Linguistische Forschungsdaten

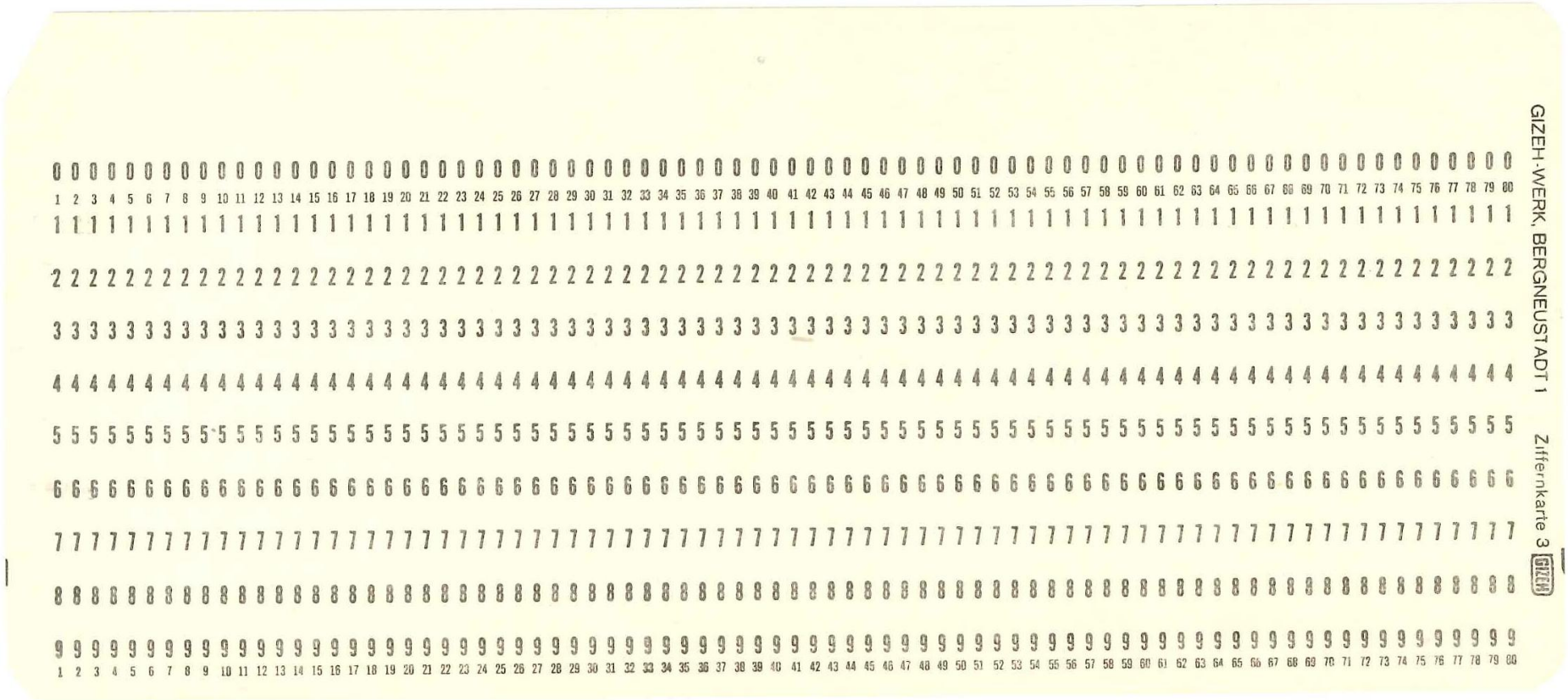
Andreas Witt

Institut für Deutsche Sprache (IDS), Mannheim

45 Jahre und mehr… IDS, Korpora, Verfügbarkeit, rechtliche Fragen (1)

