DGS-Korpus-Arbeitspapier AP03-2018-01

Öffentliches DGS-Korpus: Annotationskonventionen

Autoren: Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, Anja Regen

DOI: https://doi.org/10.25592/uhhfdm.822

Version 0: Dezember 2018
Version 1: Februar 2019: Abschnitt zu Komposita überarbeitet sowie Hyperlinks in das öffentliche DGS-Korpus ergänzt
Version 2: September 2019: Änderung der Zuordnung lexikalisierter Formen (Subtypes) von Fingeralphabet-, initialisierten oder PMS-Gebärden zu Types
Version 3: September 2020: Anpassung an veränderte Darstellungen und neue Inhalte in Release 3 (s.a. Hanke et al. 2020) und Ergänzungen (Doppel-Token-Tags, Glossierungskonventionen, Komposita)
Version 4: Februar 2022: Differenzierung von Namensgebärden in Personennamen ($NAME, $NAME-…) und Organisationsnamen ($ORG-…); Auflistung der Glossierung (Suffixe) für Tokens von Fremdgebärdensprachen

Abstract

Dieses Arbeitspapier beschreibt die Annotations- und Glossierungskonventionen, die im Öffentlichen DGS-Korpus verwendet werden. Diese stimmen in weiten Teilen mit den Annotationsrichtlinien des DGS-Korpus-Projekts überein, unterscheiden sich jedoch in einigen Details, da z.B. die Einteilung von Tokens in verschiedene Wortformen eines Types mithilfe von Modifikatoren nicht veröffentlicht werden. Diese Tokens werden lediglich als von der Zitatform des Types abweichend gekennzeichnet.

Neben den allgemeinen Regeln zur Annotation der Übersetzungen und Mundbilder, den Segmentierungs- und Lemmatisierungsregeln werden Besonderheiten der Annotation (doppelte Glossierung, Doppel-Token-Tags) erklärt und die im Öffentlichen DGS-Korpus verwendeten Glossierungen im Einzelnen beschrieben. Glossen und z.T. auch Tokens sind mit den Types-Einträgen oder entsprechenden Videosequenzen des Öffentlichen DGS-Korpus verlinkt. Im Anhang werden alle Symbole und Glossierungen als Überblick zusammengefasst.

Inhalt

Einleitung

Dieses Arbeitspapier beschreibt die Annotations- und Glossierungskonventionen, die im Öffentlichen DGS-Korpus verwendet werden (http://ling.meine-dgs.de). Annotationen beinhalten Übersetzungen, Glossen – diese werden verwendet, um die Types zu identifizieren, denen Tokens zugeordnet werden, vergleiche dazu Johnstons ID-Glossen (Johnston 2010) –, und Mundbilder. Mundgesten werden in einer vereinfachten Form durch das Hinzufügen des Kürzels „[MG]“ in der Mundbild/Mundgestik-Spur annotiert. Diese Annotationen werden aus iLex (Hanke/Storz 2008) exportiert, unsere lexikalische Datenbank und gleichzeitig Annotationswerkzeug. Mit Ausnahme der Mundbilder sind alle Annotationen auf Englisch und Deutsch verfügbar (siehe dazu den Sprachbutton DE|EN auf der rechten Seite der Kopfzeile). Die Dokumente zum Herunterladen enthalten beide Versionen in separaten Spuren.

Mit dem Update im September 2019 (Release 2) sind die geplanten ca. 50 Stunden Videomaterial vollständig. Das Update zum Release 4 im Dezember 2021 enthält insgesamt 52,4 Std. Videomaterial, davon 50 Std. übersetzt, 49,1 Std. übersetzt und lemmatisiert mit 374800 Tokens. Release 4 enthält alle Änderungen, die seit der ersten Veröffentlichung vorgenommen wurden; ältere Versionen von Transkripten oder Types-Einträgen bleiben jedoch weiterhin verfügbar (s. versionsunabhängige DOI: http://doi.org/10.25592/dgs.corpus).

Übersetzung

Aufbauend auf Erfahrungen anderer Korpusprojekte beginnen wir zunächst mit der Übersetzung, bevor die Segmentierung und Lemmatisierung der Tokens erfolgt. Um die Daten des Korpus für Forscherinnen und Forscher zugänglich zu machen, die weder DGS noch Deutsch beherrschen, werden die deutschen Übersetzungen ebenfalls ins Englische übersetzt.

Übersetzung ins Deutsche

Im Gegensatz zu anderen Projekten, wie z.B. dem Auslan-Korpus-Projekt, streben wir keine freie Übersetzung an. Vielmehr sollte die Übersetzung so nah an der DGS-Äußerung (Ausgangssprache) orientiert sein wie möglich.
Eine erste Rohübersetzung wurde von extern unter Vertrag genommenen Gebärdensprachübersetzerinnen und -übersetzern oder -dolmetscherinnen und -dolmetschern angefertigt. Diese Texte wurden grob den zugehörigen Timecodes der Turns von Informant A und B zugeordnet. Als nächstes wurden diese Texte von studentischen Hilfskräften in satzähnliche Äußerungen zerteilt und verzeitet (aligniert). Diese Abschnitte eines gebärdeten Textes sollten eine kohärente und verständliche Bedeutungs- oder Äußerungseinheit bilden (vgl. Johnston 2019:13). Weitere Hinweise zur Bildung von Abschnitten („Grenzsignale“) sind gebärdensprachinhärente Signale wie Offene-Hand-Geste (s.u. Gesten), Kopfnicken, Drehen des Oberkörpers (beim Rollenwechsel), Blick und Rhythmus. Die Studierenden lasen die Übersetzungen ebenfalls noch einmal Korrektur und hielten in Zweifelsfällen Rücksprache mit gehörlosen Mitarbeiterinnen und Mitarbeitern.

Die mehr oder weniger kurzen, deutschen, geschriebenen Sätze wurden anhand des DGS-Videos verzeitet und erfüllen dabei mehrere Zwecke. Sie erlauben auch Menschen ohne DGS-Kompetenz einen Zugang zum Inhalt der gebärdeten Texte und bieten für unsere (hörenden) studentischen Hilfskräfte – mit unterschiedlicher DGS-Kompetenz – eine Orientierung bei der Lemmatisierung. Nach den Wörtern dieser Sätze kann gesucht werden, und diese Sätze definieren vorläufige Sinnabschnitte, um nach dem Verwendungskontext eines Gebärden-Tokens zu suchen. Wie Johnston (2019:13) klarstellt, sind diese „Übersetzungssätze keine Versuche,

den [DGS-]Text in seine potentiellen sprachspezifischen syntaktischen oder grammatischen Einheiten zu segmentieren“ [eigene Übers.]. Auf der Website MEINE DGS (http://meine-dgs-.de) werden die deutschen Übersetzungen als Untertitel zu den Videoausschnitten angezeigt. Zur Wahrung der Anonymität werden Namen individueller Personen durch Variablen (#Name1, #Name2 etc.) ersetzt. Die Namen von öffentlichen Personen sowie bekannte Persönlichkeiten in der hörenden und/oder Gehörlosenwelt werden nicht anonymisiert.

Übersetzung ins Englische

Mit einer Übersetzung ins Englische soll der Inhalt der DGS-Videos für Menschen zugänglich gemacht werden, die weder DGS- noch Deutschkenntnisse besitzen. Es handelt sich dabei um eine freie Übersetzung, die in passenden Kontexten kürzer gefasst ist als das Original.

Segmentierung

Bevor die Lemmatisierung beginnen kann, ist es notwendig, den Start- und Endpunkt eines Gebärden-Tokens zu definieren. Man kann den Endpunkt einer Gebärde als den Beginn der darauffolgenden definieren – im natürlichen Gebärdenfluss existiert keine Pause zwischen Gebärden, ebenso wenig wie im natürlichen Redefluss – oder aber die Übergangsbewegungen nicht als Bestandteil der Form eines Tokens zählen. Wir haben uns für die zweite Möglichkeit entschieden, um das visuelle Rauschen beim Vergleichen von Tokens eines Types oder Subtypes zu minimieren. Eine Folge davon ist, dass im Annotationsraster Leerzeilen zwischen Token-Tags vorhanden sind, wenn es zwischen Gebärden eine Übergangsbewegung gibt.
Durch die Implementierung von Doppel-Token-Tags ist die Segmentierung zwangsläufig gröber, wenn es um den Endpunkt der manuellen Aktivität von Zweihandgebärden geht, als sie es mit der Verwendung separater Spuren für jede Hand wäre. Wir konzentrieren uns auf die aktive Hand und ignorieren die passive. Folglich ist das Ende von z.B. einem Hold nicht spezifiziert, im Token-Tag ist jedoch annotiert, dass ein Hold vorliegt. Für eine detaillierte Definition des Start- und Endpunkts einer Gebärde siehe AP03-2010-01 (Hanke et al. 2012).

Lemmatisierung

Bevor wir unsere Glossierungskonventionen im Detail erklären, werden wir kurz die Hauptunterschiede zwischen unserem Ansatz und dem anderer Gebärdensprach-Korpusprojekte zusammenfassen:

- Damit wir die Ikonizität der Gebärden angemessen berücksichtigen und zwischen konventionellen und produktiven Gebärde-Mundbild-Kombinationen unterscheiden können, implementierten wir eine Type-Hierarchie (doppelte Glossierung) im Datenbankmodell von iLex.
- Wir benutzen Doppel-Token-Tags in der Token-Spur anstelle von separaten Glossenspuren für die linke und die rechte Hand, um die Token-Type-Zuordnung zu erleichtern.
- Wir verwenden eine lexikalische Datenbank, in der Tokens direkt den zugehörigen Types zugeordnet werden, sodass Types-Einträge einfach ausgehend von einem Token im Transkript geöffnet und Tokens innerhalb eines Type-Eintrags mit benutzerdefinierten Listen aufgerufen und überprüft werden können.

Die Bedeutung von Mundbildern

Johnston (2010:115) hält Übersetzung und Lemmatisierung für eine „absolut minimale Voraussetzung, um ein maschinenlesbares Referenzkorpus einer Gebärdenspache aufzubauen“ [eigene Übers.]. Im Falle der DGS halten wir es für notwendig, schon im ersten Durchgang der Basistranskription auch Mundbilder zu annotieren – und erkennbare Mundgestiken zu erfassen –, da Mundbilder häufig einen wichtigen Hinweis auf die Bedeutung eines DGS-Tokens geben. Mundbilder werden verwendet, um zwischen konventionellen und produktiven Verwendungen einer Gebärden zu unterscheiden. Dieser Ansatz ist in iLex als Type-Hierarchie modelliert. In Kombination mit der Form der Gebärde können mit dem Mund artikulierte Wörter dazu verwendet werden, nach dem passenden Type zu suchen, dem ein Token zugeordnet werden soll.

