easier

intElligent Automatic
SIgn languagE tRanslation

D6.2 SPECIFICATION FOR THE HARMONIZATION OF SIGN LANGUAGE ANNOTATIONS

Revision: v1.0

Work Package	WP6
Task	T6.2
Due date	31/12/2021
Submission date	31/01/2022
Deliverable lead	Universität Hamburg
Version	1.0
DOI (latest version)	10.25592/uhhfdm.9841
Authors	Maria Kopf, Marc Schulder, Thomas Hanke, Sam Bigeard (Universität Hamburg – UHH)
Reviewers	Robin Nachtrab-Ribback (Swiss TXT – STXT), Michael Filhol (French National Centre for Scientific Research – CNRS)

Abstract	This document compiles information required for the joint use of several specific sign language corpora. It compares the annotations of 17 different corpora (some covering multiple languages) that include data for 17 signed languages. The comparison addresses annotation standards for manual and non-manual signs, describing their basic annotation format, formats for various specific phenomena and handshape coding approaches. This information is then used to formulate a strategy for the harmonization of these annotation standards that will result in a corpus-independent interchange format for the representation of all considered corpora.
Keywords	sign language annotation, annotation conventions, glossing, data format standards

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

WWW.PROJECT-EASIER.EU

Document Revision History

Version	Date	Description of change	List of contributors
V0.1	17/01/2022	First draft presented to consortium	Maria Kopf, Marc Schulder, Thomas Hanke, Sam Bigeard (UHH)
V0.2	24/01/2022	Second draft for internal review	Maria Kopf, Marc Schulder, Thomas Hanke, Sam Bigeard (UHH)
V0.3	30/01/2022	Internal Review	Robin Ribback (STXT), Michael Filhol (CNRS)
V1.0	31/01/2022	Camera-ready submission DOI: 10.25592/uhhfdm.9841	Maria Kopf, Marc Schulder, Thomas Hanke, Sam Bigeard (UHH)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the 'Intelligent Automatic Sign Language Translation' (EASIER) project's consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable	**R**	
Dissemination Level		
PU	Public, fully open, e. g., web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc

EXECUTIVE SUMMARY

This document discusses in detail commonalities and differences between annotation conventions as applied to several publicly accessible sign language corpora. More specifically, it describes basic annotation features and the annotation of manual signs in chapter 4, the annotation of various non-manual features in chapter 5, and the coding for handshapes in chapter 6. From there it concludes in chapter 7 with a proposal of an interchange format that shall be used to automatically convert corpus data to have a shared representation format for all corpus data used in EASIER.

CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

Funded by the Horizon 2020
Framework Programme of the European Union

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

LISTINGS

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

CA	Constructed Action
CD	Constructed Dialogue
CODA	Children Of Deaf Adults
HamNoSys	Hamburg Notation System for Sign Languages
POS	Part of Speech
SL	sign language

Sign Languages

ASL	American Sign Language
BSL	British Sign Language
DGS	German Sign Language / Deutsche Gebärdensprache
DTS	Danish Sign Language / Dansk tegnsprog
FinSL	Finnish Sign Language / Suomalainen viittomakieli
FinSSL	Finland-Swedish Sign Language / Finlandssvenskt Teckenspråk / Suomenruotsalainen Viittomakieli
GSL	Greek Sign Language / Ελληνική νοηματική γλώσσα (Elleniké Noematiké Glossa)
HSL	Hungarian Sign Language / Magyar Jelnyelv
ISL	Irish Sign Language / Teanga Chomharthaíochta na hÉireann
LIS	Italian Sign Language / Lingua Italiana dei Segni
LSF	French Sign Language / Langue des Signes Française
LSFB	French Belgian Sign Language / Langue des signes de Belgique francophone
NGT	Sign Language of the Netherlands / Nederlandse Gebarentaal
NZSL	New Zealand Sign Language
PJM	Polish Sign Language / Polski Język Migowy
SSL	Swedish Sign Language / Svenskt Teckenspråk
SZJ	Slovene Sign Language / Slovenski Znakovni Jezik
VGT	Flemish Sign Language / Vlaamse Gebarentaal

Resources

BSL Corpus	British Sign Language Corpus
CFinSL	Corpus of Finnish Sign Language / Corpus FinSL
Corpus VGT	Corpus Vlaamse Gebarentaal
DTS Corpus	Danish Sign Language Corpus

ECHO Corpus	European Cultural Heritage Online Corpus
HSL Corpus	Hungarian Sign Language Corpus
LIS Corpus	Italian Sign Language Corpus
Polytropon	Polytropon Parallel Corpus
SSLC	Swedish Sign Language Corpus / SSL Corpus

Funded by the Horizon 2020
Framework Programme of the European Union

1 INTRODUCTION

A main criterion for a modern language corpus is its machine-readability. This poses extra effort to sign language corpus creators, as sign languages have no commonly used writing system.

How sign language corpora can be annotated has been widely discussed by corpus creators. The annotation guidelines developed in the *ECHO project*[1] have been used as a basis for the creation of various other guidelines (Nonhebel et al., 2004a). Based on his work in the Auslan Corpus, Johnston (2010) introduced the use of ID-glosses, i. e. glosses that uniquely identify signs, adding indices. His annotation conventions constitute another basis for several conventions developed in different corpus projects. In 2015 the one-year project *Digging into Signs*[2] by University College London and Radboud University gathered a number of corpus creators from Europe for a workshop to further discuss annotation conventions. The project also suggested standard annotation protocols for glossing in sign language corpora.

Since 2010, more and more publications on annotation procedures have been released by corpus creators. The documentation of annotation conventions ranges from very elaborate descriptions to short insights. They may be available in English, the local language of the authors or both.

This report collects information on 17 corpora, two of them containing multiple sign languages, altogether covering 17 sign languages. These corpora and the languages that they cover are listed in Table 1.1. The comparison addresses annotation standards for manual and non-manual signs, describing their basic annotation format, formats for various specific phenomena and handshape coding approaches. This collection of topics represents the topics discussed in the majority of annotation conventions.

Although some common standards have been established over the years, such as time-aligned annotation and the use of ID-glosses, specific linguistic issues require specific ways of annotating. For EASIER a common interchange format for annotations within different datasets is needed. This format will enable EASIER to use the maximum number of European sign language resources. The suggested common format builds upon annotation conventions developed for the datasets presented in EASIER deliverable D6.1 (Kopf et al., 2021).

The remainder of this report is structured as follows: Chapter 2 describes the methodology followed for compiling the information on which this report is based. Chapter 3 introduces the basic structure of transcripts in the corpora. Chapter 4 describes the annotation standards for manual signs, separated by linguistic phenomena and other relevant categories. The annotation of non-manuals is presented in a similar manner in Chapter 5. Chapter 6 provides a tabular overview of how handshapes are annotated in different corpora. Chapter 7 concludes the report with a preliminary prototype for a unified representation of annotation data from the various considered corpora.

[1] http://sign-lang.ruhosting.nl/echo/

[2] https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/completed-projects/completed-projects/digging-signs/

Data	Corpus	Covered languages
	Auslan Corpus	Auslan
✓	British Sign Language Corpus (BSL Corpus)	British Sign Language (BSL)
✓	Corpus LSFB	French Belgian Sign Language (LSFB)
✓	Corpus NGT	Sign Language of the Netherlands (NGT)
✓	Corpus of Finnish Sign Language (CFinSL)	Finnish Sign Language (FinSL), Finland-Swedish Sign Language (FinSSL)
	Corpus Vlaamse Gebarentaal (Corpus VGT)	Flemish Sign Language (VGT)
	Danish Sign Language Corpus (DTS Corpus)	Danish Sign Language (DTS)
✓	DGS Corpus	German Sign Language (DGS)
✓	Dicta-Sign-LSF-v2 Corpus	French Sign Language (LSF)
✓	European Cultural Heritage Online Corpus (ECHO Corpus)	British Sign Language (BSL), Sign Language of the Netherlands (NGT), Swedish Sign Language (SSL)
	Hungarian Sign Language Corpus (HSL Corpus)	Hungarian Sign Language (HSL)
	Italian Sign Language Corpus (LIS Corpus)	Italian Sign Language (LIS)
	PJM Corpus	Polish Sign Language (PJM)
✓	Polytropon Parallel Corpus (Polytropon)	Greek Sign Language (GSL)
	SIGNOR Corpus	Slovene Sign Language (SZJ)
	Signs of Ireland	Irish Sign Language (ISL)
✓	Swedish Sign Language Corpus (SSLC)	Swedish Sign Language (SSL)

Table 1.1: *Overview of the inspected corpora and the language(s) each of them covers. A checkmark in the data column indicates that we were able to inspect annotated data for a corpus in addition to its annotation conventions and publications.*

Funded by the Horizon 2020
Framework Programme of the European Union

2 METHOD

To find a common format of annotations for EASIER we collected annotation conventions from the datasets presented in the EASIER project deliverable D6.1 (Kopf et al., 2021) and compared them with each other. Where we were not able to find published conventions, we referred to journal articles, conference papers and the presentations and posters of the 2015 *Digging into Signs* workshop. We only considered datasets that included lemma transcriptions of individual signs. Through this approach we were able to compare 16 of the 26 corpora listed in D6.1. To this selection of corpora, we added the Auslan Corpus, as many of the other corpora directly reference and build upon its annotation conventions.

A disclaimer regarding the recentness and accuracy of information: This report is primarily based on publicly available documents and data. In some cases, annotation conventions might have changed since their last public release. For example, the DTS Corpus conventions we report on are based on the preliminary guidelines presented by Kristoffersen and Troelsgård (2015) at the beginning of the project. We expect that their conventions have evolved since then but had no access to information regarding this.

After evaluating the annotation conventions of the corpora, we compared them to the actual annotations. For this, we inspected samples of annotation data where available. We could gain access to ten corpora, namely the BSL Corpus, Corpus LSFB, Corpus NGT, CFinSL, DGS Corpus, Dicta-Sign-LSF-v2, European Cultural Heritage Online Corpus (ECHO Corpus), Polytropon and the SSLC.

All the datasets presented here use either iLex[3] (Hanke and Storz, 2008) or ELAN[4] (Crasborn and Sloetjes, 2008) as annotation tools. Where we found differences between the published conventions and the annotations published, we discuss them in the respective section.

All the information compiled through this process will be used as the basis for a new data format equipped to represent the different annotation standards in a unified manner. A preliminary definition of this format is presented in the final chapter of this report.

We have endeavoured to be as complete and accurate as possible in our report. However, we are limited by the complexity of our task, the amount of data to be inspected and our sometimes limited access to information. Should you find incorrect, outdated, or missing information, please contact the authors at easier@dgs-korpus.de.

[3] https://www.sign-lang.uni-hamburg.de/ilex/

[4] https://archive.mpi.nl/tla/elan

3 ANNOTATION STRUCTURE

This section provides an overview of the general annotation structures used in the corpora. For each topic, a description of every individual corpus is given, describing how its structures are implemented and how this relates to the approaches of other corpora.

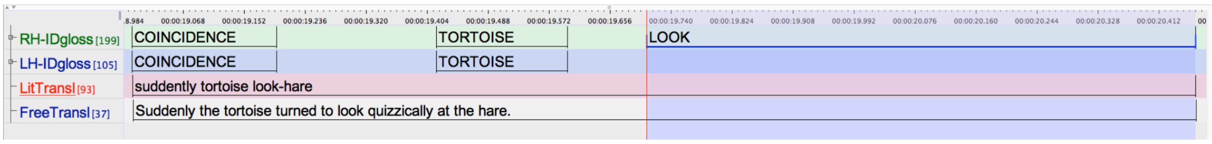
Section 3.1 describes the collected data and where to find them and elaborates on the organisation of the transcripts. The segmentation of annotations into individual units, such as sign tokens, is discussed in Section 3.2. Details of how translations are integrated are given in Section 3.3. Section 3.4 lists additional relevant aspects specific to individual corpora.

3.1 TRANSCRIPTS

All discussed datasets use ELAN or iLex for annotation and offer their transcripts either in the corresponding format or in several different formats, including online viewers with subtitles and/or transcripts. Some of the data is openly available online or for download, in other cases a registration is needed (cf. Kopf et al., 2021). The contents of individual datasets vary widely, but all transcripts contain at least glosses for manual signs and a translation into the local written language.

Auslan Corpus Transcripts in eaf format and video files in mp4 format are available via the Endangered Languages Archive (ELAR)[5] for registered users. At the time of writing we could not access the data due to a problem of the service provider. Therefore we could only look at the descriptions of the transcripts in the annotation conventions Johnston (2019).

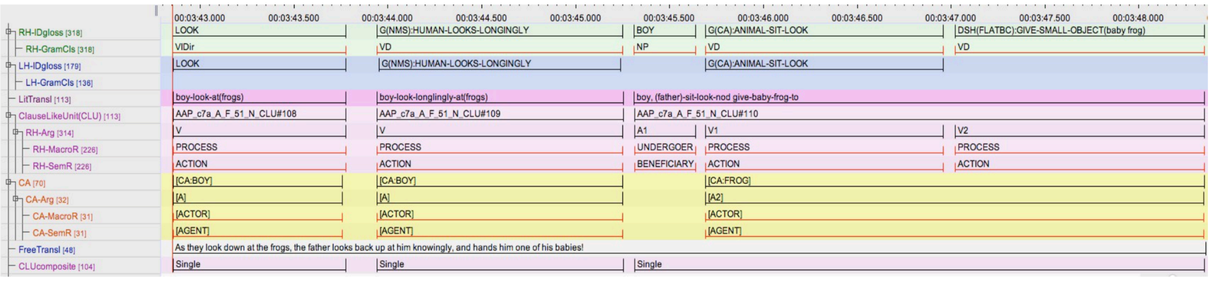
Example: Figure 3.1 shows a basic template for transcripts in the Auslan Corpus (as shown in the conventions) with tiers for both hands and the translation. Compare to Figure 3.2, an elaborate transcript with several additional tiers for different analyses.



8.984 00:00:19.068 00:00:19.152 00:00:19.236 00:00:19.320 00:00:19.404 00:00:19.488 00:00:19.572 00:00:19.656 00:00:19.740 00:00:19.824 00:00:19.908 00:00:19.992 00:00:20.076 00:00:20.160 00:00:20.244 00:00:20.328 00:00:20.412 00

RH-IDgloss [199]	COINCIDENCE	TORTOISE	LOOK
LH-IDgloss [105]	COINCIDENCE	TORTOISE	
LitTransl [93]	suddently tortoise look-hare		
FreeTransl [37]	Suddenly the tortoise turned to look quizzically at the hare.		

Figure 3.1: *A basic Auslan Corpus transcript with glosses and translations. (Johnston, 2019, p. 8)*



00:03:43.000 00:03:43.500 00:03:44.000 00:03:44.500 00:03:45.000 00:03:45.500 00:03:46.000 00:03:46.500 00:03:47.000 00:03:47.500 00:03:48.000

RH-IDgloss [318]	LOOK	G(NMS):HUMAN-LOOKS-LONGINGLY	BOY	G(CA):ANIMAL-SIT-LOOK	DSH(FLATBC):GIVE-SMALL-OBJECT(baby frog)
RH-GramCls [318]	VIDir	VD	NP	VD	VD
LH-IDgloss [179]	LOOK	G(NMS):HUMAN-LOOKS-LONGINGLY		G(CA):ANIMAL-SIT-LOOK	
LH-GramCls [136]					
LitTransl [113]	boy-look-at(frogs)	boy-look-longlingly-at(frogs)	boy, (father)-sit-look-nod give-baby-frog-to		
ClauseLikeUnit(CLU) [113]	AAP_c7a_A_F_51_N_CLU#108	AAP_c7a_A_F_51_N_CLU#109	AAP_c7a_A_F_51_N_CLU#110		
RH-Arg [314]	V	V	A1	V1	V2
RH-MacroR [226]	PROCESS	PROCESS	UNDERGOER	PROCESS	PROCESS
RH-SemR [226]	ACTION	ACTION	BENEFICIARY	ACTION	ACTION
CA [70]	[CA:BOY]	[CA:BOY]		[CA:FROG]	
CA-Arg [32]	[A]	[A]		[A2]	
CA-MacroR [31]	[ACTOR]	[ACTOR]		[ACTOR]	
CA-SemR [31]	[AGENT]	[AGENT]		[AGENT]	
FreeTransl [48]	As they look down at the frogs, the father looks back up at him knowingly, and hands him one of his babies!				
CLUcomposite [104]	Single	Single	Single		

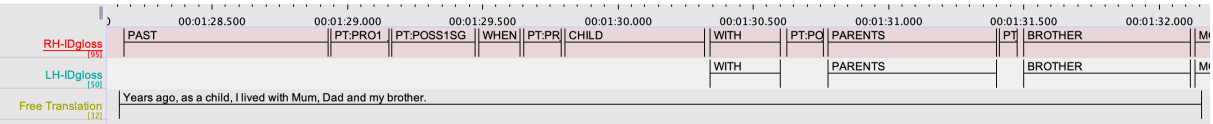
Figure 3.2: *A complex Auslan Corpus transcript with tiers for additional analyses. (Johnston, 2019, p. 74)*

[5] http://hdl.handle.net/2196/00-0000-0000-0000-D7CF-8

Funded by the Horizon 2020 Framework Programme of the European Union

BSL Corpus Transcripts in eaf format and video files in mov format are available via the CAVA repository.[6] A subset is openly available, for other files one has to register. The downloadable versions contain three tiers: **RH-IDgloss** for annotations of the right hand, **LH-IDgloss** for annotations of the left hand and **Free Translation** with a translation into written English.

Example: Figure 3.3 shows a typical part of a transcript in the BSL Corpus.



Years ago, as a child, I lived with Mum, Dad and my brother.

Figure 3.3: *Example of a BSL Corpus transcript. (Schembri et al., 2017, BF6n.eaf, 00:01:28.158–00:01:32.158, available at `https://digital-collections.ucl.ac.uk/R/DR2IJ6STT1P679IFR4HSQYKIRKEHQS4C7FPRVL94XBCXHSUUPL?local_base=BSLCP`),*

Corpus LSFB Transcripts and videos are available for registered users in an online movie player and for download, the transcripts in eaf format and the video files in mp4 format. In the online view glosses for both hands of both signers are visible. The downloadable files go beyond that by providing tiers for the annotation of a translation into French, negation as well as comments.

Example: Online view of a transcript shown in Figure 3.4 compared to the ELAN file shown in Figure 3.5:

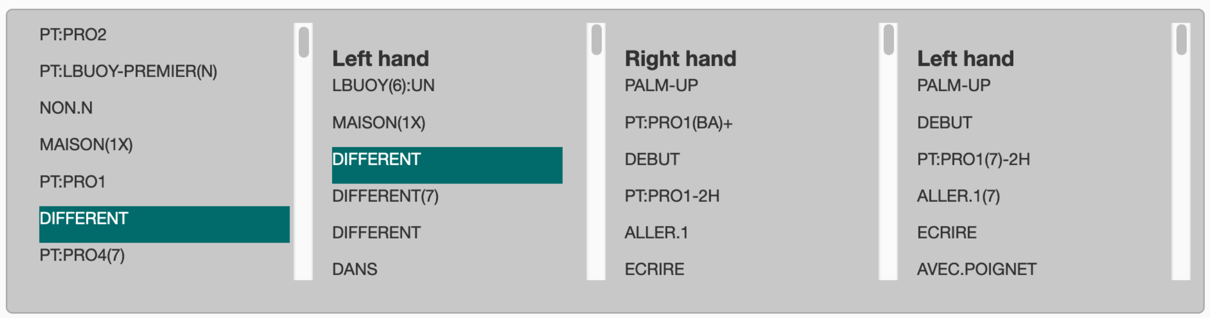
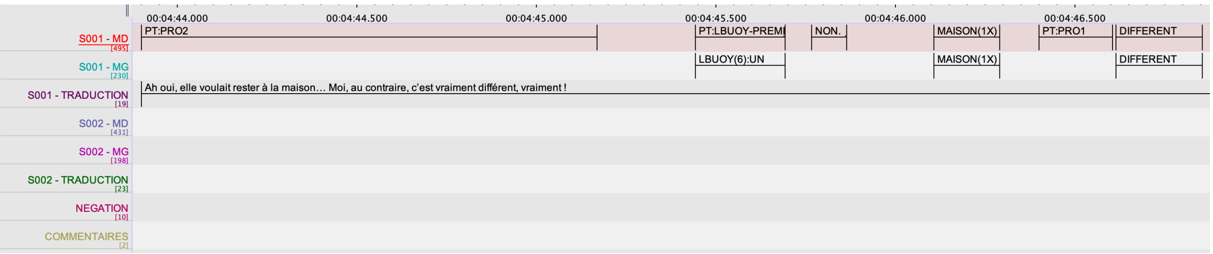


Figure 3.4: *Online view of a transcript from Corpus LSFB. (Meurant, 2015, Session 01 Task 03 - Childhood memory, available for registered users at `https://www.corpus-lsfb.be/pilote.php?session=01&tache=03&from=eaf&search=a:2:s:6:"action";s:4:"find";s:6:"filtre";s:10:"eafANDtrad";`)*



Ah oui, elle voulait rester à la maison… Moi, au contraire, c'est vraiment différent, vraiment !

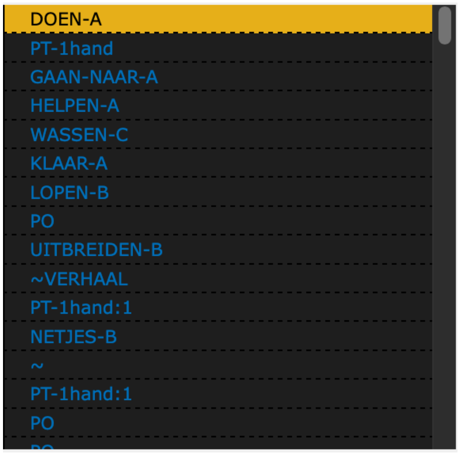
Figure 3.5: *Transcript of the Corpus LSFB in ELAN. (Meurant, 2015, CLSFBI0103.eaf, 00:04:43.902–00:04:46.862)*

[6] `https://digital-collections.ucl.ac.uk/R/DR2IJ6STT1P679IFR4HSQYKIRKEHQS4C7FPRVL94XBCXHSUUPL-04303?func=search-simple&local_base=BSLCP`

Funded by the Horizon 2020
Framework Programme of the European Union

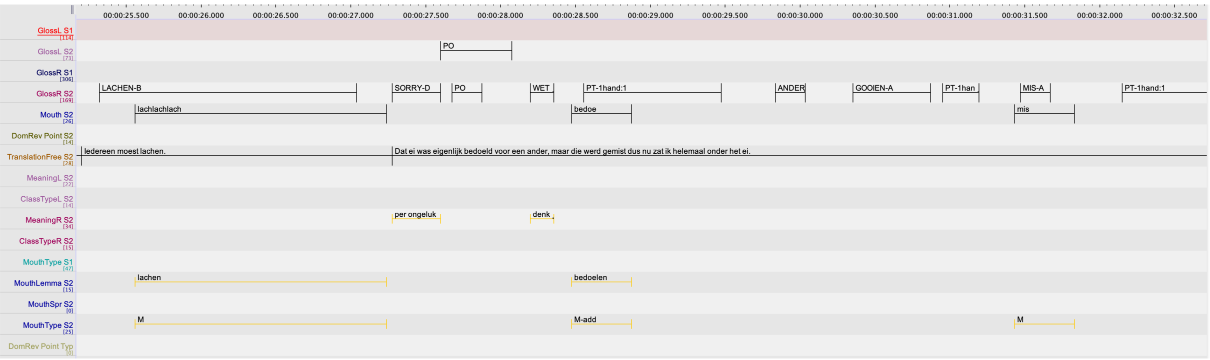
Corpus NGT Release 3 of the Corpus NGT is available online via the Language Archive.[7] It contains video files in mpg (MPEG 1) format, transcripts in eaf format (note that these are not the latest version of transcripts available) and voice-over translations into spoken Dutch in wav format. Transcripts of release 4 are available via the Radboud University homepage[8] and not yet stored in the Language Archive at the time of writing.[9] The transcripts can also be viewed online on the Corpus NGT Homepage.[10] The downloadable transcripts contain several tiers for glosses, meaning descriptions, mouthings and translations. Not all transcripts are openly available. Internally each ELAN file contains more than 200 tiers. (Crasborn et al., 2020, p. 9)

Example: Comparison of the online view of a transcript in Figure 3.6 and the ELAN file in Figure 3.7.



DOEN-A
PT-1hand
GAAN-NAAR-A
HELPEN-A
WASSEN-C
KLAAR-A
LOPEN-B
PO
UITBREIDEN-B
~VERHAAL
PT-1hand:1
NETJES-B
~
PT-1hand:1
PO

Figure 3.6: *Online view of a transcript from the Corpus NGT. (Crasborn et al., 2008b, CNGT0004, Personal recollection, available at https://www.corpusngt.nl/corpusvideo/177)*



GlossL S1 [114]
GlossL S2 [73] PO
GlossR S1 [306]
GlossR S2 [169] LACHEN-B SORRY-D PO WET PT-1hand:1 ANDER GOOIEN-A PT-1han MIS-A PT-1hand:1
Mouth S2 [26] lachlachlach bedoe mis
DomRev Point S2 [14]
TranslationFree S2 [28] Iedereen moest lachen. Dat ei was eigenlijk bedoeld voor een ander, maar die werd gemist dus nu zat ik helemaal onder het ei.
MeaningL S2 [22]
ClassTypeL S2 [14]
MeaningR S2 [34] per ongeluk denk
ClassTypeR S2 [15]
MouthType S1 [47]
MouthLemma S2 [15] lachen bedoelen
MouthSpr S2 [0]
MouthType S2 [25] M M-add M
DomRev Point Typ [0]

Figure 3.7: *Transcript of Corpus NGT in ELAN. (Crasborn et al., 2008b, CNGT0004.eaf, 00:00:25.200–00:00:31.200, available at https://hdl.handle.net/1839/00-0000-0000-0009-2D5C-5)*

Corpus FinSL From the Corpus FinSL a subset of video files in mp4 format, annotations in eaf

[7] https://hdl.handle.net/1839/b0c69aeb-222a-41df-b7c5-979001b635b3

[8] https://www.ru.nl/publish/pages/1013556/cngt_r4_public.zip

[9] Note that the persistent identifier given for each example will send you to the release 3 version, which makes it necessary to download the newer versions of transcripts separately via the address given above.

[10] https://www.corpusngt.nl/

format and metadata in IMDI format is available via the Language Bank of Finland[11]. Downloadable transcripts come with a range of tiers and daughter tiers for both hands: **ID_1/2_oik** for the right hand and **ID_1/2_vas** for the left hand, both with daughter tiers for additional entries, called **@_1/2_vas/oik**, tiers for comments, called **ID_huomioita_1/2**, tiers for the translation into written Finnish, called **Käännös_1/2** and for comments on the translation, called **Käännöshuomioita_1/2**. (Salonen et al., 2019, p. 4)

Example: Figure 3.8 shows a typical part of a transcript in the Corpus FinSL. The tiers for right and left hand seem to be duplicated into tiers with **-cp** appended, e. g. **ID_2_vas-cp** contains a copy of **ID_2_vas**. Several daughter tiers are collapsed for space reasons.

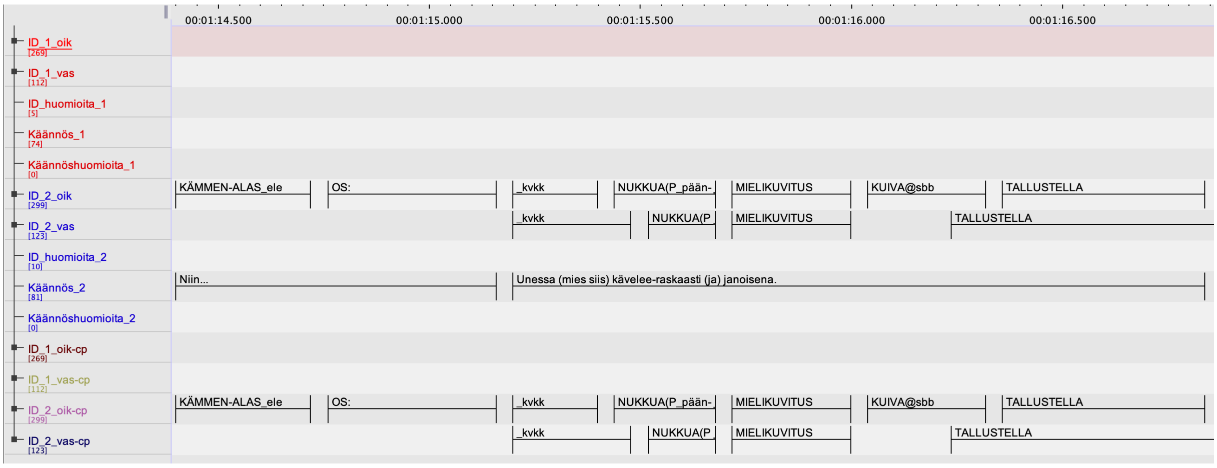


Figure 3.8: *Example of a Corpus FinSL transcript. (Jantunen, 2018, CFINSL2014_011_03.eaf, 00:01:14.400–00:01:16.840, available at `https://korp.csc.fi/download/cfinsl/elicit/cfinsl-elicit-3.zip`)*

Corpus VGT Due to registration issues we could not gain access to the transcripts and annotation conventions of the Corpus VGT. Annotations are done in ELAN with 26 tiers, although the main focus lies on the manual tiers **GlosRH/LH i1/2** and the tier for translations **Vertaling i1/2**. All information displayed here is taken from Verstraete et al. (2015).

Example: Figure 3.9 shows part of a Corpus VGT transcript.

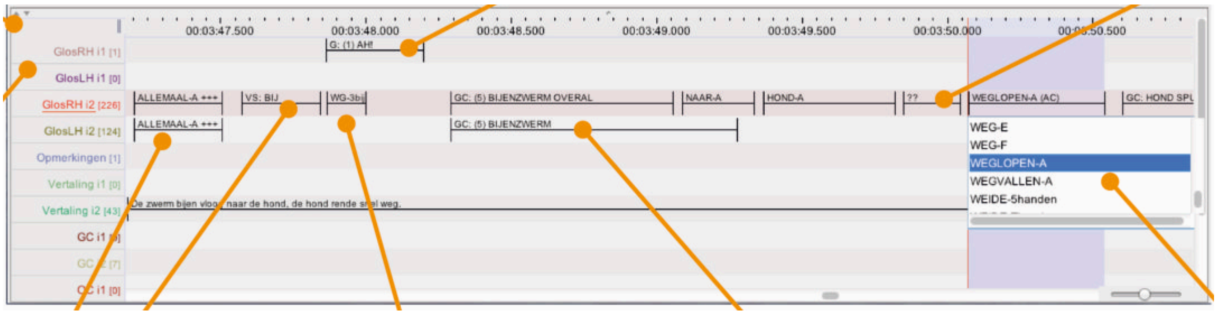


Figure 3.9: *Transcript of Corpus VGT as shown in Verstraete et al. (2015, picture)*

DTS Corpus The DTS Corpus is not yet available. The transcripts are done in iLex.

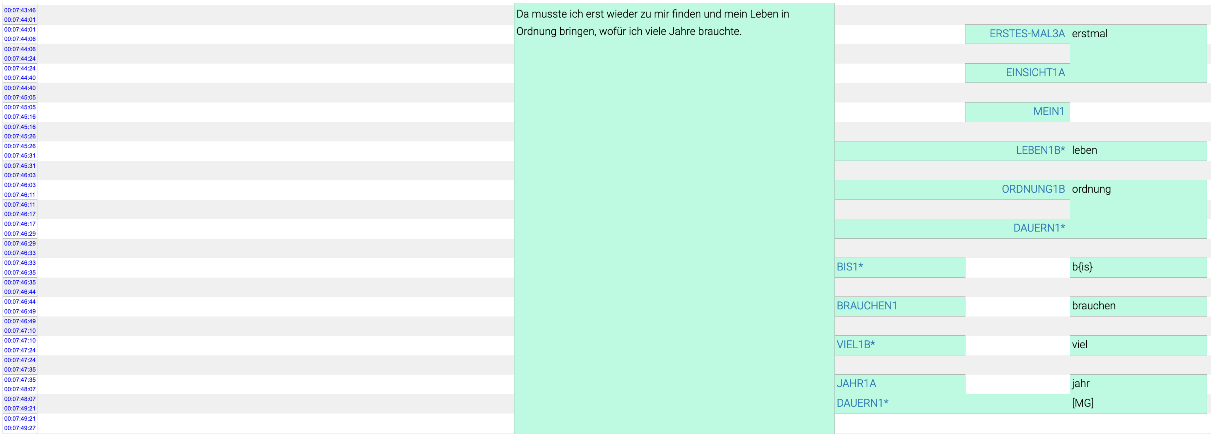
[11] `http://urn.fi/urn:nbn:fi:lb-2019092711`

Funded by the Horizon 2020
Framework Programme of the European Union

DGS Corpus A subset of the DGS Corpus, called the Public DGS Corpus, is publicly available online. It provides videos in mp4 format, transcripts in eaf and ilex format, subtitles in srt format, metadata in CMDI format and OpenPose in JSON format. The whole public corpus, every transcript and each type have their own set of DOIs, one for each release and a release-independent one. The transcripts and type information can also be viewed online on the Public DGS Corpus website.[12]

The downloadable transcripts contain several tiers for glosses in German and English, mouthings and translations into German and English. The iLex version differs from the ELAN version in that both hands are annotated in one tier with separate slots for right and left hand instead of two separate ones. More information on this approach, called 'double-token tags', will be given in Section 4.3. In order to differentiate between the online and offline (or internal) versions of the transcript we use the term 'Public DGS Corpus' to refer to the online versions and 'DGS Corpus' to the internal iLex versions (not to be confused with the downloadable iLex versions of the Public DGS Corpus). The Public DGS corpus is also available via a public access page[13] where the video files can be watched with German subtitles.

