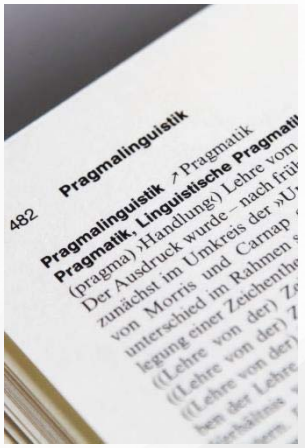
INSTITUT FÜR DEUTSCHE SPRACHE



Thomas Schmidt

VOM ANALOGEN ARCHIV ZUM DIGITALEN FORSCHUNGSDATENZENTRUM

Aktuelle Herausforderungen im Archiv für Gesprochenes Deutsch (AGD)

Mitglied der Leibniz-Gemeinschaft

GLIEDERUNG

Vom analogen Archiv zum digitalen Forschungsdatenzentrum

– Aktuelle Herausforderungen im Archiv für Gesprochenes Deutsch (AGD)

1. Historisches
2. Aktuelles
3. Zukünftiges

DEUTSCHES SPRACHARCHIV (DSAV)

- Gegründet **1932** als "Deutsches Spracharchiv" von Eberhard Zwirner (Vorbilder: Phonogrammarchive in Wien und Zürich): Sprachaufnahmen auf **Schallplatte** zur phonometrischen Untersuchung der Struktur gesprochener Sprache
- Nach **1945**: Neugründung in Braunschweig, ab 1955 Erhebung des Korpus „Deutsche Mundarten“ (aka Zwirner-Korpus): Aufnahmen auf **Tonband** mit Rundfunktechnik, ca. 10000 Aufnahmen bis 1970, vergleichbare Erhebung in der DDR
- 1971 Übernahme ins Institut für Deutsche Sprache, dort seit **1965** Erhebung von Gesprächskorpora („Grundstrukturen“, „Dialogstrukturen“) auf **Kompaktkassette**
- Ab ca. **1990**: Digitalisierung der Bestände, damals mit **DAT + CD**

INSTITUT FÜR DEUTSCHE SPRACHE

AUFGABEN DES DSAV 1982

Das Deutsche Spracharchiv

im

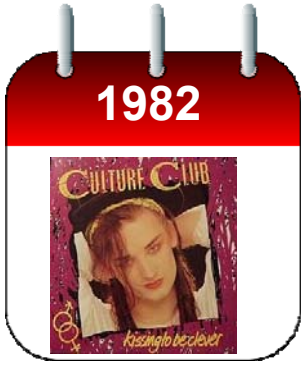
Institut für deutsche Sprache

MITTEILUNGEN 6

DEUTSCHES SPRACHARCHIV

Edeltraud Knetschke / Margret Sperlbaum

Mannheim 1983



AUFGABEN DES DSAV 1982

III. SERVICELEISTUNGEN

1. Kopien (Tonband, Protokoll, Text)

„[neben Forschung und Dokumentation] gleichrangig: die [Archivbestände] jederzeit **der Forschung in Form von Kopien zur Verfügung zu stellen**. […] Dabei [wird] auf den äußersten **Schutz der Person** […] geachtet.

Inzwischen sind insgesamt über 40.000 Tonbandkopien an einzelne Wissenschaftler und Forschungsinstitute abgegeben worden. Größere Korpora sind seit 1974 […] kopiert worden [für]:

[…]

Dr. Jörg Bergmann, Universität Konstanz

[…]

Margret Selting, Universität Bielefeld

Dr. Jan Wirrer, Universität Bielefeld

[…, insgesamt **37 Einträge** in der Liste]“

AUFGABEN DES DSAV 1982

2. Clearing / Information

„[…] weitere Aufgabe […], jede Art von Anfrage zum Quellenmaterial und zur Methodik seiner Auswertung sowie **Fragen zur Dokumentation und Archivierung gesprochener deutscher Sprache** zu beantworten.