Type-Hierarchie (Doppelte Glossierung)

In der Gebärdensprachliteratur werden (lexikalische) Gebärden als Wörter behandelt, d.h. als Einheiten der jeweiligen Gebärdensprache, die in einem Wörterbuch beschrieben werden sollen. Ein Prinzip, das dabei zum Tragen kommt, ist das der Idiomatizität (Johnston/Schembri 1999). Die Dinge stellen sich jedoch anders dar, wenn man folgende Tatsache berücksichtigt: Sprecherinnen und Sprecher einer Gebärdensprache haben, bedingt durch die visuell-gestische Modalität von Gebärdensprachen, die Möglichkeit, Bedeutung zu visualisieren, indem sie eine direkte Verbindung zwischen der Form der Gebärde und der visuell wahrnehmbaren Welt herstellen. Der Einfluss der Ikonizität spiegelt sich auch im Lexikon wider, das sich in einigen Punkten vom Lexikon einer Lautsprache unterscheidet:

- Viele Gebärden sind ikonisch motiviert.
- Ein typisches Muster beim Gebärden ist die Verwendung von lexikalischen Gebärden, die etwas bezeichnen (*sagen*), gefolgt von sogenannten produktiven Gebärden, die die Äußerungsabsicht anschaulich darstellen (*zeigen*).
- Viele lexikalische Gebärden können graduell modifiziert oder delexikalisiert werden (Reikonisierung im Sinne von *zeigen*).
- Gebärden, zumindest in der DGS und anderen europäischen Gebärdensprachen, werden oft mit (lautlos) artikulierten Wörtern (sog. Mundbildern) kombiniert. Dies trägt dazu bei, dass Gebärden häufig eine Vielzahl unterschiedlicher Bedeutungen abdecken.

Ignoriert man die Rolle der Ikonizität und wendet die Regel *gleiche Form, gleiche Bedeutung => gleicher Type* an – und konsequenterweise *gleiche Form, unterschiedliche Bedeutung => unterschiedlicher Type* –, dann führt die Lemmatisierung, die die wahrnehmbaren Mundbilder mit berücksichtigt, dazu, dass das Lexikon der Lautsprache auf das der Gebärdensprache übertragen wird, ein unangemessenes und unbefriedigendes Ergebnis (König et al. 2008).
Die funktionale Perspektive nach Ebbinghaus/Heßmann hilft, das Zusammenspiel von Gebärden und Wörtern als wechselseitige Kontextualisierung zu verstehen. „Einige dieser Kombinationen treten mit häufigerer Regelmäßigkeit auf als andere und können als simultane Kollokationen betrachtet werden“ (Ebbinghaus/Heßmann 2001:134; eigene Übers.). Zum Beispiel wird die DGS-Gebärde VIERECK1^ (symmetrische Zweihandgebärde, mit den Zeigefingern ein vertikales Quadrat in die Luft zeichnend) im DGS-Korpus häufig für konventionalisierte Bedeutungen wie ‚Quadrat‘, ‚Seite‘, ‚Brief‘, ‚Rezept‘ oder ‚Karte‘ verwendet – die manuelle Gebärde wird in der Regel von den zugehörigen Mundbildern begleitet. Allerdings wird VIERECK1^ ebenfalls in Kombination mit Mundbildern wie „zeitung“, „visa“, „fernseher“ oder „stola“ verwendet. All diese Wörter bezeichnen Konzepte, die zum ikonischen Gehalt

der Gebärde, visualisiert durch das Zeichnen eines Quadrats in einer vertikalen Ebene, passen. Es sind jedoch keine konventionalisierten Bedeutungen dieser Gebärde.

> Lexikalische Untersuchungen zur DGS sollten sich ausgehend von einem strikten Kriterium der sprachlichen Form auf die Ermittlung von Gebärden, d.h. konventionellen Handzeichen, konzentrieren sowie typische Kollokationen von Gebärde und Wort beschreiben. (Ebbinghaus/Heßmann 1995:60)

Um in diesem Sinne die lexikographische Beschreibung zu unterstützen und die Auswirkungen der Ikonizität abzubilden, implementierten wir eine Type-Hierarchie in iLex und führten eine doppelte Glossierung ein. Die Types-Einträge in der Datenbank sind in einer Eltern-Kind-Beziehung miteinander verbunden. Der Eltern-Type (im Folgenden: Type) wird durch eine Zitatform spezifiziert. Jeder Kind-Type (im Folgenden: Subtype) steht für eine konventionalisierte Form-Bedeutungs-Beziehung. In den meisten Fällen entsprechen diese Beziehungen typischen Gebärde-Mundbild-Kombinationen. Ein Subtype erbt die Zitatform und den ikonischen Gehalt vom Eltern-Type. Type-Glossen sollten idealerweise einen Hinweis auf den ikonischen Gehalt der Gebärde geben, während Subtype-Glossen, wie Stichwörter, einer Kernbedeutung entsprechen.
Im ersten Schritt der Token-Type-Zuordnung werden Tokens konventioneller Gebärde-Mundbild-Kombinationen den passenden Subtypes zugeordnet – der Token-Tag im Annotationsraster zeigt die Subtype-Glosse – und Tokens produktiver Gebärde-Mundbild-Kombinationen den Types – das Token-Tag zeigt die Type-Glosse. Type-Glossen werden durch ein Zirkumflex am Ende gekennzeichnet wie zum Beispiel VIERECK1^. Alle anderen Glossen (ohne Zirkumflex) repräsentieren Subtypes.
Am Ende des Lemmatisierungsprozesses ist die Bandbreite an Bedeutungen, die eine Gebärde abdecken kann, in einer strukturierten Form dokumentiert, indem Tokens der verschiedenen konventionalisierten Verwendungen voneinander getrennt sind sowie von produktiven Verwendungen. Diese Art der Vorsortierung unterstützt die lexikalische Beschreibung von Gebärden-Types und ihren Bedeutungen. So werden zum Beispiel in der Type-Hierarchie für den Eltern-Type VIERECK1^ die folgenden Subtypes als Kinder-Types aufgelistet, die unterschiedliche konventionalisierte Bedeutungen, angezeigt durch Mundbilder, ausdrücken. All diese Bedeutungen sind motiviert durch das der Gebärdenform zugrunde liegende Bild: BESCHEINIGUNG2, BILD2B, BILDSCHIRM1, BLATT-PAPER1, BRIEF2, FENSTER6, FORMULAR1, HANDTUCH2, KISSEN1, PAPIER4, PLAN5, POSTER2, SCHILD1, SPIEGEL2, URKUNDE1, VIERECK1, ZETTEL1, ZEUGNIS4.

Doppel-Token-Tags

Die Lemmatisierung von Gebärdensprachtexten muss sich damit auseinandersetzen, dass Gebärden mit einer oder mit beiden Händen artikuliert werden. Zweihandgebärden können in zweihändige symmetrische und nicht-symmetrische Gebärden unterteilt werden, im Gegensatz zu komplexen Gebärdenkonstruktionen, bei denen jede Hand eine unterschiedliche Gebärde artikuliert. In iLex entschieden wir uns für eine Token-Spur, die es uns erlaubt, für jede Hand einen Type einzutragen, um den Zeitaufwand für den Annotationsprozess zu verringern. Zweihandgebärden werden entweder im Eingabefeld für die rechte oder linke Hand eingetragen: Im Falle von Einhand- und asymmetrischen Gebärden wird das Eingabefeld der aktiven Hand verwendet. Für symmetrische Gebärden wird entweder das Eingabefeld der Hand verwendet, die beginnt oder sich oberhalb der anderen Hand bewegt (z.B. im Fall von punktsymmetrischen Gebärden wie SPIELEN2, wo eine Hand in einer höheren Position beginnt als die andere) oder, wenn es keinen Unterschied in der Höhe gibt, das Eingabefeld der rechten Hand

(Standard-Eingabefeld). (iLex sperrt automatisch die Möglichkeit, in das zweite Eingabefeld einen Type/Subtype einzutragen, sobald ein Eingabefeld bereits einen Type/Subtype enthält, dessen HamNoSys-Notation eine Zweihandgebärde ausdrückt.) Ob eine gebärdende Person links-, rechts-, oder beidhändig ist, kann im Nachhinein festgestellt werden. Man muss nicht schon im Voraus eine Einschätzung zur Händigkeit der gebärdenden Person geben.
Im Online-Transkript ist die Subtype/Type-Spur („Lexem/Gebärde“) in zwei Spalten aufgeteilt. Bei einer rechtshändig ausgeführten einhändigen Gebärde steht die Glosse linksbündig in der linken Spur. Wird die Gebärde zweihändig ausgeführt, dann sind beide Spuren farbig markiert. Bei linkshändig ausgeführten Gebärden ist die Glosse in der rechten Spur rechtsbündig eingetragen. Bei zweihändigen Gebärden, bei denen die linke Hand die dominante ist, sind ebenfalls beide Spuren farbig markiert (s. Abb. 1).

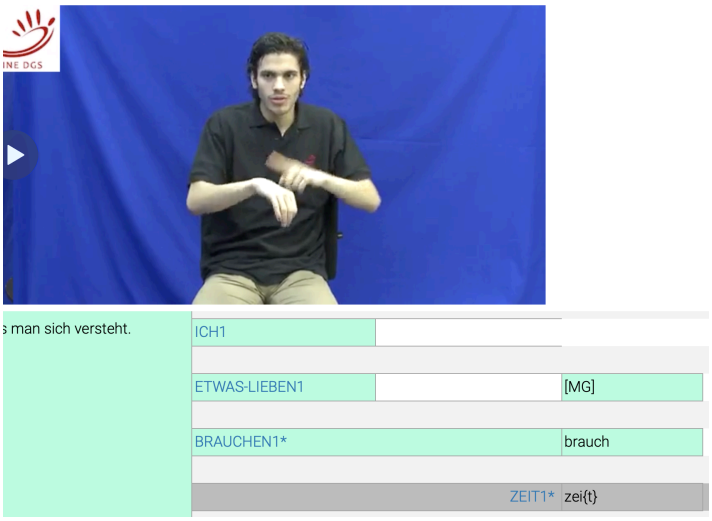

Abbildung 1: Glossen für rechts- und linkshändig ausgeführte Gebärden im Online-Transkript

In der KWIC-Konkordanz der Vorkommen eines Types/Subtypes steht bei rechtshändigen Gebärden die Glosse oben in der Zeile für die rechte Hand (r) und füllt bei zweihändig ausgeführten Gebärden die Zellen für die rechte und linke Hand aus, bei linkshändigen Gebärden steht die Glosse unten in der Zeile für die linke Hand (l) und füllt bei zweihändigen Gebärden ebenfalls beide Zellen aus (s. Abb. 2)

Abbildung 2: Ein Vorkommen von ZEIT1 mit KWIC-Konkordanz in der Types-Liste

Die ELAN-Dateien zum Herunterladen wurden aus unserer iLex-Datenbank exportiert. Sie folgen Johnstons (2019:14-17) Ansatz, Tokens auf je eine Spur pro Hand zu verteilen. Gemeinsam mit der Type-/Subtype-Unterscheidung ergibt das vier Spuren pro Informant: zwei Spuren für die linke und die rechte Hand mit Subtype- oder Type-Glossen (Lexem_Gebärde_r_A, Lexem_Gebärde_l_A; „Lexem“ entspricht dem Subtype, „Gebärde“ dem Type) ergänzt durch zwei Spuren mit den Type-Glossen (Gebärde_r_A, Gebärde_l_A). Zusätzlich enthalten die Dateien die entsprechenden Spuren für die englischen Glossen (Lexeme_Sign_r_A, Lexeme_Sign_l_A, Sign_r_A, Sign_l_A).