1964 Das Institut für Deutsche Sprache wird gegründet

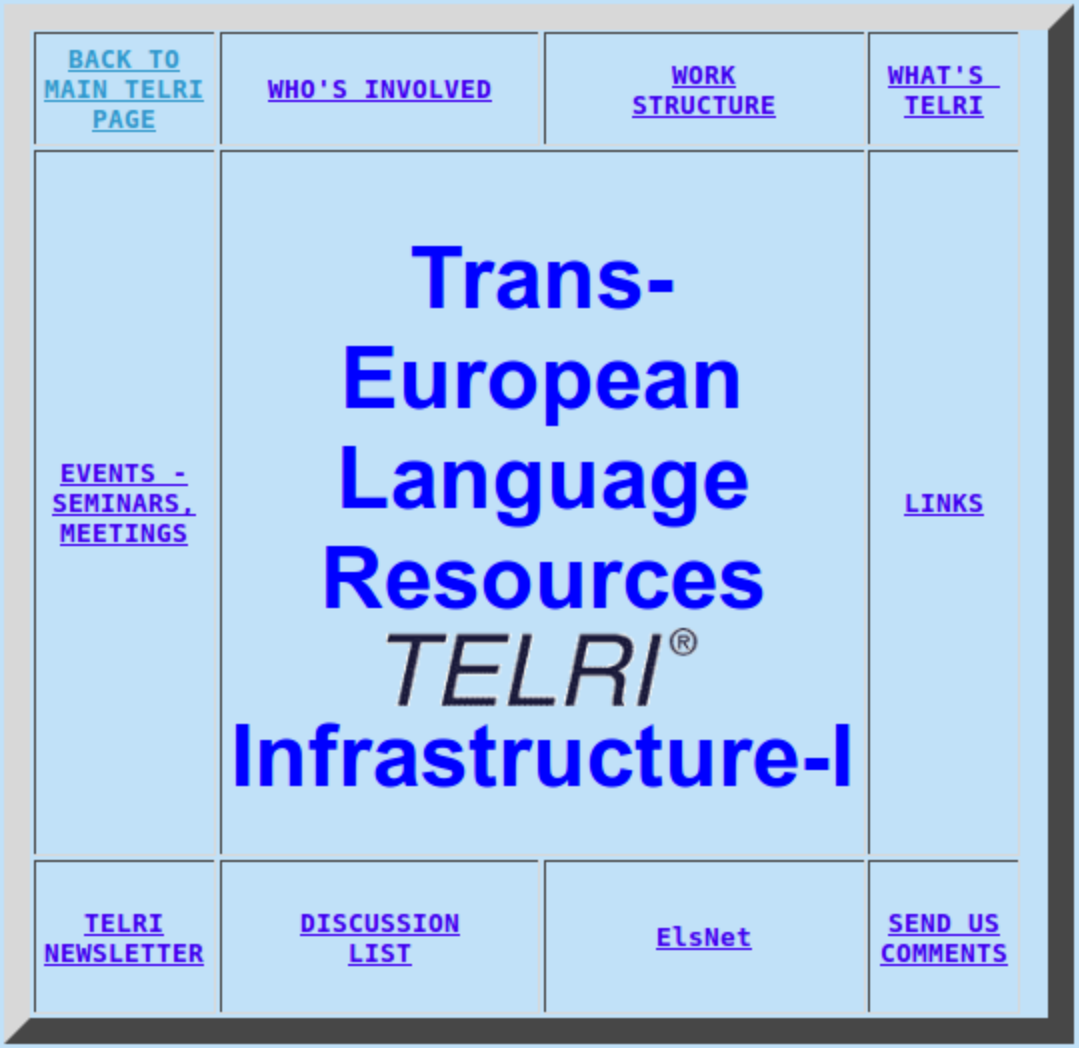
1967 Erstes elektronisches Korpus (MK1) auf Lochkarten: 2.2 Millionen Worte