Example: Figures 3.10 to 3.12 provide a comparison between the view of a transcript in the online, ELAN and iLex view, respectively. In the example the signer starts with one handed signs on the left hand, followed by two-handed signs and then changes their hand dominance and continues with one-handed signs on the right hand.

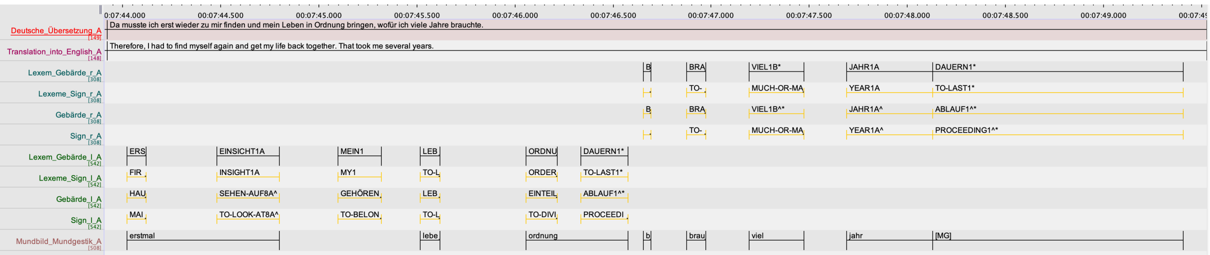


00:07:43:46 00:07:44:01		Da musste ich erst wieder zu mir finden und mein Leben in Ordnung bringen, wofür ich viele Jahre brauchte.			
00:07:44:01 00:07:44:06				ERSTES-MAL3A	erstmal
00:07:44:06 00:07:44:24					
00:07:44:24 00:07:44:40				EINSICHT1A	
00:07:44:40 00:07:45:05					
00:07:45:05 00:07:45:16				MEIN1	
00:07:45:16 00:07:45:26					
00:07:45:26 00:07:45:31				LEBEN1B*	leben
00:07:45:31 00:07:46:03					
00:07:46:03 00:07:46:11				ORDNUNG1B	ordnung
00:07:46:11 00:07:46:17					
00:07:46:17 00:07:46:29				DAUERN1*	
00:07:46:29 00:07:46:33					
00:07:46:33 00:07:46:35			BIS1*		b{is}
00:07:46:35 00:07:46:44					
00:07:46:44 00:07:46:49			BRAUCHEN1		brauchen
00:07:46:49 00:07:47:10					
00:07:47:10 00:07:47:24			VIEL1B*		viel
00:07:47:24 00:07:47:35					
00:07:47:35 00:07:48:07			JAHR1A		jahr
00:07:48:07 00:07:49:21			DAUERN1*		[MG]
00:07:49:21 00:07:49:27					

Figure 3.10: *Online transcript of the Public DGS Corpus. (Konrad et al., 2020a, dgskorpus_ber_08 – Experience of Deaf Individuals, 00:07:43:46–00:07:49:27, available at https://doi.org/10.25592/dgs.corpus-3.0-text-1418889)*

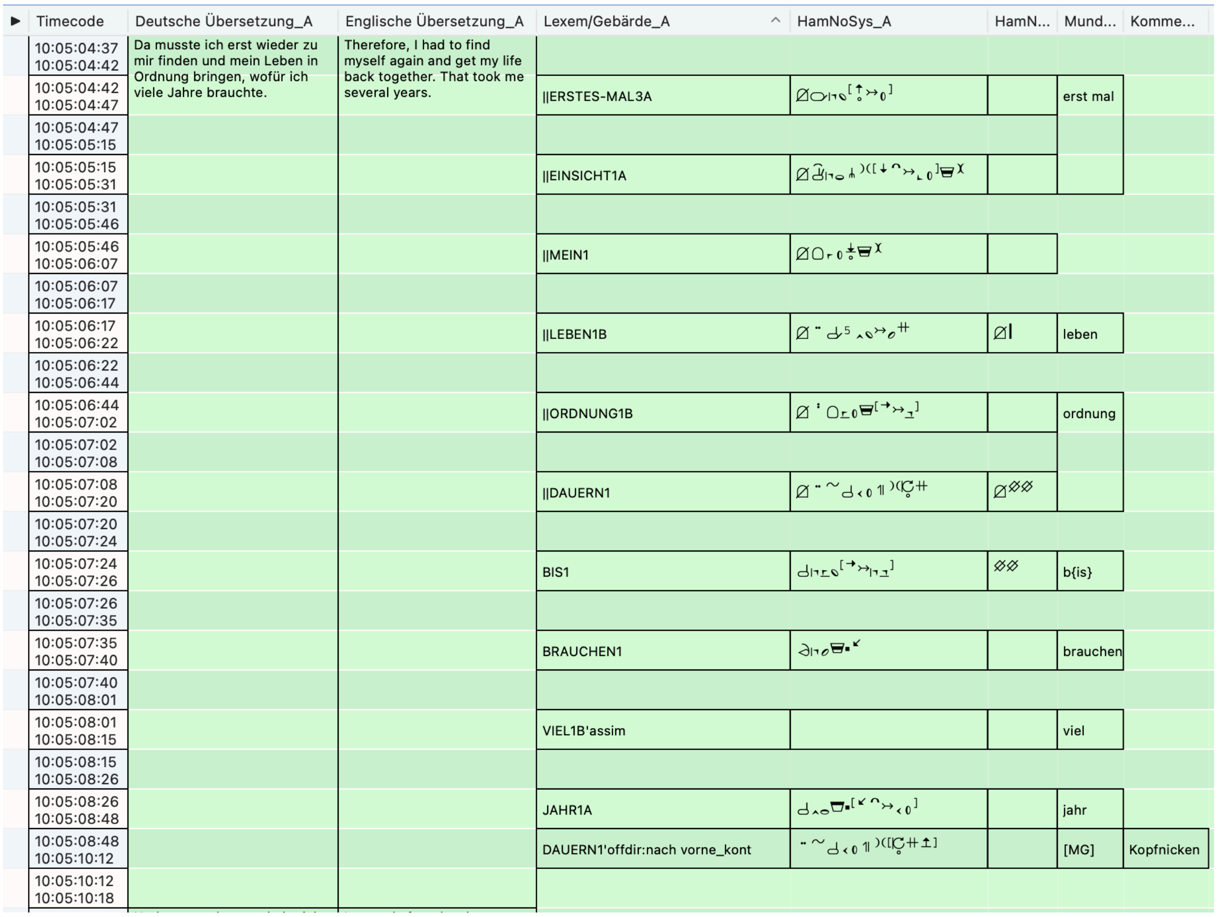
[12] http://ling.meine-dgs.de/

[13] http://meine-dgs.de



	0:07:44.000	00:07:44.500	00:07:45.000	00:07:45.500	00:07:46.000	00:07:46.500	00:07:47.000	00:07:47.500	00:07:48.000	00:07:48.500	00:07:49.000	00:07:4		
Deutsche_Übersetzung_A [149]	Da mussste ich erst wieder zu mir finden und mein Leben in Ordnung bringen, wofür ich viele Jahre brauchte.													
Translation_into_English_A [148]	Therefore, I had to find myself again and get my life back together. That took me several years.													
Lexem_Gebärde_r_A [308]							B	BRA	VIEL1B*	JAHR1A	DAUERN1*			
Lexeme_Sign_r_A [308]								TO-	MUCH-OR-MA	YEAR1A	TO-LAST1*			
Gebärde_r_A [308]							B	BRA	VIEL1B^*	JAHR1A^	ABLAUF1^*			
Sign_r_A [308]								TO-	MUCH-OR-MA	YEAR1A^	PROCEEDING1^*			
Lexem_Gebärde_l_A [542]	ERS	EINSICHT1A	MEIN1	LEB	ORDNU	DAUERN1*								
Lexeme_Sign_l_A [542]	FIR	INSIGHT1A	MY1	TO-L	ORDER	TO-LAST1*								
Gebärde_l_A [542]	HAU	SEHEN-AUF8A^	GEHÖREN	LEB	EINTEIL	ABLAUF1^*								
Sign_l_A [542]	MAI	TO-LOOK-AT8A^	TO-BELON	TO-L	TO-DIVI	PROCEEDI								
Mundbild_Mundgestik_A [508]	erstmal	lebe	ordnung	b	brau	viel	jahr	[MG]						

Figure 3.11: *ELAN transcript of the Public DGS Corpus. (Konrad et al., 2020a, 1418889.eaf, 00:07:43.920–00:07:49.540, available at https://doi.org/10.25592/dgs.corpus-3.0-text-1418889)*



Timecode	Deutsche Übersetzung_A	Englische Übersetzung_A	Lexem/Gebärde_A	HamNoSys_A	HamN...	Mund...	Komme...
10:05:04:37 10:05:04:42	Da musste ich erst wieder zu mir finden und mein Leben in Ordnung bringen, wofür ich viele Jahre brauchte.	Therefore, I had to find myself again and get my life back together. That took me several years.					
10:05:04:42 10:05:04:47			||ERSTES-MAL3A	[illegible]		erst mal	
10:05:04:47 10:05:05:15							
10:05:05:15 10:05:05:31			||EINSICHT1A	[illegible]			
10:05:05:31 10:05:05:46							
10:05:05:46 10:05:06:07			||MEIN1	[illegible]			
10:05:06:07 10:05:06:17							
10:05:06:17 10:05:06:22			||LEBEN1B	[illegible]	[illegible]	leben	
10:05:06:22 10:05:06:44							
10:05:06:44 10:05:07:02			||ORDNUNG1B	[illegible]		ordnung	
10:05:07:02 10:05:07:08							
10:05:07:08 10:05:07:20			||DAUERN1	[illegible]	[illegible]		
10:05:07:20 10:05:07:24							
10:05:07:24 10:05:07:26			BIS1	[illegible]	[illegible]	b{is}	
10:05:07:26 10:05:07:35							
10:05:07:35 10:05:07:40			BRAUCHEN1	[illegible]		brauchen	
10:05:07:40 10:05:08:01							
10:05:08:01 10:05:08:15			VIEL1B'assim			viel	
10:05:08:15 10:05:08:26							
10:05:08:26 10:05:08:48			JAHR1A	[illegible]		jahr	
10:05:08:48 10:05:10:12			DAUERN1'offdir:nach vorne_kont	[illegible]		[MG]	Kopfnicken
10:05:10:12 10:05:10:18							

Figure 3.12: *iLex transcript of the DGS Corpus. (Konrad et al., 2020a, dgskorpus_ber_08: Erfahrungen als Gehörloser, 10:05:04:37–10:05:10:18, available at https://doi.org/10.25592/dgs.corpus-3.0-text-1418889)*

Dicta-Sign-LSF-v2 Corpus Video files in mp4 format, elicitation material in wmv and ppt format, annotations in csv format and preprocessed signer representations in npy format are available via ORTOLANG[14]. The transcripts are split up into two files: the first contains a mapping from gloss IDs to their textual gloss representation in French; the second contains the annotation of the transcript (file reference, time stamp, handedness, etc.) and refers to specific glosses only

[14] https://hdl.handle.net/11403/dicta-sign-lsf-v2/v1

by their ID. Annotations were made with iLex using four tiers, three for the manual signs (**LH** for left hand, **RH** for right hand, **2H** for two hands) and a tier **Translation** for the translation into French. (Braffort, 2019, p. 3)

Example: Figure 3.13 shows excerpts of the corpus data files. The gloss mapping is shown in Figure 3.13a and the transcript annotation in Figure 3.13b.

```
id;name
43889;POUR_RIEN
43913;BLOND:VAR
42368;CONTENT1/JOIE
43893;CABINE (BATEAU)
43894;COMPRIS
43895;RELIGION
42385;ENSEIGNER
43896;MONUMENTS
44454;CRU2
43858;CLANDESTIN
45046;VALLEE2
43899;OFFICIEL
43900;PEUT-ETRE2
```

(a) *ID-to-gloss mapping (Dicta-Sign-LSF_ID.csv)*

```
Video,Loc,Track,Start,End,Cat,Value
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,33,36,ID,42495
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,42,48,ID,42856
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,55,60,ID,42093
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,130,136,ID,42159
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,141,145,ID,42856
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,165,170,ID,43122
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,174,179,ID,42503
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,240,244,G,
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,278,284,ID,42319
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,354,359,ID,43614
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,438,447,ID,43403
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,453,457,ID,42319
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,462,468,ID,43470
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,511,523,ID,43470
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,546,551,ID,43131
DictaSign_lsf_S2_T1_A11_front.mp4,A11,2H,554,561,ID,43101
```

(b) *Annotation file (Dicta-Sign-LSF_Annotation.csv)*

Figure 3.13: *Excerpts of data files from the Dicta-Sign-LSF-v2 Corpus, available at* *https://www.ortolang.fr/market/corpora/dicta-sign-lsf-v2/v1?path=/data/annotations*. *(LIMSI, 2020)*

Digging into Signs The standard annotation guideline of the *Digging into Signs* project is not based on one corpus but on an extensive comparison and adaption of the annotation guidelines from Corpus NGT and BSL Corpus. The comparison focused on manual action only. In addition to the exchange between the two Corpora teams a discussion with various researchers from the field contributed to the development of these guidelines.

Both corpora use ELAN in combination with SignBank as lexical database for annotations. Separate tiers for the left and right hand are used. (Crasborn et al., 2015a, pp. 1–4)

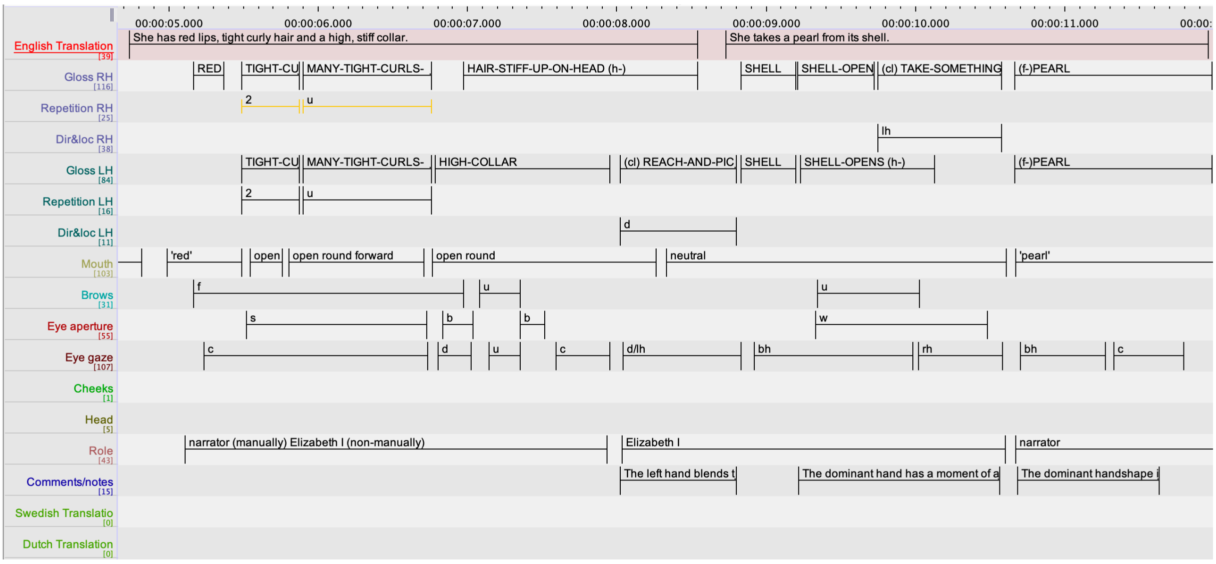
As there is no Digging into Signs corpus there are no transcripts that we can use as examples here. Some of the conventions described can be seen as examples in the sections on the BSL Corpus and the Corpus NGT, respectively.

ECHO Corpus Transcripts in eaf format and video files in mpeg format are available via The Language Archive[15]. Additionally, metadata in CMDI format and some additional material on the elicitation methods can be downloaded. The transcripts have several tiers to annotate left and right hand, non-manual features, and translations. (Nonhebel et al., 2004a, p. 2)

Example: Figures 3.14 to 3.16 show basic templates for the transcripts of the ECHO Corpus, one per language. There are minor differences in the naming of the tiers, e. g. **Gloss RH**

[15] https://hdl.handle.net/1839/00-0000-0000-0001-4892-C

English, Brows in the NGT dataset, **Gloss English RH, Brow** in the SSL dataset and **Gloss RH, Brows** in the BSL dataset.



00:00:05.000 00:00:06.000 00:00:07.000 00:00:08.000 00:00:09.000 00:00:10.000 00:00:11.000 00:00:

English Translation [39]: She has red lips, tight curly hair and a high, stiff collar. | She takes a pearl from its shell.

Gloss RH [116]: RED | TIGHT-CU | MANY-TIGHT-CURLS- | HAIR-STIFF-UP-ON-HEAD (h-) | SHELL | SHELL-OPEN | (cl) TAKE-SOMETHING | (f-)PEARL

Repetition RH [25]: 2 | u

Dir&loc RH [38]: lh

Gloss LH [84]: TIGHT-CU | MANY-TIGHT-CURLS- | HIGH-COLLAR | (cl) REACH-AND-PIC | SHELL | SHELL-OPENS (h-) | (f-)PEARL

Repetition LH [16]: 2 | u

Dir&loc LH [11]: d

Mouth [103]: 'red' | open | open round forward | open round | neutral | 'pearl'

Brows [31]: f | u | u

Eye aperture [55]: s | b | b | w

Eye gaze [107]: c | d | u | c | d/lh | bh | rh | bh | c

Cheeks [1]

Head [5]

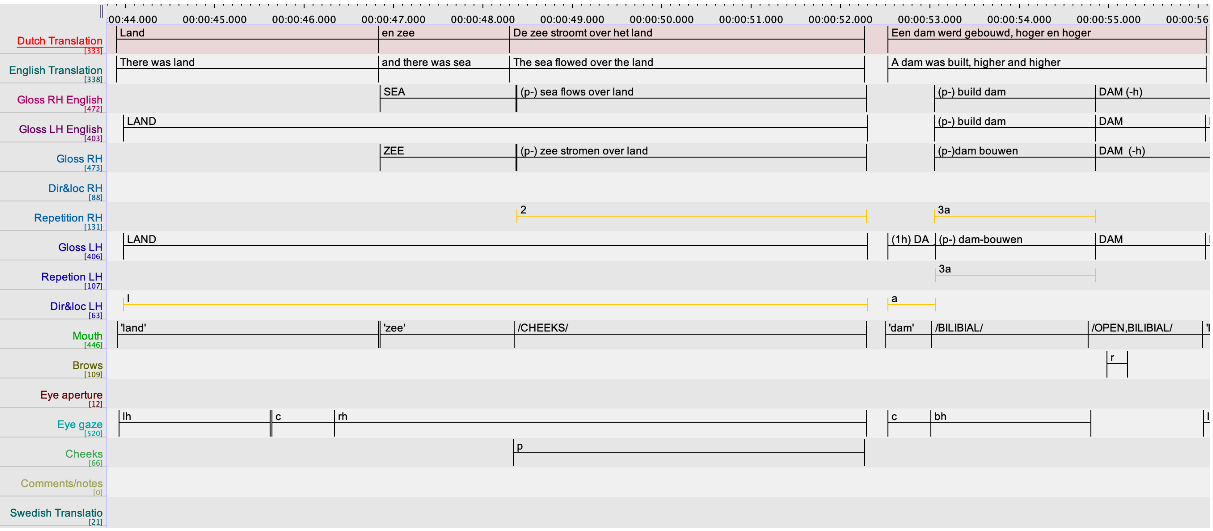
Role [43]: narrator (manually) Elizabeth I (non-manually) | Elizabeth I | narrator

Comments/notes [15]: The left hand blends t | The dominant hand has a moment of a | The dominant handshape i

Swedish Translatio [0]

Dutch Translation [0]

Figure 3.14: *A BSL transcript of the ECHO Corpus. (Woll et al., 2004, BSL_PS_poem3.eaf, 00:00:04.740–00:00:11.970)*



00:44.000 00:00:45.000 00:00:46.000 00:00:47.000 00:00:48.000 00:00:49.000 00:00:50.000 00:00:51.000 00:00:52.000 00:00:53.000 00:00:54.000 00:00:55.000 00:00:56

Dutch Translation [333]: Land | en zee | De zee stroomt over het land | Een dam werd gebouwd, hoger en hoger

English Translation [338]: There was land | and there was sea | The sea flowed over the land | A dam was built, higher and higher

Gloss RH English [472]: SEA | (p-) sea flows over land | (p-) build dam | DAM (-h)

Gloss LH English [403]: LAND | (p-) build dam | DAM

Gloss RH [473]: ZEE | (p-) zee stromen over land | (p-)dam bouwen | DAM (-h)

Dir&loc RH [88]

Repetition RH [131]: 2 | 3a

Gloss LH [406]: LAND | (1h) DA | (p-) dam-bouwen | DAM

Repetion LH [107]: 3a

Dir&loc LH [63]: l | a

Mouth [446]: 'land' | 'zee' | /CHEEKS/ | 'dam' | /BILIBIAL/ | /OPEN,BILIBIAL/ | '

Brows [109]: r

Eye aperture [12]

Eye gaze [520]: lh | c | rh | c | bh | l

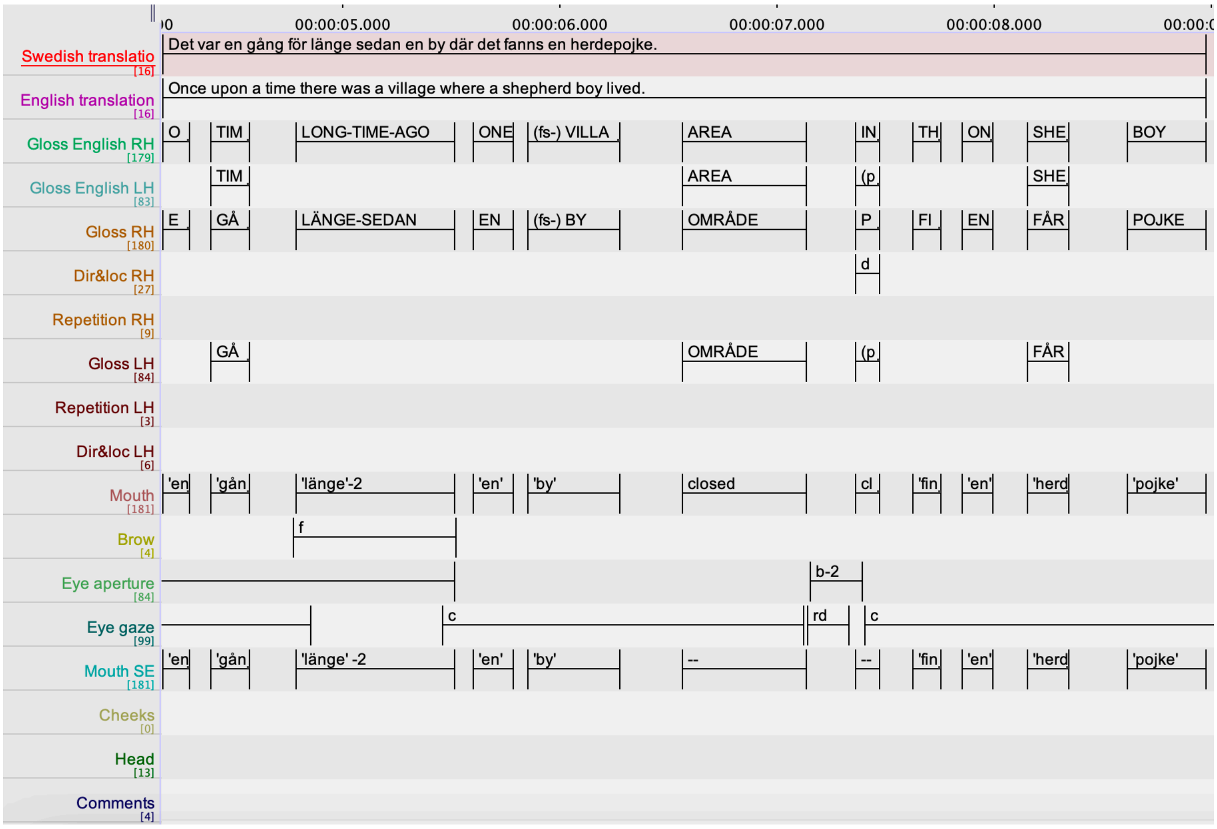
Cheeks [66]: p

Comments/notes [0]

Swedish Translatio [21]

Figure 3.15: *A NGT transcript of the ECHO Corpus. (Crasborn et al., 2004, NGT_WE_poems.eaf, 00:00:43.900–00:00:56.090)*

Funded by the Horizon 2020
Framework Programme of the European Union



Tier	Annotations
Swedish translatio [16]	Det var en gång för länge sedan en by där det fanns en herdepojke.
English translation [16]	Once upon a time there was a village where a shepherd boy lived.
Gloss English RH [179]	O, TIM, LONG-TIME-AGO, ONE, (fs-) VILLA, AREA, IN, TH, ON, SHE, BOY
Gloss English LH [83]	TIM, AREA, (p, SHE
Gloss RH [180]	E, GÅ, LÄNGE-SEDAN, EN, (fs-) BY, OMRÅDE, P, FI, EN, FÅR, POJKE
Dir&loc RH [27]	d
Repetition RH [9]	
Gloss LH [84]	GÅ, OMRÅDE, (p, FÅR
Repetition LH [3]	
Dir&loc LH [6]	
Mouth [181]	'en, 'gån, 'länge'-2, 'en', 'by', closed, cl, 'fin, 'en', 'herd, 'pojke'
Brow [4]	f
Eye aperture [84]	b-2
Eye gaze [99]	c, rd, c
Mouth SE [181]	'en, 'gån, 'länge' -2, 'en', 'by', --, --, 'fin, 'en', 'herd, 'pojke'
Cheeks [0]	
Head [13]	
Comments [4]	

Figure 3.16: *An SSL transcript from the ECHO Corpus. (Bergman and Mesch, 2004, SSL_JI_fab1.eaf, 00:00:04.170–00:00:08.970)*

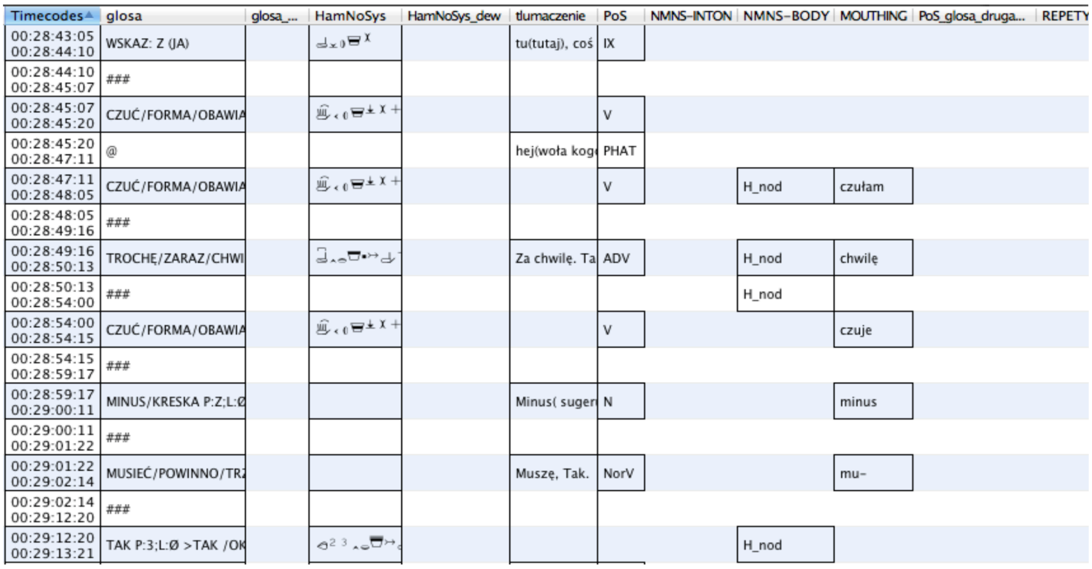
Hungarian Sign Language Corpus (HSL Corpus) The HSL Corpus is not available at the time of writing. For the annotation three ELAN templates have been created: sociolinguistic, grammatical, and lexical. The templates are based on Johnston (2013) and consist of 140 tiers per person, including the fieldworkers as well as the informants. The tiers cover all linguistic levels from phonetics to pragmatics and the translation. The annotation conventions differ for the three templates. Due to limited resources the translations were written into a Word chart and will be entered into ELAN in the future. So far five sociolinguistic interviews already were checked and transferred to ELAN. (Bartha et al., 2016, p. 4).

LIS Corpus The LIS Corpus is not available at the time of writing. For each signer the first 100 words have been annotated in ELAN. A lot of information is annotated on separate tiers instead of in the ID-glosses, therefore the ELAN template is more complex than compared to others. (Santoro and Geraci, 2015, slides 6, 25)

PJM Corpus The PJM Corpus is not available online at the time of writing. The annotation is done in iLex. Several tiers are used for glosses for the dominant and non-dominant hand (each on a separate tier), Hamburg Notation System for Sign Languages (HamNoSys) transcriptions, translations and further tags for non-manual movements and Part of Speech. (Rutkowski et al., 2015, slide 8)

Example: Figure 3.17 shows part of an iLex transcript for the PJM Corpus.

Funded by the Horizon 2020
Framework Programme of the European Union



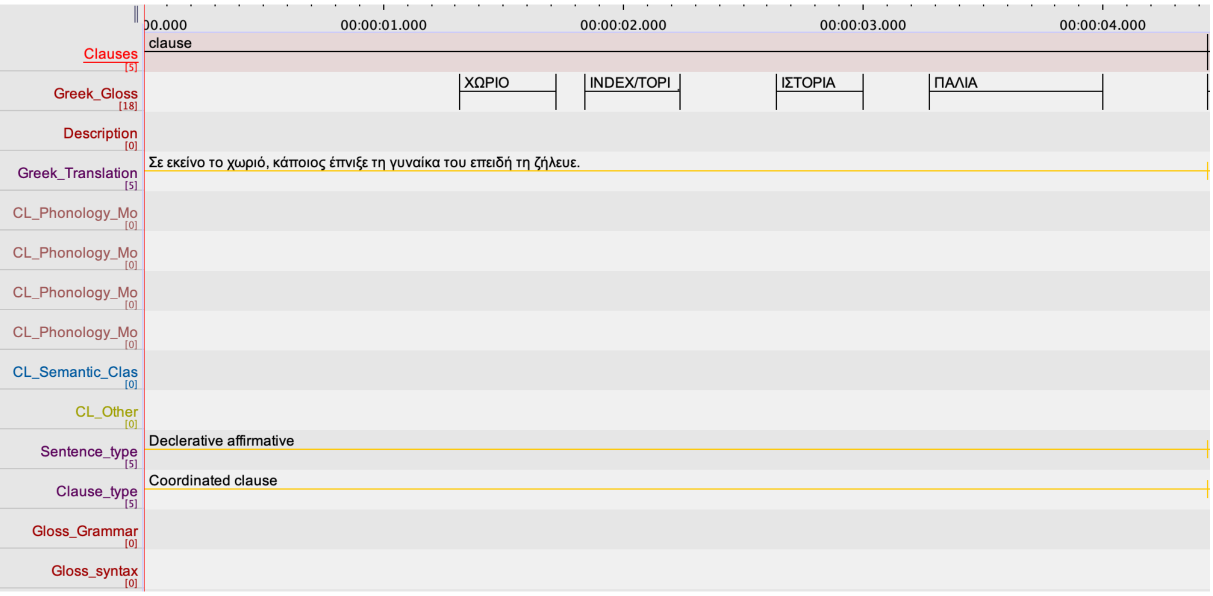
Timecodes	glosa	glosa_...	HamNoSys	HamNoSys_dew	tlumaczenie	PoS	NMNS-INTON	NMNS-BODY	MOUTHING	PoS_glosa_druga...	REPETY
00:28:43:05 00:28:44:10	WSKAZ: Z (JA)				tu(tutaj), coś	IX					
00:28:44:10 00:28:45:07	###										
00:28:45:07 00:28:45:20	CZUĆ/FORMA/OBAWIA					V					
00:28:45:20 00:28:47:11	@				hej(woła kog	PHAT					
00:28:47:11 00:28:48:05	CZUĆ/FORMA/OBAWIA					V		H_nod	czułam		
00:28:48:05 00:28:49:16	###										
00:28:49:16 00:28:50:13	TROCHĘ/ZARAZ/CHWI				Za chwilę. Ta	ADV		H_nod	chwilę		
00:28:50:13 00:28:54:00	###							H_nod			
00:28:54:00 00:28:54:15	CZUĆ/FORMA/OBAWIA					V			czuje		
00:28:54:15 00:28:59:17	###										
00:28:59:17 00:29:00:11	MINUS/KRESKA P:Z;L:Ø				Minus(suger	N			minus		
00:29:00:11 00:29:01:22	###										
00:29:01:22 00:29:02:14	MUSIEĆ/POWINNO/TRZ				Muszę, Tak.	NorV			mu-		
00:29:02:14 00:29:12:20	###										
00:29:12:20 00:29:13:21	TAK P:3;L:Ø >TAK /OK							H_nod			

Figure 3.17: *An iLex transcript for the PJM Corpus. (Rutkowski et al., 2015, slide 8)*

Polytropon Transcripts in eaf format and video files in mp4 format are available to download for registered users via clarin:el[16]. The original annotation work was done in iLex, the data was then transferred to ELAN *(E. Efthimiou, personal communication, January 24, 2022)*. The downloadable transcripts contain several tiers for the gloss and the translation as well as different levels of analysis like clause annotations, phonology, semantic and grammar annotations, and others, but – in contrast to many other corpora – only one tier for both hands.

Polytropon differs from the other corpora in that it is one signer signing prepared sentences which are based on the Polytropon lexicon.