[…] Einzelforschern und studentischen Gruppen die Möglichkeit bieten […], sich mit […] **methodischen Verfahren** […] vertraut zu machen. […] Dazu gehören […] Informationen über die Herstellung der Typoskripte auf mit phonetischer Tastatur ausgerüsteten IBM-Maschinen. [Diese Schreibmaschine] ist **nach Einführung des API-Systems** […] von Mitarbeitern des DSAv in Zusammenarbeit mit IBM entwickelt worden.

[…] sind dem DSAv [zwei weitere **Korpora] zur Archivierung und Dokumentation übergeben** worden.“

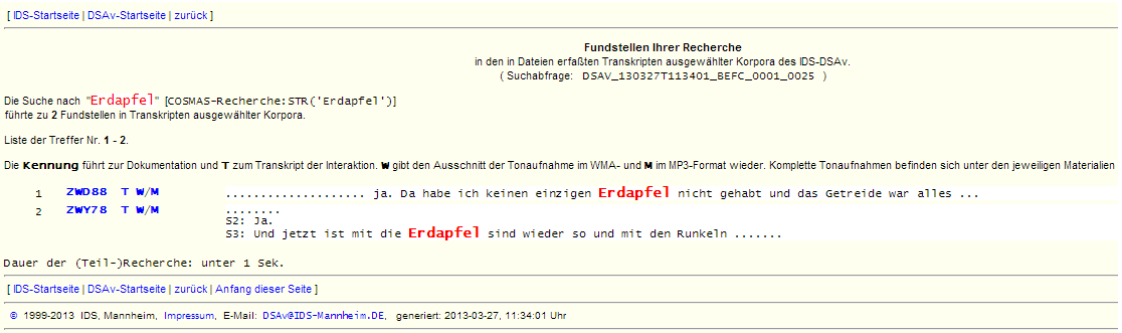
SCHRIFTVERKEHR MICHAEL CLYNE /DSAV 1966

Einige unserer "Postgraduate Students" und ich beabsichtigen, im Laufe der kommenden Ferien (Weihnachten bis Anfang Februar) einige Tonbandaufnahmen von den wenigen noch vorhandenen zweisprachigen Einwohnern der ehemaligen deutschen Siedlungen Westvictorias zu machen. Tonbänder und eventuelle Ergebnisse dieser kleinen Forschung würden wir Ihnen gerne zur Verfügung stellen. Sollten Sie irgendwelche Vorschläge oder Wünsche über die Durchführung dieser Untersuchung haben, so würde ich mich freuen, sie zu hören.

Wie ich Ihnen im Februar schon schrieb ist unser Archiv fast ausschließlich aus Aufnahmen mit 19cm/sec. in Vollspur zusammengestellt. Wir wären Ihnen außerordentlich dankbar, wenn wir dann von Ihren Aufnahmen, die in der gleichen Weise aufgenommen sind, Kopien haben dürften.

INSTITUT FÜR DEUTSCHE SPRACHE

ARCHIV FÜR GESPROCHENES DEUTSCH (AGD)



- Digitalisierung von Transkripten über **OCR** ab späte 1990er Jahre
- Automatisches **Alignment** von digitalen Tonaufnahmen und digitalen Transkripten ab späte 1990er Jahre
- Computergestützter **Transkriptionstechnologie** (DIDA) ab späte 1990er Jahre
- Digitale Erfassung des Katalogs (**Metadatendokumentation**) ab späte 1990er Jahre
- **Webbasierter Zugriff** über die Datenbank Gesprochenes Deutsch ab 1997
- 2004 Restrukturierung, Umbenennung in „Archiv für Gesprochenes Deutsch“
- Ab ca. 2008: Entwicklung der DGD2, Aufbau Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

