# Entwicklung eines korpusbasierten, elektronischen Wörterbuchs Deutsche Gebärdensprache – Deutsch

## Langzeitprojekt der Akademie der Wissenschaften in Hamburg

Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, Christian Rathmann
Kontakt: dgs-korpus@sign-lang.uni-hamburg.de

### Das Projekt

Das DGS-Korpus-Projekt hat zwei Ziele:

1. Die Sammlung und Aufbereitung gebärdensprachlicher Daten in einem **annotierten Korpus**
2. Die Erstellung eines **korpusbasierten elektronischen Wörterbuchs** DGS – Deutsch

Im Rahmen des Projekts werden Sprachdaten in verschiedenen Regionen Deutschlands gefilmt und in einem Korpus zusammengestellt. Dieses wird von einem Team aus gehörlosen und hörenden Mitarbeitern am Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser transkribiert, annotiert und ausgewertet. Ein Teil der Filme wird der Gehörlosengemeinschaft und allen an der DGS Interessierten öffentlich zugänglich gemacht. Das Korpus bietet u.a. Sprachwissenschaftlern langfristig die Möglichkeit, unterschiedliche Aspekte der DGS empirisch zu untersuchen.

Für eine visuelle Sprache bietet sich eine elektronische Publikation in besonderer Weise an. Gebärden können mithilfe von Filmen und/oder Avataren bestmöglich dargestellt werden. Darüber hinaus ist effektives Suchen innerhalb eines Sprachsystems ohne festgelegte alphabetische Reihenfolge für einen gezielten Zugriff möglich.

### DGS-Korpus

- ca. 350–400 Stunden
- ca. 300 Informanten
- ca. 2,25 Mio. Tokens
- Referenzkorpus, (repräsentativ) ausgewogen
- in Größe und Umfang vergleichbar mit großen Korpora gesprochener Sprache
- Korpusdesign: Alltagskommunikation, verschiedene Textsorten, Grundwortschatz (25 Sachthemen)
- soziolinguistische Variablen: Alter, Geschlecht, sozialer Status
- Verwendungsmöglichkeiten:
  - Grundwortschatz (frequenzbasiert),
  - Grammatik (Detailtranskription)
  - Identifizierung versch. Bedeutungen und Kollokationen einer Gebärde
  - soziolinguistische Vergleichsstudien (s. Lucas et al. (2001, Schembri/Johnston (2004)).

### Öffentliches, annotiertes Teilkorpus

- repräsentative Teile des Korpus (ca. 50 Stunden)
- englische Übersetzung des Inhalts, der Glossen und der Metadaten
- mit geeigneten Austauschformaten (z.B. ELAN, IMDI)
- online verfügbar

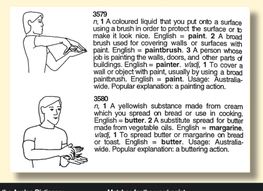
### Stand der Gebärdensprachlexikographie

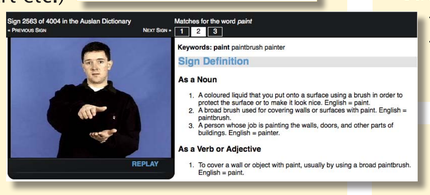
bisher v.a. viele pragmatisch orientierte Gebärdensammlungen:

- überwiegend introspektiv, nicht korpusbasiert
- zweisprachig mit einseitiger Benutzungsrichtung Deutsch→DGS
- ausgehend von deutscher Wortliste, oft keine „echten“ Gebärdeneinträge (Gebärdenlemmata fehlen)
- Mikrostruktur: wenig Informationen zur Gebärde, meist nur Gebärdenform (Zeichnung oder Film)
- Zugriff: über deutsche Wörter, evtl. Sachthemen, selten auch nach wenigen formalen Aspekten wie z.B. Handform oder Anzahl der Hände

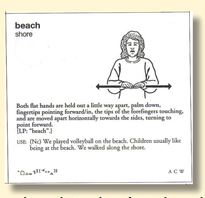
wissenschaftlich-linguistisch orientierte Gebärdensprachlexikographie:

- ASL-Dictionary (Stokoe/Casterline/Cronenberg 1976)
- AUSLAN-Dictionary (Johnston 1998)
  - teilnehmende Beobachtung; wird nachträglich durch Auswertung eines Korpus überarbeitet
  - “echte” Gebärdeneinträge
  - ausdifferenzierte Mikrostruktur mit verschiedenen Angaben zur Gebärde (Bedeutung, Wortart etc.)
  - Makrostruktur (Zugriff): Einträge sortiert nach Parametern der Gebärdenform (Handform, Ausführungshöhe…)