Example: Figure 3.18 shows an example of the basic template that is used for all transcripts.



Clauses [5]: clause

Greek_Gloss [18]: ΧΩΡΙΟ, INDEX/TOPI, ΙΣΤΟΡΙΑ, ΠΑΛΙΑ

Description [0]

Greek_Translation [5]: Σε εκείνο το χωριό, κάποιος έπνιξε τη γυναίκα του επειδή τη ζήλευε.

CL_Phonology_Mo [0]

CL_Phonology_Mo [0]

CL_Phonology_Mo [0]

CL_Phonology_Mo [0]

CL_Semantic_Clas [0]

CL_Other [0]

Sentence_type [5]: Declerative affirmative

Clause_type [5]: Coordinated clause

Gloss_Grammar [0]

Gloss_syntax [0]

Figure 3.18: *A public ELAN transcript of Polytropon. (ILSP Athena Research Center, 2018, Clarin_pnigw_kapoion_ex_2_HD_SVQ1.eaf)*

[16] http://hdl.handle.net/11500/ATHENA-0000-0000-4C77-6

SIGNOR Corpus Transcripts of the SIGNOR Corpus are not available at the time of writing. Annotations are done in iLex. All examples are taken from Jerko and Vintar (2015).

Signs of Ireland The Signs of Ireland Corpus is not available at the time of writing. Annotations are done in ELAN. From the provided picture it seems as there is only one tier for both right and left hand, which contrasts with a lot of other transcript organisations but is the same as in the POLYTROPON Corpus. (Matthews and Sheridan, 2015, slide 5).

Example: Figure 3.19 shows a part of a transcript.

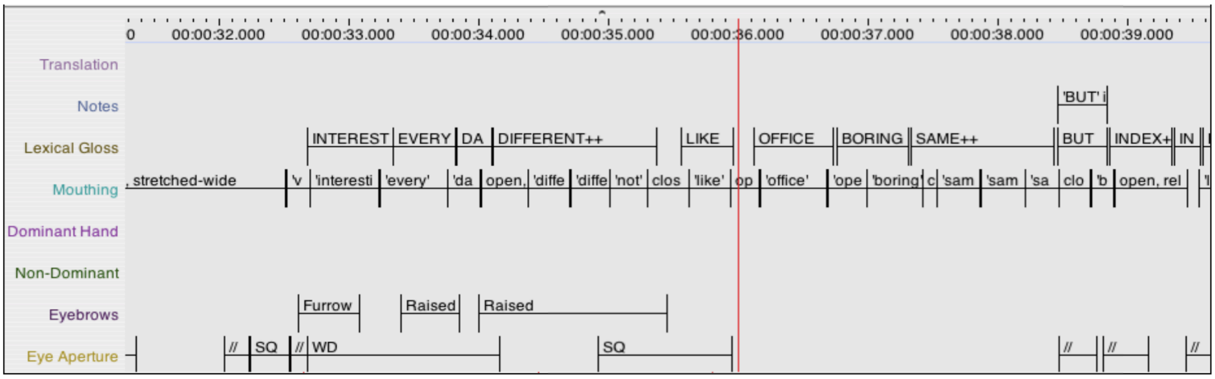


Figure 3.19: *Transcript of the Signs of Ireland Corpus as shown in Matthews and Sheridan (2015, slide 5)*

SSL Corpus Video files are available for download in mp4 format via the university webpage[17]. Transcripts are available in an online viewer on the corpus homepage[18]. The annotation is done in ELAN, but eaf files are not available for download at the time of writing. The online viewer shows six tiers, two for each signer (dominant and non-dominant hand) called **Glosa_DH S1/S2** and **Glosa_NonDH S1/2** and two for the translation of each signer called **Översättning S1/S2**. In ELAN more tiers are used.

Example: Figure 3.20 shows a transcript in ELAN view as shown in Wallin and Mesch (2018, p. 1), while Figure 3.21 shows a transcript as seen in the online viewer of the corpus.

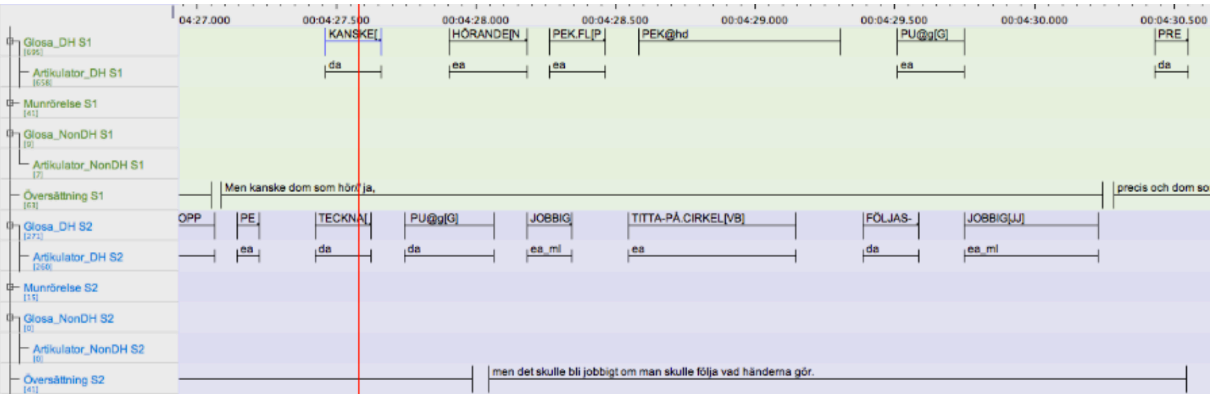
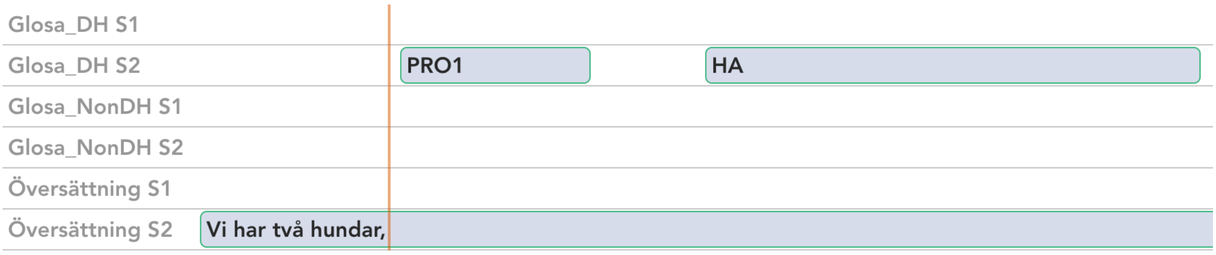


Figure 3.20: *Example of a transcript of the SSL Corpus in ELAN. (Wallin and Mesch, 2018, p. 1)*

[17] https://ling33.ling.su.se/sslc/video/
[18] https://teckensprakskorpus.su.se



Glosa_DH S1		
Glosa_DH S2	PRO1	HA
Glosa_NonDH S1		
Glosa_NonDH S2		
Översättning S1		
Översättning S2	Vi har två hundar,	

Figure 3.21: *Example of a transcript of the SSL Corpus in its online viewer. (Mesch et al., 2012, https://teckenspraakskorpus.su.se/#/video/sslc01_045.eaf)*

3.2 SEGMENTATION

Before a gloss annotation can be done the sign stream must be separated into single signs, defining start and end of each sign token. There are different conventions to define the length of a sign, e. g. some researchers see the transitional movement as a part of the sign while others do not. Even where the transitional movement is considered part of the sign, most corpora leave at least one video frame between sign segment due to technical reasons. The process of separating the sign stream into individual signs is also found under the term ‘parsing’ (Cormier et al., 2017, p. 5).

In this section no examples are given as the pictures in the subsequent sections contain a full range of different segments.

Auslan Corpus Segmentation is done rule-based with transitional movements seen as part of the signing and therefore possibly short breaks in between signs, although at least one frame between glosses is recommended for technical reasons. (Johnston, 2019, pp. 46–47)

BSL Corpus Segmentation is done rule-based with transitional movements seen as part of the signing, like in the Auslan Corpus. For two-handed signs the end is determined on basis of the dominant hand and the non-dominant is aligned. Between signs there are small gaps of two frames due to historical (technical) reasons. (Cormier et al., 2017, p. 5)

Corpus LSFB From the available transcripts it seems as the segmentation is done with transitional movements not seen as part of the sign. Start and end of a sign seem to be defined by parameter changes, for most two-handed signs start and end are determined on the dominant hand and the non-dominant hand is aligned but holds of the non-dominant hand and complex constructions are segmented separately for each hand. (Sinte et al., 2015)

Corpus NGT Segmentation is done rule-based for each hand separately with transitional movements not seen as part of the sign. (Crasborn et al., 2020, p. 15)

Corpus FinSL Segmentation is done rule-based according to the definitions of Jantunen (2013), who defines the sign as a relatively long unit containing the transitional movements. At least one video frame is left between the annotation cells in ELAN. (Salonen et al., 2019, p. 6)

Corpus VGT From the picture in Verstraete et al. (2015) it seems like there being at least small gaps, longer than one frame, between glosses. It becomes not apparent if start and end of two-handed signs are time-aligned or not.

DTS Corpus No information available.

DGS Corpus Segmentation is done rule-based with the transitional movement not seen as part of a sign. For two-handed signs the end is determined on basis of the dominant hand and the non-dominant is aligned. (Konrad et al., 2020b, pp. 4–6) Detailed information on the segmentation is given in Hanke et al. (2019).

Dicta-Sign-LSF-v2 Corpus Segmentation is done rule-based, start and end are determined based on parameter changes not counting the transitional movement as part of the sign. (Braffort, 2019, pp. 3–4)

Digging into Signs No information available.

Funded by the Horizon 2020 Framework Programme of the European Union

ECHO Corpus From the available transcripts it seems as the segmentation of the BSL dataset is done with the transitional movement seen as part of a sign. Start and end of signs seem to be determined for each hand separately. For the NGT dataset it seems that annotation is done in the same way but with start and end of signs determined on the dominant hand and the non-dominant hand gets aligned. But holds of the non-dominant hand and complex constructions are segmented separately for each hand. For the SSL dataset it seems as the transitional movement is not seen as part of the sign. Start and end of a sign are determined on the dominant hand, as with the NGT dataset, but glosses may be directly lined up without a frame in between. (Woll et al., 2004; Crasborn et al., 2004; Bergman and Mesch, 2004)

HSL Corpus No information available.

LIS Corpus Segmentation is done rule-based following Brentari (1998), with parameters as indicators. The more proximal movement is used as a reference for complex signs. (Santoro and Geraci, 2015, slide 13)

PJM Corpus From the picture in Rutkowski et al. (2015, slide 8) (see Figure 3.17) it seems as the transitional movements are not part of the sign but glossed in the same way as indecipherable, invisible, unclear, or doubtful signs with the gloss ### with no frames left between the glosses.

POLYTROPON Segmentation is done rule based. The start and end point of each sign is indicated as accurately as possible. In order to determine the boundaries of a sign Crasborn et al. (2015b) and Crasborn et al. (2020) was followed. For two-handed signs the end point is determined on basis of the dominant hand and the non-dominant is aligned. Transitional movements are not considered as part of the sign. There is at least one video frame between signs due to technical reasons. *(E. Efthimiou, personal communication, January 24, 2022)*

From the available transcripts it seems as the segmentation is done with transitional movements not seen as part of the sign. Start and end of a sign seem to be defined by parameter changes. (ILSP Athena Research Center, 2018)

SIGNOR Corpus 3,000 utterance boundaries have been annotated, but no further information on the rules for segmentation is given. (Jerko and Vintar, 2015, 'Segmentation into utterances')

Signs of Ireland From the picture in Matthews and Sheridan (2015, slide 5) (see Figure 3.19) it seems as transitional movements are seen as part of the sign. While some glosses seem to have one frame in between others are linked together.

SSL Corpus Segmentation is done rule based; transitional movements are not part of the gloss. (Wallin and Mesch, 2018, pp. 2–3)

3.3 TRANSLATION

The translation into a written form of a spoken language is a basic property of each corpus. Several aspects on translation can vary between corpora: (a) the *kind of translation* can vary between narrow/literal and free (while narrow/literal translations stay close to the source language and do not necessarily result in a fluent text, free translations lead to a fluent text adapted to the target language; various degrees in between these two extremes are possible), (b) the *number of languages* into which the signing stream is translated (all corpora presented here provide a translation into the dominant spoken language of the region, some add additional languages like English), (c) the *modalities* of offered translations (most corpora provide a translation into a written form of a spoken language; Corpus NGT also offers a translation into spoken Dutch), and (d) the *length of translation units* can vary from short chunks to sentences of different length and up to whole paragraphs.

Auslan Corpus A free translation into written English is added to the tier **Free Translation** (abbreviated to **FreeTransl**) and time aligned with chunks of signing. The length of these chunks can be determined by meaning (a stretch of signing that aligns with approximately an English sentence) or delivery (based on pauses, head nods, changes in visual-gestural intonation and rhythm). Most English translation units span several Auslan clauses. A literal translation is made based on clauses and added to the tier **LitTransl**. This translation tries to capture the meaning of the signs and can be quite creative. For example, the gloss sequence *COINCIDENCE TORTOISE LOOK*, is translated into 'suddently [sic] tortoise look-hare' as the literal translation, compared to the free translation of 'Suddenly the tortoise turned to look quizzically at the hare.'. As visible in the pictures provided in the conventions the literal translation is often separated into more chunks than the free translation. (Johnston, 2019, pp. 13, 71–74, 85–86)

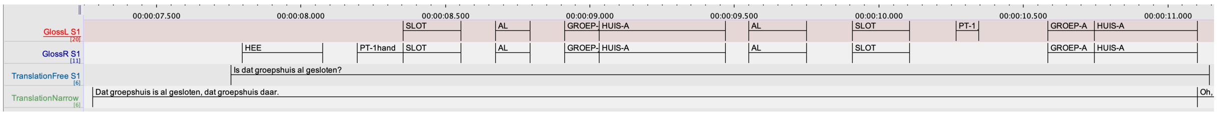
BSL Corpus A free translation into English is given in the tier **Free Translation**. The translation is quite close to BSL and in a plain and simple style (not too formal or informal). A literal translation is planned for the future. The translation is segmented into English sentences where possible and aligned with the signs that are relevant to the translation in question. Periods with no signing and false starts are not taken into account. Beginning and end of the translation segment are roughly time aligned to start and end point of the first and last sign of the translated utterance, respectively. (Cormier et al., 2017, p. 15)

Example: The gloss sequence *PAST THINK FS:OCTOBER(OCT)* is translated into 'Recently, I think it was October.' Every translation unit in the transcript consists of a relatively short English sentence. (Schembri et al., 2017, BF8n.eaf, 00:00:03:288–00:00:05.091)

Corpus LSFB A translation into French is given in the tier **Signernumber – TRADUCTION**. From the data it becomes apparent that the translation units are relatively long, sometimes containing several sentences. (Meurant, 2015; Sinte et al., 2015)

Corpus NGT Different kinds of translations are added to the transcripts: a free translation in form of a smooth running text with relatively long sentences, a narrow translation with short sentences staying close to the source text, a voice-over and a written translation of the voice-over. However, it is noted, that the narrow translation often is more like a free translation and that the voice-over and the transcript of it are only available for a few files. Therefore, the workflow and goals of the translation should be revised. If multiple alternative translations are given for one stretch of signing, they are separated by two slashes. (Crasborn et al., 2020, p. 12)

Example: Figure 3.22 shows a stretch of signing with a free and narrow translation. The signer asks if the clubhouse has been closed yet. The Dutch free translation in the figure could be translated to English as: 'Is that clubhouse closed yet?' while the Dutch narrow translation could be translated to English as: 'That clubhouse is already closed, that clubhouse over there.'. The voice-over (only available as audio) is: 'Hee, het clubhuis is dat al gesloten of nog niet?' which could be translated into English as: 'Hey, the clubhouse is that closed already or not yet?'. In general the narrow translation units are longer than the free translation units, but both contain a maximum of one full stop; several commas can be used in relatively long Dutch sentences though.



GlossL S1 [20] SLOT AL GROEP- HUIS-A AL SLOT PT-1 GROEP-A HUIS-A

GlossR S1 [11] HEE PT-1hand SLOT AL GROEP- HUIS-A AL SLOT GROEP-A HUIS-A

TranslationFree S1 [6] Is dat groepshuis al gesloten?

TranslationNarrow [6] Dat groepshuis is al gesloten, dat groepshuis daar. Oh,

Figure 3.22: *Free and narrow translation in the Corpus NGT. (Crasborn et al., 2008b, CNGT0694.eaf, 00:00:07.280–00:00:11.120, available at `https://hdl.handle.net/1839/00-0000-0000-0009-2D5C-5`)*

Corpus FinSL A free translation into written Finnish is given in the tier **Käännös_1/2**. The translation takes the way of expressing things in the source language, both manually and non-manually, into account. The sign stream is separated into meaningful sentences by the translator's intuition. The manual contains quite elaborate instructions that should help in making the translations consistently. (Salonen et al., 2020, p. 200; Salonen et al., 2019, pp. 35–39)

Example: The gloss sequence *KÄMMEN-YLÖS_ele KÄMMEN-YLÖS_ele TULLA OMA:minun* on the left hand and the simultaneous gloss *_kvkt* are translated into 'rosvo-osoittaa-aseella: "Tule!"': Or in English the gloss sequence could be translated into *PALM-UP_ele PALM-UP_ele BECOME OWN:my* and the translation to 'with a robber-pointing gun: "Come on!"'. In general the translation is segmented into relatively short Finnish sentences. Within the sentences a lot of information is written in between brackets.
(Jantunen, 2018, CFINSL2014_012_03.eaf, 00:03:49.600–00:03:51.200)

Corpus VGT A translation into Dutch is given in the tier **Vertaling i1/2**. (Verstraete et al., 2015, picture)

DTS Corpus No information available.

DGS Corpus In the DGS Corpus a German translation as close to DGS as possible is given in the tier **Deutsche Übersetzung_A/B**. The German translation is translated again into English, to be found in the tier **Englische Übersetzung_A/B**. In the Public DGS Corpus, the English version of the online transcripts shows the English translation, the German version the German translation, respectively. The downloadable files contain both tiers. Individual persons names are represented with variables like '#name1, #name2' for anonymisation reasons.

Translations were first aligned to the corresponding turns of informants and then further split into units, following the rules of the Auslan Corpus that we described previously. These units contained more or less short written German sentences. (Konrad et al., 2020b, p. 3)

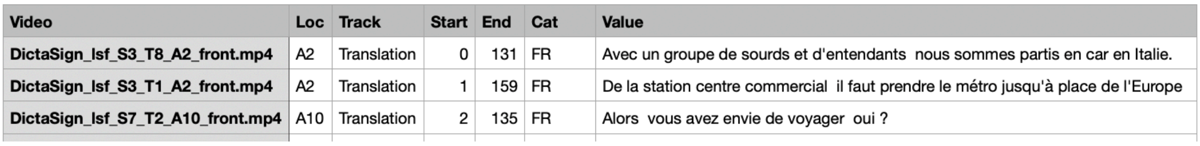
Example: The gloss sequence *$INDEX1* VERHALTEN2 CHARAKTER2A PERSON1* GUT3* ER-SIE-ES2** on the right hand of the signer is translated into German as 'Sein Verhalten und Charakter sind gut.' and into English as 'He has a great attitude and personality.'. In general most translation units contain only one full sentence, but some contain several. (Konrad et al.,

Funded by the Horizon 2020
Framework Programme of the European Union

2020a, Public DGS Corpus, dgskorpus_ber_01: Experience of Deaf Individuals, 00:04:37:07–00:04:39:35, https://doi.org/10.25592/dgs.corpus-3.0-text-1413451-11105600-11163240)

Dicta-Sign-LSF-v2 Corpus A translation into French is given in the tier **Translation**. The translations are listed under the category *FR*. The translation was done by an interpreter, first orally and then transcribed into written French. Therefore the style is close to oral French. (Braffort, 2019, p. 3)

Example: In general the translation units contain between one and several full sentences, as well as parts of sentences separated with several periods, e. g. *Les technologies ... l'avion ça pollue énormément. C'est très ...*, meaning 'Technologies ... the plane pollutes a lot. It's very ...' represents one translation unit. Figure 3.23 shows further examples of the translation into French:



Video	Loc	Track	Start	End	Cat	Value
DictaSign_lsf_S3_T8_A2_front.mp4	A2	Translation	0	131	FR	Avec un groupe de sourds et d'entendants nous sommes partis en car en Italie.
DictaSign_lsf_S3_T1_A2_front.mp4	A2	Translation	1	159	FR	De la station centre commercial il faut prendre le métro jusqu'à place de l'Europe
DictaSign_lsf_S7_T2_A10_front.mp4	A10	Translation	2	135	FR	Alors vous avez envie de voyager oui ?

Figure 3.23: *Example of translations in the Dicta-Sign-LSF-v2 Corpus (LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)*

Digging into Signs The existence of sentence-level translations into English or Dutch on separate tiers is mentioned but not further discussed. (Crasborn et al., 2015a, p. 3)

ECHO Corpus All three datasets are translated into English. Tiers to translate the content in the other spoken languages (Dutch for the SSL dataset, Swedish for the NGT dataset and both for the BSL dataset) are created as child tiers of the English translation tier, but not used so far.

Example: The two glosses *(p-) build dam* and *DAM (-h)* are translated to Dutch 'Een dam werd gebouwd, hoger en hoger' and English 'A dam was built, higher and higher'. (Crasborn et al., 2004, NGT_WE_poems.eaf, 00:00:52.530–00:00:56.090) In general the translation is segmented into parts of one or several English, Dutch or Swedish sentences, but the length varies for different styles of signing, with poems having sometimes only single words in the translation units.

HSL Corpus Translations into written Hungarian were done by Children Of Deaf Adultss (CODAs) or interpreters respected by the members of the deaf community. The translations were written into Word charts and are planned to be checked and transferred to the ELAN files. (Bartha et al., 2016, p. 4)

LIS Corpus No information available.

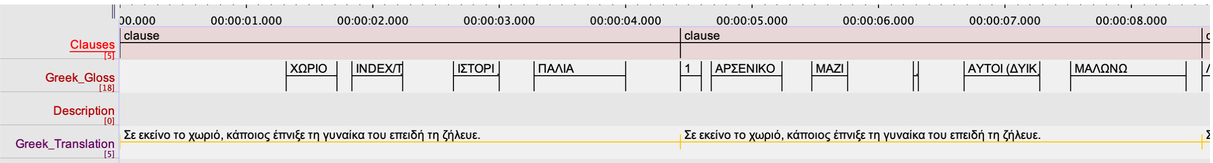
PJM Corpus Pieces of translation are visible on the pictures provided, but no further information is given.

POLYTROPON The annotated sentences were first translated to a strongly sign-influenced initial version by the annotator (a coda GSL expert). A second version with Greek sentences judged to be fully acceptable with regard to naturalness and grammaticality was created by a language expert of Greek. (Efthimiou et al., 2018, p. 41)

Funded by the Horizon 2020
Framework Programme of the European Union

Example: The gloss sequence *ΤΡΩΩ/ΦΑΓΗΟ INDEX/TOPIC ΑΗΔΙΑΖΩ* (English: *I EAT FOOD INDEX / TOPIC LOATHE*) is translated to Greek 'Αυτό το φαγητό με αηδιάζει.' (English: 'This food disgusts me.').
(ILSP Athena Research Center, 2018, Clarin_aidiazw_ex_1_HD_SVQ1)

Form the data it comes apparent, that the translation is always given for the whole transcript. But as the translation glosses are aligned with the clause annotation, the same translation segment gets repeated over several clause units. See Figure 3.24 for an iterated translation tag.



Σε εκείνο το χωριό, κάποιος έπνιξε τη γυναίκα του επειδή τη ζήλευε.

Σε εκείνο το χωριό, κάποιος έπνιξε τη γυναίκα του επειδή τη ζήλευε.

Figure 3.24: *Example of how in POLYPTROPON the same translation segment is repeated multiple times as it gets aligned with the clause units. (ILSP Athena Research Center, 2018, Clarin_pnigw_kapoion_ex_2_HD_SVQ1.eaf, 00:00:00.000–00:00:08.560)*

SIGNOR Corpus No information available.

Signs of Ireland No information available.

SSL Corpus A translation into Swedish is given in the tier **Översättning**. The Swedish text is divided into phrases or sentences suitable for Swedish. (Wallin and Mesch, 2018, p. 28)

Example: The gloss sequence *PRO1 HA TVÅ STYCK HUND* in English *PRO1 HAVE TWO DOGS* is translated into Swedish 'Vi har två hundar,' meaning 'We have two dogs,' and is only a small part of a large sentence split into several phrases. (Mesch et al., 2012, Fritt – husdjur, `https://teckenspraakskorpus.su.se/#/video/sslc01_045.eaf`)

Funded by the Horizon 2020
Framework Programme of the European Union

3.4 OTHER PROPERTIES

This section describes additional aspects of the annotation structure specific to individual corpora.

Auslan Corpus The Transcription Conventions of the Auslan Corpus (Johnston, 2019) are the most comprehensive guidelines in this comparison. They also contain information on the annotation of clause like units, the grammatical classification of signs, the lexical status of signs and more.

BSL Corpus No other properties.

Corpus LSFB No other properties.

Corpus NGT No other properties.

Corpus FinSL No other properties.

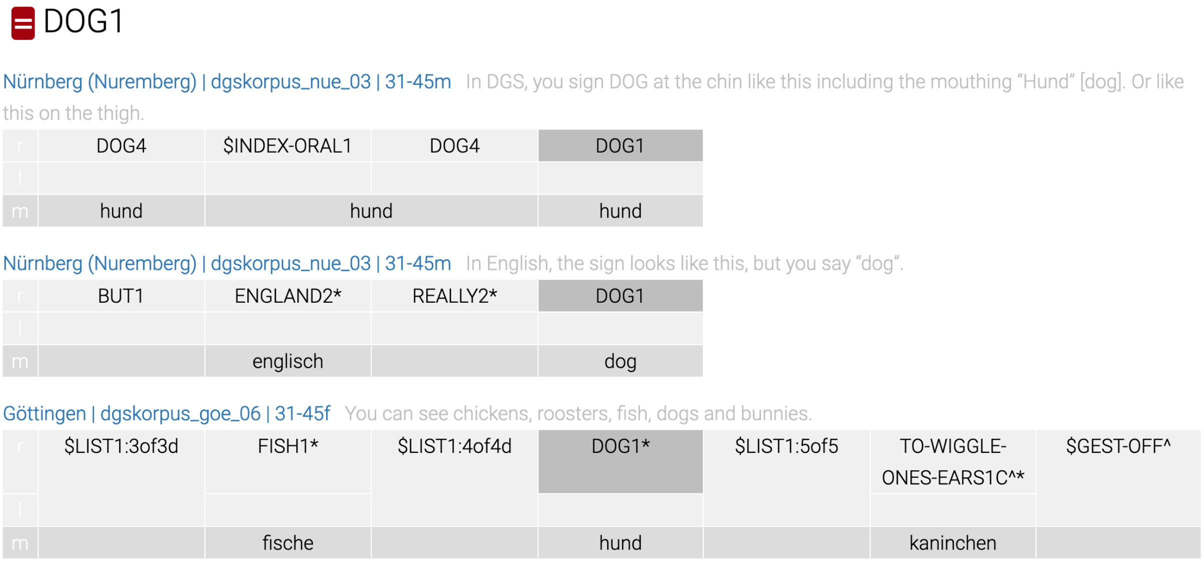
Corpus VGT No other properties.

DTS Corpus No other properties.

DGS Corpus For every type in the Public DGS Corpus a list of tokens including a concordance view is available via the corpus homepage. The type entry also provides a HamNoSys transcription of its citation form and, where available, a studio recording of the citation form and links to other lexical resources with entries for the same type.

In the DGS Corpus a lot of additional information is given in the comment tier, as head shaking, nodding, false-starts, but also uncertainties regarding the annotation. The transcripts contain several other tiers for information on e. g. the meaning of tokens, signs up for anonymisation, key-words and feedback. Still under investigation are reference tracking, argument structure and clause boundaries. (*R. Konrad, personal communication, January 24, 2022*)

Example: Figure 3.25 shows an example for the online concordance view with a maximum of three tokens on the left and the right.



DOG1

Nürnberg (Nuremberg) | dgskorpus_nue_03 | 31-45m In DGS, you sign DOG at the chin like this including the mouthing “Hund” [dog]. Or like this on the thigh.

r	DOG4	$INDEX-ORAL1	DOG4	DOG1
l				
m	hund	hund		hund

Nürnberg (Nuremberg) | dgskorpus_nue_03 | 31-45m In English, the sign looks like this, but you say “dog”.

r	BUT1	ENGLAND2*	REALLY2*	DOG1
l				
m		englisch		dog

Göttingen | dgskorpus_goe_06 | 31-45f You can see chickens, roosters, fish, dogs and bunnies.

r	$LIST1:3of3d	FISH1*	$LIST1:4of4d	DOG1*	$LIST1:5of5	TO-WIGGLE-ONES-EARS1C^*	$GEST-OFF^
l							
m		fische		hund		kaninchen	

Figure 3.25: *List of tokens for DOG1 in the Public DGS Corpus, given as a concordance view. (Konrad et al., 2020a, HUND1ˆ/DOG1ˆ, https://doi.org/10.25592/dgs.corpus-3.0-type-17423)*

Dicta-Sign-LSF-v2 No other properties.

Digging into Signs No other properties.

ECHO Corpus No other properties.

HSL Corpus No other properties.

LIS Corpus No other properties.

PJM Corpus From the picture in Rutkowski et al. (2015, slide 8) it comes apparent that the PJM Corpus does Part of Speech tagging. No further information could be found on the approach.

Polytropon In the Polytropon clause boundaries are defined in the tier **Clauses** and categorised in the tier **Clause_type**. The two main categories are main clauses, which also contain coordinated constructions and subordinate clauses, with clear subordination markers like *BECAUSE*. Both clause types are further categorised in the tier **Sentence_type** into Declarative-Affirmative, Declarative-Negative, Interrogative (Yes/No, Wh), Rhetoric Q&A, Imperative and Exclamation. (Efthimiou et al., 2018, pp. 41–42)

Example: The four signs *ΧΩΡΙΟ INDEX/TOPIC ΙΣΤΟΡΙΑ ΠΑΛΙΑ*, in English *VILLAGE INDEX/TOPIC HISTORY OLD* are annotated as one clause. The translation of the whole sequence is given for each clause: ‘Σε εκείνο το χωριό, κάποιος έπνιξε τη γυναίκα του επειδή τη ζήλευε.’ in English ‘In that village, someone drowned his wife because he was jealous of her.’. The Clause is categorised as *Coordinated clause* and the sentence type *Declerative [sic] affirmative* is annotated. (ILSP Athena Research Center, 2018, Clarin_pnigw_kapoion_ex_2_HD_SVQ1.eaf, 00:00:00.000–00:00:04.440)

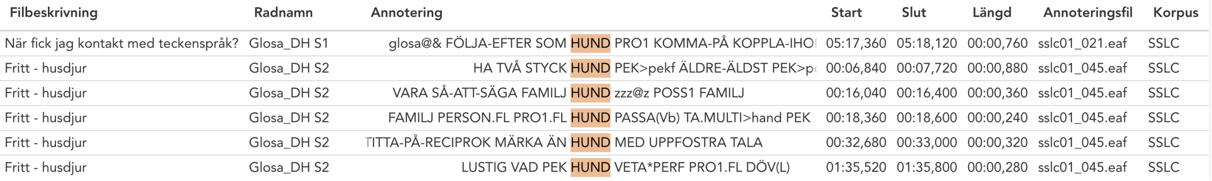
SIGNOR Corpus No other properties.

Signs of Ireland No other properties.

SSL Corpus In the SSLC signs are tagged for Part of Speech (POS). 14 different classes are used for tagging; abbreviations for the classes are suffixed to the gloss in square brackets, e. g. *TECKNA[VB]* for the verb ‘subscribe’, *PÅ[PP]* for the preposition ‘on’ or *MAMMA[NN]* for the substantive ‘mother’. (Wallin and Mesch, 2018, p. 41)

The search function on the SSLC homepage shows results for tokens in a concordance view linked to the transcripts.

Example: Figure 3.26 shows an example for the online concordance view.



Filbeskrivning	Radnamn	Annotering			Start	Slut	Längd	Annoteringsfil	Korpus
När fick jag kontakt med teckenspråk?	Glosa_DH S1	glosa@& FÖLJA-EFTER SOM	HUND	PRO1 KOMMA-PÅ KOPPLA-IHO	05:17,360	05:18,120	00:00,760	sslc01_021.eaf	SSLC
Fritt - husdjur	Glosa_DH S2	HA TVÅ STYCK	HUND	PEK>pekf ÄLDRE-ÄLDST PEK>p	00:06,840	00:07,720	00:00,880	sslc01_045.eaf	SSLC
Fritt - husdjur	Glosa_DH S2	VARA SÅ-ATT-SÄGA FAMILJ	HUND	zzz@z POSS1 FAMILJ	00:16,040	00:16,400	00:00,360	sslc01_045.eaf	SSLC
Fritt - husdjur	Glosa_DH S2	FAMILJ PERSON.FL PRO1.FL	HUND	PASSA(Vb) TA.MULTI>hand PEK	00:18,360	00:18,600	00:00,240	sslc01_045.eaf	SSLC
Fritt - husdjur	Glosa_DH S2	TITTA-PÅ-RECIPROK MÄRKA ÄN	HUND	MED UPPFOSTRA TALA	00:32,680	00:33,000	00:00,320	sslc01_045.eaf	SSLC
Fritt - husdjur	Glosa_DH S2	LUSTIG VAD PEK	HUND	VETA*PERF PRO1.FL DÖV(L)	01:35,520	01:35,800	00:00,280	sslc01_045.eaf	SSLC

Figure 3.26: *Search result for the gloss HUND in the SSL Corpus, given as a concordance view. (Mesch et al., 2012, https://teckenspraakskorpus.su.se/#/?q=hund)*

Funded by the Horizon 2020
Framework Programme of the European Union

4 MANUAL SIGNS

This section addresses the annotation of manual signs. It outlines the annotation practices of different corpora, listing each corpus individually and positioning its approach relative to that of the others. Where possible, the description of annotation conventions is accompanied with an example from the actual data.