INSTITUT FÜR DEUTSCHE SPRACHE

ARCHIV FÜR GESPROCHENES DEUTSCH

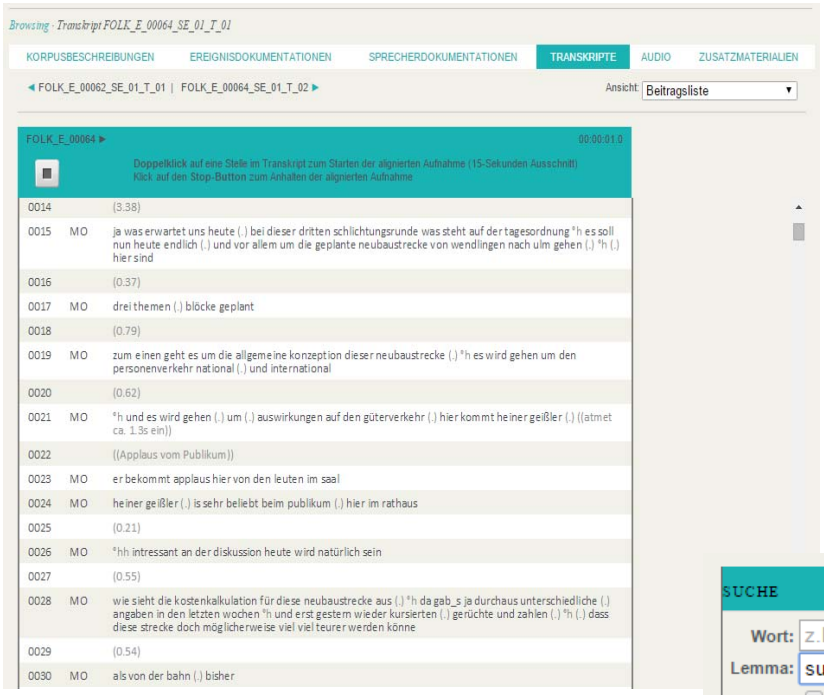
```
1803  <annotationGrp xmlns="http://standoff.proposal" who="#HBG2" start="#TLI_20"
1804      end="#TLI_22">
1805      <u xmlns="http://www.tei-c.org/ns/1.0" xml:id="c19"><w xml:id="w17">sup<anchor
1806              synch="#TLI_21"/>er</w></u>
1807      <spanGrp xmlns="http://www.tei-c.org/ns/1.0" type="lemma">
1808          <span from="#w17" to="#w17">super</span>
1809      </spanGrp>
1810      <spanGrp xmlns="http://www.tei-c.org/ns/1.0" type="pos">
1811          <span from="#w17" to="#w17">ADJD</span>
1812      </spanGrp>
1813  </annotationGrp>
```

Noch gar nicht so lange her (2005-2012):

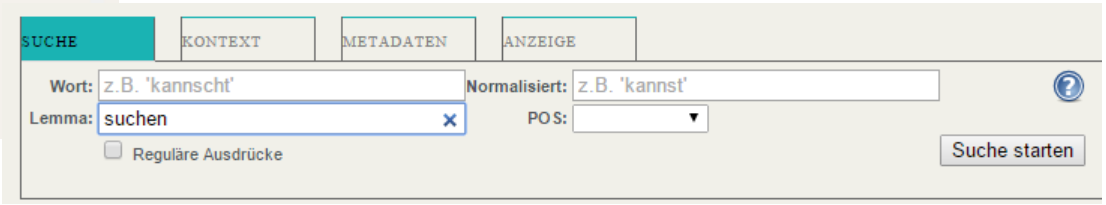
- Ablage von Audio- und Videodaten auf **RAID-Platten** mit Backup (vorher: CD)
- Umstellung des gesamten Managements von Textdaten (Transkripte und Metadaten) auf **XML** (vorher: SGML / TXT / HTML)
- **Datenbanksystem** (Oracle mit XML-Unterstützung) zur Verwaltung der Daten im Backend (vorher: Perl + MS-DOS)
- **HTML5** zur Integration audiovisueller Daten in Webseiten (vorher: ???)
- Archivstandards, Hardware-Kapazitäten für **Video** (vorher: Ø)