Lemmatisieren mit iLex

Die Zuordnung von Tokens zu Types oder das Identifizieren von (lexikalischen) Gebärden sind idealerweise ein reiner Top-Down-Prozess. Hat man jedoch keine mehr oder weniger vollständige lexikalische Ressource der DGS zur Verfügung, gehen Lemmatisierung und Aufbau einer lexikalischen Ressource Hand in Hand in einem steten Wechsel zwischen Top-Down- und Bottom-Up-Prozessen (König et al. 2010). Obwohl die Basisannotation von Gebärdensprachtexten so theorieneutral wie möglich gehalten werden sollte, kann sie nicht ohne theoretische Annahmen durchgeführt werden. Eine Annahme ist die Unterscheidung von drei Gebärdenkategorien: lexikalische Gebärden (vgl. Johnston/Schembri 2010: *vollständig lexikalische Gebärden*; „fully-lexical signs"), produktive Gebärden (vgl. Johnston/Schembri 2010: *teilweise lexikalische Gebärden*; „partly-lexical signs") und sonstige (vgl. Johnston 2019: 41-46: *nicht-lexikalische Gebärden*; „non-lexical signs"). Die im Folgenden aufgeführten Glossierungskonventionen erlauben die Kategorisierung und Sortierung von Tokens aus jeder Gebärdenkategorie sowie die Unterscheidung weiterer Untergruppen.
iLex unterstützt die Lemmatisierung durch eine lexikalische Datenbank. Types und Subtypes sind separate Einheiten der Datenbank mit eindeutigen Ids, denen Tokens zugeordnet werden. Ein Eintrag eines Types/Subtypes wird mindestens durch eine Glosse und eine HamNoSys-Notation der Gebärdenform bestimmt. In jedem Type-/Subtype-Eintrag werden die ihm zugeordneten Tokens mit weiteren Informationen aufgelistet. In iLex implementierte Beschränkungen garantieren, dass alle Glossennamen eindeutig sind. Da Glossen Etiketten für Gebärden-Types/-Subtypes sind und für eindeutige Types-Einträge in der lexikalischen Datenbank stehen, ist es nicht notwendig, den Begriff „ID-Glossen" (Johnston 2010) zu verwenden, da jede Glosse in unserem System die Kriterien für ID-Glossen erfüllt. Das heißt, dass die Glossen im Online-Transkript oder in den ELAN-Dateien als ID-Glossen angesehen werden können. Dies gilt auch für die englischen Glossen. Deutsche Glossen haben auf Type- und Subtype-Ebene ebenfalls voneinander verschiedene und eindeutige englische Glossen. Mit den englischen Glossen werden ebenso lexikalische und phonologische Varianten gruppiert (s. u. Glossierungskonventionen), genauso wie mit den deutschen Glossen. Mit Ausnahme von produktiven Gebärden können die Glossen im öffentlichen DGS-Korpus eindeutigen Types-Einträgen in unserer lexikalischen Datenbank zugeordnet werden. Da die Einteilung produktiver Gebärden in iLex vorläufig ist und weiterer Detailtranskription bedarf, sind alle Tokens produktiver Gebärden mit $PROD glossiert.

Glossierungskonventionen

Eine Glosse ist ein deutsches Wort, das auf Subtype-Ebene eine Kernbedeutung (Schlagwort) einer Gebärde wiedergibt. Sollten dafür mehrere Wörter notwendig sein, werden diese durch einen Bindestrich getrennt. Glossen werden in Großbuchstaben geschrieben, um zu signalisieren, dass sie Gebärden repräsentieren und nicht deutsche Wörter.
Die Wahl des Glossennamens ist abhängig von der Type-Ebene. Glossennamen für Types orientieren sich an dem zugrunde liegenden Bild der Gebärde, Glossennamen für Subtypes geben eine stark konventionalisierte Bedeutung dieser Gebärde wieder. Zum Beispiel visualisiert der Type TRINKEN1^, wie ein Trinkgefäß festgehalten und zum Mund geführt wird, eine Handlung, die typischerweise beim Trinken aus einem Gefäß vollzogen wird. Dazugehörige Subtypes sind u.a. BIER4, WEIN4 oder COLA1. Damit ist in vielen Fällen das Mundbild identisch mit Glossennamen des Subtypes (oder eine Abkürzung desselben). Dieses Vorgehen soll auch eine effektive Suche nach einer passenden Gebärde bei der Token-Type-Zuordnung ermöglichen. In iLex können Gebärden nach Form (HamNoSys) und Bedeutung (lautsprachliches

Wort) gesucht werden. Das gesuchte Wort kann entweder im Glossennamen oder in den Bedeutungen eines Subtypes verwendet werden. Uns ist bewusst, dass durch dieses Vorgehen in Einzelfällen auch stark diskriminierende Wörter wie z.B. NEGER1 als Glossennamen Verwendung finden. Bei der Korpusarbeit steht die Sprachdokumentation im Vordergrund. Entsprechend müssen wir dies hier aus Konsistenzgründen hinnehmen.[1]

Wenn ein neuer Type oder Subtype in der lexikalischen Datenbank angelegt wird, endet der Name der Glosse normalerweise mit einer Zahl. Unterschiedliche Zahlen werden zur Unterscheidung lexikalischer Varianten verwendet. Das sind synonymer Gebärden, die sich in der Form unterscheiden, aber die gleiche oder ähnliche Bedeutung haben und in bestimmten Kontexten austauschbar sind. Beispiele sind FRAU5 und FRAU8. Da wir für die Glossennamen deutsche Wörter verwenden, haben wir es auch mit dem Problem deutscher Polyseme zu tun. Zahlen dienen daher in einigen Fällen auch dazu, zwischen Gebärden zu unterscheiden, die sowohl eine verschiedene Form als auch Bedeutung haben, deren Bedeutungen aber durch dasselbe polyseme deutsche Wort ausgedrückt werden kann wie zum Beispiel ZU3^ (zusammendrücken, zusammengedrückt und damit verschlossen), ZU7 (geschlossen – z.B. eine Tür, die zugeht), ZU9 (zu etwas oder jemandem hin; LBG-Gebärde). Diese Gebärden haben denselben Glossennamen ZU, da das deutsche Wort „zu“ diese verschiedenen Bedeutungen ausdrücken kann. Bei der Übersetzung des deutschen Glossennamens ins Englische haben wir die Bedeutungsunterschiede berücksichtigt und die Glossennamen SQUEEZED3^, CLOSED7 und TOWARDS9 gewählt. Die Problematik deutscher Polyseme betrifft weiterhin die Übersetzung deutscher Glossennamen ins Englische. In vielen Fällen haben wir uns für eine Kombinationen englischer Wörter entschieden wie z.B. GAP-OR-DISTANCE1 (deutsche Glosse: ABSTAND1) und DISTANCE-OR-RANGE1 (deutsche Glosse: ENTFERNUNG1). Damit soll erreicht werden, dass auch bei der Suche nach Glossen in der englischen Types-Liste die entsprechenden Types gefunden werden.

Ebenso wie lexikalische Varianten sind phonologische Varianten synonyme Gebärden, die jedoch formähnlich sind – sie unterscheiden sich i.d.R. nur in ein oder zwei Parametern und haben dasselbe zugrunde liegende Bild. Phonologische Varianten haben denselben Glossennamen und dieselbe Nummer, zusätzlich jedoch verschiedene Buchstaben, zum Beispiel FRAU2A, FRAU2B, FRAU2C und FRAU2D.

Englische Glossen haben die gleichen Zahlen und Buchstaben zur Unterscheidung lexikalischer und phonologischer Varianten wie die entsprechenden deutschen Glossen. Diese Zahlen und Buchstaben werden nicht geändert, um Lücken zu schließen, weder in der lexikalischen Datenbank, wenn Types gelöscht oder Glossennamen geändert werden, noch bei den für das öffentliche Korpus exportierten Glossen. In einigen Fällen stimmen phonologische Varianten auf Subtype-Ebene nicht mit den phonologischen Varianten auf Type-Ebene überein, sodass zum Beispiel die Subtypes BEGLEITEN1C und BEGLEITEN1D beide dem Type BEGLEI-

---

[1] Aus demselben Grund haben wir uns entschieden, auch in den deutschen Übersetzungen die durch das Mundbild vorgegebene Bedeutung zu übernehmen. Die Wahl einer nicht diskriminierenden Bezeichnung würde nicht wiedergeben, was tatsächlich geäußert wurde. Sie würde ebenso eine Antwort auf die Frage vorwegnehmen, ob die Gebärde (inkl. Mundbild) in diskriminierenderer Absicht verwendet wurde. Anders als in der Korpuslinguistik ist es in der Lexikografie sinnvoll und üblich, auf diskriminierenden Sprachgebrauch hinzuweisen. Im Digitalen Wörterbuch der Deutschen Gebärdensprache (DW-DGS) (http://dw-dgs.meine-dgs.de), in dem die Einträge nicht durch Glossen benannt werden, besteht die Möglichkeit, im Eintrag auf eine negative konnotative Bedeutung oder diskriminierenden Sprachgebrauch hinzuweisen.

TEN1A^ zugeordnet werden.

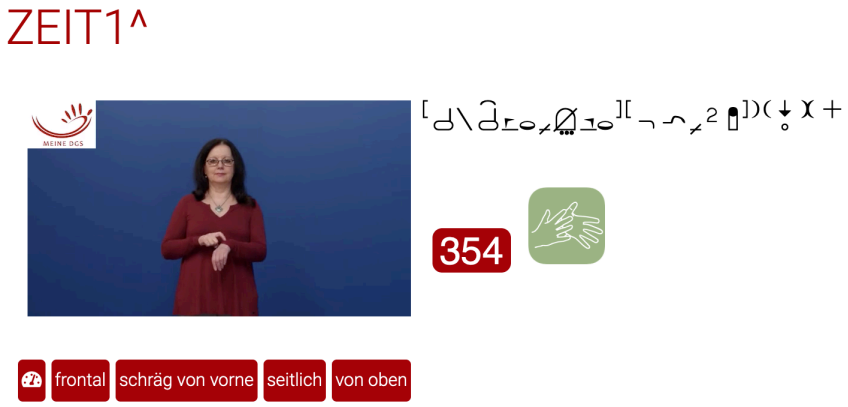

Abbildung 3: Beginn des Eintrags ZEIT1^ aus der Types-Liste

Durch das Klicken auf eine Glosse in einem Transkript öffnet man den entsprechenden Eintrag in der Types-Liste, in dem alle Vorkommen des Types und der dazugehörigen Subtypes aufgelistet werden.
Release 3 zeigt diese Vorkommen in einer KWIC-Konkordanz mit drei linken und rechten Nachbarn sowie der Übersetzung aus dem entsprechenden Übersetzungs-Tag, die neben dem Link auf die Stelle im Transkript steht (s. Abb. 2). Umgekehrt kommt man durch das Klicken auf eines dieser Tokens zur entsprechenden Zeile im Transkript, in dem man die dazugehörige Videosequenz anschauen kann.
Gibt es bereits Vorauseinträge des Wörterbuches DGS–Deutsch, dann werden Film und HamNoSys der Grundform sowie Verweise auf das Wörterbuch (hier: 354) sowie weitere lexikografische Ressourcen gegeben (s. Abb. 3; Müller et al. 2020). Die HamNoSys-Notationen sind nicht-redigierte Arbeitsversionen, bei denen bestimmte Aspekte nicht explizit notiert sind wie z.B. die Stellung des Daumens bei den Types ARBEITEN1^ und FISCH1^. In beiden Fällen ist nur die Grund-Handform (Faust bzw. Flachhand) notiert. Bei ARBEITEN1^ wäre ein angelegter oder abgespreizter Daumen anatomisch nicht sinnvoll, bei FISCH1^ ist die Stellung des Daumens variabel, aber irrelevant für die konventionalisierte Bedeutung und den ikonischen Gehalt der Gebärde.
Tokens, die von der Zitatform des Subtypes/Types abweichen, werden mit einem Sternchen in der Subtype/Type-Spur sowie in der Types-Liste gekennzeichnet, zum Beispiel FLACH1*. Tokens mit Glossen ohne Sternchen sollten der Zitatform des Types oder Subtypes entsprechen. Zum jetzigen Stand wurde nur ein Teil der fast 356.000 Tokens einer Lemmarevision (Konrad/Langer 2009) unterzogen.