In some cases we provide English translations of content to support the reader's understanding of the provided examples. These translations are to be seen only as an aid in the given context of this report and may not be ideal translations of e. g. , gloss names.

4.1 BASIC GLOSSES

In most cases, glosses mainly consist of a spoken language word associated with the sign or a possible meaning of the sign. A gloss name cannot reflect the whole meaning of a sign and should therefore never be seen as a full translation. The chosen keyword is pre- and suffixed with different symbols, names, and labels to further specify the sign. This section collects the basic gloss format for lexical signs within the different corpora. The following sections will explain different special cases.

Auslan Corpus ID-glosses are written in English in uppercase and collected in the *Auslan Signbank*[19]. If more than one distinct word is needed a hyphen is placed between the words, e. g. *GROW-UP*. (Johnston, 2019, p. 17)

If the citation form of a sign or a modification of it is used can be tagged in the tier called **citation modification or variation (abbreviated to ModOrVar)**. (Johnston, 2019, pp. 65–66)

If no gloss is available a temporary gloss is created and appended with the initials of the annotator. The meaning of the sign is annotated in the tier **meaning**. The same tier is used for meanings of signs that are not yet recorded in the lexical database, i. e. newly detected or missing meanings. (Johnston, 2019, p. 19)

BSL Corpus ID-glosses are written in English in uppercase and collected in the *BSL Signbank*[20]. If more than one distinct word is needed a hyphen is placed between the words, e. g. *PULL-APART*. (Cormier et al., 2017, p. 5)

If no gloss is available a new gloss called *ADD-TO-SIGNBANK* is created and the suggested name for the gloss is added in brackets, or in the case that the sign is not known to the annotator *(UNKNOWN)* is added, e. g. *ADD-TO-SIGNBANK(RED3)*. (Cormier et al., 2017, p. 5)

Corpus LSFB ID-glosses are written in French in uppercase and collected in the *Lex-LSFB*[21]. (Sinte et al., 2015, 'Lex-LSFB')

[19] https://www.auslan.org.au/
[20] https://bslsignbank.ucl.ac.uk/
[21] https://www.corpus-lsfb.be/lexique.php

Corpus NGT ID-glosses are written in Dutch in uppercase and collected in the *Global Signbank*[22]. If more than one distinct word is needed a hyphen is placed between the words. Neutral choices are preferred for the Dutch words building the glosses, meaning unmarked form, singular and infinitive are preferred. The exact meaning of a sign in context is specified on an extra tier called **Meaning**. If the meaning of the sign is not clear, it is glossed as *%* (Crasborn et al., 2020, pp. 14, 16, 27)

Corpus FinSL ID-glosses are written in Finnish in uppercase and collected in the *Finnish Signbank*[23]. If it is a FinSL or a FinSSL sign is visible in the signbank, not in the transcripts. The glossary can contain several words, also with different inflections of the same word. If more than one distinct word is needed a hyphen is placed between the words, e. g. *MIES, PALLO, MENNÄ-NUKKUMAAN, OLE-HYVÄ*. Some categories are suffixed to the gloss using an underscore, e. g. *SIX-YEAR_num, PALM-UP_ele*, grammatical features can be suffixed to the gloss or annotated in a separate tier using the symbol @, e. g. *@neg, @toisto, @y*. (Salonen et al., 2019, pp. 8–11; Salonen et al., 2020, p. 199)

Corpus VGT Annotation conventions of the Corpus VGT are based on the Auslan Corpus and the Corpus NGT. ID-Glosses are written in uppercase in Dutch and collected in a lexical database in the form of a Google sheet which is linked to the ELAN files. In comparison to the majority of corpora, the Corpus VGT has spaces inside the glosses. (Verstraete et al., 2015, 'General comparison with annotation guidelines of the BSL and NGT corpora', 'Lexical database with ID-glosses (ECV)')

DTS Corpus ID-glosses are written in uppercase in Danish and collected in the integrated lexical database in iLex. If more than one distinct word is needed, a hyphen is placed between the words. Approximately 2,200 lemmas of the DTS Dictionary[24] were used as a basis for the lexical database, signs from older dictionaries and sign lists were added and new signs were added while annotating the corpus. The sign vocabulary in iLex is structured as a hierarchy on three levels: the sign level, the type level and the meaning level. (Kristoffersen and Troelsgård, 2015, '1. Basic gloss', 'Reuse of dictionary ID-glosses', '2. Two-handed signs'; Troelsgård and Kristoffersen, 2018, pp. 195–196)

DGS Corpus DGS Corpus implements a type hierarchy in the database model of iLex to take the iconicity of signs into account; this approach is called 'double glossing'. Types are linked to each other by a parent-child relation, whereby each child-type, or subtype, stands for a conventionalised form-meaning relationship. The child types often occur with specific mouthings. Citation form and iconic value are inherited from the parent type. For example, the sign sketching a vertical square is used for conventionalised meanings such as 'square', 'page', 'letter', 'recipe' and much more. All glosses are written in uppercase in German, parent types are marked with a caret at the end, e. g. *SQUARE1^*, children types are not marked, e. g. *SQUARE1*, *PIECE-OF-PAPER1*, *LETTER-MAIL2*, *PICTURE2B*. Parent and child types are stored separately in the lexical database integrated in iLex with unique IDs[25]. If more than one distinct word is needed, a hyphen is used to connect the words. (Konrad et al., 2020b, pp. 4–6, 7)

Dicta-Sign-LSF-v2 Corpus Each manual unit is assigned to one of the seven mutually exclusive categories: lexical signs, illustrative structures, holds, pointing signs, numbers, fingerspelling

[22] https://signbank.cls.ru.nl/datasets/NGT

[23] https://signbank.csc.fi/

[24] www.tegnsprog.dk

[25] An exported version of the lexical database is available in the Public DGS Corpus at https://www.sign-lang.uni-hamburg.de/meinedgs/ling/types_en.html

and gesture. The category for lexical signs, called *ID*, uses numbers as sign identifiers, whereas only form is taken into account, not meaning. The identifying numbers are associated with glosses in French.[26] The glosses are written in uppercase with hyphens or underscores in between words if more than one distinct word is needed. To indicate different possible meanings a list of words is given, separated by slashes, but without the goal of giving a complete list, e. g. *COMME/MÊME/AUSSI* for the tree meanings 'like', 'same', 'also'. If the sign is a variant of the citation form *VAR* is suffixed after the gloss following a colon, e. g. *OUI: VAR*. It is noted, that the rules are not applied systematically. (Braffort, 2019, pp. 4–5, 8)

Example: The non-systematic approach becomes apparent when looking at the gloss list: ID *43900* is associated with *PEUT-ETRE2*, ID *43889* is associated with the gloss *POUR_RIEN*, ID *43026* is associated with the gloss *ANNONCER/PUBLIQUE* and ID *43893* is associated with the gloss *CABINE (BATEAU)*. In the case of *PEUT-ETRE2* and *POUR_RIEN* more than one distinct French word is needed, one separated with a hyphen, the other with an underscore. In the case of *ANNONCER/PUBLIQUE* and *CABINE (BATEAU)* the second French word seems to specify the meaning, once separated with a slash, once written in brackets. (LIMSI, 2020, Dicta-Sign-LSF_ID.csv)

In the gloss list the variants are marked, but their citation forms are not listed, e. g. *ENTRER:VAR* is listed, but the citation form of *ENTRER* is missing. (LIMSI, 2020, Dicta-Sign-LSF_ID.csv)

Digging into Signs The *Digging into Sign* guidelines recommend the usage of ID-glosses written in uppercase, collected in a lexical database. (Crasborn et al., 2015a, p. 4)

ECHO Corpus Glosses are written in Swedish, Dutch or English in uppercase. For the SSL and NGT dataset the glosses are translated into English in an extra tier **Gloss RH/LH English**. If more than one distinct word is needed a hyphen is used between words. (Nonhebel et al., 2004a, p. 2)

HSL Corpus For certain linguistic types controlled vocabularies have been created, that describe elements of handedness, movement and non-manual elements. There are no ID-glosses but the actual meaning of the signs is used to annotate the data. ID-glosses are planned for the future. (Bartha et al., 2016, pp. 4–5)

For the dictionary work a simplified template was used and 209 predefined expressions were annotated on six levels (Hungarian translation equivalent, type of the sign (one handed, two-handed, mirror-symmetrical, etc.), handshape of the dominant hand, handshape of the non-dominant hand, region or location of the signing, type of movement). For this annotation instead of a controlled vocabulary a virtual keyboard with pictograms was used. (Bartha et al., 2016, p. 5)

LIS Corpus ID-glosses are written in uppercase in Italian. Infinitival word forms are used, adjectives are always in masculine singular form, nouns always in singular, e. g. *GUIDARE*. If more than one distinct word is needed, a hyphen is used between words. The glosses are not collected in a lexical database. (Santoro and Geraci, 2015, slides 6, 14)

PJM Corpus ID-glosses are written in Polish in uppercase and collected in the integrated lexical database in iLex. Each gloss is suffixed with a number and a handshape marker for the right and left hand, *P* marking the right hand, *L* marking the left hand. If more than one sign uses the same handshape, additional information is given in brackets, e. g. *MAMA 1.1 P:N;L:Ø*

[26] A list of all sign IDs and names can be downloaded as registered user at https://hdl.handle.net/11403/dicta-sign-lsf-v2/v1.

(CHEEK), MAMA 1.2 P:N;L:Ø (CHEEK-CHIN). Signs are considered subtypes of the same parent-type if they differ in not more than one parameter. If more than one distinct word is needed, a hyphen is used to separate the words.

The PJM Corpus uses glosses with more than one word, to mark the ambiguity of signs by listing different meanings separated by a slash, e. g. *JEŚĆ/JEDZENIE/POSIŁEK/JADALNIA/ŚNIADANIE/KOLACJA P:E;L:Ø* or in English *EAT/FOOD/MEAL/DINING/BREAKFAST/DINNER P:E;L:Ø*.(Rutkowski et al., 2015, slides 9, 16, 17; Rutkowski et al., 2013, p. 305)

POLYTROPON Glosses are written in Greek in uppercase and collected in the POLYTROPON lexical database. (Efthimiou et al., 2018, pp. 39–41) Glosses can have more than one Greek word as name. If the words are separated by a slash, the represent different translations, if a second word is given in brackets it represents a precision. As there is no standard format, these conventions may differ across the transcripts. *(E. Efthimiou, personal communication, November 26, 2021)*

Example: Different approaches can be found in the transcripts: *ΘΗΛΥΚΟ, ΑΥΤΟΙ (ΔΥΙΚΟΣ), ΛΟΓΩ-ΑΙΤΙΑ, ΑΥΤΟΣ/ΑΥΤΗ/ΑΥΤΟ* or in English *FEMALE, THEY (TWO), REASON-CAUSE, HE / SHE / THIS*.
(ILSP Athena Research Center, 2018, Clarin_pnigw_kapoion_ex_2_HD_SVQ1.eaf)

SIGNOR Corpus ID-glosses are written in uppercase in Slovenian, e. g. *VRTEC* for 'kindergarden'. If more than one word is needed a space is left in between words, *IME V KRETNJI* for 'name in gesture'. Basis for the glosses were previous sign language dictionaries. (Jerko and Vintar, 2015, picture, 'Uncertainties')

Signs of Ireland The annotation scheme is based on Nonhebel et al. (2004a). Glosses are written in uppercase in English, if more than one distinct word is needed a hyphen is put between words, e. g. *ALL-AROUND*. Some signs have a handshape coding suffixed. There is no use of ID-glosses or a lexical database. (Matthews and Sheridan, 2015, slides 4, 9, 11, 23)

SSL Corpus ID-glosses are written in Swedish in uppercase, the ID numbers align with the SSL Dictionary[27]. The Swedish words used are written in their basic form, some exceptions exist like the Swedish word for 'parents' 'föräldrar', which is always in plural, so the gloss is also in plural *FÖRÄLDRAR*. If more than one distinct word is needed a hyphen is used between words, e. g. *FÅ-SYN-PÅ* for 'get-sight-on' meaning 'notice'. Different signs that use the same gloss are specified with a suffixed handshape coding, e. g. *ENSAM(B), ENSAM(L)* two signs meaning 'alone', one signed with the B-handshape, the other with the L-handshape. (Wallin and Mesch, 2018, pp. 4–7)

[27] https://teckensprakslexikon.su.se/

4.2 VARIANTS

As spoken language words are used for glosses, there are different cases that need extra attention. Different signs that express the same or a similar meaning are called 'lexical variants' or 'synonyms'. In most cases they use the same keyword which is specified with labels in the form of letters, numbers, or handshape codes (see also Chapter 6). The same sign occurring with small differences is often separated into different phonological variants and marked in some way. Some corpora distinguish between lexical and phonological variants by means of differing parameters.

One sign can also have several different meanings. Different corpora have different approaches for homonymous signs, they can be collected under the same gloss or divided onto different glosses, with their homonymic relationship sometimes marked in the lexical databases.

Auslan Corpus If several signs are associated with the same English word, each gloss is specified with a word, a handshape letter code or a number after a period, to give a hint at the form or meaning of the sign, e. g. *FINISH.GOOD* and *FINISH.FIVE* both with handshape codes (see Chapter 6). (Johnston, 2019, p. 18)

BSL Corpus Signs with a similar or related meaning are seen as lexical variants if they differ in at least two parameters. The glosses are separated by numbers, whereas the first one is without a number (*1* does not exist), e. g. *BROWN, BROWN2, BROWN3*. Phonological and morphological variants are marked with small letters *b, c, d*, whereas the first one is without a letter (*a* does not exist), e. g. *BOOKEDb, LOVELY2b*. Both conventions were introduced later into the annotation scheme. (Cormier et al., 2017, p. 6)

Example: From the data it seems that the way of numbering variants has been changed from using one digit to using two digits: *DOG03* (Schembri et al., 2017, BF1n.eaf, 00:00:26.104–00:00:26.365).

Morphological variants could only be found as suggestions for new glosses:
ADD-TO-SIGNBANK(FLOWERc) (Schembri et al., 2017, BF8n.eaf, 00:00:19.548–00:00:20.088).

Corpus LSFB Lexical variants are identified by a memory aid that is added to the gloss following a period, e. g. *APRIL.NOSE, APRIL.WEBSITE*, similar to the form/meaning hints of the Auslan Corpus (see above). In some cases a handshape code is added to the gloss, e. g. *LOOK(U)*. (Sinte et al., 2015, '1. Basic gloss', '4. Lexical variants', 'Lex-LSFB')

Example: There are two ways of identifying lexical variants in the transcripts of Corpus LSFB: *NOM(1), WEBCAM.CAMERA* (Meurant, 2015, CLSFBI0301.eaf, 00:00:30.945–00:00:31.075 and 00:01:38.365–00:01:38.825).

Corpus NGT Lexical variants are glossed with the same gloss with alphabetic markers suffixed, e. g. *SIGN-A, SIGN-B*. Homonyms are linked with each other in the Signbank. (Crasborn et al., 2020, pp. 15–16)

Example: The signer uses two variants for 'nice' glossed as *LEKKER-A* and *LEKKER-B*. (Crasborn et al., 2008b, CNGT0696.eaf, 00:00:31.560–00:00:32.440, 00:00:34.040–00:00:34.320)

Corpus FinSL Lexical variants are defined as signs that differ in at least two to three structural units; they are glossed as separate lexemes with a code suffixed in brackets. Movements are coded with the suffix *L_* followed by a description of the movement, orientation is labelled with a suffixed *O_* followed by a description of the orientation, location is coded with a suffixed *P_* and a description of the location, handshapes are coded according to the codes listed in Chapter 6, e. g. *AUSTRALIA(2o), AUSTRALIA(VI)* coding different handshapes, *JÄRVI(L_ympyrä), JÄRVI(L_loittoneva)* coding different movements, the first a circle movement, the second a fading movement, *EI(O_ylös), EI(O_eteen)* coding different orientations, the first one up, the second forward and *HAUSKA(P_leuka), HAUSKA(P_rinta)* coding different locations, the first at the chin, the second at the chest.

Phonetic variants are defined as signs with the same meaning that differ in one or two structural units; they are combined in one gloss, e. g. instead of using different glosses with a handshape coding suffixed in brackets, like *ISÄ(Ax), ISÄ(AI), ISÄ(G)*, they are all combined in the gloss *ISÄ*. (Salonen et al., 2019, pp. 12–13)

Example: Two lexical variants of the sign for 'same' are used in the transcript: *SAMA(L_alas)* marked for downwards movement and *SAMA(L_lähenevä)* marked for convergent movement. (Jantunen, 2018, CFINSL2014_019_03.eaf, 00:03:28:520–00:03:28.760 and 00:03:39:920–00:03:40.040)

Corpus VGT Lexical variants are suffixed with an information on the form or meaning of the sign, e. g. *DARE-Ahands, LOSE-game*, similar to the Auslan Corpus (see above). The convention on lexical variants has changed, before glosses were suffixed with letters, e. g. *WEG-E, WEG-F*, as in the Corpus NGT (see also above). (Verstraete et al., 2015, 'Lexical variants')

DTS Corpus Signs with the same meaning but two or more differing parameters are seen as lexical variants and suffixed with a tilde symbol and a number, e. g. *SIGN˜1, SIGN˜2*. Signs with the same meaning but only one parameter different are seen as phonological variants and suffixed with a tilde symbol and small case letters, e. g. *SIGN˜2˜a, SIGN˜2˜b*. (Kristoffersen and Troelsgård, 2015, '4. Lexical variants')

DGS Corpus Lexical variants are marked with numbers suffixed to the gloss, e. g. *WOMAN5, WOMAN8*. In a few cases, numbers are also used to differentiate between different signs that use homonymous German words for the gloss, e. g. *ZU7, ZU9* where the first means 'closed' and the second 'towards'. In contrast to the BSL Corpus, the first variant is suffixed with the number 1.

Phonological variants are glossed with letters following the lexical variant numbers, e. g. *WOMAN2A, WOMAN2B, WOMAN2C, WOMAN2D*. Tokens that differ from the citation form are marked with an asterisk, e. g. *PLANE1**. (Konrad et al., 2020b, pp. 7–8)

Example: The signer uses two different lexical variants of the sign for 'life': *LEBEN1C, LEBEN4* (Konrad et al., 2020a, DGS Corpus, dgskorpus_ber_01: Experience of Deaf Individuals, 11:16:52:13–11:16:52:21 and 11:11:05:46–11:11:06:11, `https://doi.org/10.25592/dgs.corpus-3.0-text-1413451-11105600-11163240`)

Dicta-Sign-LSF-v2 Corpus Lexical variants are separated by numbers following the gloss, e. g. *NON1, NON2*, same approach as in the DGS Corpus. (Braffort, 2019, p. 8)

Example: For the concept 'page' three glosses are listed: *PAGE1, PAGE2, PAGE3*. (LIMSI, 2020, Dicta-Sign-LSF_ID.csv)

Digging into Signs The conventions of BSL Corpus and Corpus NGT are described without further discussion. (Crasborn et al., 2015a, p. 5)

ECHO Corpus No information available.

HSL Corpus No information available.

LIS Corpus Lexical and phonological variants are not specially coded. (Santoro and Geraci, 2015, slide 14)

PJM Corpus Lexical variants are marked with consecutive numbers after the gloss but before the articulation info, e. g. *AUTOBUS 1 P:O;L:O*. Phonological variants are defined as signs that differ in no more than one parameter and are distinguished on the subtype level with additional running numbers as markers, e. g. the type *TATA 1 P:N;L:Ø/P:Z;L:Ø* has several subtypes as *TATA 1.1 P:N;L:Ø (CZOŁO-BRODA), TATA 1.2 P:N;L:Ø (CZOŁO-BRODA), TATA 1.4 P:N;L:Ø (POLICZEK)*. (Rutkowski et al., 2015, slides 9, 12, 20, 21)

POLYTROPON Phonological variants are indicated by a suffixed number *-2, -3* on the gloss of the most frequent sign. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus All signs with their variants are numbered, including the main lexical form, e. g. *POTEM1, POTEM5* both for 'then'. (Jerko and Vintar, 2015, 'Variants')

Signs of Ireland Lexical variants are suffixed with a handshape code or with *-m* for male signs and *-f* for female signs. (Matthews and Sheridan, 2015, slide 13)

SSL Corpus Lexical variants and different signs that use the same gloss (homonymous Swedish words) are specified with a suffixed handshape coding, e. g. *ENSAM(B), ENSAM(L)* two signs meaning 'alone', one signed with the B-handshape, the other with the L-handshape. If the same handshape is used an articulation coding can be used: *(ea)* for a one-handed sign, *(da)* for signs with both hands active and *(ml)* for signs with one hand as basis, e. g. two signs meaning 'still' both articulated with a B-handshape but different movement are glossed as *FORTFARANDE(da), FORTFARANDE(ml)*. (Wallin and Mesch, 2018, pp. 6–7)

Example: Two signers use different lexical variants for the sign for 'deaf': Signer 1 uses *DÖV(J)*, Signer 2 uses *DÖV(L)* and once *DÖV(Jvt)*. (Mesch et al., 2012, När lär man sig teckna, `https://teckenspraakskorpus.su.se/#/video/sslc01_121.eaf?q=döv`)

4.3 TWO-HANDED SIGNS

When using two hands for signing, they can be coordinated in several ways: (a) *symmetrical signs* are two-handed signs where the handshape, location and movement of both hands are mirrored, (b) *asymmetrical signs* are two-handed signs where the handshape, location or movement of the hands differ, and (c) a *complex construction* in which the hands can be used simultaneously to perform independent signs, one with each hand.

Additional special cases are when a one-handed sign is performed with two hands or vice versa. When a sign is two-handed in its citation form but signed with only one hand, it is called a 'weak hand drop'. When a sign with a one-handed citation form is signed with both hands, it is called a 'weak hand prop'. Some corpora mark these cases directly in the transcripts, others postpone it to later analysis.

In general the approach to glossing both hands differs between corpora: Some use one tier for both hands with individual spots for each hand, others use individual tiers for each hand. Some align start and end of a sign for both hands, other set them individually. The naming scheme of the hands also differs. In some corpora they are called right and left hand, while in other corpora they are labelled as dominant and non-dominant hand.

Auslan Corpus One tier per hand, labelled as left and right hand tiers: **RH/LH-IDgloss**. In the metadata and the file name one can find the hand dominance of the signer. One handed signs are glossed in the respective tier, two-handed signs are glossed in both tiers. If the hands sign different signs at the same time, the according glosses are made in the respective tiers. Deviations from the citation form are suffixed in the gloss with *-2H* for weak prop and *-1H* for weak drop, e. g. *LOOK-2H* and *OWL-1H*. (Johnston, 2019, pp. 20–21)

Shadowing, anticipation and perseveration are not annotated if they are not meaningful and seen as minor activity. (Johnston, 2019, pp. 47–48)

BSL Corpus One tier per hand, labelled as left and right hand: **RH/LH-IDgloss**. Two-handed signs are annotated on both hand tiers, with start and end point determined by the dominant hand. If different signs are signed parallel, they are annotated in the appropriate tier. In the first release only the dominant hand of two-handed signs was annotated, this approach can still be found in the data as not all transcripts are changed to the new convention yet. (Cormier et al., 2017, pp. 4–5)

Example: Figure 4.1 shows the two-handed sign 'hoover' followed by two signs produced separately but in parallel.

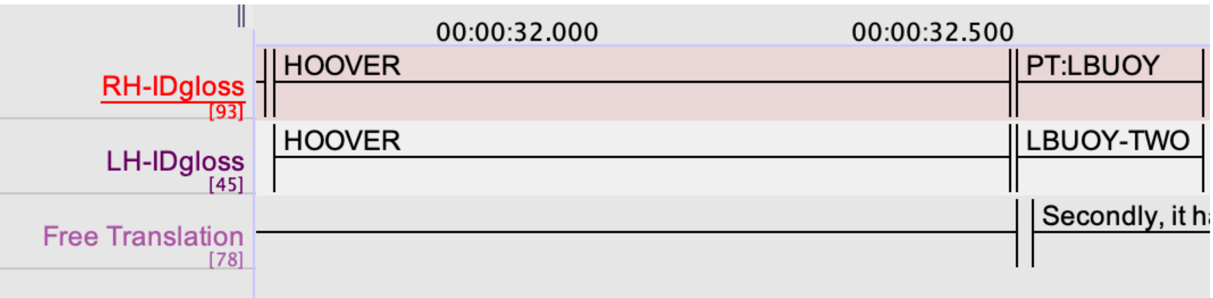


Figure 4.1: *Annotation of two-handed and simultaneous signs in the BSL Corpus. (Schembri et al., 2017, BF1n.eaf, 00:00:31.708–00:00:32.828)*

Corpus LSFB Same conventions as BSL Corpus. The tiers are labelled as **S001 - MD** for the right hand (main droite) and **S001 - MG** for the left hand (main gauche), where the beginning is the code for each signer. (Sinte et al., 2015, '2. Two-handed signs')

Example: It seems from the data, that beginning and end of two-handed signs is determined by the dominant hand, while simultaneous distinct signs are segmented independently. An example of this can be seen in Figure 4.2.

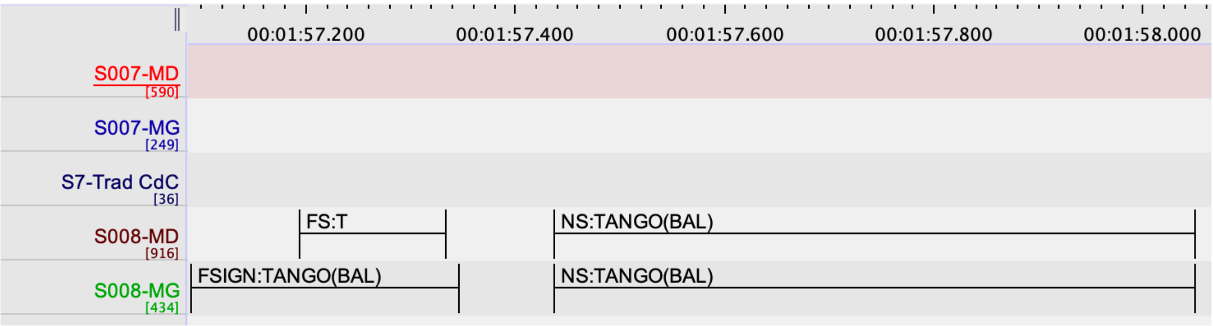


Figure 4.2: *A pair of simultaneous signs and a two-handed sign in the Corpus LSFB. (Meurant, 2015, CLSFBI0301.eaf, 00:01:57.090–00:01:58.053)*

Corpus NGT Same conventions as BSL Corpus. The tiers are labelled as **GlossL S1/2** for the left hand and **GlossR S1/2** for the right hand, with two sets – one for each signer. (Crasborn et al., 2020, pp. 14–15)

Weak drop is not specially marked, but the sign is glossed only at the according hand tier. (Crasborn et al., 2020, p. 15)

Hand dominance is annotated in great detail on five different tiers with a range of special codes. This procedure is explained in more detail in Crasborn and Sáfár (2016). (Crasborn et al., 2020, pp. 33–37)

Example: Figure 4.3 shows a two-handed sign with the same start and end time for both hands, followed by two different signs signed simultaneously with different start and end times followed by a one-handed sign. In the tier **DomRev Point S1** a dominance reversal from first right to left (gloss *RL*) and then from left to right (*LR*) is annotated.

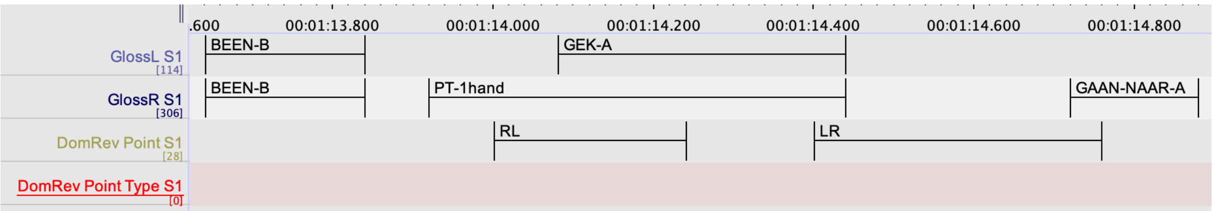


Figure 4.3: *Annotation of two-handed, simultaneous, and one-handed signs with a dominance reversal in Corpus NGT. (Crasborn et al., 2008b, CNGT0004.eaf, 00:01:13.640–00:01:14.880, `https://hdl.handle.net/1839/00-0000-0000-0009-2D5C-5`)*

Corpus FinSL Both hands are annotated separately. For independently produced signs, as well as for two-handed signs, the duration of the annotation may be different for each hand. (Salonen et al., 2019, p. 7)

Example: In the example in Figure 4.4 the two-handed sign for 'happen' has been formed on the right hand earlier than on the left hand.

Funded by the Horizon 2020 Framework Programme of the European Union

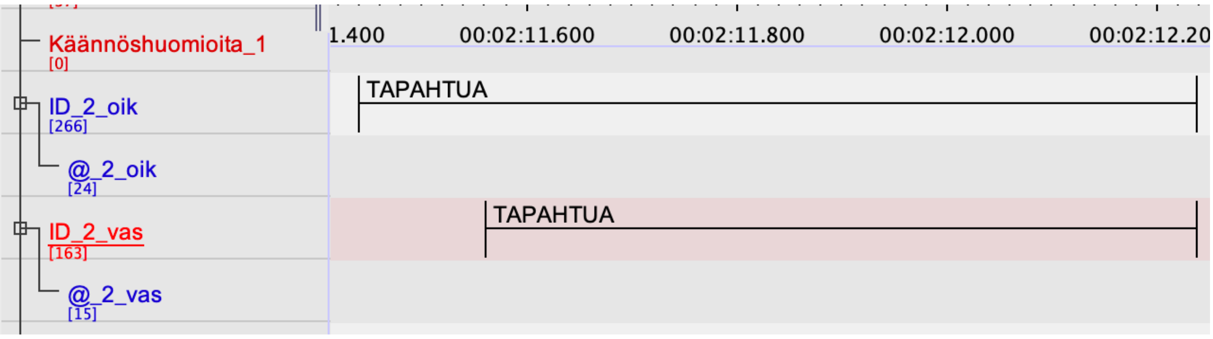


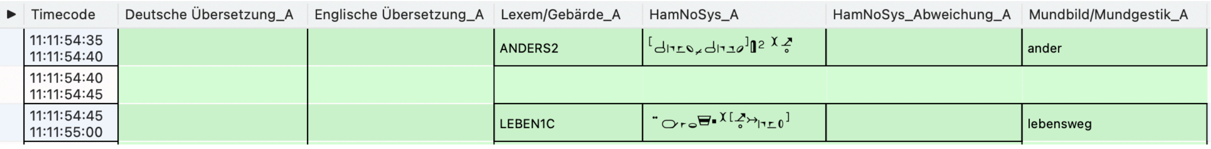
Figure 4.4: *Annotation of a two-handed sign in the Corpus FinSL. Start and end times are determined independently for each hand. (Jantunen, 2018, CFINSL2014_019_03.eaf, 00:02:11.440–00:02:12.240)*

Corpus VGT Each hand is annotated in separate tiers called **GlosRH/LH i1/2**. Start and end of two-handed signs are determined independently for each hand, same as in the BSL Corpus. Signs with one active and one passive hand are marked with the suffix *AC* in brackets in the active hand tier, e. g. *WEGLOPEN-A (AC)*. (Verstraete et al., 2015, picture, 'Two-handed signs')

DTS Corpus Start and end of the sign is determined independently for each hand. Weak drop and weak prop are not tagged separately but can be detected by comparing the annotation with the information on the sign in the sign base. DTS Corpus is considering working with the double token-tag approach from the DGS Corpus. (Kristoffersen and Troelsgård, 2015, '2. Two-handed signs')

DGS Corpus The DGS Corpus uses double token-tags. As already mentioned in Section 3.1 this means that the tier for right and left hand, called **Lexem/Gebärde_A/B** (English: **Lexeme/Sign_A/B**), contains a slot for the right hand and a slot for the left hand. Two-handed signs can be annotated in the slot for the right or the left hand, depending on which hand is active in one-handed and asymmetric signs. For symmetric signs the hand that starts or moves higher is chosen. If there is no difference between the hands the right hand slot is used as default. Start and end of the sign are determined according to the active hand. (Konrad et al., 2020b, pp. 4, 6)

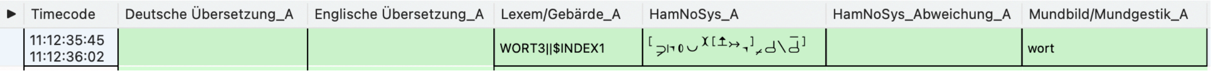
Example: Figure 4.5 shows two different two-handed signs, the first asymmetric and the second symmetric. To distinguish the two kinds of two-handed sign, their HamNoSys notation has to be checked. Figure 4.6 provides an example of a complex construction in which a sign for 'word' and a pointing gesture are performed at the same time.



