

Offener Diskurs: “Sprachdaten in Deutschland”

**Prof. Andreas Witt
(Institut für Deutsche Sprache)**

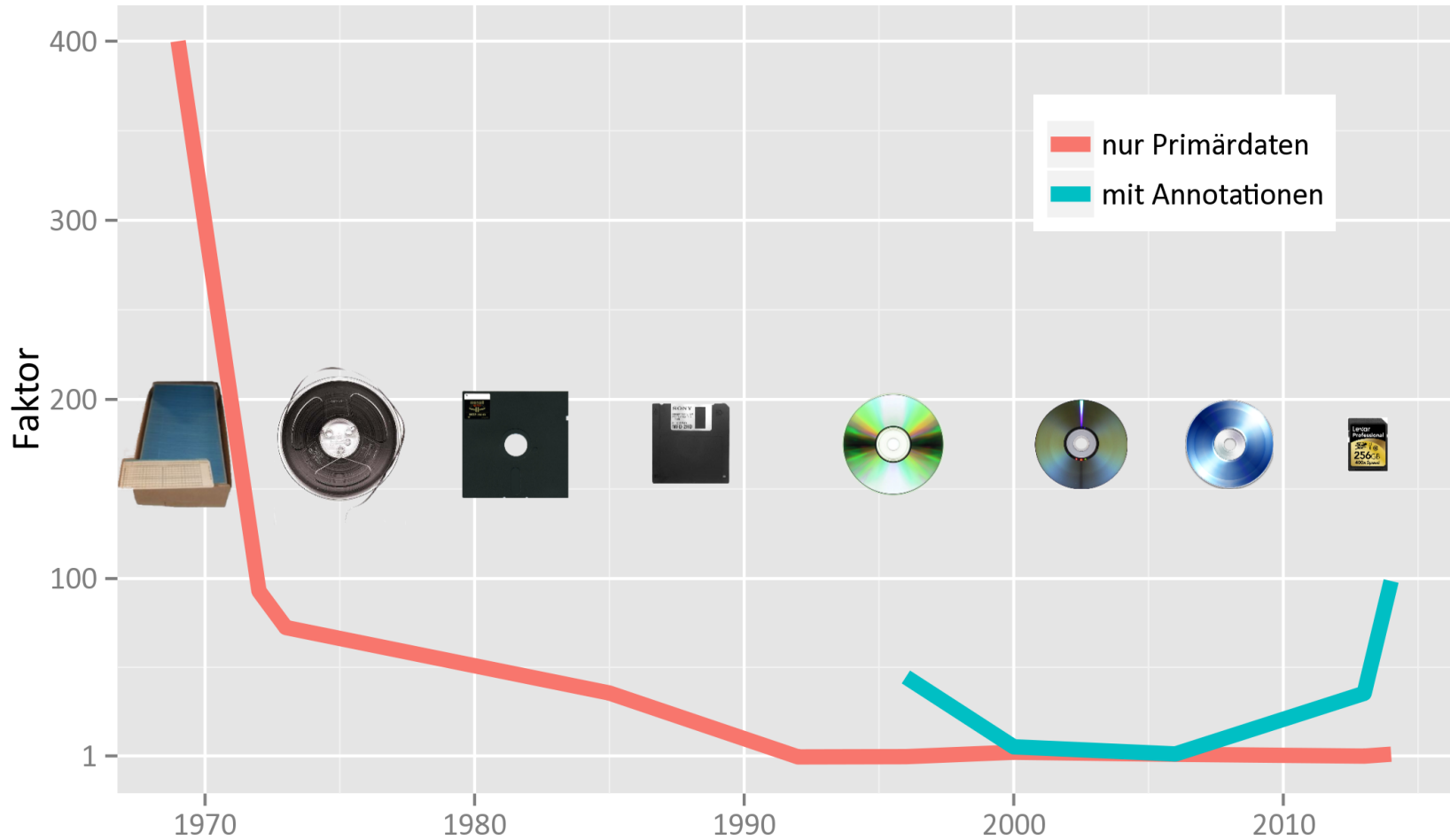
- Das Institut für Deutsche Sprache
 - **ist die zentrale außeruniversitäre Einrichtung zur Erforschung und Dokumentation der deutschen Sprache** in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte
 - besteht seit 1964 in Mannheim
 - wird IDS je zur Hälfte vom Bund und dem Land Baden-Württemberg (unter Beteiligung aller Länder) finanziert
 - ist Mitglied der Leibniz-Gemeinschaft
 - hat 210 Mitarbeiterinnen und Mitarbeiter





- Sprachdaten am IDS umfassen
 - Wörterbücher
 - Korpora gesprochener Sprache
 - Grammatikinformationssysteme
 - Textkorpora (geschriebene Sprache)

- 1969 1. Korpus am IDS veröffentlicht:
Mannheimer Korpus I (ca. 2 Mio. Wörter auf Lochkarten)
- 1983 1. Recherchesoftware in Betrieb: REFER
- 1992 1. Analysesoftware online:
COSMAS I (IDS-Korpora: ca. 20 Mio. Wörter)
- 1999 Start des BMBF-Verbundprojekts »DeReKo«
- 2004 »Korpusausbau« als Dauerprojekt etabliert
(Dr. Marc Kupietz, Dr. Harald Lungen)