Lexikalische Gebärden

Da wir ein korpusbasiertes DGS-Wörterbuch erstellen wollen, liegt der Fokus des DGS-Korpus-Projekts auf lexikalischen Gebärden, d.h. Lexemen als Einträge eines Gebärdensprachwörterbuchs oder einer lexikalischen Ressource. Durch den Sprachgebrauch verbinden sich Form und Bedeutung lexikalischer Gebärden zu stark konventionalisierten Einheiten. Deshalb werden sie auch als *etablierte („established")* oder *gefrorene („frozen")* Gebärden bezeichnet (Brennan 1992). Sie sind „über verschiedene Kontexte hinweg relativ stabil und konsistent" (Johnston 2019:15; eigene Übers.). Kontextunabhängig werden entweder die Bedeutung – fragt man einen native Signer nach der Bedeutung einer Gebärde – oder die Form – fragt man nach einer Gebärde für eine Bedeutung, z.B. ‚Alkohol' – spontan miteinander ver-

knüpft. Das bedeutet, dass lexikalische Gebärden eine Zitatform und mindestens eine Kernbedeutung haben.
Johnston/Schembri (1999) legen den Schwerpunkt auf die Idiomatizität lexikalischer Gebärden. Obwohl die Form vieler lexikalischer Gebärden ikonisch motiviert ist – ihre Parameter Handform, Orientierung, Lokation und Bewegung drücken eine allgemeine Bedeutung aus (erste Ebene der Konventionalisierung), z.B. steht eine flache Hand für ‚etwas Flaches' –, ist die Bedeutung lexikalischer Gebärden idiomatisch, d.h. sie ergibt sich nicht allein dadurch, dass man die Bedeutungen ihrer Komponenten miteinander kombiniert: Diese Gebärden haben eine zweite Ebene der Konventionalisierung durchlaufen. Beispielsweise sind bei der Gebärde TRINKEN1^ alle Parameter bedeutungstragend und können als „etwas mit einer zylindrischen Form halten und seinen Inhalt in den Mund gießen" beschrieben werden. Diese Gebärde, in Verbindung mit einer wiederholten Bewegung, hat die konventionalisierte Bedeutung ‚Alkohol'. Diese Bedeutung des Lexems ALKOHOL2 passt zur allgemeinen Bedeutung, aber es gibt keine Regel, die ausgehend von der allgemeinen Bedeutung zur Bedeutung ‚Alkohol' führen würde. Die Gebärde muss als lexikalisiertes und stabiles Form-Bedeutungs-Paar gelernt werden. Im Unterschied zu produktiven Gebärden (siehe unten), werden sie von Johnston/Schembri (2010) als *vollständig lexikalische Gebärden* („fully-lexicalized signs") bezeichnet.
Es existiert eine kleine Anzahl an Gebärden, bei denen die allgemeine Bedeutung mit der konventionalisierten Bedeutung übereinstimmt, so z.B. bei FLACH1^ (Subtype FLACH1: ‚flach') oder SCHLAGEN1^ (Subtype SCHLAGEN1: ‚schlagen'). Johnston/Schembri (1999:133-134) nennen diese Gebärden „allgemeine Gebärden" („general signs"), da das Kriterium der Idiomatizität nicht zutrifft (siehe dazu auch Fenlon et al. 2015:191). Da wir den Schwerpunkt nicht auf die Idiomatizität, sondern auf die Konventionalität legen und die Ikonizität der Gebärden mit berücksichtigen (siehe oben), sind diese Gebärden hochkonventionalisierte Form-Bedeutungs-Einheiten und werden genauso wie andere lexikalische Gebärden behandelt.
Bei der Verwendung ikonisch motivierter lexikalischer Gebärden ist der ikonische Gehalt ihrer Bestandteile deaktiviert. Dieser kann jedoch reaktiviert werden (Delexikalisierung), um zusätzliche Bedeutungsaspekte auszudrücken. Zum Beispiel kann der Parameter Bewegung der Gebärde ALKOHOL1^ modifiziert werden, um auszudrücken, dass jemand ‚übermäßig viel Alkohol trinkt', möglicherweise noch begleitet durch entsprechende (expressive) Mimik sowie Körperbewegung oder veränderte Haltung des Oberkörpers.

Innerhalb der lexikalischen Gebärden werden die folgenden Untergruppen durch die Verwendung von Prä- und Suffixen zum Glossennamen unterschieden. Präfixe werden durch das vorangestellte Dollarzeichen ($) gekennzeichnet, damit diese bei der alphabetischen Sortierung der Glossen gruppiert werden.

Namensgebärden ($NAME)

Gebärdennamen für Personen werden durch $NAME gekennzeichnet, eine Sammelglosse (Subtype) für alle Tokens von Eigennamen privater Personen und nicht konventionalisierter Ortsnamen. Zur Wahrung der Anonymität von Personen wird im öffentlichen DGS-Korpus keine weitere Differenzierung vorgenommen. Handelt es sich um Personen des öffentlichen Lebens oder um bekannte Personen in der Welt der Hörenden und/oder Gehörlosen, folgt dem Präfix der Name der Person, z.B. $NAME-ANGELA-MERKEL1.
Namen von Organisationen werden mit dem Präfix $ORG gekennzeichnet wie z.B. politische Parteien ($ORG-GRÜNE1A) und Organisationen ($ORG-STASI1A), Fußballvereine ($ORG-BAYERN-MÜNCHEN1A) oder Firmen ($ORG-VW1). Auch konventionalisierte Gebärden von Produkten werden so glossiert ($ORG-POLO1, $ORG-TRABANT1).

Unbekannte (regionale) Gebärden ($KANDIDAT)

Taucht eine neue Gebärde auf, die niemandem in unserem Team bekannt ist und bei der nicht genügend Daten vorhanden sind, um ihre Konventionalität zu bestätigen, wird sie als Kandidat für eine lexikalische Gebärde markiert. Eine regelmäßige Verwendung dieser Gebärde von einer/einem oder mehreren Informantinnen und Informanten aus einer Region kann auf eine (regionale) Variante hinweisen. Solche möglicherweise lexikalischen Gebärden erhalten als Glosse ein Schlagwort, das die vermutete Bedeutung der Gebärde wiedergibt, und das Suffix $KANDIDAT, gefolgt vom Kode für die Region, in der die Daten erhoben wurden, sowie eine fortlaufende Zahl, z.B. AUGUST-$KANDIDAT-MST05. Durch dieses Vorgehen werden die Gebärden in der Types-Liste alphabetisch sortiert und zusätzlich als Kandidaten für einen Eintrag als lexikalische Gebärde ausgewiesen. Ergänzende Korpusdaten, eine Verifizierung außerhalb des Korpus oder eine genauere Analyse der Daten kann zu einer Neubewertung als lexikalische Gebärde führen, was zur Folge hätte, dass das Suffix $KANDIDAT entfernt werden würde.

Gebärden für (gebundene) Morpheme im Deutschen ($WORTTEIL)

Einige Gebärden werden ausschließlich dazu verwendet, gebundene Morpheme deutscher Wörter auszudrücken, so zum Beispiel das Suffix „-in“ als morphologische Markierung des weiblichen Geschlechts, z.B. in „Lehrerin“. Diese Gebärden werden durch das Präfix $WORTTEIL gekennzeichnet, z.B. $WORTTEIL-IN1.

In der DGS wie auch z.B. in Auslan (Johnston 2001) können viele lexikalische Gebärden die Funktion eines Nomens, Verbs oder einer anderen Wortart übernehmen ohne zusätzliche formale morphologische Markierung (Schwager/Zeshan 2008). Um den Unterschied zwischen beispielsweise ‚unterrichten‘ und ‚Lehrerin‘ deutlich zu machen, kann nach der DGS-Gebärde UNTERRICHTEN1 die Gebärde PERSON1 ausgeführt werden, die deutlich macht, dass es sich um die unterrichtende Person und nicht um den Akt des Unterrichtens handelt (Token: UNTERRICHTEN1). Im Lemmatisierungsprozess wurden diese Fälle als zwei Tokens segmentiert und den entsprechenden Types/Subtypes zugeordnet. Ob das Vorkommen von PERSON1 in diesen Kontexten als Morphem oder als grammatikalische Gebärde (ähnlich wie Funktionswörter) behandelt werden sollte, muss später geprüft werden.

Fremdsprachliche Gebärden

Wenn Informantinnen und Informanten fremdsprachliche Gebärden verwenden, z.B. bei direkter Rede (Constructed Dialogue) oder bei einer Diskussion über Unterschiede zwischen den Gebärden verschiedener Gebärdensprachen, werden diese Tokens mit Glossennamen der zugehörigen gesprochenen Sprache und einer Kennzeichnung der jeweiligen Gebärdensprache versehen wie z.B. NO-ASL1. Glossen für internationale Gebärden sind markiert durch den Zusatz INTS, z.B. GERMANY-INTS1.

Im Öffentlichen DGS-Korpus sind Tokens von insgesamt acht Fremd-Gebärdensprachen (inkl. Internationale Gebärden) enthalten. Die Glossen enthalten folgende Suffixe:

- -ASL: Amerikanische Gebärdensprache (American Sign Language)
- -AUSLAN: Australische Gebärdensprache (Australian Sign Language)
- -BSL: Britische Gebärdensprache (British Sign Language)
- -INTS: Internationale Gebärden (International Signs)
- -LIS: Italienische Gebärdensprache (Lingua Italiana Dei Segni)
- -LSM: Mexikanische Gebärdensprache (Lengua de Señas Mexicana)
- -NZSL: Neuseeländische Gebärdensprache (New Zealand Sign Language)
- -PJM: Polnische Gebärdensprache (Polski Język Migowy).

Einige lexikalische Gebärden werden regelmäßig mit einer Mundgestik ausgeführt, sogenannte Spezialgebärden (engl.: multi-channel signs). Diese Gebärden werden nicht durch Präfixe oder Suffixe extra gekennzeichnet, ebensowenig wie Komposita. Sie werden genauso segmentiert wie andere lexikalische Gebärden. Bei Komposita dagegen gibt es besondere Segmentierungsregeln.