Timecode	Deutsche Übersetzung_A	Englische Übersetzung_A	Lexem/Gebärde_A	HamNoSys_A	HamNoSys_Abweichung_A	Mundbild/Mundgestik_A
11:11:54:35 11:11:54:40			ANDERS2	[illegible]		ander
11:11:54:40 11:11:54:45						
11:11:54:45 11:11:55:00			LEBEN1C	[illegible]		lebensweg

Figure 4.5: *Two-handed signs in the DGS Corpus. The first sign is asymmetric and the second symmetric, which is indicated by their HamNoSys notation. (Konrad et al., 2020a, DGS Corpus, dgskorpus_ber_01: Experience of Deaf Individuals, 11:11:54:35–11:11:55:00, https://doi.org/10.25592/dgs.corpus-3.0-text-1413451-11105600-11163240)*

Funded by the Horizon 2020
Framework Programme of the European Union



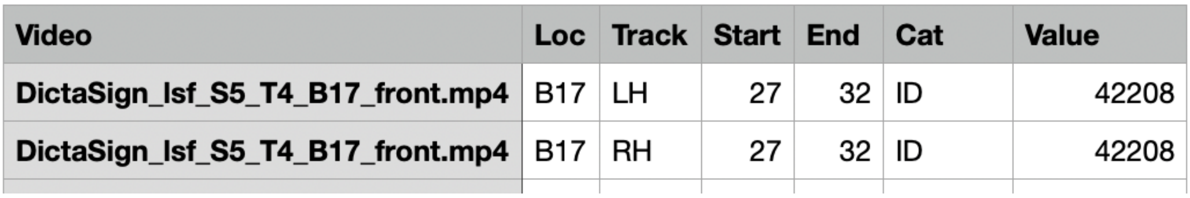
▶	Timecode	Deutsche Übersetzung_A	Englische Übersetzung_A	Lexem/Gebärde_A	HamNoSys_A	HamNoSys_Abweichung_A	Mundbild/Mundgestik_A
	11:12:35:45 11:12:36:02			WORT3||$INDEX1	[illegible]		wort

Figure 4.6: *A complex construction with a sign for 'word' on the right hand and a pointing sign on the left hand. (Konrad et al., 2020a, DGS Corpus, dgskorpus_ber_01: Experience of Deaf Individuals, 11:12:35:45–11:12:36:02, `https://doi.org/10.25592/dgs.corpus-3.0-text-1413451-11105600-11163240`)*

Dicta-Sign-LSF-v2 Corpus Two-handed signs are glossed on a separate tier **2H**. Weak prop is glossed in the same tier; weak drop is glossed in the according tier for the hand signing **LH** or **RH**. (Braffort, 2019, p. 3)

Example: The sign *41759* with the gloss *DEUX HEURES* is annotated three times in the tier **RH** and one time in the tier **2H**, which indicates that the latter could be a weak prop. (LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)

In the data we also find occurrences where the same sign is annotated on the left and right hand with the same start and end time. An example of this can be seen in Figure 4.7. These might be annotations of a weak hand prop in deviation from the annotation conventions, or the two hands may represent separate entities at the same time. Given the information available to us, we were unable to clarify this question further.



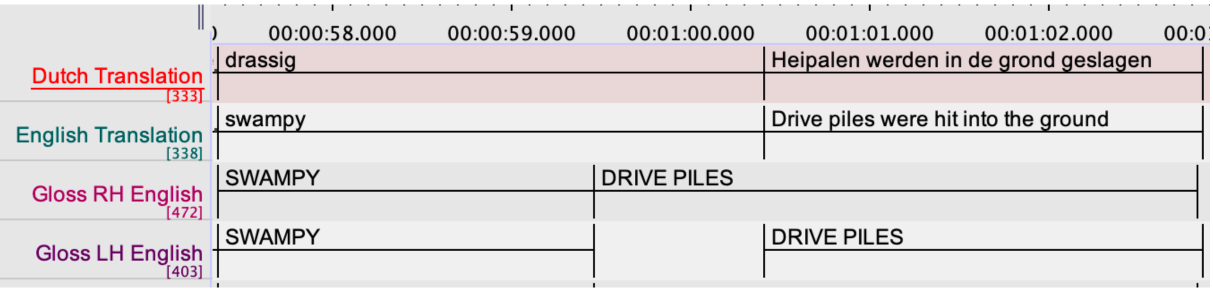
Video	Loc	Track	Start	End	Cat	Value
DictaSign_lsf_S5_T4_B17_front.mp4	B17	LH	27	32	ID	42208
DictaSign_lsf_S5_T4_B17_front.mp4	B17	RH	27	32	ID	42208

Figure 4.7: *An annotation in the Dicta-Sign-LSF-v2 Corpus that is either a non-standard gloss for a two-handed sign or glosses for a weak prop sign. (LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)*

Digging into Signs Recommends the Corpus NGT conventions. (Crasborn et al., 2015a, p. 4)

ECHO Corpus Each hand is annotated in a separate tier, the right hand in **Gloss RH**, the left hand in **Gloss LH**. Two-handed signs are annotated on both tiers, same as in the BSL Corpus. Weak prop is glossed with the prefix *2h* in brackets, weak drop is prefixed with *1h* in brackets, similar to the Auslan Corpus. (Nonhebel et al., 2004a, p. 2)

Example: In the transcript shown in Figure 4.8 one can see that for the two-handed sign *SWAMPY* the start and end points are the same on both tiers, while for the two-handed *DRIVE PILES* the right-hand gloss begins considerably earlier than the left-hand gloss and also stops slightly sooner.



00:00:58.000 00:00:59.000 00:01:00.000 00:01:01.000 00:01:02.000 00:0

Dutch Translation [333]: drassig | Heipalen werden in de grond geslagen

English Translation [338]: swampy | Drive piles were hit into the ground

Gloss RH English [472]: SWAMPY | DRIVE PILES

Gloss LH English [403]: SWAMPY | DRIVE PILES

Figure 4.8: *Annotation of two two-handed signs in the ECHO Corpus. (Crasborn et al., 2004, NGT_WE_poems.eaf, 00:00:57.360–00:00:59.460)*

HSL Corpus Each hand is annotated in a separate tier, labelled as right and left hand based on a data-driven approach. For the focused grammatical research the labels active and passive hand were used, these are planned to be transferred to the right/left hand system in the future. As already mentioned in Section 4.1 is the handedness annotated in more detail for the dictionary research. (Bartha et al., 2016, pp. 4–5)

LIS Corpus Each hand is annotated in a separate tier. (Santoro and Geraci, 2015, slide 10)

PJM Corpus Two-handed signs are only glossed on the dominant hand tier. Dominant hand change is not marked explicitly; if a gloss appears only in the tier for the non-dominant hand and the dominant hand stays empty it is assumed that a change in the dominant hand happened. If different signs are signed on both hands they are glossed according to the signing hand. The start and end time are annotated for each hand separately. (Rutkowski et al., 2015, slides 15, 17, 18)

POLYTROPON Citation forms of two-handed signs are labelled the same as one-handed signs, but handedness can be determined by inspecting the HamNoSys transcription of the sign, which can be found in the lexical database. Variation, like weak prop, is annotated by a suffixed asterisk. As there is only one occurrence of a weak drop in the POLYTROPON there is no label for it. *(E. Efthimiou, personal communication, January 24, 2022)*

Example: In the data we can see that one tier is used for all signs, no matter if one- or two-handed: The gloss sequence *ΣΚΟΝΗ*2 ΝΕΡΟ ΤΙΠΟΤΑ / ΚΑΘΟΛΟΥ*, in English *DUST*2 WATER NOTHING / NOT-AT-ALL*[28] represents a two-handed sign that exists also in a one-handed version, followed by a one-handed sign and a two-handed sign. (ILSP Athena Research Center, 2018, Clarin_stegnos_ex_1_HD_SVQ1.eaf, 00:00:06.558–00:00:08.364)

SIGNOR Corpus There is no distinction made between the two hands. As there is no further information it remains unclear whether the double token-tag approach of the DGS Corpus is adopted or not. (Jerko and Vintar, 2015, ‘Two-handed signs’)

Signs of Ireland Two-handed signs are prefixed with *2h-*, one handed signs are not labelled. (Matthews and Sheridan, 2015, slide 12)

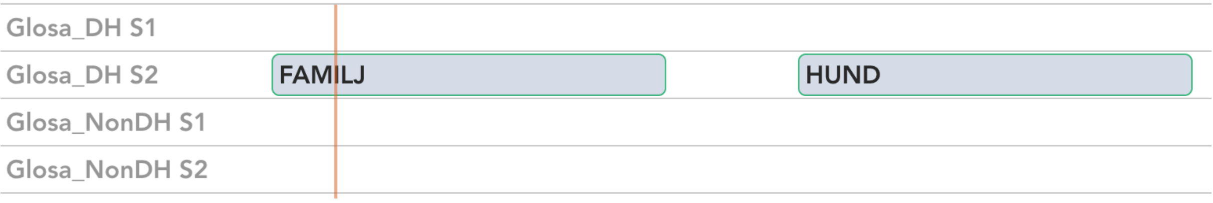
SSL Corpus Two-handed signs are glossed only in the tier for the dominant hand. The dominant hand is identified by looking at the performance of one-handed signs, as they normally are signed with the dominant hand of the signer. (Wallin and Mesch, 2018, pp. 3–5)

[28] Note that the English translation is provided by us as an aid to the reader. Hyphens in the English gloss indicate a Greek single-word gloss that required multiple words to be expressed in English.

Funded by the Horizon 2020 Framework Programme of the European Union

In the tiers **Artikulator_DH** and **Artikulator_NonDH** signs are further categorised into one-handed signs glossed with *ea*, signs with one hand as basis are glossed as *ea_ml* and signs with both hands active as *da*. (Wallin and Mesch, 2018, p. 21)

Example: Figure 4.9 shows a two-handed sign for 'family' followed by a one-handed sign for 'dog', both annotated in the tier for the dominant hand.



Glosa_DH S1		
Glosa_DH S2	FAMILJ	HUND
Glosa_NonDH S1		
Glosa_NonDH S2		

Figure 4.9: *A two-handed sign followed by a one-handed sign in the SSL Corpus. (Mesch et al., 2012, `https://teckensprakskorpus.su.se/#/video/sslc01_045.eaf`)*

Funded by the Horizon 2020
Framework Programme of the European Union

4.4 BUOYS

Buoys (also known as 'holds') are instances where a handshape, usually on the non-dominant hand, is held in the signing space used as a point of reference, while the other hand continues signing.

Depending on the method of annotating two-handed signs (see Section 4.3), buoys are identified in the gloss itself or via another tier with further descriptions. A commonly used categorisation for buoys used in annotation schemes is based on the description from Liddell (2003): list buoys, fragment buoys, theme buoys, and pointer buoys.

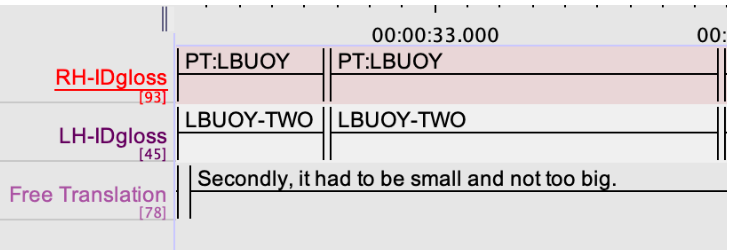
With list buoys it is common that the dominant hand points towards the non-dominant hand. Conventions on how to annotate this pointing are listed below if defined.

Apart from actual buoys, one can also encounter instances where the articulation of a sign is continued on one of the hands, not conveying any further meaning or being a point of reference, but just a very slow relaxation or return to the rest position. Whether these perseverations are glossed or not varies between corpora.

Auslan Corpus Gloss begins with a label identifying the type of buoy (according to Liddell (2003)), followed by a label of the handshape being used in brackets (cf. Chapter 6) and, after a colon, a short description of what the buoy stands for, e. g. *LBUOY(1):FIRST*. (Johnston, 2019, pp. 39–41)

BSL Corpus Buoys are categorised into the four classes mentioned above (list, pointer, fragment and theme) and glossed as: *LBUOY, PBUOY, FBUOY, TBUOY*. List buoys can incorporate number signs; if the do so the conventions for number incorporation are used, namely the number sign is suffixed to the buoy gloss, e. g. *LBUOY-TWO* (see also Section 4.8). (Cormier et al., 2017, pp. 7–8; Sinte et al., 2015, 'Comparison', *K. Cormier, personal communication, January 31, 2022*)

Example: Figure 4.10 shows a case where the list buoy incorporates a number sign to show a list of two; this is signed on the left hand. At the same time the right hand is pointing towards the buoy, annotated according to the conventions for points (see Section 4.7). The annotation leaves open to which finger of the left hand buoy the right hand is pointing. In the video recording one can see that the first point is to the index finger, while the second point it towards the middle finger:



00:00:33.000 00:

RH-IDgloss [93]: PT:LBUOY | PT:LBUOY

LH-IDgloss [45]: LBUOY-TWO | LBUOY-TWO

Free Translation [78]: Secondly, it had to be small and not too big.

Figure 4.10: *Buyos in the BSL Corpus. The annotations are more detailed than specified in the annotation conventions. (Schembri et al., 2017, BF1n.eaf, 00:000:32.605–00:00:33.437)*

Corpus LSFB Same conventions as BSL Corpus with a small deviation regarding fragment buoys: perseverations are only glossed as fragment buoys when the other hand points at the hold hand

Funded by the Horizon 2020 Framework Programme of the European Union

or when the hand is kept in the signing space during two or more signs signed by the other hand. (Sinte et al., 2015, ‘Comparison’, ‘Fragment Buoys’; Gabarro-Lopez and Meurant, 2014, sec. 4.1.1)

Example: In the data it comes apparent that conventions may have changed. Apart from fragment and list buoys, we also find unspecified glosses *BUOY*, points towards glosses *PT:BUOY* and more specified glosses as *PT:BUOX-DEUX*. Figure 4.11 shows an example of such an unspecified gloss. In this case the gloss specifies that it is a buoy, but not which type. Tier **S007-MD** is the right hand of the signer coded as S007, **S007-MG** is the left hand of the same signer.

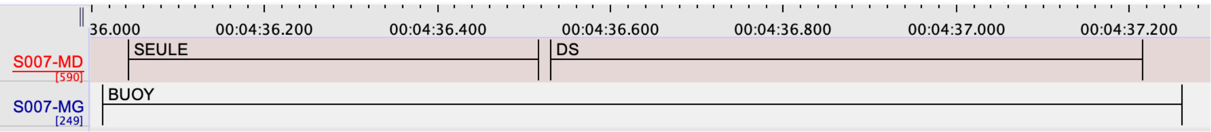


Figure 4.11: *A buoy in Corpus LSFB. This gloss does not match the naming scheme from the annotation conventions. (Meurant, 2015, CLSFBI0301.eaf, 00:04:36.013–00:04:37.266)*

Corpus NGT The duration of the perseveration is marked on a separate tier called **MoveHold**. List buoys are glossed as *TELHAND* and suffixed with information on the extended fingers, e. g. *TELHAND+1-A*. If the other hand points towards the counting hand, this is glossed as a point and gives information on the finger(s) that are pointed at. If the point is not to a specific finger, but in a sweeping motion across all fingers two glosses can be used (for both hands): *ENZOVOORTS-A*, *ENZOVOORTS-C*. (Crasborn et al., 2020, p. 18)

Example: Figure 4.12 shows a list buoy on the left hand of the signer glossed as *TELHAND-3-A* and a pointing sign towards the index finger (‘wijsvinger’ in Dutch) of the list buoy followed by the sign for ‘eating’ on the right hand of the signer.

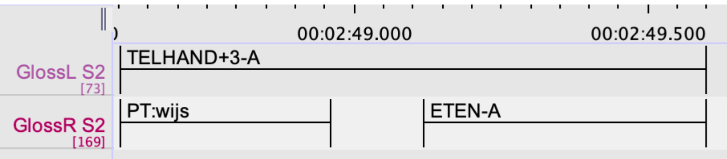
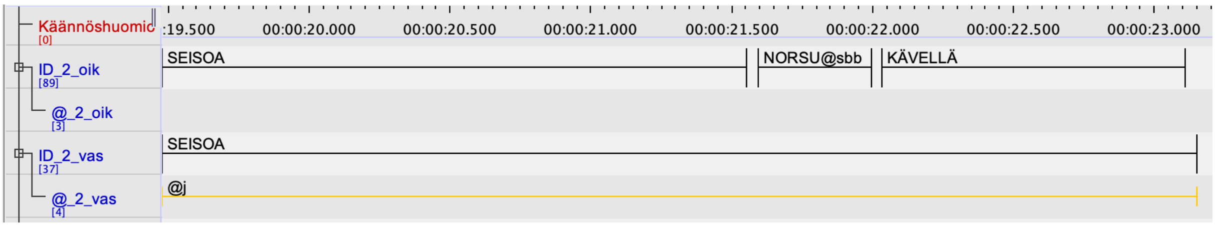


Figure 4.12: *A list buoy TELHAND-3-A in Corpus NGT. (Crasborn et al., 2008b, CNGT0004.eaf, 00:02:48.600–00:02:49.600, `https://hdl.handle.net/1839/00-0000-0000-0009-2D5C-5`)*

Corpus FinSL The gloss is extended for the whole duration of the perseveration of the hand and the gloss *@j* is added on the daughter tier **@_1/2_vas/oik**. List buoys are glossed as *@poijul*. (Salonen et al., 2019, pp. 28–29, 33–34; Salonen et al., 2020, p. 199)

Example: Figure 4.13 shows a perseveration in the left hand. Tier **ID_2_oik** is the right hand, **ID_2_vas** is the left hand and *@j* a in **@_2_vas** marks the perseveration.

Funded by the Horizon 2020 Framework Programme of the European Union



Käännöshuomio [0]
:19.500 00:00:20.000 00:00:20.500 00:00:21.000 00:00:21.500 00:00:22.000 00:00:22.500 00:00:23.000

ID_2_oik [89] — SEISOA | NORSU@sbb | KÄVELLÄ

@_2_oik [3]

ID_2_vas [37] — SEISOA

@_2_vas [4] — @j

Figure 4.13: *A perseveration in Corpus FinSL, marked by @j. (Jantunen, 2018, CFINSL2014_020_03.eaf, 00:00:19.480–00:00:23.160)*

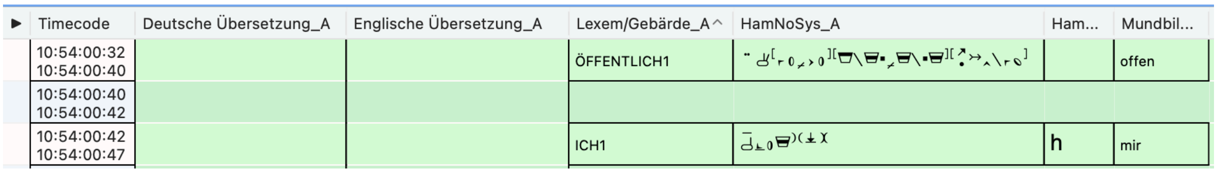
Corpus VGT Only list buoys are glossed as such, other buoys are not specially glossed. The glosses for the non-dominant hand are extended for the whole duration of the perseveration of the hand. It is not mentioned how glosses for list buoys look. (Verstraete et al., 2015, 'Buoys')

DTS Corpus Glosses for the non-dominant hand are extended for the whole duration of the perseveration of the hand. (Kristoffersen and Troelsgård, 2015, 'Buoys')

DGS Corpus Buoys in the Public DGS Corpus are glossed with a suffixed asterisk on the following sign, to mark that this sign deviates from it's citation form, as the other hand is also involved. In the DGS Corpus instead of an asterisk an *h* for 'hold' is added to the succeeding sign in the tier **HamNoSys_Abbreviations_A/B**. In both cases the hold is a perseveration of the preceding sign. This perseverations are not further specified, with the exception of list buoys. This approach was chosen because it proved to be less time consuming compared to more detailed annotations of perserverations.

Similar to the Corpus NGT, only list buoys are glossed separately as *$LIST1* (for signs where the index finger pointing out the number of the list) and *$LIST2* (for signs where the flat hand points out the number of the list). Several further glosses describe the list buoy in more detail: *$LIST-TO-LIST*, *$LIST-TO-REMOVE*, *$LIST-TOGETHER*, *$LIST1:2of4*, etc. All list buoys are suffixed with two numbers, the first representing the list item the second one the total number of shown list items, e. g. *$LIST1:2of4*. This approach accounts for both, the Public DGS Corpus and the DGS Corpus. (Konrad et al., 2020b, pp. 4, 15; *R. Konrad, personal communication, January 24, 2022*)

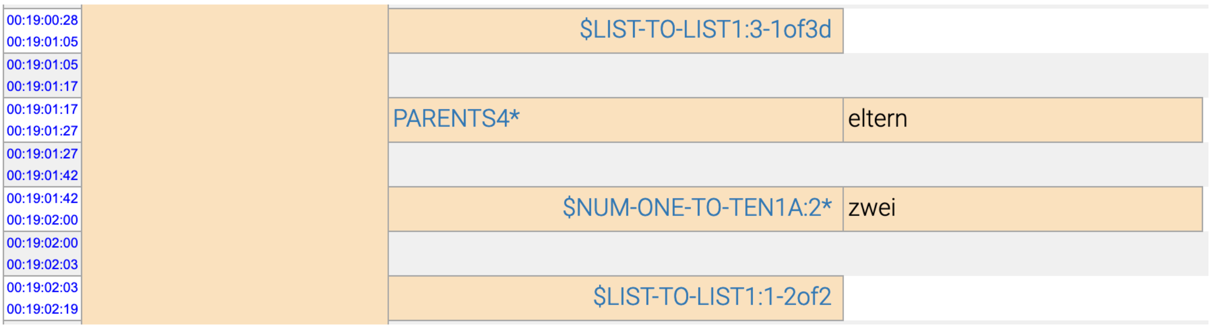
Example: Figure 4.14 shows how in DGS Corpus holds are annotated in the tier **HamNoSys_Abweichung_A** with the code *h*. In the example the left hand holds the sign *ÖFFENTLICH1*, in English 'public' in the signing space, while the right hand signs *ICH1*, in English 'I'. Figure 4.15 shows an annotation in the Public DGS Corpus in which the signer first shows a list of three, followed by a list of two, both with a sweeping motion across all listing points on the left hand.



▶	Timecode	Deutsche Übersetzung_A	Englische Übersetzung_A	Lexem/Gebärde_A ^	HamNoSys_A	Ham...	Mundbil...
	10:54:00:32 10:54:00:40			ÖFFENTLICH1	[illegible]		offen
	10:54:00:40 10:54:00:42						
	10:54:00:42 10:54:00:47			ICH1	[illegible]	h	mir

Figure 4.14: *Sign for 'I' with a hold on the left hand as annotated in the DGS Corpus. (Konrad et al., 2020a, DGS Corpus, dgskorpus_lei_12: Experience of Deaf Individuals, 10:54:00:42–10:54:00:47, https://doi.org/10.25592/dgs.corpus-3.0-text-1584617)*

Funded by the Horizon 2020 Framework Programme of the European Union

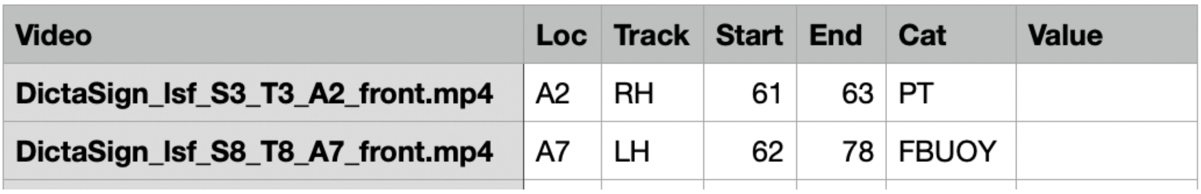


00:19:00:28 00:19:01:05	$LIST-TO-LIST1:3-1of3d	
00:19:01:05 00:19:01:17		
00:19:01:17 00:19:01:27	PARENTS4*	eltern
00:19:01:27 00:19:01:42		
00:19:01:42 00:19:02:00	$NUM-ONE-TO-TEN1A:2*	zwei
00:19:02:00 00:19:02:03		
00:19:02:03 00:19:02:19	$LIST-TO-LIST1:1-2of2	

Figure 4.15: *Two different list buoys in the Public DGS Corpus (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_lei_12: Experience of Deaf Individuals, 00:19:00:28–00:19:02:19, https://doi.org/10.25592/dgs.corpus-3.0-text-1584617)*

Dicta-Sign-LSF-v2 Corpus All holds are annotated with the category *FBUOY*. This category has no value (ID-numbers), so no further distinction between different holds is made. The gloss begins right after the preceding two-handed sign. (Braffort, 2019, p. 4)

Example: In Figure 4.16 a pointing sign on the right hand, glossed *PT*, is annotated in the right hand tier **RH** with start time 61 and end time 63, while the hold on the left hand is annotated as *FBUOY* in the tier **LH** with start time 62 and end time 78.



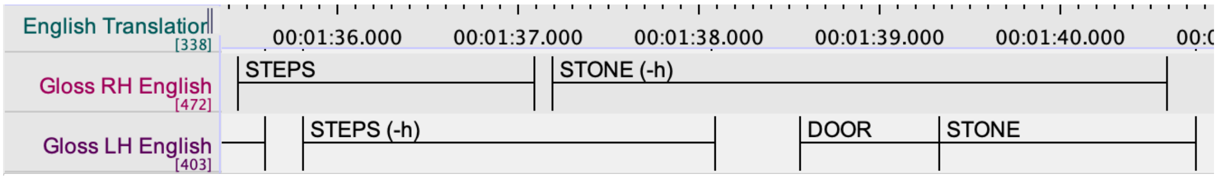
Video	Loc	Track	Start	End	Cat	Value
DictaSign_lsf_S3_T3_A2_front.mp4	A2	RH	61	63	PT	
DictaSign_lsf_S8_T8_A7_front.mp4	A7	LH	62	78	FBUOY	

Figure 4.16: *Overlapping pointing sign and buoy in the Dicta-Sign-LSF-v2 Corpus. (LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)*

Digging into Signs Corpus NGT Convention was chosen as standard. It is noted that both projects will move towards doing the segmentation based on the true start and end of the signs, with information on holds on separate tiers. (Crasborn et al., 2015a, pp. 3–5)

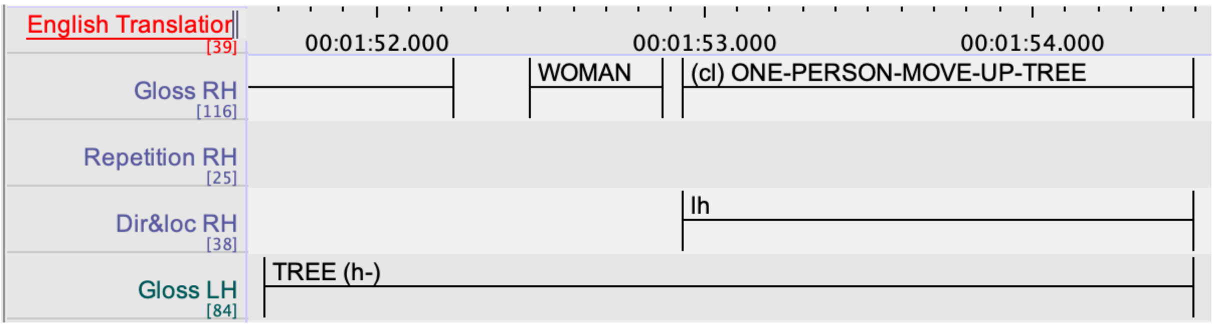
ECHO Corpus All buoys are marked with *-h* in brackets, suffixing the gloss of the perseverated sign, e. g. *RUN (-h)*. (Nonhebel et al., 2004a, p. 3)

Example: Figure 4.17 shows an example of two buoys in the NGT dataset of the ECHO Corpus and Figure 4.18 shows a buoy in its BSL dataset.



English Translation [338]	00:01:36.000	00:01:37.000	00:01:38.000	00:01:39.000	00:01:40.000	00:0
Gloss RH English [472]	STEPS	STONE (-h)				
Gloss LH English [403]	STEPS (-h)	DOOR	STONE			

Figure 4.17: *Two buoys in the NGT dataset of the ECHO Corpus. (Crasborn et al., 2004, NGT_WE_poems.eaf, 00:01:35.440–00:01:40.760)*



English Translation [39]

00:01:52.000 00:01:53.000 00:01:54.000

Gloss RH [116]: WOMAN | (cl) ONE-PERSON-MOVE-UP-TREE

Repetition RH [25]

Dir&loc RH [38]: lh

Gloss LH [84]: TREE (h-)

Figure 4.18: *A buoy in the BSL dataset of the ECHO Corpus.* (Woll et al., 2004, *BSL_PS_poem3.eaf, 00:01:51.650–00:01:54.500)*

HSL Corpus No information available.

LIS Corpus One gloss for all types of buoys: *IX-LOC*; additional information is entered on a separate tier. (Santoro and Geraci, 2015, slide 15)

PJM Corpus Glosses for the non-dominant hand are extended for the whole duration of the perseveration of the hand. List buoys are glossed separately with the base number of the list followed by *-PUNKT* and information on the handshapes, e. g. *PIĄTY-PUNKT 1 P:Z;L:5*, 'PIĄTY' meaning 'fifth'. (Rutkowski et al., 2015, slide 19)

Polytropon Within the POLYTROPON no rules on buoys are needed, as they don't appear in the data. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus Classifiers with descriptions are used. No distinction is made between motion and form classifiers. As there is no further information, the exact procedure remains unknown. (Jerko and Vintar, 2015, 'Buoys')

Signs of Ireland One gloss for all types of buoys: *BUOY*. If the dominant hand points to the non-dominant hand the gloss for pointing is suffixed with a colon and *BUOY*, e. g. *PT:BUOY* (Matthews and Sheridan, 2015, slides 12, 17)

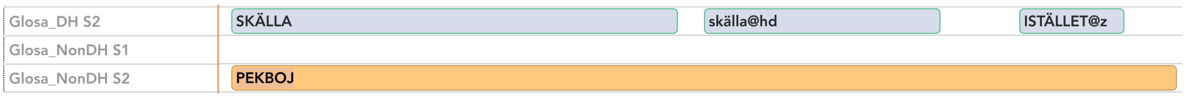
SSL Corpus SSLC combines different theoretical approaches on buoys, categorising them into four types: *LISTBOJ* (list buoy), *PEKBOJ* (pointer buoy), *TEMABOJ* (theme buoy) and *PUNKTBOJ (no translation given)*. (Wallin and Mesch, 2018, pp. 23–26)

List buoys are suffixed with the base numbers one to five, describing the used handshape: *LISTBOJ.EN, LISTBOJ.TVÅ, LISTBOJ.TRE, LISTBOJ.FYRA, LISTBOJ.FEM*. To gloss to which finger of the list buoy the other hand is pointing, the gloss on the pointing hand is suffixed by a greater-than sign and the name of the finger, e. g. *GLOSS>pekf*/långf/ringf/lillf/tumf. The shortforms for the individual fingers are *pekf* (index finger), *långf* (middle finger), *ringf* (ring finger), *lillf* (pinky) and *tumf* (thumb). If the dominant hand points to more than one finger, the finger names are listed with a hyphen in between; if they are pointed at one after another, an underscore is used.

PUNKTBOJ is suffixed with (L) for when the index finger is outstretched and (J) for when a flat hand is used. (Wallin and Mesch, 2018, pp. 23–26)

In the tier **Glosa_DH_extra** (see Section 4.19) holds are annotated with the suffix *@hd*. (Wallin and Mesch, 2018, p. 31)

Example: Figure 4.19 shows a pointer buoy on the non-dominant hand while three different signs are glossed on the dominant hand.



Glosa_DH S2	SKÄLLA	skälla@hd	ISTÄLLET@z
Glosa_NonDH S1			
Glosa_NonDH S2	PEKBOJ		

Figure 4.19: *A pointer buoy on the non-dominant hand, held over a longer stretch of signing, in the SSL Corpus. (Mesch et al., 2012, https://teckensprakskorpus.su.se/#/video/sslc01_045.eaf?q=*aboj&t=101.560)*

4.5 NAME SIGNS

Name signs are signs denoting persons or other entities in a (more or less) unique way. Name signs can be categorised by the entities they are naming, e. g. persons, organisations, geographical locations, etc. Some name signs are based on finger spellings, others use lexical signs, still others use unique signs. The different corpora encode more or less of this information within the glosses.

