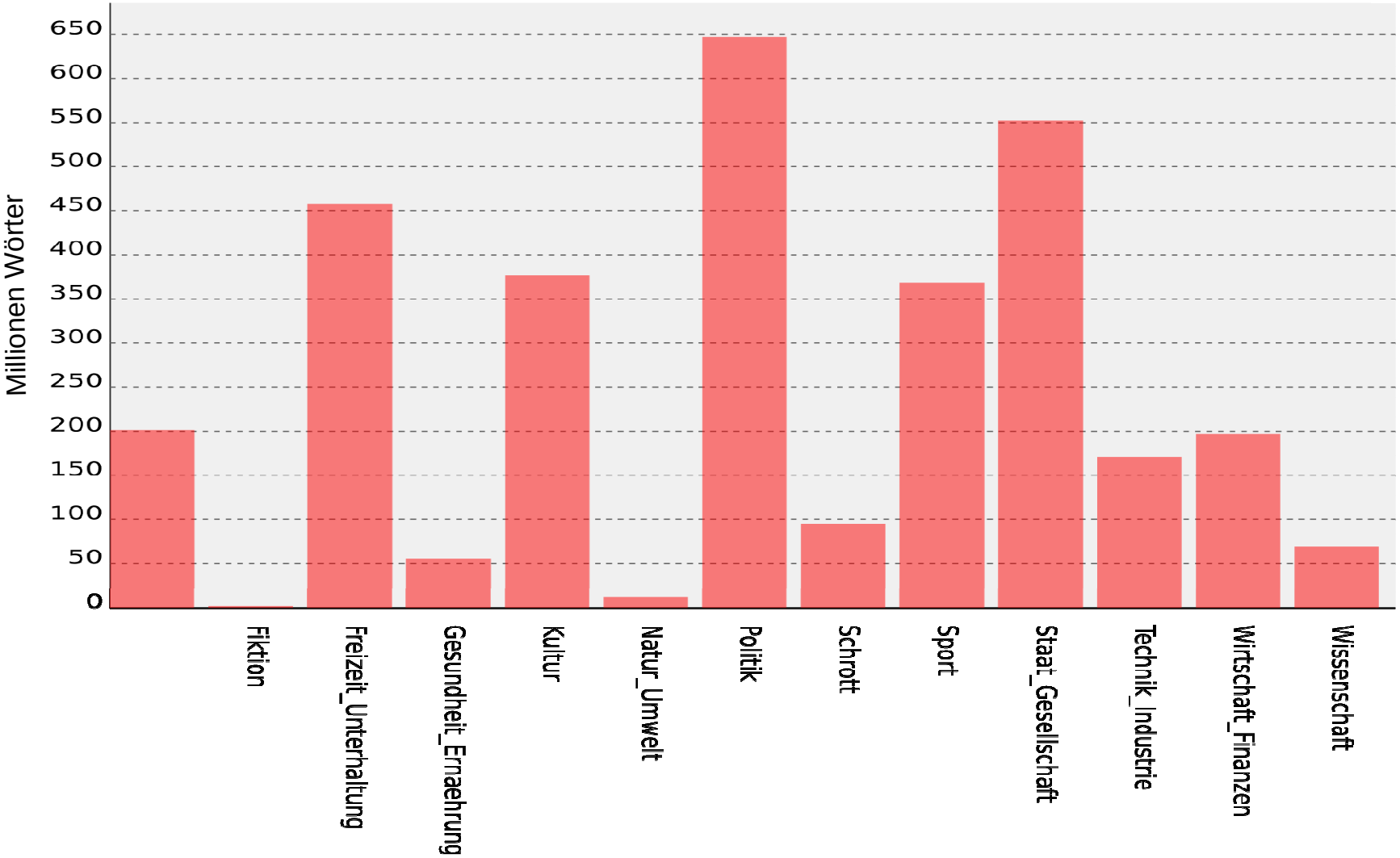
Neu im Kino:

Diese Woche frisch

Brandzeichen – Momente der Rebellion: Doku über den Kampf gegen die Neoliberalisierung in Argentinien **Das verordnete Geschlecht:** Interviews mit Hermaphroditen **Die Blume des Bösen:** Claude Chabrol seziert die französische Provinzbourgeoisie und deren Kellerleichen **Früchte der Liebe:** Ein schwuler Pianistengott, sein jugendlicher Liebhaber und dessen Mutter im Dreieck **Natürlich blond 2:** Lustig gemeinter Blondinenfilm **Planet der Kannibalen:** Düsterer Sciencefiction, schwarzweiß mit einem kleinen Lichtstreif am Horizont **Raumpatrouille Orion – Rücksturz ins Kino:** Das Weltraumabenteuer unserer Eltern jetzt endlich im Kino **Sindbad – Herr der 7 Meere:** Der Held aus 1001 Nacht als cooler Slacker **The Gathering:** Horror mit Christina Ricci **Vampire Hunter D:** Zeichentrickfassung des beliebten japanischen Vampircomics, ganz ohne Bisse

NEU KINO WOCHE FRISCH BRANDZEICHEN MOMENTE REBELLION DOKU KAMPF NEOLIBERALISIERUNG ARGENTINIEN VERORDNETE GESCHLECHT INTERVIEWS HERMAPHRODITEN BLUME BSEN CLAUDE CHABROL SEZIERT FRANZSISCHE PROVINZBOURGEOISIE DEREN KELLERLEICHEN CHTE LIEBE SCHWULER PIANISTENGOTT SEIN JUGENDLICHER LIEBHABER DESSEN MUTTER DREIECK NATRLICH BLOND _ LUSTIG GEMEINTER BLONDINENFILM PLANET KANNIBALEN DSTERER SCIENCEFICTION SCHWARZWEI KLEINEN LICHTSTREIF HORIZONT RAUMPATROUILLE ORION RCKSTURZ INS KINO WELTRAUMABENTEUER UNSERER ELTERN JETZT ENDLICH KINO SINDBAD HERR _ MEERE HELD _ NACHT COOLER SLACKER THE GATHERING HORROR CHRISTINA RICCI VAMPIRE HUNTER D ZEICHENTRICKFASSUNG BELIEBTEN JAPANISCHEN VAMPIRCOMICS GANZ OHNE BISSE

Zusammensetzung nach Thema



Nachvollziehbarkeit und Replizierbarkeit

- für jedes Fragestellung ein virtuelles Korpus
- permanente Erweiterung des Archivs

→ sind Forschungsergebnisse replizierbar?

- Lösungsansatz:
  - ISO-Standard für persistente Identifikation von Sprachressourcen (*Citation of Electronic Resources "CitER"*– im Entwurfsstadium)
    - jedes virtuelle Korpus erhält eine persistente ID
  - Korpusdaten in Versionierungssystem (subversion)
    - Archivzustände wiederherstellbar (seit 2/2007)

Linguistische Annotation

- (automatische) Auszeichnung von Wortarten, Satzteilen, etc.
- für einige wenige Korpora bereits vorhanden
- vollständige Annotation geplant für 2008/2009
- Probleme:
  - nicht unmittelbar beobachtete Daten, sondern Interpretationen
  - gerade bei „interessanten“ Phänomenen geringe Präzision
  - → nicht die Sprache wird untersucht, sondern das Annotationswerkzeug
- Lösungsansatz: Multiple, konkurr. Annotationen, ...

Viele Daten → viel Verantwortung

- nicht nur in Bezug auf Qualität
- sondern auch in Bezug auf Verfügbarkeit
- in Zukunft auch über APIs *(application programming interfaces)* verfügbar
- Eingliederung in deutsche und internationale eScience und Infrastruktur-Initiativen:
  - CLARIN (*Common Language Resources and Technology Infrastructure*)
    - viele Korpora auf einmal recherchierbar
  - TextGrid

Vielen Dank!

korpuslinguistik@ids-mannheim.de