Spezialgebärden (sogenannte idiomatische Gebärden oder multi-channel signs)

Einige Gebärden werden regelmäßig mit einer oder mehreren Mundgestiken kombiniert und selten oder nie mit einem Mundbild. Einige dieser sogenannten „multi-channel signs“ (wörtlich: Mehrkanal-Gebärden; Brennan 1992:128, Johnston/Schembri 1999:154-155) werden verwendet, um kontextabhängige Bedeutungen auszudrücken und können nur schwer durch eine 1:1-Entsprechung in die umgebende Lautsprache übersetzt werden. Sie werden dahingehend übersetzt, dass ihre Bedeutung im Kontext umschrieben wird. Die Existenz dieser Gebärden wird als ein besonderes Merkmal angesehen, das Gebärdensprachen von den sie umgebenden Lautsprachen unterscheidet. In den Sprachgemeinschaften der DGS, Deutschschweizerischen Gebärdensprache (DSGS) und Österreichischen Gebärdenspräche (ÖGS) entstand des Bedürfnis, diesen Gebärden einen eigenen Namen zu geben, *Spezialgebärden* (so verwendet seit den frühen 90er Jahren). Später wurden diese Gebärden auch als ‚idiomatische‘ Gebärden bezeichnet (vgl. Konrad 2014). Aus sprachinterner Sicht handelt es sich bei diesen Gebärden jedoch um reguläre lexikalische Gebärden, die keine eigenständige Gebärdenkategorie bilden (König et al. 2012:142). Im öffentlichen DGS-Korpus werden diese sogenannten Spezialgebärden deshalb nicht als besondere Kategorie behandelt, sondern wie jede andere lexikalische Gebärden auch, z.B. FADEN-VERLIEREN1^.

Komposita und Mehrwortausdrücke

Becker (2003) geht davon aus, dass Komposition ein unbedeutender und nicht aktiver Prozess der Lexikonerweiterung in der DGS ist. Sequenzen von Gebärden, die in ihrer Abfolge die Bestandteile deutscher Komposita wiedergeben, werden in der Lemmatisierung nicht als feste Einheiten, sondern als Abfolge individueller Gebärden behandelt. Das Gleiche gilt für Sequenzen von Gebärden, die potentiell DGS-eigene Komposita bilden. Ob diese Sequenzen als Lehnübersetzungen, Kollokationen, (sequentielle) DGS-Komposita oder etwas anderes bestimmt werden sollten, bleibt weiteren Analysen vorbehalten. Dies gilt auch für alle weiteren Formen von Mehrwortausdrücken.
Im Gegensatz dazu werden simultane Konstruktionen als eigenständige Types geführt, unabhängig davon, ob sie als simultane Komposita betrachtet werden können, bei denen Formelement von zwei Gebärden kombiniert und gleichzeitig ausgeführt werden wie z.B. bei initialisierten Gebärden (s.u. Initialisierung) oder bei ZUHAUSE1A^ (Handform von SITZEN1A^ kombiniert mit HAUS1A^), oder als Blends, bei denen sich die Formen mischen und es zu einer Reduktion der Form einer der beiden Gebärden kommt. Zum Beispiel fällt bei HOFFNUNG-KLOPFEN1^ die wiederholte Bewegung der Faust hin zum Kinn (HOFFEN1B^) weg, die Gebärde beginnt beim Kinn.

Produktive Gebärden ($PROD)

Im Gegensatz zu Lautsprachen besteht ein gebärdensprachlicher Diskurs nicht nur aus einer Abfolge lexikalischer Einheiten, sondern vielmehr aus einem Zusammenspiel von lexikalischen Gebärden (die in erster Linie etwas benennen: *sagen*) und produktiven Gebärden (die in erster Linie etwas veranschaulichen: *zeigen*). Diese produktiven Gebärden – auch bekannt als

Klassifikator- oder polymorphemische Gebärden – sind vollständig ikonisch motiviert. Sie haben eine allgemeine Bedeutung, zu der der ikonische Gehalt jeder Komponente beiträgt. Nach Johnston/Schembri (1999) haben diese Gebärden die erste Ebene der Konventionalisierung durchlaufen und bestehen aus konventionellen und nicht-konventionellen Elementen, weshalb sie als *teilweise lexikalische Gebärden* („partly-lexicalized signs“) bezeichnet werden (Johnston/Schembri 2010). Ihre Bedeutung kann nur im Kontext interpretiert werden. Eine Veränderung der Gebärdenform bedeutet immer auch eine Veränderung ihrer Bedeutung. Produktive Gebärden haben die zweite Ebene der Konventionalisierung nicht durchlaufen. Das bedeutet umgekehrt, dass produktive Gebärden keine vollständig spezifizierte Zitatform besitzen. Alle Vorkommen produktiver Gebärden erhalten die Sammelglosse $PROD (Abkürzung für produktive Gebärde).

Zeigegebärden ($INDEX)

Nach Johnston's (2019) Klassifizierung handelt es sich bei Zeigegebärden um *teilweise lexikalische Gebärden*. Sie haben ein konventionelles Element, die Handform, sowie stark durch den Kontext bestimmte Elemente: Orientierung und Bewegung. Allerdings haben Zeigegebärden ihren Ursprung höchstwahrscheinlich in Zeigegesten und drücken, im Gegensatz zu produktiven Gebärden, ein begrenztes Spektrum an Bedeutungen aus: Sie weisen auf einen Referenten oder eine Lokation (oder beides) hin und erfüllen mehrere Funktionen.
Zeigegebärden werden mit der Glosse $INDEX gekennzeichnet. Variationen in Handform (und Orientierung) werden durch zusätzliche Zahlen unterschieden:

- $INDEX1: Index-Handform,
- $INDEX2: flache Hand mit Handfläche nach oben und Fingerspitzen, die auf etwas zeigen,
- $INDEX4: Daumen.

Um unseren Informantinnen und Informanten im Studio Stimuli und Erklärungsvideos zu den Aufgaben zu zeigen, verwendeten wir jeweils einen Monitor, der auf dem Boden vor jeder Informantin/jedem Informanten stand. Dies führte dazu, dass die Informantinnen und Informanten beim Gebärden häufig auf den Monitor zeigten. Um diese Tokens getrennt zählen zu können, wurden sie mit $INDEX-MONITOR1 lemmatisiert.
Zeigegebärden mit der Index-Handform, die auf den Mund zeigen und eine pragmatische Funktion haben – den Gesprächspartner auf das Lippenlesen aufmerksam machen –, wurden dem Subtype $INDEX-ORAL1 zugeordnet.
Zeigegebärden mit der flachen Hand (Handfläche nach unten), die sich kreisförmig im neutralen Gebärdenraum bewegen, um einem Referenten eine spezifische Lokation im Gebärdenraum zuzuweisen, wurden der Glosse $INDEX-BEREICH1 zugeordnet. Diese Gebärden können als eine Kombination einer Zeigegebärde mit einer skizzierenden Bewegung analysiert werden (Langer 2005:265).
Zeigegebärden können auch lexikalisieren. Wir entschieden uns für die lexikalischen Einheiten ICH1 (auf die gebärdende Person zeigen mit Kontakt auf der Brust in der Bedeutung ‚ich') und DU1 (vom Körper weg in Richtung Adressat/Adressatin zeigen in der Bedeutung ‚du'), da es sich dabei um feste Form-Bedeutungs-Einheiten handelt (vgl. dagegen Johnston 2019:25-31, der diese Gebärden nicht aus der Gruppe der Zeigegebärden herausnimmt).
Das Zeigen auf ein Körperteil kann eine produktive Verwendung einer Zeigegebärde sein, um auszudrücken, was man meint, allerdings gibt es auch hier lexikalisierte Gebärden für bestimmte Körperteile wie zum Beispiel NASE1A oder HERZ1A.

Sonstige

Das Potential der Korpuslinguistik liegt darin, dass sprachliche Einheiten im Kontext des natürlichen Sprechens/Gebärdens analysiert werden können. Das bedeutet, dass Gebärden, die vor und nach einer zu untersuchenden Gebärde artikuliert werden, wichtige Hinweise für verschiedene Forschungsfragen geben. Deshalb sollte die Lemmatisierung kontinuierlich und ohne Lücken durchgeführt werden, um „laufende" Textwörter zu erhalten, oder, bezogen auf die Annotation von Gebärdensprache, „laufende" Glossen. Neben lexikalischen, produktiven und Zeigegebärden gibt es weitere linguistische Phänomene in Gebärdensprachen, für die ebenfalls Glossierungskonventionen bereitgestellt werden müssen:

Fingeralphabet ($ALPHA)

In der DGS ist das Fingeralphabet durch die Verwendung des einhändigen Fingeralphabets (siehe Anhang 2) konventionalisiert mit Varianten für die Handformen der Buchstaben D, F, G, J, K, T und ß. Tokens des Fingeralphabets sind mit $ALPHA gekennzeichnet. Fingeralphabet-Sequenzen werden anhand der dadurch ausgedrückten Bedeutung als ein Token segmentiert. In iLex verwenden wir die Type-Hierarchie mit Modifikatoren („qualifiers"; Konrad et al. 2012) und einem offenen Vokabular, um gefingerte Einheiten zu lemmatisieren. In der Types-Liste des öffentlichen DGS-Korpus werden alle Vorkommen des Fingeralphabets durch die Type-Glosse $ALPHA^ zusammengefasst. Durch verschiedene Subtype-Glossen unterscheiden wir verschiedene Arten des Fingerns:

- $ALPHA1: einhändiges Fingeralphabet,
- $ALPHA2: zweihändiges Fingeralphabet, bei dem die Form des Buchstabens mit zwei Händen nachgebildet wird, z.B. mit der kleinen C-Hand für linke und rechte Hand und Kontakt der Fingerspitzen von Zeigefinger und Daumen wird der Buchstabe S nachgebildet. Diese Art, das S darzustellen, ist lexikalisiert für die Bedeutung ‚September' (SEPTEMBER2A).
  Diese Art des zweihändigen Fingeralphabets ist zu unterscheiden vom zweihändigen Fingeralphabet der Britischen Gebärdensprache (BSL), glossiert mit $ALPHA-BSL^. Die Fingeralphabete weiterer Gebärdensprachen werden entsprechend glossiert, z.B. $ALPHA-NZSL^.
- $ALPHA-SK: mit dem Zeigefinger die Form des Buchstabens in die Luft zeichnen (mit Ausnahme des Buchstabens Z, da das In-die-Luft-Zeichnen dieses Buchstabens die standardisierte Form im einhändigen Fingeralphabet ist und dementsprechend mit $ALPHA1 lemmatisiert wird).