Auslan Corpus The proper name is prefixed with *NS:* and if needed suffixed with a hint regarding the form or a identical lexical sign in brackets, e. g. *NS:MISSKENTWORTH(HAIR-BUN)*, *NS:PETER(P-shake)* (Johnston, 2019, p. 24)

BSL Corpus The proper name (or *UNKNOWN* if the name is not known to the annotator) is prefixed with *SN:* and, if the name is identical to a lexical sign suffixed with this sign in brackets. If the name is fingerspelled, the fingerspelling is added (according to the conventions for fingerspelling) in brackets. If fingerspelling and lexical signs are combined, they are both written in brackets separated by a caret, e. g. *SN:MISS-JENKINS(HAIR-BUN), SN:JOHN-KING(FS:J-JOHN^KING), SN:PETER(FS:P-PETER)*, the latter being an initialised name sign. It is noted, that the conventions on first and last name are not coherent and will be made more consistent in the future. (Cormier et al., 2017, p. 7)

Example: Name sign of two combined fingerspellings: *SN:MARY-HARE(FS:M-MARY^FS:H-HARE)* (Schembri et al., 2017, BF2n.eaf, 00:03:12.740–00:03:13.576)

Corpus LSFB Name signs are prefixed with *NS:* followed by the proper name and a hint regarding the form in brackets e. g. *NS:LAURENCE(HAIRCLIP)*. This approach is very similar to the Auslan Corpus, behalf of not making a connection to identical lexical signs. (Sinte et al., 2015, '15. Sign names')

Example: The name 'Simone' extended with the hint 'baptism': *NS:SIMONE(BAPTEME)*. (Meurant, 2015, CLSFBI0103.eaf, 00:03:53.927–00:03:54.316)

Corpus NGT Name signs are not labelled but glossed with first and last name of the person separated by a hyphen, e. g. *FIRST_NAME-LAST_NAME*. If the name sign is identical to a lexical sign this is marked in the Signbank, not in the transcript. If the name of the person is unknown to the annotator the gloss *NAAMGEBAAR* (in English 'NAME-SIGN') is used. (Crasborn et al., 2020, p. 18)

For the second release of the Corpus NGT an anonymisation protocol was developed to ensure the limitation of personal information in the annotations. All glosses for name signs referring to participants or other non-public persons have been replaced by the gloss **NAAMGEBAAR* in the tier for the hands and with **eigennaam* in the tiers for mouth and translations. Glosses in the tier for meaning revealing the name have been removed. (Crasborn and Bank, 2015)

Example: The name sign for a person is glossed as *[NAAM]* in the tiers **GlossL S1, GlossR S2** and as *[naam]* in the tiers **MeaningL S2, Meaning R S2**. The glosses deviate from the above mentioned anonymisation protocol, which implies that some changes has been made for newer releases. (Crasborn et al., 2008b, CNGT0013.eaf, 00:02:19.680–00:02:20.040, `https://hdl.handle.net/1839/00-0000-0000-0009-2D71-F`)

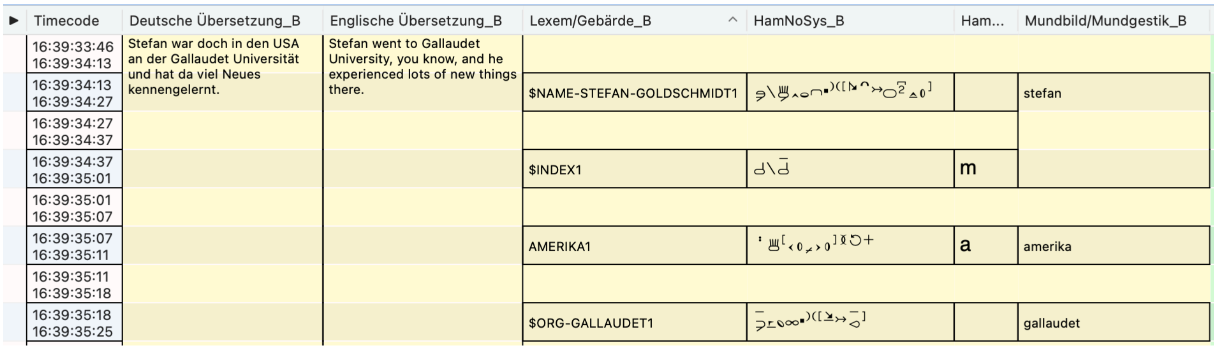
Corpus FinSL The only information found on name signs is that nicknames are written only in the form of a nickname for security reasons. (Salonen et al., 2019, p. 13)

Corpus VGT Name signs are prefixed with *NG*, which stands for 'naamgebaar', followed by the proper name (first and last name if both are known to the annotator), e. g. *NG: SVEN-VERSTRAETE*. (Verstraete et al., 2015, 'Name signs')

DTS Corpus Name signs are glossed with separate ID-glosses without special labels and no information on identical lexical signs, e. g. *WILLIAM-STOKOE, THE-PARLIAMENT*, similar to the Corpus NGT. For unknown or partially unknown name signs there are no rules yet. (Kristoffersen and Troelsgård, 2015, '15. Sign names')

DGS Corpus All name signs for private individuals are glossed collectively with the gloss *$NAME* as the names of private individuals are anonymised in the corpus. Names of persons who are public figures in general or in the deaf community are not anonymised. In this case their full name is suffixed to the gloss, e. g. *$NAME-ANGELA-MERKEL1*. Names are categorised as either person names, glossed as *$NAME*, or as names for organisations, glossed as *$ORG*.

Example: In Figure 4.20 the signer talks about Stefan Goldschmidt, a well known member of the German deaf community, who attended Gallaudet University. His name sign is glossed as *$NAME-STEFAN-GOLDSCHMIDT1* and the sign for Gallaudet University as *$ORG-GALLAUDET1*.



Timecode	Deutsche Übersetzung_B	Englische Übersetzung_B	Lexem/Gebärde_B	HamNoSys_B	Ham...	Mundbild/Mundgestik_B
16:39:33:46 16:39:34:13	Stefan war doch in den USA an der Gallaudet Universität und hat da viel Neues kennengelernt.	Stefan went to Gallaudet University, you know, and he experienced lots of new things there.				
16:39:34:13 16:39:34:27			$NAME-STEFAN-GOLDSCHMIDT1	[illegible]		stefan
16:39:34:27 16:39:34:37						
16:39:34:37 16:39:35:01			$INDEX1	[illegible]	m	
16:39:35:01 16:39:35:07						
16:39:35:07 16:39:35:11			AMERIKA1	[illegible]	a	amerika
16:39:35:11 16:39:35:18						
16:39:35:18 16:39:35:25			$ORG-GALLAUDET1	[illegible]		gallaudet

Figure 4.20: *Annotation of name signs for persons and organisations in the DGS Corpus. (Konrad et al., 2020a, DGS Corpus, dgskorpus_hb_04: Deaf Events, 16:39:33:46–16:39:35:25,* `https://doi.org/10.25592/dgs.corpus-3.0-text-1181027`*)*

Dicta-Sign-LSF-v2 Corpus No information is given in the manual.

Example: In the data we find proper names labelled with *NS*: *NAPOLEON:NS, PORTUGAL1:NS, AIRBUS:NS*. (LIMSI, 2020, Dicta-Sign-LSF_ID.csv)

Digging into Signs Both conventions from BSL Corpus and Corpus NGT are described without further discussion. (Crasborn et al., 2015a, p. 6)

ECHO Corpus No information available.

HSL Corpus No information available.

LIS Corpus All name signs are glossed as *SEGNO-NOME*. (Santoro and Geraci, 2015, slides 18, 24)

PJM Corpus Name signs are glossed as name and surname in brackets and with *ID:* as prefix, e. g. *ID: (LECH WAŁESA)*. (Rutkowski et al., 2015, slide 31)

Funded by the Horizon 2020 Framework Programme of the European Union

POLYTROPON Name signs are assigned a respective gloss when available, e. g. big cities like Athens. Whenever no name sign is available, concept representations are made via fingerspelling. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus One ID-gloss is used for all name signs. (Jerko and Vintar, 2015, 'Sign names of persons')

Signs of Ireland Name signs are prefixed with *SN:* followed by the proper name and a description or an information that the name is initialised, e. g. *SN: WENDY(W), SN: SARAH(describe), SN:PAT MATTHEWS(PM)* (note that it is not clear if the irregular use of spaces between the prefix and the name are due to layout issues or for other reasons). This approach is similar to Auslan Corpus. (Matthews and Sheridan, 2015, slide 16)

SSL Corpus Name signs and other names as organisation names or proper names are suffixed with *@en*. If fingerspelling is used for names *@b* is used as described in 4.6, e. g. *USA@b@en*. If the name consists of more than one sign the parts are separated with a hash symbol, e. g. *LARS-WALLIN@en, KURT@b#WALLANDER@b@en*. (Wallin and Mesch, 2018, pp. 8, 16)

Example: The Swedish town 'Örebro' is glossed as *ÖREBRO@en*. (Mesch et al., 2012, Sportaktiviteter, `https://teckenspraakskorpus.su.se/#/video/sslc01_005.eaf?q=%2a%40en&t=193.120`)

4.6 FINGERSPELLING

The manual alphabet is a way of representing a written alphabet with different handshapes. Some sign languages (SLs) use a one-handed manual alphabet, others, like BSL, use a two-handed one. Some signs combine a handshape from the manual alphabet, usually the first letter of the corresponding written word, with a lexical sign, these combinations are called 'initialisations'.

In the annotation conventions there are differences with regard to the segmentation of fingerspellings – is the whole spelled word glossed as one unit or as separate letters – and the gloss names – how are the letters written or how are they separated and if the target word is spelled indifferently of the received spelling or if the received spelling is annotated containing transpositions, omissions and mistakes.

Auslan Corpus The target word is prefixed with *FS:* and, if noteworthy, suffixed with the actual spelled letters in brackets. Each spelled word is glossed separately, even if it is spelled in a continuous movement, e. g. *FS:WORD(WRD), FS:MISS FS:KENTWORTH*. Initialised signs, that are not lexicalised (yet) are glossed with the single or double letter and the word for which it stands, e. g. *FS:M-MONTH, FS:M-MILE, FS:GG-GARAGE*. (Johnston, 2019, pp. 45–46)

BSL Corpus Same procedure as Auslan Corpus, but the actual spelled letters are always added to the gloss, e. g. *FS:WORD(WRD), FS:F-FORTUNE*, the latter is an example for an initialised sign: the fingerspelled letter is followed by a hyphen and the word it represents. If the target word is not clear *INDECIPHERABLE* is used, e. g. *FS:INDECIPHERABLE(GTH)*. (Cormier et al., 2017, pp. 13–14)

Example: In this example the target word is spelled fully, so no further information is needed: *FS:PUPPY*. (Schembri et al., 2017, BF1n.eaf, 00:00:02.002–00:00:03.065)

Corpus LSFB The gloss is prefixed with *FS:* followed by the letters actually spelled and the target word in brackets e. g. *FS:WRD(WORD)*. This approach is similar to the BSL Corpus but information on the target word and actually spelled-out letters are switched. (Sinte et al., 2015, '17. Fingerspelling')

Example: The word discord ('descornet') partly spelled: *FS:DESCORNTET(DESCORNET)* and the word comic ('bande dessinee') spelled with only the first letters 'B' and 'D': *FS:BD(BANDE-DESSINEE)* (Meurant, 2015, CLSFBI0301.eaf, 00:00:56.957–00:01:00.384 and 00:08:47.904–00:08:48.061)

Corpus NGT Fingerspelling is coded with a hash in front of the gloss, followed by the actually spelled letters. Each spelled word is glossed separately, e. g. *#JOHAN #ROS*. (Crasborn et al., 2020, p. 19)

Example: The fingerspelling for 'auditief' is glossed as *#AUDYT* in the tier **GlossR S1** with the actual meaning annotated in the tier **MeaningR S1**. (Crasborn et al., 2008b, CNGT0128.eaf, 00:00:12.560–00:00:13.480, `https://hdl.handle.net/1839/00-0000-0000-0009-2D6F-3`)

Corpus FinSL The actually spelled letters are written in small case separated by a hyphen followed by the suffix *_sa*, e. g. *t-o-m-i_sa*. Interrupted fingerspelling is marked with *@kes* in the daughter tier. Information on the meaning of the fingerspelling can be added at the translation tier. If the fingerspelling is a lexicalised sign or signs are initialised, this is annotated in the daughter tier with the code *@sv*. (Salonen et al., 2019, pp. 21, 32; Salonen et al., 2020, p. 199)

Example: No example for fingerspelling has been found in the data. The lexicalised sign for 'okay' *OKEI* appears with the code *@sv* in the associated daughter tier. (Jantunen, 2018, CFINSL2014_019_03.eaf, 00:01:35.160–00:01:35.520)

Corpus VGT The spelled word or perceived letter are prefixed with *VS*, which stands for 'vingerspelling', e. g. *VS: BIJ*. (Verstraete et al., 2015, 'Fingerspelling')

DTS Corpus The target word spelled is written in conventional spelling suffixed with *(H)*, e. g. *Maria(H)*. For incorrectly or partially spelled words there are no rules yet. (Kristoffersen and Troelsgård, 2015, '17. Finger-spelling')

Initialised signs are prefixed by *INITIALISED-* followed by the signed letter, e. g. *INITIALISED-C, INITIALISED-F*. As there is one ID-gloss for every letter in the Danish alphabet there are 29 glosses for initialised signs. A tier to annotate the meaning in context of the initialised sign is in consideration. (Kristoffersen and Troelsgård, 2015, '17.c Initialised signs')

DGS Corpus Fingerspellings are annotated as one sequence and glossed with *$ALPHA*. Three different kinds of fingerspelling are distinguished: one-handed manual alphabet glossed as *$ALPHA1*, two-handed manual alphabet glossed as *$ALPHA2* and a sketching of the form of the letter in the air with the index finger are glossed as *$ALPHA-SK*. Finger alphabets from other sign languages have their own glosses: *$ALPHA-BSL^* for BSL and *$ALPHA-NZSL^* for New Zealand Sign Language (NZSL).

Fingerspelled letters are annotated with uppercase letters, separated by hyphens. The DGS finger alphabet uses single letters where German uses several letters, e. g. the German unit of 'SCH' is fingerspelled with the 5-hand moving fast for- and backwards from the wrist, double letters as 'FF' are spelled with one F-Handshape moving sideways, these units are not separated. The German word for 'ship' therefore would be annotate as *$ALPHA1:SCH-I-FF*. If letters are skipped in the fingerspelling they are not annotated, e. g. the name 'Alfredo' spelled without the 'r' is glossed as *$ALPHA1:A-L-F-E-D-O*. Variations in handshapes are coded with number suffixed to the letters with an underscore in between, e. g. *$ALPHA1:T_2*. Fingerspelling that is not meant to spell a word but shows the act of fingerspelling is glossed as *$ALPHA1:#*.

The DGS Corpus defines initialised signs that use single letters from the finger alphabet and combine them with a simple movement, as sidewards movements or circles. These signs are spontaneously used and not conventionalised. Initialised signs are prefixed with *$INIT* followed by four possible movement types: *-STRAIGHT1^*, *-HAND-WRIST1^*, *-CIRCLE1^*, *-CIRCLE2^*.

Combinations of conventional signs with handshapes from the finger alphabet (the above mentioned understanding of initialised signs) is glossed with the lexicalised sign and the qualifier 'quantity', e. g. *GRUPEE3A-$SAM'1:T* for the sign meaning 'group' signed with a T-handshape.

Single letters or combinations of letters that are lexicalised have their own subtype but are listed under the parent type *$ALPHA^*. (Konrad et al., 2020b, pp. 12–13; *T. Hanke, personal communication, January 11, 2022*; *R. Konrad, personal communication, January 27, 2022*)

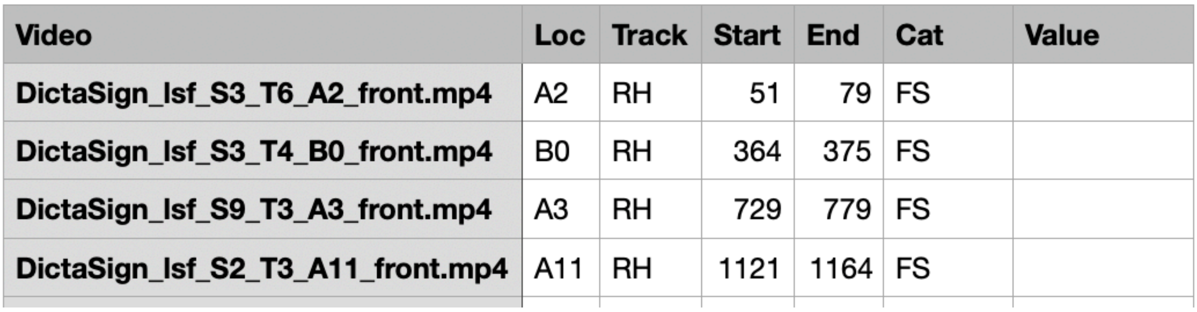
Example: The signing of the dish 'Leipziger Allerlei' is glossed as *LEIPZIG1A* $ALPHA1:A-L-E-R-L-I*. The signer used a deviation form of the citation form of *LEIPZIG* followed by a fingerspelling missing two letters. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_lei_11: Regional Specialities, 00:15:08:34–00:15:08:43, https://doi.org/10.25592/dgs.corpus-3.0-text-1584545)

From the data it becomes apparent that the hash symbol is also used for indistinguishable finger spelling: The signer first draws an 'S' into the signing space, then fingerspells some letters not legible followed by an 'A' some more illegible letters and a 'T', all together glossed as *$ALPHA-SK:S $ALPHA1:#-A-#-T*. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_mue_05: Experience Report, 00:00:26:05–00:00:26:43, https://doi.org/10.25592/dgs.corpus-3.0-text-1210208)

An occurrence where the signer initialises the place name 'Leipzig' in a straight movement is glossed as *$INIT-STRAIGHT1^**. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_lei_11: Regional Specialities, 00:12:20:17–00:12:20:23, https://doi.org/10.25592/dgs.corpus-3.0-text-1584545)

Dicta-Sign-LSF-v2 Corpus Fingerspelling is annotated with the category *FS*. In the manual it is noted that this category has no value (ID-numbers), but also that 'The value here corresponds to the letters produced.'. Therefore, it stays unclear if and where the values can be found. (Braffort, 2019, p. 5)

Example: As shown in Figure 4.21, all fingerspellings are categorised as *FS* with no further value.



Video	Loc	Track	Start	End	Cat	Value
DictaSign_lsf_S3_T6_A2_front.mp4	A2	RH	51	79	FS	
DictaSign_lsf_S3_T4_B0_front.mp4	B0	RH	364	375	FS	
DictaSign_lsf_S9_T3_A3_front.mp4	A3	RH	729	779	FS	
DictaSign_lsf_S2_T3_A11_front.mp4	A11	RH	1121	1164	FS	

Figure 4.21: *All fingerspellings are categorised as FS in the Dicta-Sign-LSF-v2 Corpus. (LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)*

Digging into Signs The conventions of BSL Corpus and Corpus NGT are described without further discussion. (Crasborn et al., 2015a, p. 7)

ECHO Corpus The target word is written in uppercase and prefixed with *fs-* in brackets, e. g. *(fs-) BUICK*. *fs-* is not used to gloss initialised signs, but the convention for initialised signs is not described. (Nonhebel et al., 2004a, p. 2)

Example: The fingerspelled Swedish word 'by' meaning 'village' is glossed in the tier for Swedish glosses as *(fs-) BY* and in the tier for English gloss as *(fs-) VILLAGE*, although the letters spelled are 'B' and 'Y'. (Bergman and Mesch, 2004, SSL_JI_fab1.eaf, 00:00:05.850–00:00:06.270)

HSL Corpus No information available.

LIS Corpus Each fingerspelled letter is written in uppercase, hyphens are put in between the letters, e. g. *C-I-A-O*. (Santoro and Geraci, 2015, slides 19, 23)

PJM Corpus All instances of fingerspelling are glossed as subtypes of the type *LITEROWANIE*, polish for 'fingerspelling'. The spelled letters are written in uppercase separated by dots, e. g. *U.R.A.Z.Y.* (Rutkowski et al., 2015, slide 32)

POLYTROPON Fingerspelling is prefixed with *FS-* followed by the spelled letters. *(E. Efthimiou, personal communication, January 24, 2022)*

Example: A fingerspelling in the POLYTROPON glossed as *FS-Π*. (ILSP Athena Research Center, 2018, Clarin_Pi_SVQ1.eaf, 00:00:01.440–00:00:02.560)

SIGNOR Corpus Each letter of a spelled word is glossed with a separate ID-gloss. (Jerko and Vintar, 2015, 'Finger-spelling')

Signs of Ireland The fingerspelled letters are separated with periods, followed by a hashtag and the target word e. g. *c.l.u.b # club G.a.l.w.a.y # Galway*. (Matthews and Sheridan, 2015, slide 16)

SSL Corpus Fingerspellings are suffixed with *@b*, e. g. *LIRARE@b*. Only the fingerspelled word 'Ja', meaning 'Yes' has two different occurrences of spelling that are glossed as *JA@b* and *JA@ub*.

Example: The signer tells how his foster parents were telling him to practice saying 'mama' and 'papa', both words being spelled out by the signer and glossed as *PAPPA@b, MAMMA@b*. (Mesch et al., 2012, När lär man sig teckna, https://teckenspraakskorpus.su.se/#/video/sslc01_121.eaf?q=pappa%40b&t=206.362)

4.7 POINTING SIGNS

Pointing signs can be highly contextual as well as lexicalised like signs for body parts that are in fact pointings to the body part, e. g. the sign for nose in different sign languages. Some corpora hardly differentiate between different pointing signs, while others categorise them for example into personal pronouns, other lexicalised signs and context-dependent pointings. The degree of description of each individual aspect also varies widely.

Auslan Corpus Pointing signs are prefixed with *PT:* or may in the first stage of annotation be only glossed as *PT*. The gloss can be further expanded with its function or role and the used handshape, e. g. *PT:PRO1SG(B)* for 'I/me' with a flat handshape. (Johnston, 2019, p. 25)

In total eleven major types of points, three points to buoys and two points that are buoys themselves are distinguished in the Auslan Corpus. (Johnston, 2019, pp. 27–29)

BSL Corpus Pointing signs are prefixed with *PT:* and further specified as pronoun, determiner, locative, possessive (all in singular or plural and where appropriate first, second or third person), point towards a body part or a buoy and ambiguous point. In total 21 different points are distinguished, e. g. *PT:PRO1SG, PT:LOCPL, PT:POSS2PL, PT:LBUOY*. If a number sign is incorporated into the point (only possible with plural points) it is suffixed to the gloss as a written word after a hyphen, e. g. *PT:PRO1PL-TWO*.

(Cormier et al., 2017, pp. 8–10).

Example: A list of different points in one transcript of the BSL Corpus: *PT:PRO1SG, PT:POSS1SG, PT:PRO1PL-TWO, PT:LOC, PT:PRO3SG* (Schembri et al., 2017, BF8n.eaf)

Corpus LSFB Pointing signs are prefixed with *PT:* followed by a grammatical classification of the point and sometimes a handshape code (see Chapter 6, e. g. *PT:PRO1, PT:PRO6, PT:POSS3(B)*. (Sinte et al., 2015, '18. Pointing signs')

Example: Different points in the Corpus LSFB: *PRO1(BENTB)_2H*, *PT:PRO1+++++*, *PT:PRO1(7)-2H*, *PT:LOC-2H*. (Meurant, 2015, CLSFBI0103) It becomes apparent that further information can be added to the gloss: if the point is with two hands (*-2H*), information on the handshape with handshape codes, repetitions with a plus symbol. The occurrence of both *-2H* and *_2H* also shows that some conventions are not used consistently.

Corpus NGT Pointing signs are glossed as *PT* with information on the handshape, movement and/ or the location of the point suffixed, e. g. *PT-Bhand:B* is a point with the B-hand to towards the signer, *PT:arc* is a point to several locations in one sweeping motion. The combination of a palm-up gesture with a point is glossed as *PALM-UP+PT*, or in the Dutch gloss as *PO+PT*. The referent of the point is specified in the tier **Referent**. (Crasborn et al., 2020, pp. 19–20)

Example: Different points in the Corpus NGT: *PT-1hand*, *PT-1hand:1*, *PT:mid*, *PT:wijs*, *PT:pink*, *PT:down*, *PO+PT*, whereas the latter is a combination of the palm-up gesture with a point. (Crasborn et al., 2008b, CNGT0814.eaf, `https://hdl.handle.net/1839/00-0000-0000-0009-2D6F-3`)

Corpus FinSL Pointing signs are glossed as *OS:*. Points towards the signer are glossed as *OS:minä* and points with a B-handshape as *OS(B):*. Pointings towards list buoys are an exception and described further in Section 4.4. (Salonen et al., 2019, p. 13; Salonen et al., 2020, p. 199)

Example: In addition to the aforementioned pointing signs *OS:* and *OS:minä*, we encounter *HAISTAA(OS:nenä)* which could be translated as *SMELL(OS:nose)* in the data. (Jantunen, 2018, CFINSL2014_011_03.eaf)

Corpus VGT Pointings are prefixed with *WG*, which stands for ‘wijsgebaar’. Person numbers are suffixed written as digits, e. g. *WG-1* means ‘I’, *WG-2* means ‘you’, etc. For the third person information on the referent is added to the gloss, e. g. *WG-3bij*, if the referent is not clear or known only *WG* is used. Pointings to the weak hand are glossed by giving this information as suffix, e. g. *WG-Lhand*. If the pointing is not performed with the 1-hand, a handshape code in brackets is added to the gloss after a space, e. g. *WG-1 (B)*. (Verstraete et al., 2015, ‘Pointing signs’)

DTS Corpus Points towards the first person singular location (near the body or on the chest) are glossed as *I*. All other points are glossed as *POINT*. In the tier **locus** information on the direction and location of the point can be added. (Kristoffersen and Troelsgård, 2015, ‘18. Pointing signs’)

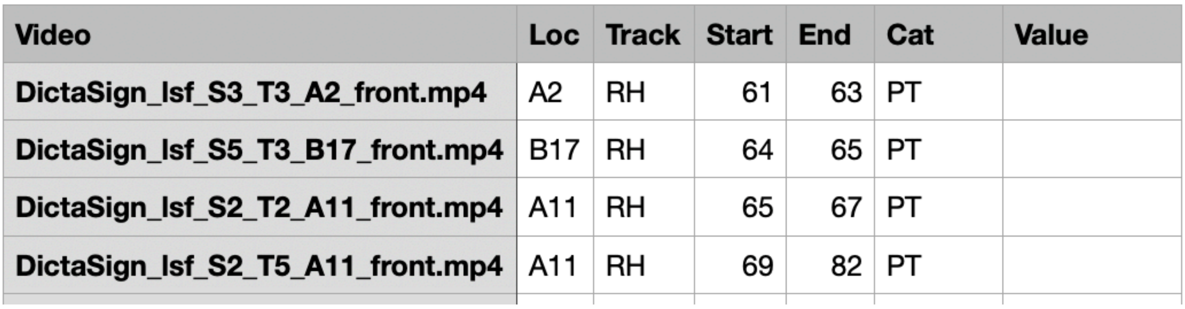
DGS Corpus Pointing signs are glossed with *$INDEX*, suffixed numbers code variation in handshape and orientation: *$INDEX1* for points with the index handshape, *$INDEX2* for points with the flat hand, palm up and finger tips pointing, *$INDEX4* for points with the thumb. Lexicalised pointings have separate glosses, like *I1, YOU1, NOSE1A, HEART1A*.

Three special cases of pointing signs receive their own glosses: *$INDEX-ORAL1* for points to the mouth to attract the addressee’s attention to lip-reading, *$INDEX-TO-SCREEN1* for pointings to the actual screen in the room for elicitation reasons and *$INDEX-AREA1* for indicating or locating a referent to a specific area in the signing space. (Konrad et al., 2020b, pp. 11–12)

Example: Different indices used by one signer in one transcript: *$INDEX-ORAL1, $INDEX1, $INDEX2, $INDEX4* (Konrad et al., 2020a, DGS Corpus, dgskorpus_fra_05: Experience of Deaf Individuals, `https://doi.org/10.25592/dgs.corpus-3.0-text-1212176`)

Dicta-Sign-LSF-v2 Corpus Pointings are annotated with the category *PT* without a value. For the future it is planned to add a value describing the handshape and the element pointed at. (Braffort, 2019, p. 5)

Example: As Figure 4.22 shows, all pointing signs are categorised as *PT* with no further value.



Video	Loc	Track	Start	End	Cat	Value
DictaSign_lsf_S3_T3_A2_front.mp4	A2	RH	61	63	PT	
DictaSign_lsf_S5_T3_B17_front.mp4	B17	RH	64	65	PT	
DictaSign_lsf_S2_T2_A11_front.mp4	A11	RH	65	67	PT	
DictaSign_lsf_S2_T5_A11_front.mp4	A11	RH	69	82	PT	

Figure 4.22: *All pointing signs are categorised as PT in the Dicta-Sign-LSF-v2 Corpus. (LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)*

Digging into Signs It is recommended to annotate all pointing signs with *PT* and to add all additional information on separate tiers, such as **GrammClass** or **Referent**. (Crasborn et al., 2015a, p. 7)

ECHO Corpus All pointing signs are glossed as *IND* (short for Index) in the NGT and BSL dataset and *PEK* in the SSL dataset. (Nonhebel et al., 2004a, p. 2)

Example: Next to a few *IND* glosses in the NGT dataset, there is also one gloss called *IND STERREN* or in English *IND STARS* which seems to be more specified than the other pointings. (Crasborn et al., 2004, NGT_WE_poems.eaf, 00:23:03.980–00:23:04.980)

HSL Corpus No information available.

LIS Corpus All pointing signs are prefixed with *IX-*. Information on pronouns and location is added, e. g. *IX-1, IX-2, IX-LOC, IX-POSS-number*. (Santoro and Geraci, 2015, slides 16 , 23)

PJM Corpus Pointing signs are prefixed with *WSKAZ:* meaning 'point', followed by information on the hand configuration and localization, e. g. *WSKAZ: 1 (JA), WSKAZ: 1 + 1 (DLA SIEBIE).* (Rutkowski et al., 2015, slide 33)

POLYTROPON Pointing signs are glossed as *INDEX/TOPIC* with no further information on location and repetition. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus Pointing signs are not labelled in a special way but glossed as pointing pronouns. (Jerko and Vintar, 2015, 'Pointing signs')

Signs of Ireland The approach of the BSL Corpus is adapted and added with a description, e. g. *PT:PRO1, PT:LOC, PT:BUOY*. Although it is not fully clear what is meant by this description and in the picture there appears a gloss called *INDEX+f/fl/sl*/...* which could be a pointing sign as well, while glosses with *PT:* are missing. (Matthews and Sheridan, 2015, slides 9, 17)

SSL Corpus Pointing signs are annotated in several different ways, most of them with the base *PEK*.

PEK is used for points directed at different locations. The meaning is not annotated. Several pointings are distinguished: *PEK (B)* for a point with the thumb, *PEK.MULTI* for two or more consecutive points, *PEK.FL* for a sweeping motion.

For absolute pointings *PEK>* is used followed by the object/person/bodypart that is pointed at, e. g. *PEK>person* for a point to a person in the room. For pointings with the index and middle finger the gloss *PEK.TVÅ* is used. The same pointing is sometimes signed with a twist in orientation, in that case it is glossed as *PEK(V)>*.

Pointing signs towards the signer themself are glossed as *PRO1*. If the point is to the signer themself and other people, in the sense of 'us two' it is annotated as *PRO1.TVÅ*, 'us three' as *PRO1.TRE* and 'us four' as *PRO1.FYRA*. A unspecified 'we' is glossed as *PRO1.FL*.

To specify different locations that are indicated with the pointing sign several glosses can be used: *PEK.PLATS.DÄR, PEK.PLATS.HÄR, PEK.PLATS.UPP, PEK.PLATS.NER* corresponding approximately to 'there', 'here', 'up' and 'down'. If the point traces a path for the sake of showing that exact path the gloss *PEK.BANSTRÄCKA* is used. If the pointing finger has contact with the flat palm of the other hand, meaning that something is written this is glossed as *PEK.STÅ-SKRIVET*. Pointings toward list buoys are glossed as *PEK.REL*. (Wallin and Mesch, 2018, pp. 11–14, 25–26)

Example: Different pointing signs in one transcript of the SSLC: *PRO1, PEK.MULTI@z, VARA*PEK, PEK>person, PEK, PEK.FL*. One of the pointings is marked for being uncertain (*@z*), another as a blend (*). (Mesch et al., 2012, När lär man sig teckna, https://teckenspraakskorpus.su.se/#/video/sslc01_121.eaf)

4.8 NUMBERS

The annotation of numbers varies in different respects: Are the numbers written as spelled out words or with digits, are number sequences glossed as one unit or as separate ones, is some kind of label used and are ordinal numbers and other specific number uses glossed individually? Some signs allow numbers to be incorporated, e. g. in most sign languages signs to express time like *WEEK*, *O'CLOCK* in Auslan can be combined with number signs to express the exact amount, e. g. 'two weeks' or 'two o'clock' is signed with one sign.

Auslan Corpus Numbers are glossed as spelled out words, number sequences are glossed as one sign with hyphens between the different parts, e. g. *EIGHT*, *NINETEEN-EIGHTY-SEVEN*. Incorporated numbers are suffixed to the base sign which incorporates the number in brackets, e. g. *WEEK(TWO)*, *O'CLOCK(TWO)*. (Johnston, 2019, p. 23)

BSL Corpus Numbers are glossed as spelled out words and all unique number signs are stored in the Signbank. Number sequences are glossed as one sign with carets between the different parts, e. g. *NINETEEN^EIGHT^NINE* Incorporated numbers are suffixed at the end of the gloss, separated by a hyphen, e. g. *AGE-FOURTEEN*. (Cormier et al., 2017, p. 6)

Example: The sign for the number *two* is glossed *TWO* (Schembri et al., 2017, BF1n.eaf, 00:00:26.796–00:00:27.058)

Corpus LSFB Numbers are glossed as digits, variants are suffixed with a period and a running number. Number sequences are glossed as one sign also written as digits, incorporated numbers are added to the basic gloss separated by a hyphen and ordinal numbers are suffixed with *E*, e. g. *6*, *6.2*, *1989*, *HOUR-5*, *1E*, *2E*, *3E*. (Sinte et al., 2015, '11. Numbers', '12. Bumber sequences', '13. Number incorporation', '14. Ordinal numbers')

Example: Different number glosses in the Corpus LSFB: *3*, *2*, *FOIS-1*, *SEMAINE-1* (Meurant, 2015, CLSFBI3318)

Corpus NGT Numbers are glossed as digits, number sequences are glossed as one sign. Incorporated numbers are suffixed to the base sign and separated with a plus character. Ordinal numbers are suffixed with *.ORD*, e. g. *1-A*, *1-B*, *182-A*, *UUR+1*, *UUR+4-A*, *1.ORD*, whereas the suffixed letters stand for lexical variants (see Section 4.2). The signs for one million and one billion receive their own glosses in written words: *MILJOEN*, *MILJARD*. As mentioned in Section 4.4, lists on the hand are glossed as *TELHAND*. (Crasborn et al., 2020, pp. 17–18)

Example: Some number glosses in the Corpus NGT: *2-A*, *2.ORD*, *3.ORD-A*, *5*, *6-A*, *8-A*, *9-A*, *10-B*, *30-A*. Note that the number five has no lexical variants, as it is always signed with all fingers of the hand. (Crasborn et al., 2008b, CNGT0117.eaf, https://hdl.handle.net/1839/00-0000-0000-0009-2D6C-D)

Corpus FinSL Numbers are glossed as spelled out words suffixed with *_num*. Number sequences are glossed as one sign with a hyphen between the different words and in the basic order in which they are referenced, e. g. *TUHAT-YKSI_num*, *VIISIKYMMENTÄ-KOLME_num*, *TUHAT-YHDEKSÄNSATAA-KAHDEKSANKYMMENTÄ-KAKSI_num* in English 'one thousand', 'fifty three' and 'thousand nine hundred eighty two'. Incorporated numbers are glossed with the number in words first followed by a hyphen and the base sign, e. g. *KUUSI-VUOSI_num* for 'six years' and *NELJÄ-VIIKKO_num* for 'four weeks'. Some basic numbers and incorporated numbers are stored in the Finnish Signbank, all others are annotated by manually typing the number without contacting the Signbank. (Salonen et al., 2020, p. 199; Salonen et al., 2019, pp. 14–15)

Example: Number two in the CFinSL: *KAKSI_num@sbb* (Jantunen, 2018, CFINSL2014_011_03.eaf, 00:00:47.720–00:00:48.080)

Corpus VGT Numbers are written in digits and prefixed with *C:*, e. g. *C: 10*.