GIZEH-WERK, BERGNEUSTADT 1
Ziffernkarte 3
GIZEH

45 Jahre und mehr… IDS, Korpora, Verfügbarkeit, rechtliche Fragen (2)

1992 IDS-Korpora öffentlich über das Internet zugänglich (30 Millionen Worte)

1995 TELRI nimmt Arbeit auf (koordiniert am IDS)



45 Jahre und mehr… IDS, Korpora, Verfügbarkeit, rechtliche Fragen (3)

1999 erstes Rechtsgutachten (im Rahmen des Projekts DEREKO I)

2001 bedeutender Rechtsstreit mit einer Verwertungsgesellschaft für digitale Medien

2004 Göttinger Erklärung zum Urheberrecht für Bildung und Wissenschaft von der WGL unterzeichnet

2008 CLARIN/D-SPIN nimmt Arbeit auf

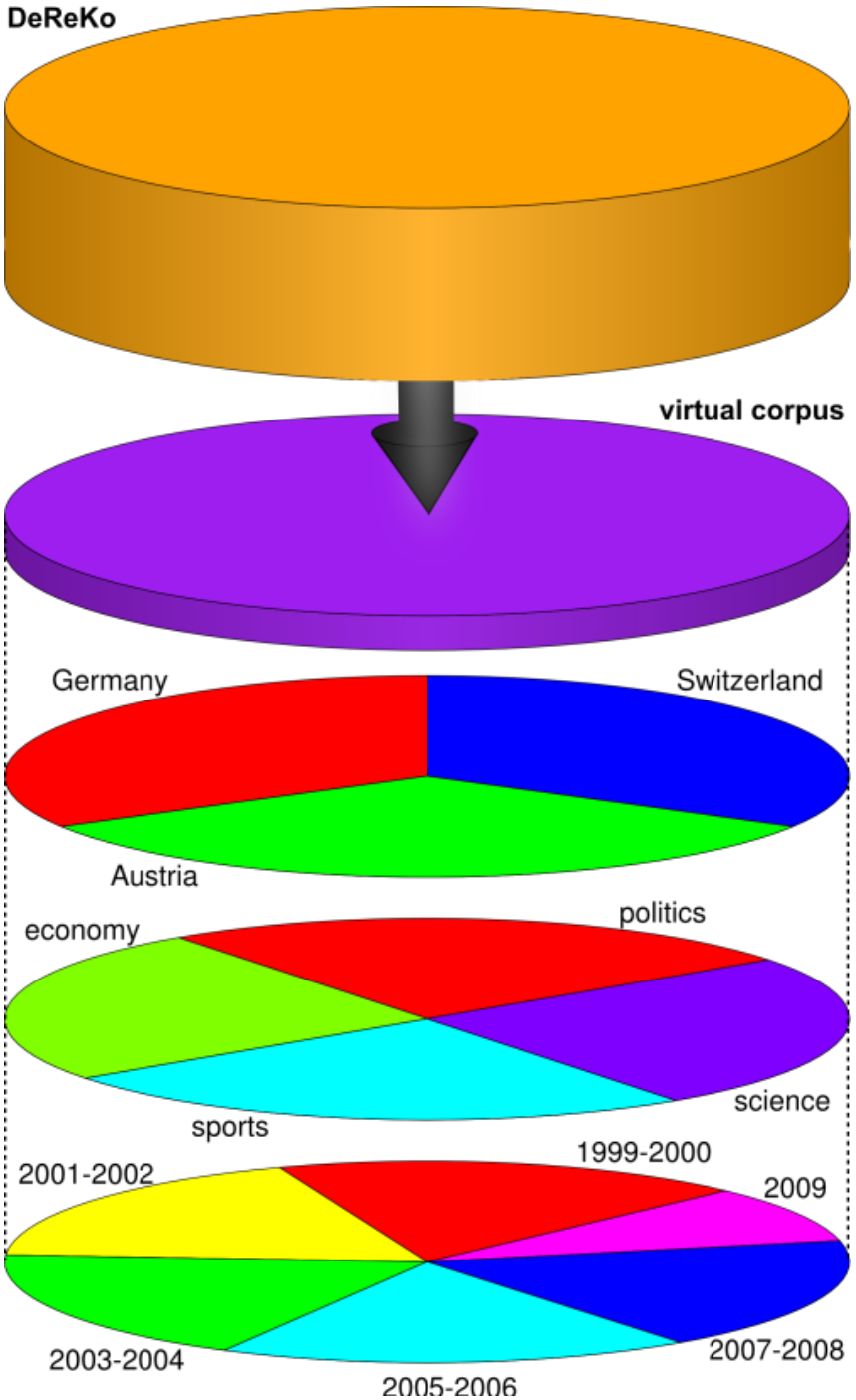
Korpora für geschriebenes Gegenwartsdeutsch am IDS (1)