Zusätzlich zum Glossennamen ergänzen wir die gefingerten Buchstaben, um einen Hinweis auf die dadurch ausgedrückte Bedeutung zu geben, z.B. $ALPHA1:B-U-S. Diese Buchstaben geben die tatsächlich gefingerten Buchstaben wieder und können von dem Wort oder der intendierten Bedeutung abweichen, z.B. $ALPHA1:A-L-F-E-D-O (gemeinter Name: Alfredo). Werden doppelte Buchstaben, nicht einzeln gefingert, sondern mit einer leichten seitlichen Bewegung der Handform, dann steht kein Trennstrich zwischen den Buchstaben, z.B. $ALPHA1:J-E-NN-Y. Bei Variation der Handform wird der Buchstabe durch eine Zahl ergänzt, z.B. $ALPHA1:T_2. Falls die gebärdende Person absichtlich keine Buchstaben fingert, sondern die Finger willkürlich bewegt, um lediglich anzudeuten, dass an dieser Stelle etwas gefingert werden müsste, werden diese Tokens mit $ALPHA1:# lemmatisiert.
Lexikalisierte Formen des ein- oder zweihändigen Fingeralphabets, bei denen lediglich ein Buchstabe eines Wortes gefingert wird, wie z.B. EX2 oder WUPPERTAL2, werden in der Types-Liste durch entsprechende Subtype-Glossen beim Type $ALPHA^ am Ende aufgelistet.

Lexikalisierte Kombinationen aus mehreren Buchstaben sind als eigenständige Type-Einträge angelegt wie z.B. BIO1^.

Initialisierung ($INIT)

In der Gebärdensprachlinguistik bezeichnet man als initialisierte Gebärden lexikalische Gebärden, die mit einer Handform des Fingeralphabets kombiniert werden. Häufig repräsentiert die Handform den ersten Buchstaben des dazugehörigen (geschriebenen) Worts wie z.B. bei CHANCE1^. Diese Gebärden erhalten, wenn es konventionelle Gebärden sind, im DGS-Korpus eine eigene Glosse wie jede andere lexikalische Gebärde auch.
In einigen Fällen existiert zu einer konventionellen Gebärde eine phonologische Variante als initialisierte Gebärde wie z.B. REGEL1C. Diese Formen werden in der Types-Liste als Subtype des übergeordneten Types, hier REGEL1D^, angezeigt. Die vorangestellte Tilde (~) zeigt an, dass diese Form durch Änderung eines (oder mehrerer) Parameter aus der übergeordneten Form abgeleitet wurde, in diesen Fällen durch die Ersetzung der Handform durch eine Handform des einhändigen Fingeralphabets.
Was wir als initialisierte Gebärden bezeichnen und mit dem Präfix $INIT kennzeichnen, sind spontan produzierte Gebärden mit einer Handform, die dem einhändigen Fingeralphabet entnommen ist, sowie einer einfachen geraden oder kreisförmigen Bewegung oder einem Wackeln im Handgelenk. Oft werden diese Gebärden für Eigennamen oder Fachbegriffe, für die die gebärdenende Person keine konventionalisierte Gebärde kennt, verwendet. In der Regel repräsentiert die Handform dabei den ersten Buchstaben des Namens oder des Fachbegriffs.
Abhängig von der Art der Bewegung werden die Tokens den folgenden Types zugeordnet:

- $INIT-GERADE1^,
- $INIT-HANDGELENK1^,
- $INIT-KREIS1^,
- $INIT-KREIS2^.

Zahlgebärden ($NUM)

Da es bekanntlich eine große Variation bei den Zahlgebärden der DGS gibt, war es unser Ziel, jedes Vorkommen einer Zahlgebärde zu dokumentieren. Deshalb wurde jedes Token einzeln segmentiert, auch bei zusammengesetzten Zahlen. Zum Beispiel ist die Gebärdensequenz zum Ausdruck der Zahl 1989 in drei verschiedene Tokens segmentiert: $NUM-TEEN1:9 (neunzehn), $NUM-EINER1A:9 (neun), $NUM-ZEHNER1:8d (achtzig). (In der DGS gibt es wie im Deutschen die Besonderheit, dass die letzte Ziffer vor der vorletzten artikuliert wird. Im Deutschen wird das Beispiel als „neunzehn|(hundert)|neun|und|achtzig“ ausgedrückt.)
Ebenso wie beim Fingeralphabet werden in iLex zur Lemmatisierung von Zahlgebärden Modifikatoren verwendet. Im öffentlichen DGS-Korpus beginnen Glossen für Zahlgebärden mit dem Präfix $NUM. Zahlgebärden werden anhand ihrer Form und des Zahlenraums, den sie abdecken, gruppiert:

- 1-10 ($NUM-EINER1A etc.),
- 11-19 ($NUM-TEEN1 etc.),
- 10, 20…90 ($NUM-ZEHNER1 etc.),
- 100, 200…900 ($NUM-HUNDERTER1),
- 1000, 2000…10 000 ($NUM-TAUSENDER1).

Die Wurzel dieser Gebärden ist eine konventionalisierte Bewegung und Handflächenorientierung. Ebenso wie bei Zahlen inkorporierenden Gebärden drückt die Handform/drücken die Handformen die Ziffer aus, die benötigt wird, um die entsprechende Zahl auszudrücken. Va-

rianten, die sich in Bewegung und/oder Handflächenorientierung unterscheiden, werden wie bei lexikalischen Gebärden mit zusätzlichen Zahlen und Buchstaben im Glossennamen versehen, z.B. $NUM-TEEN1, $NUM-TEEN2A, $NUM-TEEN2B und $NUM-TEEN2C.
Wir kodieren weiterhin im Glossennamen die Handform, die für eine Ziffer steht und zusammen mit der Wurzel eine bestimmte Zahl ausdrückt, z.B. $NUM-TEEN1:3d („dreizehn“). Handformvarianten werden mit einem zusätzlichen Kleinbuchstaben gekennzeichnet („d“ steht für Daumen, wenn der Daumen zur Wiedergabe der Ziffer verwendet wird, „f“ steht für die F-Handform zur Darstellung der Ziffer 3, „w“für die W-Handform zur Darstellung der Ziffer 3).
Konventionalisierte Gebärden gibt es auch für Schnapszahlen: $NUM-DOPPEL1A usw.
Eine tippende Bewegung des Zeigefingers oder Zeige- und Mittelfingers auf den Daumen wird verwendet, um die Zahlen 11 und 12 auszudrücken. Die Glossierung verweist auf diese typische Bewegung: $NUM-TIPPEN1.
Eine schnipsende Bewegung wird verwendet, um die Zahlen 11 bis 19 zu gebärden. Diese Zahlen sind mit $NUM-SCHNIPS1 glossiert.
Die folgenden Zahlgebärden erlauben keine Zahleninkorporation:

- $NUM-HUNDERT1 etc.,
- $NUM-TAUSEND1,
- $NUM-MILLION1.

Andere Wurzeln, die nur mit Zahleninkorporation vorkommen, erhalten ebenfalls das Präfix $NUM:

- $NUM-UHR1A etc.,
- $NUM-MARK1,
- $NUM-KLASSE1 (z.B. Schule),
- $NUM-SCHULNOTE1A etc.,
- $NUM-PERSONENZAHL etc.,
- $NUM-ZEIT-NACH-JETZT1^,
- $NUM-ZEIT-VOR-JETZT1^,
- $NUM-WOCHE-NACH-JETZT1,
- $NUM-WOCHE-VOR-JETZT1,
- $NUM-JAHR-NACH-JETZT1,
- $NUM-JAHR-VOR-JETZT1,
- $NUM-LEHRJAHR1 etc.,
- $NUM-ODER-ZWISCHEN1,
- $NUM-VON-BIS1,
- $NUM-VERHÄLTNIS1 etc.,
- $NUM-NENNER1 (math.),
- $NUM-ZÄHLER1 (math.).

Wenn lexikalische Gebärden Zahlen inkorporieren, wird die Handform der Zitatform durch die entsprechende Handform für die Ziffer ersetzt. In iLex wird diese Art der Modifikation durch das Hinzufügen des Modifikators „Anzahl“ in Kombination mit einem Kode für die jeweilige Handform erfasst, im öffentlichen Korpus hat die Glosse ein Sternchen, da die Form von der Zitatform des Types/Subtypes abweicht, z.B. JAHR1A*. Lexikalische Gebärden, die Zahlen inkorporieren, sind: ALT8B, AUZÄHLEN1B, AUZÄHLEN1C, EINIGE1, HALB6, JAHR1A, JAHR1B, JAHR3A, JAHR3B, MAL3, MONAT1, PFENNIG1, STOCKWERK1, STUNDE1B, STUNDE2A, STUNDE2B, STUNDE2C, TAG-VOR1B^, TAG-ZU-

RÜCK1C^, TAG2, VERGLEICH2, VOR1E, WOCHE1A, WOCHE1B, WOCHE1C, ZUSAMMEN3B.
Wenn römische Zahlen gebärdet werden, verwenden wir die Glosse $NUM-RÖM1.[2]
Wie bei Kardinalzahlen gibt es auch für Ordinalzahlen Gebärden mit einer Wurzel und Handformen für die entsprechende Ziffer. Diese Gebärden sind mit $NUM-ORD1 und $NUM-ORD2 glossiert.

Listen-Bojen ($LIST)

Eine weitere Möglichkeit der Aufzählung besteht darin, mit dem Zeigefinger (oder der Flachhand) auf Finger der passiven Hand zu zeigen, die eine Art gedachte Liste darstellt. Diese Listen-Bojen („list buoys"; Liddell 2003:223-242) werden als Ordinal- oder Kardinalzahlgebärden benutzt oder erfüllen eine diskursstrukturierende Funktion.
Sie werden mit $LIST glossiert. In iLex werden Listen-Bojen wie Zahlgebärden mit Modifikatoren und einem geschlossen Vokabular lemmatisiert. Im öffentlichen DGS-Korpus steht $LIST1 für Gebärden, bei denen der Zeigefinger auf eine Nummer der Liste zeigt, und $LIST2 für Gebärden, bei denen diese Aufgabe die Flachhand übernimmt.
Über das Anzeigen eines bestimmten Elements auf einer vorgegebenen Liste hinaus kann die Liste auch erweitert werden, indem nacheinander auf mehrere Listenelemente gezeigt wird ($LIST-AUFZÄHLEN1 etc.), können Elemente der Liste entfernt ($LIST-ENTFERNEN1A etc.) oder zusammen gruppiert werden ($LIST-ZUSAMMEN1C). Die Gruppierung kann mithilfe der manipulativen Technik (Langer 2005, König et al. 2008, Ebling et al. 2015; $LIST-ZUSAMMEN1C, durch unmittelbar aufeinanderfolgendes Zeigen mit dem Zeigefinger auf mehrere Elemente ($LIST-ZUSAMMEN3) oder dadurch realisiert werden, dass die Zahlhandform für die Zahl der zu gruppierenden Elemente in die zeigende Hand mit inkorporiert werden (vgl. Liddells TWO-LIST buoy; $LIST-ZUSAMMEN2).
Die Qualifier-Werte von Listen-Bojen beinhalten immer zwei Zahlen: Die erste Zahl repräsentiert den Listenplatz des angezeigten Elements (die intendierte Bedeutung), und die zweite Zahl steht für die Anzahl aller angezeigten Listenelemente (Handform der passiven Hand) wie zum Beispiel bei $LIST1:2of4. Wenn Elemente durch Zahleninkorporation gruppiert werden, wird zunächst eine weitere Zahl vorangestellt, die für die Anzahl der gruppierten Elemente steht (Handform der aktiven Hand), gefolgt von den anderen beiden Zahlen, z.B. $LIST-ZUSAMMEN2:2:1-2of4.