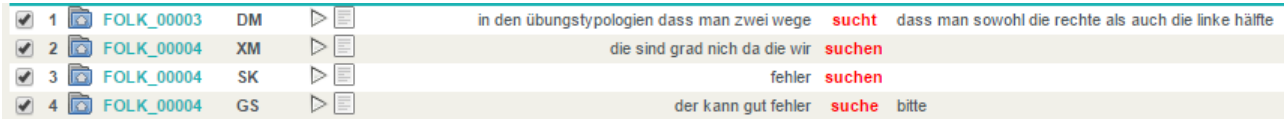
INSTITUT FÜR DEUTSCHE SPRACHE

DATENBANK FÜR GESPROCHENES DEUTSCH



- 23 Korpora aus dem AGD
- 10.000 Datensätze, 3000 Stunden Audio, 8.5 Millionen Tokens
- Online Browsing
- Online Query
- Download





SAME, SAME…

Datenübernahmen und -aufbereitung

- Korpus „Australiendeutsch“
- Korpus „Ausbildung im Bergbau“
- Korpus „Mehrsprachige Kita-Kinder“
- Korpus „Jugendkommunikation“
- Korpus „Deutsch in Ozeanien“
- …

Aufbau eigener Korpora

- Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)
- Korpus „Deutsch Heute“

Bereitstellung von Daten

- Datenbank für Gesprochenes Deutsch (DGD2)
- Persönlicher Archiv-Service

SAME, SAME…

Entwicklung von Korpustechnologie / Standards

- DGD
- FOLKER, OrthoNormal, EXMARaLDA
- ISO/TEI-Standard „Transcription of Spoken Language“

Kooperation

- Aufbau des GeWiss-Korpus
- Aufbau des Korpus „Unserdeutsch“

Beratung

- Gesprächsanalytisches Informationssystem (Online-Plattform GAIS)
- Best-Practice-Richtlinien: DFG-Handreichungen
- Transkriptionsschulungen

SAME, SAME… BUT DIFFERENT

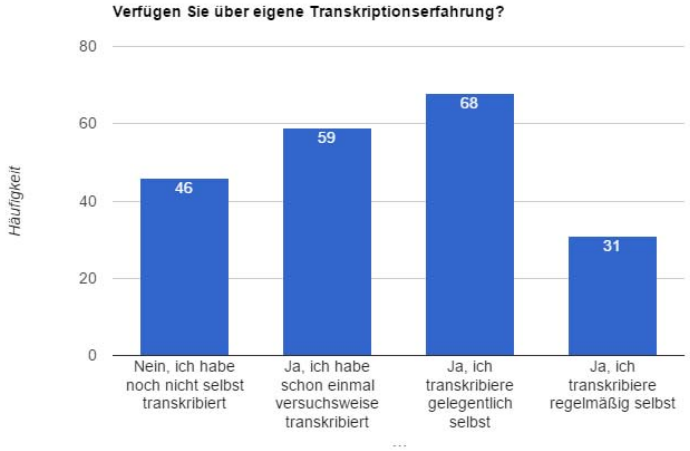
Technologie

- am Horizont: automatische Spracherkennung

Reichweite, Nutzerzahlen, Nutzertypen, Korpustypen

- 4000 registrierte DGD2-Nutzer in drei Jahren (vgl. 37 „Großabnehmer“ 1974-1982)
- Über 1000 Registrierungen jährlich für FOLKER/OrthoNormal
- Korpusarbeit als Normalfall für Forschungsprojekte
- „Ökologische Kleinbetriebe“ – Studentische, Dissertationsprojekte
- Innovative Korpustypen: Kombination mündlicher Daten mit Bilddaten, schriftlichen Daten, Fragebogendaten, …
- Diversifizierte Nutzerbasis:
  - Dialektologen, Soziolinguisten, Gesprächsforscher
  - Korpuslinguisten, Computerlinguisten, Sprachvermittler (DaF, DaZ), Sprachtechnologen
  - Qualitative Sozialforscher, Psychologen, Historiker, …