- DeReKo Charakteristika:
 - enthält über 28 Milliarden Wörter (DeReKo 2015-II)
 - größte linguistisch motivierte Sammlung deutschsprachiger Texte ab 1956
 - belletristische, wissenschaftliche und populärwissenschaftliche Texte, Zeitungstexte, weitere Texttypen
 - systematischer Aufbau von Korpora seit 1964
 - wird laufend erweitert
 - Nutzung urheberrechtlich abgesichert
- Zugang zu den Korpora: Cosmas II, KorAP



Webseite	Dokumenten- paare	Wörter DE	Wörter EN	Gebiet
www.bnd.de	1	139	97	Geheimdienst
www.bundeskanzlerin.de	8	4040	4855	Allgemein
www.staedtetag.de	5	2043	1998	Allgemein
www.difu.de	3	869	639	Allgemein
www.bmi.bund.de	29	8811	9687	Allgemein
www.bundespraesident.de	11	7473	7814	Allgemein
www.bundesrat.de	2	6512	7792	Allgemein
www.bakoev.bund.de	8	1556	1730	Allgemein
www.baua.de	4	1567	1929	Arbeit
www.bgrci.de	4	490	523	Arbeit
www.bibb.de	95	43857	52094	Bildung
www.dipf.de	10	2316	3123	Bildung
www.obs-ev.de	1	179	154	Bildung
www.unesco.de	5	686	901	Bildung



Webseite	Dokumenten- paare	Wörter DE	Wörter EN	Gebiet
www.fueakbw.de	49	10717	13213	Bundeswehr
www.bundeswehrkranken haus-westerstede.de	13	3411	4204	Bundeswehr
www.kommando.streitkrae ftebasis.de	1	635	791	Bundeswehr
www.bkg.bund.de	3	687	1025	Daten
www.grs.de	27	7232	9203	Energie
www.bfe.bund.de	9	3932	4912	Energie
www.deutsche- finanzagentur.de	27	8964	11691	Finanzen
www.bundesfinanzminister ium.de	8	4586	4986	Finanzen
www.bundesbank.de	4	1226	1668	Finanzen
www.fmsa.de	2	267	473	Finanzen

- Bei den gefundenen Dokumenten handelt es sich nur um die Spitze des Eisbergs → die große Masse an CEF.AT relevanten Sprachdaten befindet sich IN den Organisationen
- Zusätzliche Probleme:
 - Dokumente können zwar klassifiziert werden ob es sich um direkte Übersetzungen handelt oder nicht
 - Aber: Viele der gefundenen Dokumente stimmen nicht direkt überein, d.h. es handelt sich nicht um Übersetzungen, sondern um unabhängige Texte
 - Es gibt deutlich weniger Dokumente für andere Sprachpaare, z.B. Deutsch-Französisch, Deutsch-Spanisch, Deutsch-Italienisch, Deutsch-Griechisch, Deutsch-Lettisch usw.



- Bundesministerien
 - Auswärtiges Amt
 - Bundesministerium des Inneren
 - Bundesministerium für Bildung und Forschung
 - Bundesministerium für Gesundheit
 - Bundesministerium der Finanzen
 - Bundesministerium für Arbeit und Soziales
 - Bundesministerium für Wirtschaft und Energie
 - Bundesministerium für Verkehr und Digitale Infrastruktur
 - Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit
 - Bundesministerium der Verteidigung
 - etc.



- Sonstige Regierungsstellen, Behörden etc. auf Bundesebene
 - Deutscher Bundestag
 - Bundessprachenamt
 - Bundesinstitut für Berufsbildung
 - Bundesnetzagentur
 - Bundeswehr
 - etc.
- Regierungsstellen und Ministerien auf Länderebene (insbesondere in Grenzregionen)
- Externe Sprachdienstleister (Language Service Providers LSPs)



- Sonstige wichtige Organisationen
 - Deutsches Institut für Sprache
 - Goethe-Institut
 - Rat für Sozial- und Wirtschaftsdaten
 - Deutsche Bahn
 - Deutsche Flugsicherung
 - Deutscher Sprachrat
 - Gesellschaft für Deutsche Sprache
 - Deutsche Welle
 - Deutsches Rotes Kreuz
 - etc.

Offener Diskurs: “Sprachdaten in Deutschland”



- Welche Stelle ist in Ihrer Organisation für die Erstellung von Übersetzungsdaten zuständig?
- Wie werden in Ihrer Organisation Sprachdaten erstellt?
 - Extern über Language Service Provider (LSPs)?
 - Intern (z.B. Sprachendienst, PR Abteilungen, IT,...?)
- Wer nutzt Sprachdaten in Ihrer Organisation?
- Wie werden die Übersetzungen und Sprachdaten verarbeitet, aufbereitet, gespeichert?
- Was sind die typischen Datenformate für Ihre Sprachdaten: Dokumente, Translation Memories, Dictionaries, Terminologien (mono- oder bi-lingual),...?
- Wo werden die Sprachdaten aufbewahrt, archiviert, vorgehalten (zentral? dezentral?)