Gesten ($GEST)

Genau wie bei hörenden Sprecherinnen und Sprechern treten auch bei Signerinnen und Signern einer Gebärdensprache manuelle und nicht-manuelle Gesten auf. Manuelle Gesten sind ganzheitliche und expressive Zeichen, die spontan produziert werden. Bei einer manuellen Aktivität, die weder eine konventionelle noch eine produktive Gebärde ist, handelt es sich sehr wahrscheinlich um eine Geste, insbesondere wenn diese Art der Aktivität auch bei der hörenden Mehrheit in ähnlicher oder gleicher Funktion auftritt. Viele Gesten sind innerhalb der jeweiligen Kultur verbreitet, einige sind hochkonventionell (Embleme), aber es gibt darüber hinaus auch große individuelle Unterschiede, z.B. wenn Gefühle ausgedrückt werden.
Gesten werden beim Annotieren entweder $GEST^ (unspezifischer Type für alle manuellen Gesten) oder $GEST-NM^ (unpezifischer Type für alle nicht-manuellen Gesten) zugeordnet.
Die Offene-Hand-Geste, bei der die Handfläche nach oben zeigt, auch bekannt als *palm-up-open-hand gesture* (PUOH) wird als $GEST-OFF^ glossiert (OFF steht für offene Hand).

---

[2] Für die 4 wird typischerweise die einfache Umrechnung, d.h. IIII, verwendet.

Neben diesen Sammel-Types gibt es noch einige andere Type-Einträge für Gesten, mit einer spezifischen Beschreibung der Form und Bedeutung, ähnlich wie bei lexikalischen Gebärden. Diese Types können zur Lemmatisierung verwendet werden, wenn sie passen. Die Zuordnungen sind noch vorläufig und wurden noch nicht näher überprüft. Die meisten Gesten im Öffentlichen DGS-Korpus sind der Offene-Hand-Geste zugeordnet (über 15000 Tokens), gefolgt von $GEST^ (6705 Tokens), $GEST-ABWINKEN^ (2737 Tokens), $GEST-NM-KOPFNICKEN1^ (1422 Tokens), $GEST-NM-KOPFSCHÜTTELN1^ (1036 Tokens) und $GEST-ÜBERLEGEN1^ (955 Tokens).

Mundbildgebrauch (ohne gleichzeitige manuelle Aktivität; $ORAL)

Wenn es keine nennenswerten manuellen Bewegungen gibt und eine Bedeutung allein durch den Gebrauch eines Mundbilds ausgedrückt wird, werden Tokens dieser Art der Kommunikation, die vor allem bei älteren Informantinnen und Informanten beobachtet werden kann, mit $ORAL^ glossiert, dem Platzhalter für eine rein oral stattfindende (lautlose) Artikulation (deutscher Wörter). Folgerichtig muss das der Glosse zugeordnete Mundbild-Tag ein Wort oder eine Wortfolge enthalten und nicht eine Mundgestik ([MG]).

Phonembestimmtes Manualsystem ($PMS)

In den 1970er Jahren wurde in Deutschland ein System von Handzeichen zur Visualisierung der Artikulation von Phonemen für den Unterricht gehörloser Kinder entwickelt, das so-genannte Phonembestimmte Manualsystem (PMS). Einige der PMS-Handzeichen werden auch in der DGS benutzt, ähnlich wie initialisierte Gebärden, um z.B. Namen auszudrücken, für die keine konventionelle Gebärde bekannt ist. Vorkommen von PMS-Gebrauch sind mit $PMS^ glossiert. Im Laufe der Zeit sind einige PMS-Zeichen als Gebärden lexikalisiert. Wenn die Form einem PMS-Zeichen entspricht, dann sind diese Gebärden in der Types-Liste als Subtype am Ende des $PMS^-Eintrags aufgelistet, z.B. WEISS11. In einigen Fällen bilden sich zu diesen Formen zusätzliche Varianten wie z.B. bei WENN1A (PMS-Zeichen für die Artikulation des Buchstabens N) und WENN1B (Variation der Bewegung). In diesen Fällen wie auch bei Kombinationen aus mehreren PMS-Zeichen zu einer lexikalischen Gebärde (vgl. Fingeralphabet ($ALPHA)) wurden eigenständige Type-Einträge angelegt.

Unklare Fälle ($UNKLAR)

Nicht identifizierbare manuelle Aktivität, die aber als sprachliche Aktivität eingeschätzt wird, ist mit $UNKLAR^ glossiert.

Außersprachliche manuelle Aktivität ($$EXTRA-LING-MAN)

Nichtsprachliche manuelle Aktivität wie z.B. sich die Nase reiben oder sich etwas von der Kleidung wischen wird dort, wo sie besonders auffällig ist, erfasst und mit $$EXTRA-LING-MAN^ glossiert. Dies dient dazu, die Korpus-Aufnahmen als Material für maschinelles Lernen z.B. von Handaktivitäten aufzubereiten, d.h. diese „Tokens“ identifizieren Abschnitte mit starkem visuellem Rauschen (hierbei handelt es sich jedoch nicht um „non-tokens“, vgl. Langer et al. 2016).

Annotation von Mundbildern und Mundgestik

In der DGS kommen häufig Mundbilder vor. Ein Mundbild ist ein wichtiger Hinweis auf die Bedeutung des Tokens einer DGS-Gebärde und kann beim Annotieren zusammen mit der Gebärdenform dazu verwendet werden, den richtigen Type zu finden, dem das Token zugeordnet

werden soll. Aus diesem Grund annotieren wir auch Mundbilder im Arbeitsschritt der Basisannotation.
Um den Zeitaufwand gering zu halten, segmentieren wird Mundbilder und Mundgestiken nicht unabhängig von den Gebärdentokens. Im Transkript ist die Mundbild/Mundgestik-Spur abhängig von der Subtype/Type-Spur. Die Tags der Mundbild/Mundgestik-Spur übernehmen die Tag-Grenzen der Subtype/Type-Spur und können sich über ein oder mehrere solcher Tags erstrecken. Mundbilder werden mit Kleinbuchstaben notiert, um sie von deutschen Wörtern zu unterscheiden. Im Gegensatz zum Fingeralphabetgebrauch (siehe oben), bei dem wir die gefingerten Buchstaben so, wie sie gefingert werden, erfassen, auch wenn sie vom Zielwort abweichen, konzentrieren wir uns bei den Mundbildern auf die Identifikation der artikulierten Zielwörter und nicht auf die tatsächlich ausgeführte Form der Artikulation. Das bedeutet, dass zumindest das zum Ablesen intendierte artikulierte Wort (bzw. der Wortstamm) annotiert wird. Wenn ein Mundbild unvollständig ist, wird es in geschweiften Klammern ergänzt. Keine Ergänzung wird angegeben, wenn die gebärdende Person nur einen Teil eines Wortes artikuliert, während sie nach dem richtigen Wort sucht.
Unklare Mundbilder, z.B. wenn das Wort nicht identifiziert werden konnte, sind als „??“ annotiert. Da in der DGS Mundbilder auf deutsche Wörter verweisen, unterscheiden sich die artikulierten Formen von z.B. englischen Mundbildern. Deshalb stellen wir auch keine Übersetzungen der Mundbilder zur Verfügung.
Mundgestiken sind Bewegungen der Mundregion, die keine Verbindung zu Wörtern der Lautsprache haben. Da unser Schwerpunkt vor allem auf der Erfassung lexikalischer Gebärden liegt, haben wir Mundgestiken nicht nach unterschiedlichen Formen klassifiziert. Sie sind nur in einer sehr vereinfachten Form als „[MG]“ in der Mundbild/Mundgestik-Spur annotiert. Hiervon gibt es zwei Ausnahmen:

- Wenn eine gebärdende Person nicht ein bestimmtes Wort artikulieren, sondern nur allgemein andeuten möchte, dass eine Person lautsprachlich (d.h. nicht gebärdensprachlich) gesprochen hat, werden die dazu produzierten Mundbewegungen durch ein vorangestelltes Hashtag („#“) und eine ungefähre schriftliche Annäherung an die gezeigten Artikulationsbewegungen erfasst (z.B. „#lalala“).
- Wenn eine gebärdende Person visuell Geräusche nachahmt, die von Personen (z.B. Interjektionen), Tieren oder Gegenständen produziert werden oder im Zusammenhang mit bestimmten Vorgängen entstehen wie z.B. onomatopoetische Ausdrücke, werden diese Mundbewegungen durch „[LM]“ (für Lautmalerei) annotiert, gefolgt von einer passenden orthografischen Transkription des Geräuschs wie z.B. „[LM:ähm]“, „[LM:bam bam]“ oder „[LM:miau]“.

Literatur

Becker, Claudia. 2003. *Verfahren der Lexikonerweiterung in der Deutschen Gebärdensprache*. Hamburg: Signum.

Brennan, Mary. 1992. The visual world of BSL. An introduction. In: Brien, David (Hg.): *Dictionary of British Sign Language/English*. London: Faber and Faber, S. 2-133.

Ebbinghaus, Horst / Heßmann, Jens. 1995: Formen und Funktionen von Ablesewörtern in gebärdensprachlichen Äußerungen. Teil II. In: *Das Zeichen* 31, S. 50-61.

Ebbinghaus, Horst / Heßmann, Jens. 2001. Sign language as multidimensional communication: Why manual signs, mouthings, and mouth gestures are three different

things. In: Boyes Braem, Penny / Sutton-Spence, Rachel (Hgg.): *The hands are the head of the mouth: The mouth as articulator in sign language*. Hamburg: Signum, S. 133-151.

Ebling, Sarah / Konrad, Reiner / Boyes Braem, Penny / Langer, Gabriele. 2015. Factors to Consider When Making Lexical Comparisons of Sign Languages: Notes from an Ongoing Comparison of German Sign Language and Swiss German Sign Language. In: *Sign Language Studies*, 16 (1), S. 30-56. [Online verfügbar; URL: https://www.jstor.org/stable/26191929; letzter Zugriff: 2021-12-20].

Fenlon, Jordan / Cormier, Kearsy / Schembri, Adam. 2015: Building BSL SignBank: The lemma dilemma revisited. In: *International Journal of Lexicography*, Vol. 28, No. 2, S. 169-206.

Hanke, Thomas / Storz, Jakob. 2008. iLex - A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In: Crasborn, Onno / Efthimiou, Eleni / Hanke, Thomas / Thoutenhoofd, Ernst D. / Zwitserlood, Inge (eds.): *LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Paris: ELRA, S. 64-67. [Online verfügbar; URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/08011.html; letzter Zugriff: 2021-12-20].

Hanke, Thomas / Hong, Sung-Eun / König, Susanne / Konrad, Reiner / Langer, Gabriele / Matthes, Silke /Nishio, Rie / Regen, Anja. 2012: Segmentierung. [Projektinternes Arbeitspapier AP03-2010-01, 1. Überarbeitung: Nov. 2012. Online verfügbar: DOI: 10.25592/uhhfdm.822].

Hanke, Thomas / Konrad, Reiner / Jahn, Elena / Schulder, Marc. 2020. Extending the Public DGS Corpus in Size and Depth. In: Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Hochgesang, Julie / Kristoffersen, Jette / Mesch, Johanna (eds.): *Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages. 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, 2020. Paris, France: European Language Resources Association (ELRA), S. 75-82. [Online verfügbar: URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/20016.html; letzter Zugriff: 2021-12-20].