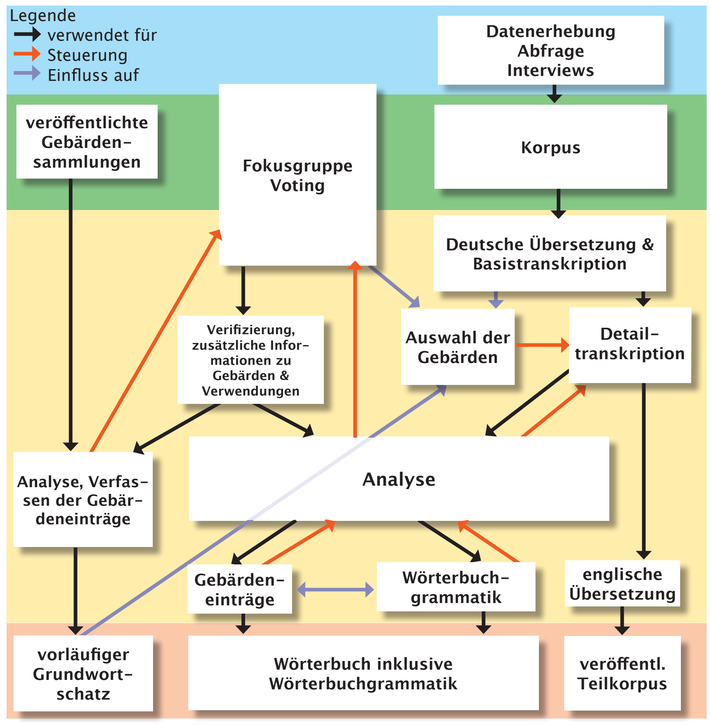



Ansätze einer korpusbasierten Gebärdensprachlexikographie:




- NZSL-Dictionary (Kennedy 1998)
  - allgemeinsprachlich,
  - teilweise korpusbasiert (kein lemmatisiertes Korpus)
  - “echte” Gebärdeneinträge
  - ausdifferenzierte Mikrostruktur mit verschiedenen Angaben zur Gebärde
- Fachgebärdenlexika des IDGS (www.sign-lang.uni-hamburg.de)
  - korpusbasiert (lemmatisiert und transkribiert)
  - Fachbegriffe, keine Alltagssprache
  - Suche nach Gebärdenform über Parameter
  - Gebärdenverzeichnis mit Gebärdeneinträgen
  - Mikrostruktur der Gebärdeneinträge: verschiedene Angaben zur Gebärde (z.B. Bedeutungen, Ikonizität, Modifikationen, Raumnutzung, Verwendung, formähnliche Gebärden etc.)


### Herausforderungen an die Gebärdensprachlexikographie

Beschreibung und Repräsentation der Form

- sprachliche Einheiten sind nicht als Lautfolgen, sondern als Bewegungen (Handzeichen/Gebärden) realisiert
  → ideales Medium zur Darstellung der Form: Film
- Gebärden werden mit Mundbewegungen, Mimik, Körperhaltungen kombiniert
- keine „Hoch“-DGS und damit keine Standardformen
  → regionale Variation berücksichtigen;
- keine Gebrauchsschrift, d.h. keine Schriftform und keine Standardorthographie

Zugriffsstrukturen: Zugriff über Form

- Filminhalte sind (noch) nicht maschinell suchbar
  → praktikable Kodierung notwendig
- es gibt kein allgemein verbreitetes, etabliertes, benutzerfreundliches Kodierungssystem mit etablierter alphabetischer Reihenfolge
- punktueller, zielgenauer Zugriff über die Gebärdenform ist schwierig
- bisher Suche über Gebärdenparameter (Handform, Orientierung der Hand im Raum, Bewegung, Ausführungsstelle etc.) wie z.B. über HamNoSys-Kodierung bei Fachgebärdenlexika, aber bei bisherigen Verfahren noch ungenau

Mikrostruktur/Angaben zur Gebärde (lexikographische Beschreibung)

- Gebärdensprachforschung bisher nicht ausreichend korpusbasiert
- Forschung in vielen Bereichen noch ohne abschließende Ergebnisse (z.B. Grammatik, Morphologie, Wortarten)
- Beschreibungskategorien entwickeln (bottom-up)
- Artikelschreibung anhand von Belegstellen in der Datenbank

Beschreibungssprache (z.B. in den Wörterbuchartikeln und Umtexten):