DTS Corpus Numbers are glossed as spelled out words, number sequences are glossed as separate glosses, e. g. *NINETEEN-HUNDRED NINE EIGHTY*. It remains unclear if the hyphen is used to indicate that 'nineteen hundred' is one sign or if it could be an incorporation of numbers. Incorporated numbers are glossed according to their meaning, the relevant number sign is typically, but not necessarily added to the gloss, e. g. *TWO-HOURS, IN-THREE-DAYS, FIRST-FLOOR*. Ordinal numbers are glossed separately without further labels. (Kristoffersen and Troelsgård, 2015, '11. Numbers', '12. Number sequences', '13. Number incorporation', '14. Ordinal numbers')

DGS Corpus Numbers are glossed as *$NUM* suffixed by one of five markers grouping numbers into their range: *-ONE-TO-TEN, -TEEN, -TENS, -HUNDREDS, -THOUSANDS* and a number and letter to specify the variant, if needed. The signed number is added with a code for digit handshapes which combined with the root indicates the intended meaning. Variation in the handshape is coded with an additional letter in small caps after the digit, e. g. *$NUM-TEEN1:3d* for thirteen with a 3-handshape using the thumb.

Number sequences are segmented into each token separately, e. g. the numer '1989' is glossed as *$NUM-TEEN1:9 $NUM-ONE-TO-TEN1A:9 $NUM-TENS1:8d* meaning 'nineteen' 'nine' 'eighty'. DGS, as German, articulates the last digit before the second-to-last if that is 2 or larger.

Some movements are glossed separately: *$NUM-DOUBLE1A* for a repeated digit, *$NUM-TAPPING1* for a tapping movement used with the numbers 11 and 12 and *$NUM-SNIP1* for a snipping movement used with the numbers 11 to 19.

Three number signs don't allow number incorporation: *$NUM-HUNDRED1,$NUM-THOUSAND1* and *$NUM-MILLION1*. Signs that require number incorporation have their own roots prefixed with *$NUM*, e. g. *$NUM-CLOCK1A, $NUM-GRADE1, $NUM-WEEK-AFTER-NOW1* etc. Lexical signs that incorporate numbers are marked with an asterisk, e. g. *YEAR1A**. In the DGS Corpus the incorporation is annotated in more detail using quantity qualifiers and handshape codes.

Numbers signed by depicting the Roman script are glossed as *$NUM-ROMAN1*. Ordinal numbers are glossed as *$NUM-ORDINAL1, $NUM-ORDINAL2*. (Konrad et al., 2020b, pp. 13–15)

Example: Different numbers signed in one transcript of the Public DGS Corpus: *$NUM-ONE-TO-TEN1D:8d, $NUM-TENS2A:3d*, $NUM-TENS2B:3d*, $NUM-ONE-TO-TEN1A:1, $NUM-ONE-TO-TEN1A:3d, $NUM-TENS2A:2d, $NUM-TEEN1:6d, $NUM-TEEN1:7d, NUM-ORDINAL1:2*, some part of a number sequence: *$NUM-ONE-TO-TEN1A:4* and *$NUM-TENS2A:7d** meaning '74'. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_fra_05 - Experience of Deaf Individuals, `https://doi.org/10.25592/dgs.corpus-3.0-text-1212176`)

Dicta-Sign-LSF-v2 Corpus Numbers from zero to ten are annotated as lexical signs. All numbers bigger than then 10 are labelled with the category *N* without any further value. Internally there seems to be values corresponding to the actual numbers, but these are not openly available. Harmonization work is planned for the future. (Braffort, 2019, p. 5)

Example: In the data it comes apparent, that the lexical numbers are suffixed with *:NUM* or *:NUM:VAR*: *DEUX1:NUM:VAR, TROIS3:NUM, VINGT:NUM, SEPT1:NUM:VAR, HUIT2:VOT_HUIT1:NUM (PAUME VERS SOI)*. With the latter containing information on the palm orientation in the gloss in brackets.

In addition, lexical number signs greater than ten are also glossed and a space is used in between words: *DIX SEPT:NUM, DIX HUIT:NUM:VAR, QUATRE VINGT:NUM, QUATRE VINGT DIX:NUM*. (LIMSI, 2020, Dicta-Sign-LSF_ID.csv)

Digging into Signs The different approaches from BSL Corpus and Corpus NGT are described without further discussion. (Crasborn et al., 2015a, p. 6)

ECHO Corpus No information available.

HSL Corpus No information available.

LIS Corpus Numbers are glossed as spelled out words, number sequences are spelled out without separation, incorporated numbers are prefixed to the basic sign as spelled out words and separated with a hyphen, e. g. *ONE, MILLENOVECENTOOTTANTANOVE, QUATTRO-ORA*. (Santoro and Geraci, 2015, slides 23, 24)

PJM Corpus Numbers are glossed as spelled out words with the prefixed *NUM:*, e. g. *NUM: DRUGI 1.2 P:V;L:Ø (BEZ OBRÓT), NUM: CZTERY © 1.2 P:4;L:Ø* being the number two without rotation and the number four. What the symbol '©' means is not described. Incorporated numbers and ordinal numbers are also prefixed with *NUM:* followed by the number and in the case of incorporation the base word, separated by a hyphen, e. g. *NUM: SZESC-MIESIECY 1 P:1;L5* for 'six months' and *NUM: DRUGI 1 P:V;L:Ø (OBRÓT) (DWA)* for 'second' with a rotation. (Rutkowski et al., 2015, slide 28–30)

POLYTROPON Numbers are glossed as digits. *(E. Efthimiou, personal communication, January 24, 2022)*

Example: Examples of glossed number signs in the data: *11, 2*. (ILSP Athena Research Center, 2018, larin_paralia_2_ex_3_HD_SVQ1.eaf)

SIGNOR Corpus Number sequences are treated as compounds and therefore glossed separately and joined in an extra tier. Ordinal numbers get an extra ID-gloss. Number incorporation is not annotated. (Jerko and Vintar, 2015, 'Number sequence', 'Ordinal numbers', 'Number incorporation')

Signs of Ireland Numbers are written as digits, number sequences as one gloss in digits, incorporated numbers with the digit followed by the basic sign, and ordinal numbers with digits and superscript 'st', 'nd' or 'rd', e. g. *1, 1989, 4-HOUR, 1st, 2nd, 3rd*. This approach is similar to the Corpus LSFB. (Matthews and Sheridan, 2015, slide 15)

SSL Corpus Numbers are written as spelled out Swedish words, number sequences are written with carets between numbers, as in the BSL Corpus, e. g. *TJUGO^FEM*, meaning 'twenty-five'. The number one occurs with three different movements glossed separately as *EN, EN.ENDA, EN.BARA*. Ordinal numbers are glossed with the prefix *ORDNING+*, e. g. *ORDNING+EN*. If the ordinal number is articulated with both hands a *(da)* is suffixed, e. g. *ORDNING+EN(da)*. If the ordinal number is used with the meaning of a list like for floors it is suffixed with *LIST* and the according number, e. g. *ORDNING.LIST+TVÅ*.

Incorporated numbers are prefixed with the base sign followed by a plus symbol and the number, this is the same approach as in the Corpus NGT. The listed base signs are for krona, time, weeks,

years, days, floors and numbers in the sense of start numbers, channel numbers, house numbers, e. g. *KRONA+FEM*, *KLOCKTID+TVÅ*, *VECKA+EN*, *ÅR+FYRA*, *DAG-DÅTID+EN*, *DAG-FRAMTID+TVÅ*, *GÅNG+TVÅ*,
NUMMER+TVÅ, *VÅNINGSPLAN+TVÅ*. (Wallin and Mesch, 2018, pp. 9–11)

Example: The signer tells about two dogs, one being three years old and the other being one year old using numbers glossed as *TRE* and *EN*. (Mesch et al., 2012, Fritt - husdjur, https://teckenspraakskorpus.su.se/#/video/sslc01_045.eaf?q=TRE&t=9.760)

Funded by the Horizon 2020
Framework Programme of the European Union

4.9 REPETITION

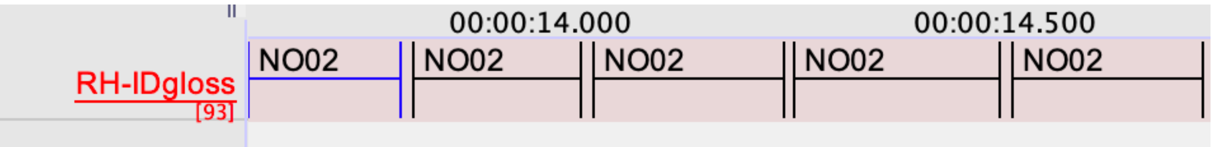
A repeated movement in signs can have different lexical status. It can be part of the citation form of the sign, e. g. the LSFB sign *CHAT.F* is signed with a repeated movement. A repetition can also express the plural form of a sign or aspect, e. g. several children, or doing something eagerly. Repetition can also simply mean that the sign is being repeated. Most corpora distinguish between these different kinds of repetitions and adjust the segmentation of the repetition accordingly.

Auslan Corpus Each instance of a repeated sign is glossed separately, e. g. *SCREAM WOLF WOLF*. If the repetition is part of a modification of the sign it is glossed as one instance and a note is made in the dedicated tier for grammatical modification. (Johnston, 2019, pp. 48–49)

BSL Corpus Each instance of a repeated sign is glossed separately, e. g. *BOY SHOUT WOLF WOLF WOLF*. (Cormier et al., 2017, p. 6)

Phonological and grammatical information, also on repetition, is annotated on extra tiers that are not publicly available. The conventions for their annotation are also not public, but largely follow those of the Auslan Corpus. (*K. Cormier, personal communication, January 31, 2022*)

Example: Figure 4.23 shows several repetitions of the sign for ‘no’.



RH-IDgloss [93]

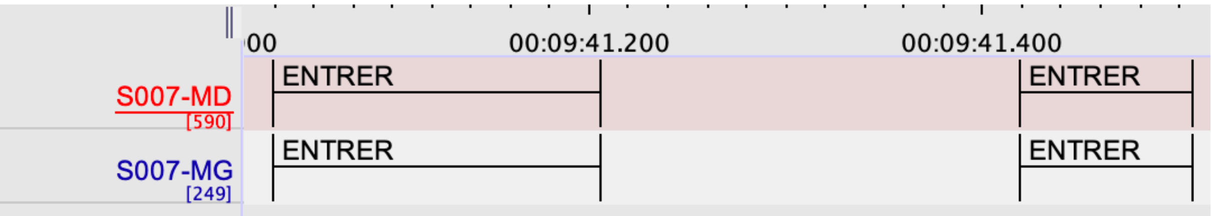
00:00:14.000			00:00:14.500	
NO02	NO02	NO02	NO02	NO02

Figure 4.23: *Annotation of repetitions in the BSL Corpus. (Schembri et al., 2017, BF1n.eaf, 00:00:13.686–00:00:14.716)*

Corpus LSFB If the repetition is distinguishable the signs are glossed separately, e. g. *WORK WORK WORK*. If the sign is made with more movements than the citation form this information can be suffixed in brackets, e. g. *WORK(3x)*. If the movement is made at one go it is glossed with one or more plus symbols, e. g. *WORK+++*. It becomes not totally clear what the exact differences between this different approaches are and which cases the suffixed information is used instead of separate glosses. If the repetition is used to mark plural, this information is also suffixed in brackets, e. g. *CHILD(pl)*. It is questioned whether different repetitions are always distinguishable. (Sinte et al., 2015, ‘6. Repetition’, ‘10.Plurality’, ‘Question’)

Example: Figure 4.24 shows a distinguishable repetition of the sign for ‘enter’. Other examples are repetitions of the signs for ‘change’ and ‘marriage’:

- *CHANGER+++++++* (Meurant, 2015, CLSFBI0301.eaf, 00:09:41.039–00:09:41.508)
- *MARIAGE(3X)* (Meurant, 2015, CLSFBI0103.eaf, 00:08:16.766–00:08:17.500)



	00	00:09:41.200	00:09:41.400
S007-MD [590]	ENTRER		ENTRER
S007-MG [249]	ENTRER		ENTRER

Figure 4.24: *Annotation of distinguishable repetitions in the Corpus LSFB. (Meurant, 2015, CLSFBI0301.eaf, 00:09:41.039–00:09:41.508)*

Funded by the Horizon 2020 Framework Programme of the European Union

Corpus NGT Repetition marking a plural form are marked with a suffixed *.PL*, e. g. *KIND.PL*. The number of movement cycles can also be annotated in a tier dedicated to the phonetic form of manual actions, in this case for repetitions, called **NOM**. (Crasborn et al., 2020, pp. 17, 31)

Example: A sign in plural form in the Corpus NGT: *DING.PL* meaning 'thing'. (Crasborn et al., 2008b, CNGT0814.eaf, 00:00:29.640–00:00:30.120, https://hdl.handle.net/1839/00-0000-0000-0008-EE8B-2)

Corpus FinSL Repeated movements can be marked in form of the coding for variants. For example, *TIETÄÄ(L_toisto)* marks the sign for 'know' as repeated. Repetition is also annotated on the daughter tier with the code *@toisto*, if it is not part of the citation form but to mark plural or aspect. The gloss itself is not written in the plural form. (Salonen et al., 2020, p. 199; Salonen et al., 2019, pp. 244, 27–28)

Example: In Figure 4.25 the two-handed sign for 'dispute' is repeated.

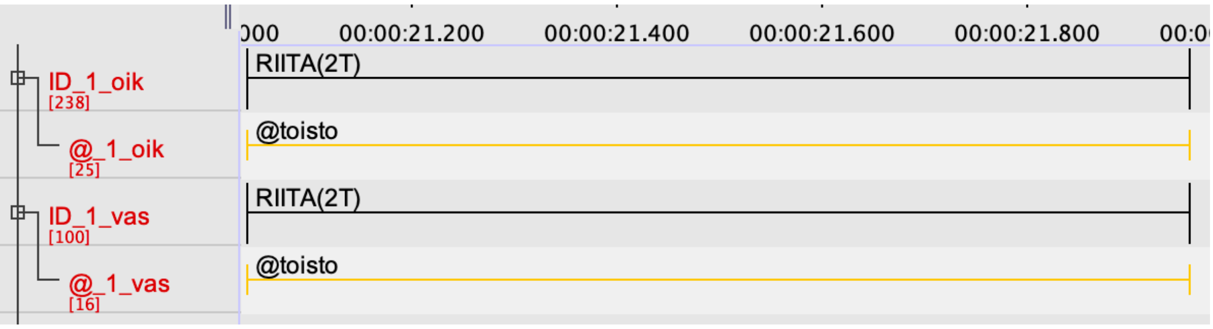


Figure 4.25: *Annotation of repetitions in Corpus FinSL. (Jantunen, 2018, CFINSL2014_019_03.eaf, 00:00:21.040–00:00:21.960)*

Corpus VGT Repetitions are not further categorised but all glossed with three plus symbols no matter how many times the sign is repeated, e. g. *ALLEMAAL-A +++*. (Verstraete et al., 2015, 'Repetition')

DTS Corpus Each instance of a repeated sign is glossed separately. If the repetition is a modification of the sign this information is added on a separate child tier, similar to the Auslan Corpus. An exception is made for lexicalised plurals which receive an ID-gloss, if the forms are not the regular plural modification or have a different meaning than the mere plural of the base sign, e. g. *CHILD, TREE* and their plurals *CHILDREN, FOREST*. (Kristoffersen and Troelsgård, 2015, '6. Repetition', '10. Plurality')

DGS Corpus In the Public DGS Corpus repetition is annotated as a form difference indicated by an asterisk. In the DGS Corpus repetition is annotated in more detail using the qualifier 'phases' with feature values to document the number of movements or an unspecified number of movements. (Konrad et al., 2012, pp. 90–92)

Example: The one-handed sign for child is used with different repetitions by different signers: once with a repeated downwards movement, once with a repeated downward movement combined with a sidewards movement. Both these occurrences are annotated in the Public DGS Corpus as *CHILD2*. In the DGS Corpus these differences are documented using additional qualifiers. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_fra_10: Experience of Deaf Individuals, 00:03:06:06–00:03:06:26, https://doi.org/10.25592/dgs.corpus-3.0-text-1245887 and Public DGS Corpus, dgskorpus_ber_12: Experience of Deaf Individuals, 00:05:41:35–00:05:41:46, https://doi.org/10.25592/dgs.corpus-3.0-text-1419797)

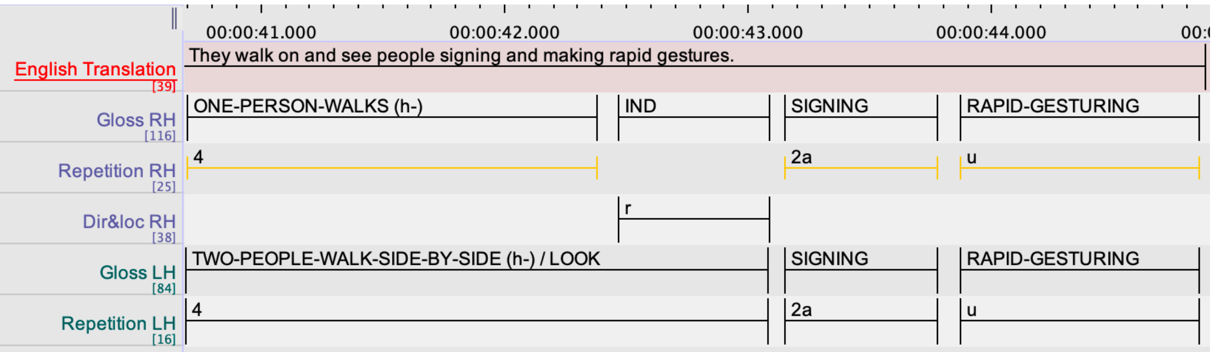
Funded by the Horizon 2020
Framework Programme of the European Union

Dicta-Sign-LSF-v2 Corpus Repeated signs are glossed as one manual unit if the transition between the sub parts is very short. If the transition is longer or at least one parameter is changed each unit is glossed separately. (Braffort, 2019, p. 4)

Digging into Signs It is recommended to gloss each repetition of a sign separately. If the repetition is marking plurality this should not be labelled in the gloss but annotated on a separate tier. (Crasborn et al., 2015a, pp. 5–6)

ECHO Corpus Each hand tier has a child tier called **Repetition RH/LH** to annotate repetition, the number of repetition is annotated using numbers (in digit format). If the repetition is uncountable it is glossed as *u*. Alternating repetitions are glossed with the number of repetitions and a suffixed *a*. (Nonhebel et al., 2004a, p. 3)

Example: Figure 4.26 shows the sign *ONE-PERSON-WALKS* with a hold and repeated four times in a symmetrical manner, followed by a one-handed point, the sign *SIGNING* repeated two times in an alternating manner and the sign *RAPID-GESTURING* with uncountable repetitions.



00:00:41.000 00:00:42.000 00:00:43.000 00:00:44.000

English Translation [39]: They walk on and see people signing and making rapid gestures.

Gloss RH [116]: ONE-PERSON-WALKS (h-) | IND | SIGNING | RAPID-GESTURING

Repetition RH [25]: 4 | 2a | u

Dir&loc RH [38]: r

Gloss LH [84]: TWO-PEOPLE-WALK-SIDE-BY-SIDE (h-) / LOOK | SIGNING | RAPID-GESTURING

Repetition LH [16]: 4 | 2a | u

Figure 4.26: *Annotation of repetitions in the ECHO Corpus. (Woll et al., 2004, BSL_PS_poem3.eaf, 00:00:40.680–00:00:44.880)*

HSL Corpus No information available.

LIS Corpus Repetition and plurality are not glossed. (Santoro and Geraci, 2015, slide 22)

PJM Corpus Grammatical repetition, such as plural or aspect is not annotated separately, but described in the HamNoSys notation. Repetition without grammatical function is glossed separately. (Rutkowski et al., 2015, slide 22)

POLYTROPON Repetitions are not annotated in the POLYTROPON. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus Information given is 'Contextual meaning if plural', which does not explain if a contextual meaning is annotated to mark plurals or if the contextual meaning is used to identify plurals. (Jerko and Vintar, 2015, 'Plurality')

Signs of Ireland Repetitions are not further categorised but all glossed with one plus symbol per repetition, e. g. *WOLF+++*, *WORK+++*, *SAME++*. Plurals are suffixed with *-PL* or the plural form of the English word is used, e. g. *CHILDREN*, *SIGN-PL*. (Matthews and Sheridan, 2015, slides 5, 13, 14)

SSL Corpus Reduplication is annotated with the suffix *@rd*, e. g. *SKRIVA@rd*, *ENVIS@rd*. If the repetition indicates a plural form the gloss is suffixed with *PL*, e. g. *BILD.PL*. (Wallin and Mesch, 2018, pp. 7, 16)

Funded by the Horizon 2020 Framework Programme of the European Union

Example: The signer uses a plural form of 'person' glossed as *PERSON.PL* and a reduplicated form of 'signing' glossed as *TECKNA-FLYT@rd* (Mesch et al., 2012, Världskongress för döva, `https://teckenspraakskorpus.su.se/#/video/sslc01_163.eaf`)

Funded by the Horizon 2020
Framework Programme of the European Union

4.10 COMPOUNDS

Compounds can be annotated as one unit with information on the separate parts within the gloss or they can be annotated separately with additional information about their occurrence in compounds in extra tiers or in the lexical database. Some corpora differentiate between sequential compounds, simultaneous compounds and blends. In some SLs bound morphemes from the surrounding spoken language are transferred into signing, they may be annotated individually too.

Auslan Corpus Lexicalised compounds are glossed as one unit, if they satisfy two conditions:
'1. the meaning of the whole is not predictable from the elements,
2. it is not possible to insert another sign between the two elements at all, or without changing the meaning of the particular utterance.' (Johnston, 2019, p. 49)

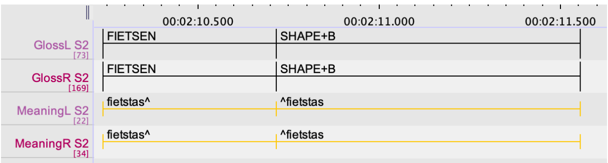
Glosses for compounds exist of one unique ID-gloss and not of the elements in the sign, e. g. *PARENTS* instead of *MOTHERˆFATHER*. If two signs appear to be a compound but there is no unique ID-gloss in the Signbank the two sign elements are used with a caret symbol in between and a comment is made in the tier **comments**, so that a new ID-gloss can be built. (Johnston, 2019, p. 49)

BSL Corpus Most compounds are found as distinct ID-glosses in the BSL Corpus, as *MOTHER, FATHER*. If the compound is not yet in the Signbank, a temporarily gloss is used with two ID-glosses separated by a caret, e. g. *FS:G-GRAPHICˆART* for 'graphic designer'. These temporarily glosses are later checked and – if in widespread use – glossed with new ID-glosses or – if not widely used – glossed with two separate signs and seen as collocations. This approach is similar to the Auslan Corpus. (Cormier et al., 2017, p. 6)

Corpus LSFB Examples for compound glosses are given on the poster by (Sinte et al., 2015, '07. Compounds'), but conventions are not defined explicitly. The four given examples suggest that compounds can be glossed as one word, two words or two words separated by a hyphen, e. g. *PARENTS, SLEEPLESS-NIGHT, THIS-IS IT*. The convention seems to follow the approach of Corpus NGT.

Corpus NGT Fixed combinations of signs that receive a specialised meaning through their combination are glossed with a separate ID-gloss. In the Signbank the constituting parts are referenced at, e. g. the fixed combination of *AUTO* for 'car' and *BOEK* for 'book' is annotated as *RIJBEWIJS-A* for 'driving licence'. Combinations that are not fixed, but still have a special meaning are glossed as single signs and an entry containing a caret is made in the meaning tier. This approach is also similar to the Auslan Corpus. (Crasborn et al., 2020, p. 16)

Example: In Figure 4.27 the combination of *FIETSEN* 'bicycle' and the classifier *SHAPE+B* is marked as having a special compound meaning 'fietstas' (English: 'bicycle bag') in the meaning tier using two separate glosses marked with carets:



GlossL S2 [73]	FIETSEN	SHAPE+B
GlossR S2 [169]	FIETSEN	SHAPE+B
MeaningL S2 [22]	fietstas^	^fietstas
MeaningR S2 [34]	fietstas^	^fietstas

Figure 4.27: *Annotation of FIETSEN and SHAPE+B resulting in the compound meaning 'fietstas' (English: 'bicycle bag'). (Crasborn et al., 2008b, CNGT0004.eaf, 00:02:10.240–00:02:11.560, https://hdl.handle.net/1839/00-0000-0000-0009-2D5C-5)*

Corpus FinSL Potential compounds are annotated as separate glosses and marked with *@y* in the comments tier. If the compound is fused together, so that the sign looks almost like a single sign or if the two signs are a common concept the compound is annotated in a single gloss, e. g. *LUMIUKKO* meaning 'snowman' or *KUUROJEN-LIITTO* meaning 'Association of the Deaf'. (Salonen et al., 2019, pp. 30–32; Salonen et al., 2020, p. 199)

Example: In Figure 4.28 the signs *TEHDÄ(BB)* meaning 'make' and *HENKILÖ(Lc)* meaning 'person' are marked as a potential compound.

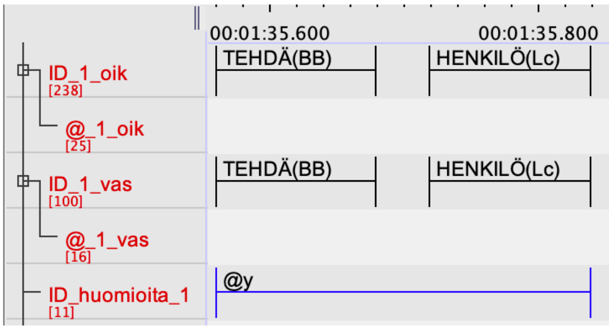


Figure 4.28: *Annotation of a potential compound in Corpus FinSL. (Jantunen, 2018, CFINSL2014_019_03.eaf, 00:01:35.560–00:01:35.840)*

Corpus VGT No information available.

DTS Corpus For each lexicalised compound in the sign base a separate ID-gloss is used. Possible compounds are glossed as consecutive signs. Signs with prefixes and suffixes, identified as calques from spoken Danish, have separate ID-glosses starting or ending with a caret, e. g. *UNˆ, ˆS-GENITIVE*. (Kristoffersen and Troelsgård, 2015, '7. Compounds', '7.b Affixes')

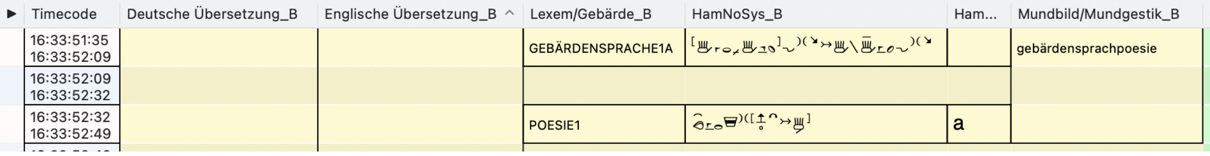
DGS Corpus Sequential compounds are glossed as two separate signs and further categorisation into compounds, loan translations or others is left to a later analysis. Sequences of signs that mirror a German compound are also glossed as separate signs. Mouthing or meaning tags can indicate possible compounds.

Simultaneous constructions are treated as single signs, e. g. the simultaneous compound sign for 'at home' glossed as *AT-HOME1A ˆ* is a combination of the handshape of *TO-SIT1A* and the movement of *HOUSE1Aˆ*; the blend *TO-KNOCK-ON-WOOD1ˆ* deletes the repeated movement of the initial part of the compound *TO-HOPE1Bˆ*. Common blends are found in negated signs, like *WISSEN-NICHT1-$SAM* that combines the signs *WISSEN2A-$SAM* and *WEG-VERLIEREN1-$SAM'hd:1*, in english 'know' and 'away-loose' combined to 'don't know'. These negated forms are glossed with the suffix *-NICHT* in German and prefixed with *DONT-* in English.

For sequential compounds, simultaneous compounds and blends an entry is made in the lexical database containing all components of the signed construction.

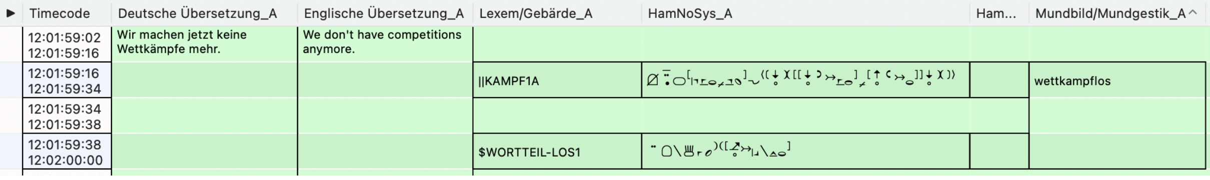
In DGS a few signs are used to express bound morphemes of German words like the suffix '-in' that express the female gender, e. g. 'Lehrerin' being a female teacher compared to 'Lehrer' being the generic or male version. Bound morpheme signs are glossed as *$MORPH*, or *$WORTTEIL* in German, followed by the meaning, e. g. *$MORPH-FEMALE1*. (Konrad et al., 2020b, pp. 10–11; *R. Konrad, personal communication, January 27, 2022*)

Example: In Figure 4.29 the signer talks about sign language poetry, which in DGS consists of two signs, but in German is expressed by the compound word 'Gebärdensprachpoesie'. For the annotation separate glosses are made but the mouthing hints at it being a compound in German. Another example in Figure 4.30 showing the use of a bound morpheme meaning 'less' combined with the stem for 'fight' meaning 'without any competitions':



Timecode	Deutsche Übersetzung_B	Englische Übersetzung_B	Lexem/Gebärde_B	HamNoSys_B	Ham...	Mundbild/Mundgestik_B
16:33:51:35 16:33:52:09			GEBÄRDENSPRACHE1A	[illegible]		gebärdensprachpoesie
16:33:52:09 16:33:52:32						
16:33:52:32 16:33:52:49			POESIE1	[illegible]	a	

Figure 4.29: *The German compound 'Gebärdensprachpoesie' glossed as two separate signs but a single mouthing in the DGS Corpus. (Konrad et al., 2020a, DGS Corpus, dgskorpus_hb_04: Deaf Events, 16:33:51:35–16:33:52:49, https://doi.org/10.25592/dgs.corpus-3.0-text-1181027)*



Timecode	Deutsche Übersetzung_A	Englische Übersetzung_A	Lexem/Gebärde_A	HamNoSys_A	Ham...	Mundbild/Mundgestik_A
12:01:59:02 12:01:59:16	Wir machen jetzt keine Wettkämpfe mehr.	We don't have competitions anymore.				
12:01:59:16 12:01:59:34			||KAMPF1A	[illegible]		wettkampflos
12:01:59:34 12:01:59:38						
12:01:59:38 12:02:00:00			$WORTTEIL-LOS1	[illegible]		

Figure 4.30: *Use of the bound morpheme '-less' (German: '-los') in the DGS Corpus. (Konrad et al., 2020a, DGS Corpus, dgskorpus_hh_01: Free Conversation, 12:01:59:02–12:02:00:00, https://doi.org/10.25592/dgs.corpus-3.0-text-1176566)*

Dicta-Sign-LSF-v2 Corpus No information available.

Digging into Signs The two approaches of BSL Corpus and Corpus NGT are described with a small note that the BSL Corpus may adopt the procedure of referencing to the constituting parts of a compound in the lexical database. (Crasborn et al., 2015a, p. 5)

ECHO Corpus No information available.

HSL Corpus No information available.

LIS Corpus Hyphens are used to separate the two or more parts, e. g. *MOTHER-FATHER*. (Santoro and Geraci, 2015, slides 16, 19)

PJM Corpus Possible compounds are not marked but receive their own gloss. Information on the constituent parts is given after the information on articulation in brackets, e. g. *KOSCIOŁ 1.1 P:ZB;L:ZB (KRZYZ+DOM)* is the sign for 'church' consisting of the signs for 'cross' and 'house'. (Rutkowski et al., 2015, slide 23)

POLYTROPON Compounds are annotated with separate glosses. Compound categories, following the classification by Sandler and Lillo-Martin (2006): classifier+classifier, sign+sign, sign+classifier, classifier+sign and ad-hoc formations are planned to be annotated in the future. (Efthimiou et al., 2018, p. 42; *E. Efthimiou, personal communication, January 24, 2022*)

SIGNOR Corpus Separate glosses are joined in a separate tier. (Jerko and Vintar, 2015, 'Compounds')

Signs of Ireland Same as Corpus NGT, e. g. *PARENTS, GRAPHIC-ART*. (Matthews and Sheridan, 2015, slide 13)

SSL Corpus Compounds are annotated as one unit with two words separated by a caret, e. g. *SOMMAR^JOBB, FOLK^HÖG^SKOLA*, in English 'summerjob' and 'community college'. (Wallin and Mesch, 2018, p. 8)

Blends are also annotated as one unit but with an asterisk between the words, e. g. *FÖRSTÅ*INTE* meaning 'do not understand'.