- Deutsches Referenzkorpus (DeReKo)
- Ziel: empirische Grundlage für die germanistische sprachwissenschaftliche Forschung
- Aufbau begann1964
- Maximale Größe, Konzept der Ur-Stichprobe

Konzept der Ur-Stichprobe

- DEREKO wurde nicht als „ausgewogenes“ Korpus konzipiert, weil Ausgewogenheit nur in Bezug auf Parameter definiert werden kann
- die Ur-Stichprobe enthält das gesamte durch das IDS bezogene Textmaterial
- die Texte sind mit Metadaten angereichert
- Wissenschaftler können ihre Korpora durch Auswahl einer Teilmenge aus der Ur-Stichprobe definieren
- das definierte Korpus wird „virtueller Korpus“ genannt
- ein definiertes virtuelles Korpus hat einen PID

Beispiel zur Definition eines virtuellen Korpus



Korpora für geschriebenes Gegenwartsdeutsch am IDS (2)

- aktuelle Größe 4,3 Milliarden Wörter (Aufwuchsrate 300 Millionen Wörter im Jahr)
- von 18.500 registrierten Benutzern verwendet

Wesentliche Merkmale von DEREKO (1)

- enthält im Wesentlichen Texte ab 1956
- wird kontinuierlich ausgebaut
- enthält
  - literarische,
  - wissenschaftliche,
  - Zeitungstexte und
  - viele andere Textsorten
- nur vollständige und unveränderte Texte, d.h. keine Verbesserung von Rechtschreibfehlern usw.

Wesentliche Merkmale (2)

- nur lizenziertes Material
- nicht zum Download verfügbar (aufgrund von Lizenz-Verträgen und Urheberrecht)
- maximale Größe, Konzept der Ur-Stichprobe
- ermöglicht das Zusammenstellen von spezialisierten Stichproben
- derzeit mit gleichzeitig drei Annotationsschichten ausgestattet

Rechtliche Aspekte

- das IDS hat Lizenzverträge mit den verschiedenen Rechteinhabern (Verlage, Autoren) vereinbart
  - keine Kommerzialisierung der Daten, Nutzung ausschließlich zu Forschungszwecken
  - Zugang nur über Software, die u.a. die Lesbarkeit von Volltexten verhindert
  - die Lizenzen sind unbefristet, können aber jederzeit gekündigt werden

    → das IDS kann die Persistenz von Texten nicht garantieren

➢ die Situation ist nicht ideal aber das IDS muss den Drahtseilakt zwischen den Interessen der Zielgruppen und der Rechteinhaber wagen

Nachnutzung und Nachhaltigkeit

- Überführung von Rohdaten
  - Pipeline von im Verlagswesen üblichen Formaten zu angepasstem XCES
    - spezialisierte Filter
  - Migration nach TEI P5 möglich
- Persistenz und Datenhaltung
  - seit 2007 Daten in Versionierungssystem
  - hauseigene Persistent Identifier
  - Digitale Langzeitarchivierung muss im IDS erfolgen
- Es IDS kann gezwungen werden, Datensätze (Artikel) zu entfernen