Johnston, Trevor. 2001. Nouns and Verbs in Australian Sign Language: An Open and Shut Case? In: *Journal of Deaf Studies and Deaf Education* 6 (4), S. 235-257. DOI: 10.1093/deafed/6.4.235.

Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. In: *International Journal of Corpus Linguistics*, 15, 1, S. 106-131. DOI: 10.1075/ijcl.15.1.05joh.

Johnston, Trevor. 2019. Auslan Corpus Annotation Guidelines. August 2019 version. [Online verfügbar; URL: https://www.academia.edu/40088269/Auslan_Corpus_Annotation_Guidelines_August_2019_version_; letzter Zugriff: 2021-12-20].

Johnston, Trevor / Schembri, Adam. 1999: On Defining Lexeme in a Signed Language. In: *Sign Language & Linguistics* 2, 2, S.115-185. DOI: https://doi.org/10.1075/sll.2.2.03joh.

Johnston, Trevor / Schembri, Adam. 2010. Variation, lexicalization and grammaticalization in signed languages. In: *Langage et société* n° 131, mars 2010, S. 19-35. DOI : 10.3917/ls.131.0019.

König, Susanne / Konrad, Reiner / Langer, Gabriele. 2008. What's in a sign? Theoretical lessons from practical sign language lexicography. In: Quer, Josep (ed.): *Signs of the time. Selected papers from TISLR 2004*. Hamburg: Signum, S. 379-404.

König, Susanne / Konrad, Reiner / Langer, Gabriele / Nishio, Rie. 2010. How Much Top-Down and Bottom-Up do We Need to Build a Lemmatized Corpus? Poster presented at the Theoretical Issues in Sign Language Research Conference (TISLR 10), Sept 30 - Oct 2, 2010 at Purdue University, Indiana, USA. [Online verfügbar; URL: https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/PosterTISLRTranskription1_R_12.pdf; letzter Zugriff: 2021-12-20].

König, Susanne / Konrad, Reiner / Langer, Gabriele. 2012: Lexikon – Der Wortschatz der DGS. In: Eichmann, Hanna / Hansen, Martje / Heßmann, Jens (eds.): *Handbuch Deutsche Gebärdensprache: Sprachwissenschaftliche und anwendungsbezogene Perspektiven*. Hamburg: Signum, S. 111-164.

Konrad, Reiner. 2014. „Where have all the Idioms Gone?“. Ergänzungen zu Ines Schütte: „‚Idiome und Redewendungen‘ in der DGS – Begriffsdefinition und Versuch einer Kategorienbildung“. In: *Das Zeichen* 97, S. 64-67.

Konrad, Reiner / Langer, Gabriele. 2009: Synergies between transcription and lexical database building: The case of German Sign Language (DGS). In: Mahlberg, Michaela / González-Díaz, Victorina / Smith, Catherine (Hgg.): Proceedings of the Corpus Linguistics Conference (CL2009). University of Liverpool, UK, 20-23 July 2009. [Online verfügbar; URL: https://www.academia.edu/47465363/Synergies_between_transcription_and_lexical_database_building_The_case_of_German_Sign_Language_DGS_; letzter Zugriff: 2021-12-20].

Konrad, Reiner / Hanke, Thomas / König, Susanne / Langer, Gabriele / Matthes, Silke / Nishio, Rie / Regen, Anja. 2012. From form to function. A database approach to handle lexicon building and spotting token forms in sign languages. In: Crasborn, Onno / Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Kristoffersen, Jette / Mesch, Johanna (eds.): *Interaction between Corpus and Lexicon. Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages. 8th International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey. Paris: ELRA, S. 87-94. [Online verfügbar; URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/12023.html; letzter Zugriff: 2021-12-20].

Langer, Gabriele. 2005. Bilderzeugungstechniken in der Deutschen Gebärdensprache. In: *Das Zeichen* 70, S. 254-270.

Langer, Gabriele / Hanke, Thomas / Konrad, Reiner / König, Susanne. 2016. „Non-Tokens“: When Tokens should not Count as Evidence of Sign Use. In: Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Hochgesang, Julie / Kristoffersen, Jette / Mesch, Johanna (eds.): *Workshop Proceedings. 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia, 28 May 2016. Paris: ELRA, S. 137-142. [Online verfügbar; URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/16013.html; letzter Zugriff: 2021-12-20].

Liddell, Scott K. 2003. *Grammar, gesture and meaning in American Sign Language*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511615054.

Müller, Anke / Hanke, Thomas / Konrad, Reiner / Langer, Gabriele / Wähl, Sabrina, 2020. From Dictionary to Corpus and Back Again – Linking Heterogeneous Language Resources for DGS. In: Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Hochgesang, Julie / Kristoffersen, Jette / Mesch, Johanna (eds.): *Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Proceedings of the 9th Workshop on the Representation and Processing of*

Sign Languages. 12th International Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, 2020. Paris, France: European Language Resources Association (ELRA), S. 157-164. [Online verfügbar: URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/20025.html; letzter Zugriff: 2021-12-20].

Schwager, Waldemar / Zeshan, Ulrike. 2008. Word classes in sign languages. Criteria and classifications. In: *Studies in Language* 32, 3, S. 509-545. DOI: 10.1075/sl.32.3.03sch.

Anhang 1: Symbole und Glossierungskonventionen (Überblick)

Types-Liste

Symbol	Erklärung	Beispiel
≙	Type-Glosse (nicht Subtype) als Überschrift für alle Tokens, die diesem Type zugeordnet sind	≙ FLACH1^
=	Subtype-Glosse (mit der gleichen Zitatform wie der übergeordnete Type (Eltern-Type))	= GRUND-BODEN3
~	Lexikalisierte Gebärde (Subtype), die aus der Grundform des Types durch Änderung einer oder mehrerer Parameter abgeleitet ist	~ WAND3

Glossen

^	Type-Glossen (im Gegensatz zu Subtype-Glossen, die kein ^ am Ende haben)	FLACH1^ vs. FLACH1
GLOSSE1, GLOSSE2 …	Lexikalische Varianten (oder verschiedene Gebärden mit unterschiedlicher Bedeutung, die aufgrund der Polysemie des als Glosse benutzten Wortes den gleichen Glossennamen haben)	FRAU5, FRAU8 (ZU3^, ZU7, ZU9)
GLOSSE1A, GLOSSE1B …	Phonologische Varianten	FRAU2A, FRAU2B
GLOSSE-ASL, -BSL …, -INTS	Gebärden einer anderen nationalen Gebärdensprache oder Internationale Gebärden	GERMANY-INTS1
-$KANDIDAT-	Unbekannte (möglicherweise regionale) Gebärde, die ein wahrscheinlicher Kandidat für den Eintrag als lexikalische Gebärde ist	AUGUST-$KANDIDAT-MS-T05
$	Steht vor dem Glossennamen, um Gebärden zu gruppieren (Untergruppen lexikalischer Gebärden, produktiver Gebärden und anderer Gebärden)	$PROD, $ALPHA1 … (siehe unten)
$$EXTRA-LING-MAN	Außersprachliche (manuelle) Aktivität	$$EXTRA-LING-MAN^
$ALPHA	Fingeralphabetgebrauch	$ALPHA1, $ALPHA2, $ALPHA-SK, $ALPHA-BSL^ $ALPHA-NZSL^
$PMS	Zeichen des Phonembestimmten Manualsystems (PMS)	$PMS^

$GEST	Gesten	$GEST^, $GEST-NM^, $GEST-OFF^, $GEST-ABWINKEN^, $GEST-NM-KOPFNICKEN1^, $GEST-NM-KOPFSCHÜT-TELN1^, $GEST-ÜBERLEGEN1^ …
$INDEX	Indexikalische Gebärden (Zeigegebärden) (mit dem Zeigefinger, Daumen oder der Flachhand/Fünfhand)	$INDEX1, $INDEX2, $INDEX4, $INDEX-BEREICH1, $INDEX-MONITOR1, $INDEX-ORAL1
$INIT	Initialisierte Gebärden (Kombination einer Fingeralphabetgebärde und einer einfachen Bewegung)	$INIT-GERADE1^, $INIT-KREIS1^, $INIT-HANDGELENK1^
$LIST	Listen-Bojen	$LIST1, $LIST2, $LIST-AUFZÄHLEN1, $LIST-ENTFERNEN1A, $LIST-ZUSAMMEN1C
$WORTTEIL	Lexikalische Gebärden, die dazu benutzt werden gebundene Morpheme des Deutschen zu visualisieren	$WORTTEIL-IN1 …
$NAME	Namensgebärden für Personen und unbekannte Ortsnamen	$NAME
$ORG	Namensgebärden für Organisationen	$ORG-GRÜNE1A $ORG-VW1
$NUM	Zahlgebärden	$NUM-EINER1A, $NUM-TEEN1, $NUM-ZEHNER1, $NUM-HUNDERTER1, $NUM-TAUSENDER1 …
$ORAL	Orale Artikulation von Wörtern, ohne Begleitung durch eine nennenswerte manuelle Aktivität	$ORAL^
$PROD	Produktive Gebärden (im Gegensatz zu lexikalischen Gebärden und anderen Gebärden)	$PROD
$UNKLAR	Nicht zu identifizierende manuelle Aktivität	$UNKLAR^

Tokens

*	Die Gebärdenform des Tokens weicht von der Zitatform des Types bzw. Subtypes ab	FLACH1*
$ALPHA1:#	Nur absichtsvolle Andeutung, dass etwas mit dem einhändigen Fingeralphabet gefingert wird (ohne dieses jedoch richtig auszuführen)	$ALPHA1:#
$ALPHA1:J-E-NN-Y	Bei Wiederholung von Buchstaben, die nicht einzeln gefingert werden, sondern mit einer seitlichen Bewegung der Handform, steht kein Trennstrich zwischen den Buchstaben.	$ALPHA1:J-E-NN-Y
_2, _3	Handformvarianten beim Fingeralphabet	$ALPHA1:T_2

1d, 2d, 3d, 3f, 3w, 4d	Handformvarianten bei Zahlgebärden und Listen-Bojen d = Daumen f = F-Handform (des einhändigen Fingeralphabets) w = W-Handform (des einhändigen Fingeralphabets)	$NUM-TEEN1:3d
$LIST1: [Zahl]of[Zahl]	Angezeigte Element, das an einer gegebenen Listen-Handform angezeigt wird	$LIST1:2of4
$LIST-ZUSAMMEN2: [Zahl]:[Zahl]- [Zahl]of[Zahl]	Anzahl der Elemente, die gleichzeitig an einer gegeben Listenhandform angezeigt werden	$LIST-ZUSAMMEN2:2:1-2of4

Mundbilder/Mundgestik

#	Absichtsvolle, nur angedeutete orale Artikulation (die deshalb nicht aus richtig lautlos artikulierten Wörter besteht)	#lalala
??	nicht erkennbares Mundbild	??
[MG]	Mundgestik	[MG]
[LM:…]	Mundaktivität, die lautlos Geräusche imitiert	[LM:bam bam]

Anhang 2: Fingeralphabet (DGS)