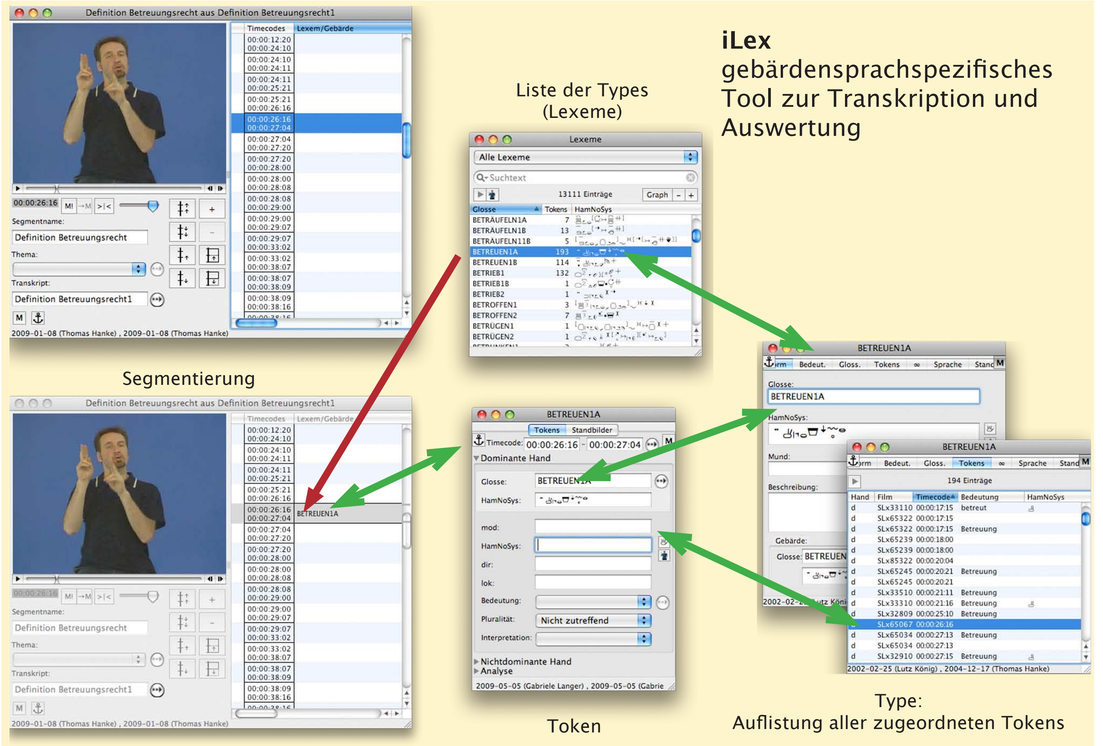
- Schriftform einer Lautsprache (z.B. Deutsch) notwendig, da
- keine Gebrauchsschrift für Gebärdensprachen verfügbar,
- Inhalte in gebärdensprachlichen Texten nicht maschinell durchsucht werden können
  - DGS-Texte flüchtig sind
  - auch deusche Gehörlose für ihre schriftliche Kommunikation geschriebenes Deutsch verwenden

### Allgemeinsprachliches, korpusbasiertes, elektronisches Wörterbuch DGS-Deutsch

- enthält ca. 6000 Gebärdeneinträge
- Informationen sind korpusbasiert
- deskriptiv
- Verwendungsbeispiele aus dem Korpus
- auf DGS konzentriert (ausdifferenzierte Gebärdeneinträge)
- Deutsch-Teil verschafft hörenden Nutzern Zugang zur DGS; gehörlose Nutzer bekommen grundlegende Informationen über das jeweilige deutsche Wort
- bidirektional benutzbar, z.B. über Suche nach Gebärdenform
- Kombination aus verschiedenen Wörterbuchtypen
- dient vor allem den folgenden Zielgruppen:
  - DGS-Lerner, z.B. hörende Eltern gehörloser Kinder, Studenten der Gebärdensprache/des Gebärdensprachdolmetschens
  - professionelle Dolmetscher DGS – Deutsch,
  - DGS-Muttersprachler: gehörlose Erwachsene, CODAs,
  - gehörlose Kinder oder Schüler, die DGS als Muttersprache lernen
  - Gebärdensprachlehrer, Linguisten u.a.