Linguistische Annotationen

- Erste Annotationen
  - 1995: Logos Tagger, 1999: Gertwol Tagger
- Neuere Ansätze: Machinese Phrase Tagger, TreeTagger, Xerox FST, Xerox XIP
  - 5 TB der Standoff-Annotationen
  - Jede Annotationsebene ist durch Lizenzvereinbarungen eingeschränkt
- DeReKo-2009-I (IDS, 2009) erstmals mit Annotationen veröffentlicht
- (die Annotationen sind jedoch nur teilweise zugänglich)

NUTZUNG VON DEREKO

- COSMAS II: Corpus Management und Analyse Tool
  - entworfen 1993
  - Zusammenstellung der virtuelle Korpora
  - bietet komplexe Suchmöglichkeiten
    - z.B. Lemmatisierung, Wordabstandsoperatoren, Suche über Satzgrenzen, logische Operatoren
  - kann komplexe (nicht-zusammenhängende) Kollokationsanalysen höherer Ordnung durchführen
  - bietet verschiedene Darstellungen für Suchergebnisse und diverse Benutzeroberflächen
- Eine langfristige Archivierung der Suchanfrage und/oder Suchergebnisse findet nicht statt

Neues Korpus-Analyse-System (1)

- „KorAP“ gestartet am 1.7.2011
- Anforderungen
  - es muss für die Durchführung von methodisch fundierter, empirischer sprachwissenschaftlicher Forschung geeignet sein
  - die untersuchte Datenbasis muss von deren Interpretation unterscheidbar sein
  - es muss große Mengen an Textdaten und Annotationen (30 Milliarden Wörter mit 20 Annotationsschichten) verarbeiten können
  - der Suchmechanismus sollte Multi-Layer-Abfragen ermöglichen
  - Abfrage-, Analyse-und Metadaten-Funktion sollten mit digitalen Infrastrukturen verknüpfbar sein

Neues Korpus-Analyse-System (2)

- Anforderungen (Fortsetzung)
  - virtuelle Korpora sollten anhand von Metadaten und textimmanenten Eigenschaften definierbar sein
  - Nutzern sollte es möglich sein, ältere Datenbestände zu bearbeiten
  - Nutzern sollte es möglich sein, sich virtuelle Korpora (oder Kollektionen) dauerhaft anzeigen zu lassen
  - Nutzern sollte es möglich sein, kumulative Annotationen zu erzeugen
  - Nutzern sollte es möglich sein, die Daten mit eigenen Programmen zu bearbeiten
  - das System muss garantieren, dass keine Lizenzbedingungen verletzt werden
- Eine langfristige Archivierung der Suchanfragen und damit die Reproduzierbarkeit der Suchergebnisse soll hierdurch möglich werden

Zusammenfassung

- Korpusausbau seit 1964
- das IDS hat trotz rechtlicher Schwierigkeiten umfangreiche Sprachressourcen aufgebaut
- Wichtigste Voraussetzungen für die Nachhaltigkeit
  - kontinuierliche Pflege
  - Nutzwert und Anwendbarkeit für der Ziel-Community
- das IDS wird neue Korpus-Management-Software entwickeln
- das IDS beteiligt sich an Infrastrukturvorhaben mit der Zielrichtung Nachhaltigkeit und Zugänglichkeit von Sprachressourcen

Die wichtigsten Faktoren für Nachhaltigkeit von DEREKO

- organisatorische Stabilität
- kontinuierliche Entwicklung (z. B. durch Einrichtung eines Langzeitprojekts)
- maximale Größe und Schichtung (Konzept der Ur-Stichprobe)
- vielseitige Nachnutzung durch virtuelle Korpora
- enge Vernetzung von Forschungs-, Entwicklungs- und Infrastrukturvorhaben
- Abschätzung künftiger Nutzeranforderungen

Planung

- Aufbau eines Archivs für
  - germanistische Forschungsprimärdaten *und*
  - Forschungspublikationen
- Diese Arbeiten beginnen in diesem Monat

Vielen Dank!