SAME, SAME… BUT DIFFERENT



Nutzerstudie „Mündliche Korpora“ (mit HZSK, GeWiss Leipzig)

- Fragebogenstudie
- Think-Aloud-Sitzungen an der DGD
- Qualitative Interviews mit „Power-Usern“

- Erwartete Ergebnisse:
  - Zugänge müssen Zielgruppen angepasst werden
  - Zugänge müssen miteinander vernetzt werden

SAME, SAME… BUT MORE

Digitale Forschungsinfrastrukturen

CLARIN-D

IDS als eines von neun **CLARIN-Zentren** in Deutschland

- Langzeitarchivierung über IDS-Repositorium
- Zentrenübergreifende Standards: CMDI, TEI/ISO (?)
- Anschluss an CLARIN-Instrumente: Virtual Language Observatory, Federated Content Search, WebLicht, CLARIN-Hepldesk, …

➔ Synergien

➔ Abstimmungsbedarf:

- Institutsintern: gemeinsamer Überbau für mündliche, schriftliche Daten
- Welche Standards?
- Welche Daten an welches Zentrum?
- Weitere Vergrößerung der Reichweite, Diversifizierung der Nutzerbasis
- Internationalisierung – CLARIN ERIC mit 16+ Mitgliedern

SAME, SAME… BUT MORE

- CLARIN: Auffinden, Auswerten, Aufbereiten und Aufbewahren
- **Aufbauen? Ausbauen?**

- Viele Nutzer finden viele Fehler
- Ca. 50% der DGD-Bestände noch nicht durch Transkription erschlossen
- Kaum Erschließung durch weitere Annotation
- Aufbau von FOLK erfordert Feldzugänge, Beiträger aus ganz Deutschland
- Kooperationspartner müssen eigene Umgebungen zum Korpusaufbau entwickeln

KRauT
Kollaborativer **Rau**m für **T**ranskription
„Crowd"-Sourcing
Digitale Plattform zum Aufbau mündlicher Korpora

WER MACHT‘S?

Interne Kompetenzen?

- (Medien-)Archivar / Medientechniker
- Korpuslinguist / Gesprächsforscher / Dialektologe
- Software-Entwickler / Texttechnologe / Sprachtechnologe / Programmierer
- Gesucht: Korpuslinguistischer Gesprächsdialektologe mit Zusatzausbildung zum digitalen Medienarchivar und Erfahrung in der Entwicklung text- und sprachtechnologischer Software-Anwendungen (mehrsprachig, mit Auslandserfahrung, promoviert)

Externe Kompetenzen und Ressourcen nutzen (siehe KRauT):

- Feldzugänge externer Projekte (Gespräche im Theater, Prüfungsgespräche, Schulunterricht) ➔ Datenspenden / „Beraten gegen Daten“
- Datenjäger
- Transkription friesischer, plattdeutscher, elsässischer Aufnahmen
- Daten für computerlinguistische / sprachtechnologische Forschungsprojekte nutzbar machen
- Entwicklung von Korpustechnologie mit Partnern abstimmen

ZUSAMMENFASSUNG

- Aufgaben ähnlich über Jahrzehnte
- Fortwährende Anpassungen an technische Entwicklungen

- Zunahme an Quantität
  - mehr Daten
  - mehr Nutzer

- Zunahme an Komplexität
  - mehr Vernetzung zwischen Daten
  - mehr Vernetzung zwischen Institutionen

- Problem: Kapazitäten und Kompetenzen müssen ausgebaut werden
- Chance: Neue Formen der Arbeitsteilung werden möglich