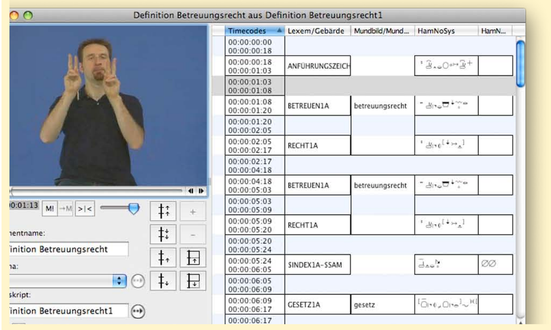
### Wörterbuchgrammatik

- korpusbasiert, mit zusätzlichen Informationen aus der Fokusgruppe und Auswertung aus dem Feedback
- Überblick über die wichtigsten grammatischen Merkmale der DGS mit Beispielen aus dem Korpus
- in einfacher Sprache geschrieben




### iLex

gebärdensprachspezifisches Tool zur Transkription und Auswertung




weitere Angaben z.B. Mundbild, Formabweichung

- kein Dokument-zentrierter Ansatz, sondern eine relationale Datenbank (PostgreSQL)
- integriertes Lexikon (iLex)
- Transkription und Erweiterung der lexikalischen Datenbank sind in iLex ein gemeinsamer Vorgang
- type-token-matching als wesentlicher Schritt der Transkription
- direkte Avatar-Ansteuerung möglich (z.B. zur Überprüfung von HamNoSys-Notationen)
- Vorteile einer relationalen Datenbank:
  - Daten müssen konsistent transkribiert werden
  - Daten können aus verschiedenen Perspektiven betrachtet werden
  - statistische Auswertung sind einfach

### Erhebung

- an 12 Orten (s. Karte), keine Dialektgrenzen
- es werden immer zwei Informanten zusammen aufgenommen
- Durchführung: regionale Kontaktperson (regionale Variation)
- die Erhebung besteht aus
  (a) standardisiertem Interview zur Erhebung von Sprach- und Sozialdaten
  (b) Konversation zu bestimmten Themen
  (c) versch. Aufgaben mit ausgewählten Stimuli


Studiosituation vs. natürlichsprachlich
nur DGS-Muttersprachler anwesend: 2 Informanten, Moderator (=Kontaktperson), Techniker

**Setting**: s. Bild

**Technik**:
7 Kameras (5 HD-, 2 Stereokameras), 12 Rechner
Grobe Schätzung: es werden ca. 575 TB für die kompletten Aufnahmen benötigt (reine Aufnahmezeit inkl. Aufgabenstellungen & Pausen)
Ablaufsteuerung (Session Director) mit automatischem Protokoll des Erhebungsablaufs

### Ziele der Datenerhebung

Ziel der Datenerhebung ist es, möglichst natürliche, spontane Sprachdaten der Alltags-DGS zu dokumentieren. Dafür werden den Teilnehmern verschiedene Aufgaben gestellt und Stimuli gezeigt, die zum Teil von anderen Korpus-Projekten übernommen und/oder angepasst und zum Teil von Projekt-Mitarbeitern des DGS-Korpus-Projekts neu entwickelt wurden.

### Rechtliche und ethische Aspekte

- Veröffentlichung der Filme (im Teilkorpus, Beispiele im Wörterbuch) → Informanten sind auf Filmen zu erkennen (Anonymisierung ist bei Gebärdensprache unmöglich)
- dauerhafte Speicherung personenbezogener Daten (Metadaten für Forschungszwecke)

→ Informanten müssen gut informiert/aufgeklärt werden (Erläuterungen: schriftlich und in DGS auf DVD)
→ schriftliche Einverständniserklärung der Informanten
→ Informanten können nach Erhebung noch einzelne Stellen von der Veröffentlichung ausschließen oder löschen lassen

### Literatur

Fachgebärdenlexika: s. http://www.sign-lang.uni-hamburg.de/projekte
Johnston, Trevor (ed.). 1998a: Signs of Australia. A new dictionary of Auslan the sign language of the Australian deaf community. Rev. ed. North Rocks, NSW: North Rocks Press.
Johnston, Trevor (ed.). 1998b: Signs of Australia on CD-ROM. A Dictionary of Auslan (Australian Sign Language). North Rocks, NSW: Royal Institute for Deaf and Blind Children. [CD-ROM].
Kennedy, Graeme / Arnold, Richard / Dugdale, Pat / Moskovitz, David. 1998: A dictionary of New Zealand Sign Language. Auckland: Auckland University Press.
Lucas, Ceil / Bayley, Robert / Valli, Clayton. 2001: Sociolinguistic Variation in American Sign Language. Washington, DC: Gallaudet University Press.
Schembri, Adam / Johnston, Trevor. 2004: Sociolinguistic variation in Auslan (Australian Sign Language). A research project in progress. In: Deaf Worlds, 20, 1, 78–90.
Stokoe, William C. / Casterline, Dorothy C. / Croneberg, Carl G. 1976: A dictionary of American Sign Language on linguistic principles. New Edition. Silver Spring, MD: Linstok Press. [1. Aufl. 1965].




Messe zur elektronischen Lexikographie, 10. März 2010, IDS Mannheim