Example: A signer using the compound 'very good' glossed as *JÄTTE^BRA* and a blend glossed as *HÖRA*INTE* (Mesch et al., 2012, Fritt - husdjur, https://teckenspraakskorpus.su.se/#/video/sslc01_045.eaf?q=j%C3%A4tte%5Ebra&t=79.800 and https://teckenspraakskorpus.su.se/#/video/sslc01_045.eaf?q=H%C3%96RA%2aINTE&t=85.800)

4.11 MANUAL NEGATIVE INCORPORATION

As with the incorporation of numbers or fingerspelling, special movements can be incorporated into signs to negate their meaning. These cases are glossed differently than signs that have a negative meaning in their citation form, e. g. the Auslan sign *KNOW-NOT* compared to *NEVER*.

Auslan Corpus The general meaning gloss is followed by a marker for negation: *-NOT*, e. g. *KNOW-NOT, WANT-NOT*. (Johnston, 2019, pp. 23–24)

BSL Corpus Same procedure as Auslan Corpus. (Cormier et al., 2017, p. 6)

Example: The negated sign 'can' in the BSL Corpus: *CAN-NOT* (Schembri et al., 2017, BF6n.eaf, 00:01:19.689–00:01:20.154)

Corpus LSFB Basic gloss is suffixed with *-NOT* or the word is glossed with a negative meaning, e. g. *LIKE-NOT, DISLIKE*. (Sinte et al., 2015, '8. Manual negative incorporation')

Example: As seen in the data, the French word *-PAS* is used: *COMMUNICATION-PAS* (Meurant, 2015, CLSFBI0301.eaf, 00:02:29.575–00:02:30.399)

Corpus NGT Same approach as in the Auslan Corpus, but with the Dutch word *-NIET* instead of *-NOT*, e. g. *WILLEN-NIET, KUNNEN-NIET-A*. (Crasborn et al., 2020, p. 16)

Example: The manually negated sign for 'not necessary' is glossed as *HOEFT-NIET-A*. (Crasborn et al., 2008b, CNGT0697.eaf, 00:00:19.560–00:00:19.840), `https://hdl.handle.net/1839/00-0000-0000-0008-F6D0-D`

Corpus FinSL Signs with a negative incorporation or meaning are glossed with their basic meaning suffixed with *-EI*, meaning 'no'. Additionally in the associated daughter tier the code *@neg* is added, e. g. *TIETÄÄ-EI*. If the sign is negated by a headshake and no manual movement, the code *@pp* for headshake is added to the daughter tier followed by the code for negation *@neg*. Negative particles in Finnish already exist of a negative element, e. g. 'nothing' is a compound of 'no' and 'anything'; for the glosses the Finnish word order is copied but separated with a hyphen, e. g. *EI-MITÄÄN* for 'no-anything' meaning 'nothing'. (Salonen et al., 2019, pp. 26–27; Salonen et al., 2020, p. 199)

Example: Figure 4.31 shows a non-manual negation followed by a manual negation.

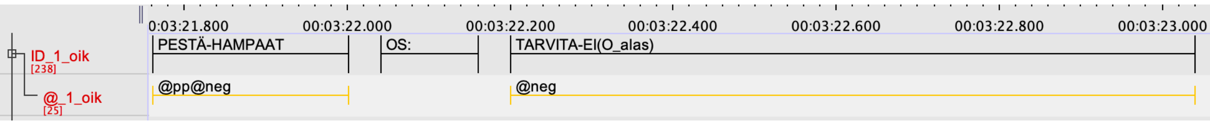


Figure 4.31: *A non-manual negation followed by a manual negation in the Corpus FinSL (Jantunen, 2018, CFINSL2014_019_03.eaf, 00:03:21.760–00:03:23.040)*

Corpus VGT No information available.

DTS Corpus Signs with manual incorporation are glossed according to their meaning and typically include the suffix *-NOT*, although not necessarily. The suffix *-NOT* can also appear in other sign types. (Kristoffersen and Troelsgård, 2015, '8. Manual negative incorporation')

DGS Corpus In the DGS Corpus, incorporated negation by changing the signs movement into a movement that looks like tracing the form of the greek letter 'α' is annotated in more detail using the qualifier 'alpha_negation'. (Konrad et al., 2012, p. 92)

In the Public DGS-Corpus these form differences, like others, are indicated by an asterisk. It is under discussion whether these forms and other forms having a negative part, like blends (see Section 4.10), are to be made available in future releases of the Public DGS Corpus. *R. Konrad, personal communication, January 27, 2022*

Dicta-Sign-LSF-v2 Corpus No information available.

Example: In the data we find glosses, that could indicate a negative incorporation, as *44605 CHER-PAS2*, but without time-aligned transcripts this question can not be answered. (LIMSI, 2020, Dicta-Sign-LSF_ID.csv)

Digging into Signs Both corpora use the same convention of adding the suffix -*NOT* to the gloss, which is also recommended here. (Crasborn et al., 2015a, p. 5)

ECHO Corpus No information available.

HSL Corpus No information available.

LIS Corpus Same approach as Auslan Corpus but with the suffix -*NEG*. (Santoro and Geraci, 2015, slide 16)

PJM Corpus Signs with manual incorporation are prefixed with *NIE*-, e. g. *NIE-BEDZIE 1.2 P:B;L:B* for the negation of ‘it will be’. (Rutkowski et al., 2015, slide 24)

POLYTROPON Signs with manual negative incorporation are annotated by prefixing *ΔΕΝ*-, meaning ‘not’, to the negated item, e. g. *ΔΕΝ-ΞΕΡΩ*, in English ‘NOT-KNOW’. The form of the negative suffix can vary, being annotated according to the rules for variants, e. g. *ΔΕΝ-2*, *ΔΕΝ-3*. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus Negative manual incorporation is not glossed. (Jerko and Vintar, 2015, ‘Manual negative incorporation’)

Signs of Ireland Same as BSL Corpus and Corpus NGT, e. g. *KNOW-NOT*. (Matthews and Sheridan, 2015, slide 14)

SSL Corpus See Section 4.10 for blends with the negative particle ‘inte’.

4.12 DIRECTIONAL VERBS

Some signs can be modified in the direction of their path movement to convey a specific meaning. If and how the directionality is annotated varies between corpora.

Auslan Corpus Auslan corpus does not annotate directional verbs in a special way in the primary processing step. In the secondary processing step however more information, such as meaning, grammatical class, sign form and others, is transcribed in the daughter tiers of the Gloss tier, called **RH/LH ID-gloss**. Directional verbs can be categorised as verbs that indicate directionality, glossed as *VIDir* in the tier for grammatical class called **RH/LH-GramCls**. Movement can additionally be coded in the tier for movement called **RH/LH-Move** where HamNoSys can be used. (Johnston, 2019, pp. 64–67)

BSL Corpus This information is annotated on extra tiers not publicly available (same as for repetitions, see Section 4.9). The approach on annotating directional verbs is described in detail in Cormier et al. (2015). *(K. Cormier, personal communication, January 31, 2022*)

Corpus LSFB Directional verbs are marked with an asterisk, e. g. *CALL*, INVITE**. (Sinte et al., 2015, '9. Directional verbs')

Example: The sign 'to inform' with a directional modification and one repetition: *INFORMER*+* (Meurant, 2015, CLSFBI3318.eaf, 00:01:26.118–00:01:26.538)

Corpus NGT Directionally modified signs are marked with the suffix *:1* if the movement is towards the signer and with the prefix *1:* if it is away from the signer, e. g. the gloss for 'ask' in citation form, towards the signer, and away from the signer: *VRAGEN, VRAGEN-A:1, 1:VRAGEN*. (Crasborn et al., 2020, p. 17)

Example: A stretch where the signer signs about him visiting someone, is glossed as *PT-1hand:1 1:BEZOEKEN-A*. (Crasborn et al., 2008b, CNGT0095.eaf, 00:01:08:440–00:01:09.120, https://hdl.handle.net/1839/00-0000-0000-0009-2D73-0)

Corpus FinSL No information found.

Corpus VGT No information available.

DTS Corpus Directional verbs are not glossed in a special way, but with normal ID-glosses. Information on grammatical status or modification can be placed on a separate tier. (Kristoffersen and Troelsgård, 2015, '9. Directional verbs')

DGS Corpus In the DGS Corpus, directionally modified tokens are annotated in more detail using the qualifiers 'source' and 'goal', or 'movement direction'. In contrast, signs modified by a different location are annotated either by using the qualifier 'location' or 'body location'. In the Public DGS Corpus these form differences, like others, are indicated by an asterisk. (Konrad et al., 2012, p. 92; *R. Konrad, personal communication, January 27, 2022*)

Dicta-Sign-LSF-v2 Corpus No information available.

Digging into Signs It is recommended to annotate information on modification or grammatical information on a separate tier. For a further discussion Fenlon et al. (2018) is recommended. (Crasborn et al., 2015a, pp. 5–6)

ECHO Corpus Each hand tier has a child tier for annotating signs modified in their direction and location called **Dir&loc RH/LH**. For the annotation the following codes are used: *r/l-90* meaning to the left or right, close to 90 degrees, *r/l* meaning to the right or left, close to 45 degrees, *lh/rh* meaning to the left or right hand. For modifications of the height the codes *u* for upward, *d* for downward, *a* for ahead, more to the front, *s* for towards or closer to the signer and *p* for toward a person that is present are used. These features can be combined. (Nonhebel et al., 2004a, p. 4)

Example: Next to different codes used as described above in the NGT dataset there is one sign *TELLEN* or *COUNT* in English with a code in the daughter tier that is further specified as *s (head)*. (Woll et al., 2004, NGT_WE_poems.eaf, 00:23:47.750–00:23:38.360)

HSL Corpus In the realm of the dictionary work the type of movement is annotated in more detail by using pictograms, as already mentioned in Section 4.1. (Bartha et al., 2016, p. 5)

LIS Corpus Directional verbs are not specially labelled. (Santoro and Geraci, 2015, slide 22)

PJM Corpus Information on directionality is given on the subtype level. There is no further explanation but an example which suggests, that the information is given by adding information on location and hand configuration at the start and end of a sign, separated by a slash: *POWIEDZIEC/PYTAC (MI) 1 P:B;L:Ø / P:M;L:Ø (POKLEPAC+RAMIE)* could mean the sign for 'say/ask' starting at the signer with a B-handshape moves or patts the arm with a M-handshape. It is noted, that additional tags on a separate tier are planned. (Rutkowski et al., 2015, slide 25)

POLYTROPON No distinct label is used for directional verbs. Based on the used handshape different glosses are assigned. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus Grammatical modification is glossed in separate ID-glosses, e. g. *UČITI* for 'to teach', *UČITI SE* for 'to learn' and *UČITI ME* for 'to teach me'. (Jerko and Vintar, 2015, 'Directional verbs')

Signs of Ireland According to the slides, almost same procedure as BSL Corpus, e. g. *ASK TAKE-OVER*. As BSL Corpus does not specify their approach it stays unclear how the approach exactly looks. (Matthews and Sheridan, 2015, slide 14)

SSL Corpus The annotation of directional verbs is not mentioned in Wallin and Mesch (2018) but in Wallin and Mesch (2015, slide 12). If the change of direction differs from the citation form the gloss is suffixed with the number 1, e. g. *ASK1*. Backwards verbs are not marked specially. The described approach can also be found in the data.

Example: The sign meaning 'get' is signed towards the signer and therefore glossed as *FÅ1*. (Mesch et al., 2012, På arbetsplats, https://teckenspraakskorpus.su.se/#/video/sslc01_004.eaf?q=%2a1&t=333.630)

4.13 CLASSIFIER CONSTRUCTIONS

The following linguistic phenomena is called 'classifiers', 'depicting signs', 'illustrative structures' (in the Dicta-Sign-LSF-v2 Corpus), 'productive signs' (in the DGS Corpus) as well as 'polysynthetic signs' (in the SSLC). Most corpora categorise these non-lexicalised signs into different types.

Auslan Corpus Classifiers are prefixed with *DS*, followed by a code identifying the subtype of classifier: *L* for locative, *M* for movement, *H* for handling, *S* for size and shape or descriptive, *G* for ground. This prefix is extended with a handshape code (and maybe an orientation code) in brackets, followed by a colon and a description of the meaning of the sign, e. g. *DSM(1):HUMAN-MOVES*. (Johnston, 2019, pp. 31–32)

BSL Corpus Classifiers are prefixed with *DS*, like in the Auslan Corpus, but followed by a different coding system identifying the subtype of classifier: *DSEW* for whole entity, *DSEP* for part entity, *DSH* for handling, *DSS* for size and shape. This prefix is extended with a handshape code (see Chapter 6) and a code identifying the movement type: *MOVE* for path movement, *PIVOT* for a change in position of a referent, *AT* for localisations, *BE* for no meaningful movement, although for DSS movement type is not specified. Examples are: *DSEW(1)-AT, DSEP(2)-PIVOT, DSS(1)*.

If the sign is ambiguous in the sense of it being a classifier or a CA, both options can be written into the gloss, separated by a slash, e. g. *DSH(FIST)-MOVE/G:CA:push-pram*.

Type-like classifiers are glossed separately following the conventions of the Auslan Corpus: *DSEW(1-VERT)-MOVE:HUMAN* (Cormier et al., 2017, pp. 10–12)

Example: An example for a whole entity classifier: *DSEW(BENT2-HORI)-MOVE:HUMAN* (Schembri et al., 2017, BF6n.eaf, 00:01:54.732–00:01:55.406).
An example for an ambiguous classifier/CA in the BSL Corpus: *DSH(5)-MOVE/G:CA:BEING-CARRIED-DOWN-STAIRS* (Schembri et al., 2017, BF6n.eaf, 00:02:05.685–00:02:07.161).

Corpus LSFB Classifiers are prefixed with *DS*, as in the Auslan Corpus and the BSL Corpus. (Sinte et al., 2015, '19. Classifier/depicting sign')

Example: A classifier used to describe three images on one sheet of paper: *DS:PHOTO PAPIER* (Meurant, 2015, CLSFBI3318.eaf, 00:00:03.320–00:00:03.752)

Corpus NGT Four types of classifiers are distinguished and glossed as: *MOVE* for a path movement of a referent, *PIVOT* for a change in position of a referent, *AT* for a localisation of a referent, *BE* for no meaningful movement. (Note that these are the same classes that are used in the BSL Corpus to tag the movement type of classifiers.) Information on the handshape is added to the classifier gloss in form of a handshape code (see Chapter 6) with a plus sign in between, e. g. *MOVE+V, PIVOT+B, AT+V, BE+S*. (Crasborn et al., 2020, pp. 20–23)

Shape classifiers are glossed separately as *SHAPE* suffixed with information on the handshape using the same handshape codes (see Chapter 6) as other classifiers, e. g. *SHAPE+B*. If the shape construction is two-handed, but one hand is hold still as a reference point, this hand is glossed *SHAPE-RP*, e. g. *SHAPE-RP+Baby_beak_open*. (Crasborn et al., 2020, pp. 25–26)

For both, classifiers and shape constructions, the actual meaning in context is annotated in the meaning tier. (Crasborn et al., 2020, pp. 20, 25)

Example: A classifier used to describe how something is falling over: *MOVE+5* (Crasborn et al., 2008b, CNGT0117.eaf, 00:00:08.520–00:00:08.720, `https://hdl.handle.net/1839/00-0000-0000-0009-2D6C-D`)

Another classifier describing a woman holding an egg in her hand: *BE+C* (Crasborn et al., 2008b, CNGT0004.eaf, 00:00:14.520–00:00:15.320, https://hdl.handle.net/1839/00-0000-0000-0009-2D5C-5)

Corpus FinSL Classifiers are categorised into six classes and glossed as *_kv*, followed by a code identifying the subtype of classifier: *kk* for whole entity, *kt* for handling, *mk* for shape and size, *ap* for time and place, *ak* for abstract points of reference, *x* for unclear or unclassified cases. (Salonen et al., 2019, pp. 15–16; Salonen et al., 2020, p. 199)

Example: Figure 4.32 shows the annotation of a description of a man walking with a swollen eye. The annotation is limited to basic glosses for gestures (*_ele*) and classifiers for shape and size (*_kvmk*).

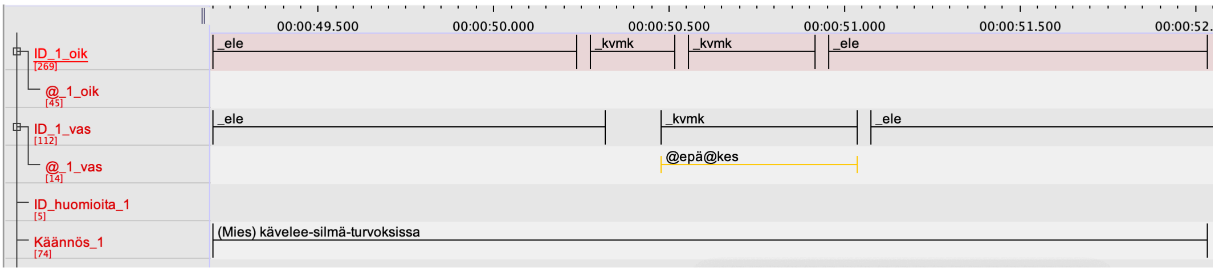


Figure 4.32: *Classifier and gesture glossing in Corpus FinSL (Jantunen, 2018, CFINSL2014_011_03.eaf, 00:00:49.200–00:00:52.040)*

Corpus VGT Classifiers are called 'sign constructions' in the Corpus VGT. They are prefixed with *GC:*, which stands for 'gebarenconstructie' followed by information on the handshape in brackets and a description of the meaning, e. g. *GC: (5) BIJENZWERM*. (Verstraete et al., 2015, 'Depicting signs')

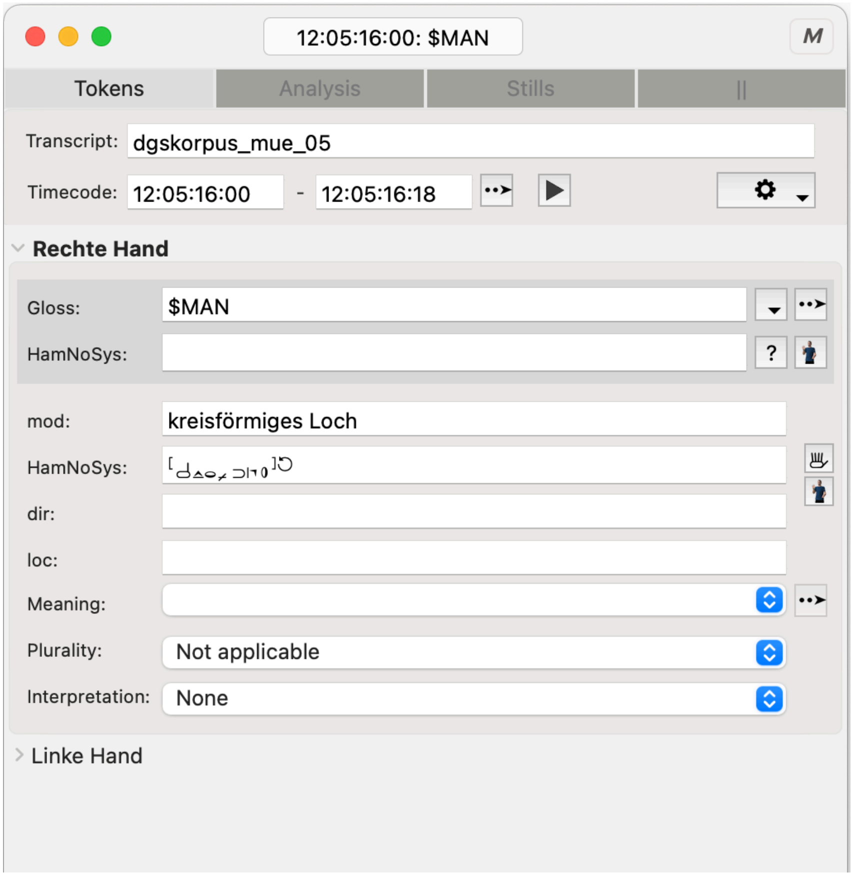
DTS Corpus Glosses for classifiers are prefixed with *PF-*, additionally a description of the meaning and movement is added to a separate child tier. Labelling the movement with category tags like *MOVE, PIVOT, AT, BE* is in consideration. Shape signs have separate ID-glosses without any label, but prefixes and a tier for sign meaning are considered to be added. Up until now, 50 classifier signs and 10 shape signs have been identified and received ID-glosses. (Kristoffersen and Troelsgård, 2015, '19. classifier/depicting signs', '20. Shape constructions')

DGS Corpus In the Public DGS Corpus all occurrences of classifiers or – in the terms of DGS Corpus – productive signs are glossed with *$PROD*. (Konrad et al., 2020b, p. 11)

In the DGS Corpus these tokens are further annotated for meaning in context and at least handshape(s) in HamNoSys. Productive signs are not in the focus of the DGS-Korpus project, as the goal is a corpus-based dictionary. Instead of using the well-known classifier categories the Hamburg research team developed a different approach in analysing any iconic sign by determining the image producing technique of each hand (see Ebling et al., 2015, p. 44). (*R. Konrad, personal communication, January 24, 2022*)

Example: A classifier used to describe holes that were made into a wall glossed as *$PROD** in the Public DGS Corpus. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_mue_05: Experience Report, 00:01:09:19–00:01:09:37, https://doi.org/10.25592/dgs.corpus-3.0-text-1210208)

Figure 4.33 shows the token view of the same classifier in the DGS Corpus with more information on the context (in english 'circular hole' and a HamNoSys notation.



12:05:16:00: $MAN

Transcript: dgskorpus_mue_05

Timecode: 12:05:16:00 - 12:05:16:18

Rechte Hand

Gloss: $MAN

HamNoSys:

mod: kreisförmiges Loch

HamNoSys:

dir:

loc:

Meaning:

Plurality: Not applicable

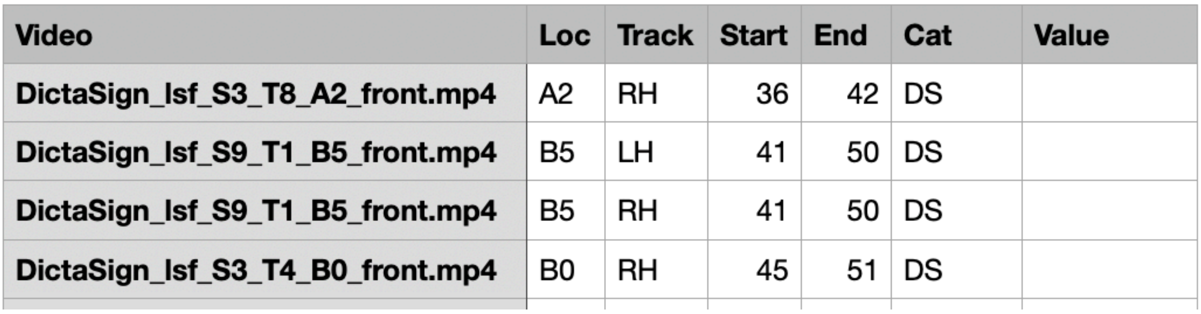
Interpretation: None

Linke Hand

Figure 4.33: *View of the token entry of a classifier in the DGS Corpus (Konrad et al., 2020a, DGS Corpus, dgskorpus_mue_05: Experience Report, 12:05:15:45–12:05:15:48, https://doi.org/10.25592/dgs.corpus-3.0-text-1210208)*

Dicta-Sign-LSF-v2 Corpus Classifiers are annotated with the category *DS*. For the future it is planned, that a value describing the linguistic roles of the hands is added. (Braffort, 2019, p. 5)

Example: All classifiers are categorised as *DS* with no further value, as seen in Figure 4.34.



Video	Loc	Track	Start	End	Cat	Value
DictaSign_lsf_S3_T8_A2_front.mp4	A2	RH	36	42	DS	
DictaSign_lsf_S9_T1_B5_front.mp4	B5	LH	41	50	DS	
DictaSign_lsf_S9_T1_B5_front.mp4	B5	RH	41	50	DS	
DictaSign_lsf_S3_T4_B0_front.mp4	B0	RH	45	51	DS	

Figure 4.34: *Classifiers are all categorised as DS in the Dicta-Sign-LSF-v2 Corpus(LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)*

Digging into Signs The conventions of BSL Corpus and Corpus NGT are described without further discussion, but the more elaborate approach of the BSL Corpus is mentioned positively. (Crasborn et al., 2015a, pp. 7–8)

Funded by the Horizon 2020 Framework Programme of the European Union

ECHO Corpus Classifiers are prefixed with *p-* in brackets, 'p' stands for 'poly' and means 'many meaning components', e. g. *(p-)vehicle-be-located*. (Nonhebel et al., 2004a, p. 2)

Example: classifier used to show how a wolf is walking in the dataset for SSL glossed as *(p-) tass-gå* or in English *(p-) claw-walk*. (Bergman and Mesch, 2004, SSL_JI_fab1.eaf, 00:01:15.290–00:01:16.090)

HSL Corpus No information available.

LIS Corpus The meaning of the classifier is suffixed with *-CL*, e. g. *PASSARE-CL*. (Santoro and Geraci, 2015, slides 15, 19, 24)

PJM Corpus Classifiers are prefixed with *$:KL:* followed by the hand configuration and an approximate meaning in brackets, e. g. *$:KL: 4 (DRZEWA)* for depicting trees. Shape constructions and type-like classifiers are also prefixed with *$:KL:*, e. g. *$:KL: B (JEZDZIC/POJAZD/AUTO/POCIAG/SAMOLOT/PARKING)* describing different vehicles and the ability to park. (Rutkowski et al., 2015, slides 34–36)

POLYTROPON Classifier constructions are annotated on a morphophonemic level and a semantic level. The morphophonemic level involves information on the handshape, whether the classifiers are one or two-handed and in case of two-handed classifiers whether both hands produce the same or two different handshapes and which type of handshape is associated with each hand. On the semantic level classifiers are grouped into six categories based on their semantic function: *[entity], [handling], [body part], [lexicalized], , [predicative] and [SASS]*, whereas the last one is further divided into *[static] and [tracing]*. (Efthimiou et al., 2018, p. 41; *E. Efthimiou, personal communication, January 24, 2022*)

Example: From the data it becomes apparent that classifiers are glossed as *$MAN* followed by a handshape code with the above described information on separate tiers. In Figure 4.35 we see the annotation for two classifiers with different handshapes (A-hand and 5C-hand), the first one signed with both hand with the same handshape, the second with one hand.

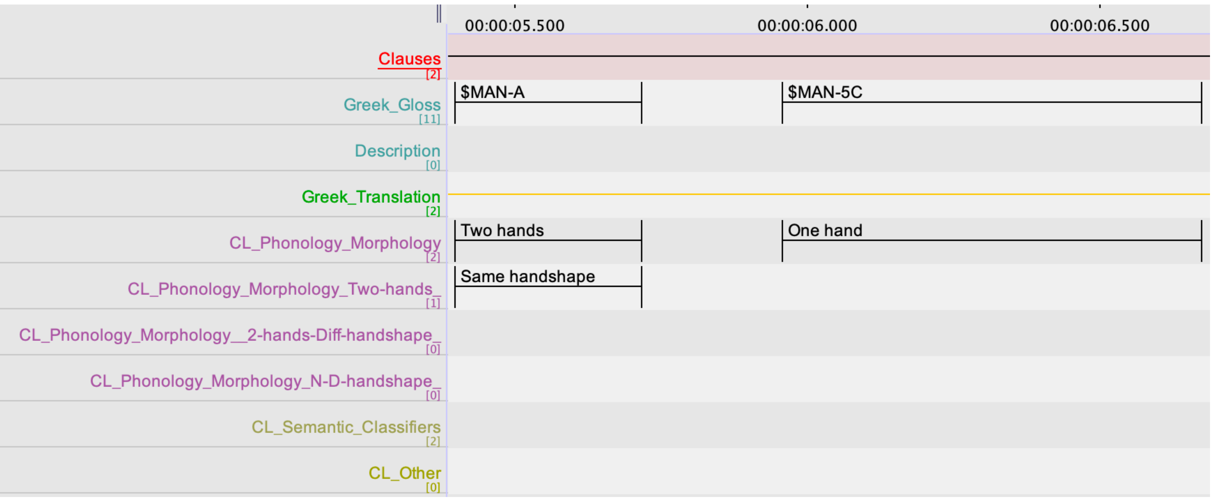

00:00:05.500
00:00:06.000
00:00:06.500
Clauses [2]
Greek_Gloss [11]
$MAN-A
$MAN-5C
Description [0]
Greek_Translation [2]
CL_Phonology_Morphology [2]
Two hands
One hand
CL_Phonology_Morphology_Two-hands_ [1]
Same handshape
CL_Phonology_Morphology__2-hands-Diff-handshape_ [0]
CL_Phonology_Morphology_N-D-handshape_ [0]
CL_Semantic_Classifiers [2]
CL_Other [0]

Figure 4.35: *Classifier handshapes in POLYTROPON. (ILSP Athena Research Center, 2018, Clarin_kampoyris_ex_3_HD_SVQ1.eaf, 00:00:05.398–00:00:06.682)*

SIGNOR Corpus Two ID-glosses *iconic movement* and *iconic shape*, both with very high frequency. (Jerko and Vintar, 2015, 'Classifier/depiciting signs Shape constructions Type-like classifier-/depiciting signs')

Signs of Ireland Classifiers are categorised into five kinds, four of which are prefixed with *CL-*: *CL-MOVE*, *CL-PIVOT*, *CL-AT*, *CL-HANDLE*. The fifth is for shape constructions and glossed as *SHAPE* followed by a handshape code. In the provided pictures there appear to be further forms of classifiers, namely *CL-ISL-L, CL-INDEX*. (Matthews and Sheridan, 2015, slides 9, 19)

SSL Corpus Classifiers, or in the terms of SSLC polysynthetic signs, are suffixed with *@p*. The basic principle to annotate them is to name the category of the classifier followed by a handshape code in brackets, e. g. *ENTITET(J)+@p*. The plus symbol indicates that the words denote semantic parts. Different classifiers can be combined by naming one after the other in one gloss, e. g. *VARELSE(Vb)+FÖRFLYTTA-FRÅN+ENTITET(J)@p* meaning 'being', 'move from' and 'entity' and describing a person jumping down from a scene. The exact meaning of the classifier is annotated in the child tier for meaning called **Betydelse_DH/NonDH**.

The categories for classifiers are: *VARELSE* for 'being', *ENTITET* for 'entities', *GREPP* for 'grip', *HAND* for 'hand', *VÄTSKA* for 'liquid', *KOLLEKTIV* for 'collective' and *FORM* for 'form'. Each of these categories has different values with which they can be combined, e. g. a still standing entity can be glossed as *ENTITET+SILLASTÅENDE*. (Wallin and Mesch, 2018, pp. 16–19)

Noun classifiers are suffixed with *@kl*, e. g. *PERSON@kl, RUND@kl, FÄLT@kl* for 'person', 'round' and 'field'. (Wallin and Mesch, 2018, pp. 15, 47)

Example: The signer uses a classifier to describe the sole of the shoe he was wearing. The classifier is glossed as *FORM(Jv)+BESKRIVNING+ENTITET(J)@p*, describing the classifier to be a form, a description and an entity, with a Jv-handshape on the dominant hand and a J-handshape on the non-dominant hand. (Mesch et al., 2012, Från Örebrotiden, https://teckenspraakskorpus.su.se//#/video/sslc01_007.eaf?q=%2a%2B%2a&t=302.270)

4.14 CONSTRUCTED ACTION

A CA is a sequence where the signer copies or quotes an action or expression of someone or something other. CA is also known as 'enactment'. Some corpora annotate the manual signs only, others use special markers or labels for these sequences.

Auslan Corpus CA and constructed dialogue are annotated on a separate tier called **CA**. Glosses are prefixed with *CA* followed by a colon and the name of the person or entity that is enacted, e. g. *CA:BOY*. Constructed Dialogue (CD) follow the same scheme, but with the prefix *CD* instead of *CA*, e. g. *CD:BOY*. (Johnston, 2019, pp. 59–62)

In the course of the tertiary processing signs that occur during a CA are tagged on the tier **CA co-occurrences**. (Johnston, 2019, p. 101)

BSL Corpus In the BSL Corpus CA is seen as a form of gesture and therefore glossed with the prefix *G:CA* followed by a short description of the meaning, e. g. *G:CA:HOLD-HANDS-UP-IN-FRIGHT*. As these tokens are not lexicalised they are not included into the Signbank. (Cormier et al., 2017, p. 13)

Example: CA in the BSL Corpus: *G:CA:PUTTING-ARMS-AROUND-SOLDIER* (Schembri et al., 2017, BF6n.eaf, 00:01:58.803–00:02:01.044)

Corpus LSFB There are no conventions on glossing CA in the found descriptions. They may be annotated as gestures. (Sinte et al., 2015, '24. Manual constructed action')

Corpus NGT CA is glossed with the generic gloss *imitatie*, meaning 'imitation'. There may be cases where some occurrences are glossed with the gloss for gestures *%* as this has not yet been double-checked. (*O. Crasborn, personal communication, January 29, 2022*)

Example: The signer tells a story about them falling into the water and being all wet, which they show using a CA glossed as *imitatie*. (Crasborn et al., 2008b, CNGT0004.eaf, 00:01:36.240–00:01:36.920, available at https://hdl.handle.net/1839/00-0000-0000-0009-2D5C-5)

Corpus FinSL CA is only mentioned in the realm of translation not regarding the glosses.

Corpus VGT No information available.

DTS Corpus No rules yet. (Kristoffersen and Troelsgård, 2015, '24. Manual constructed action')

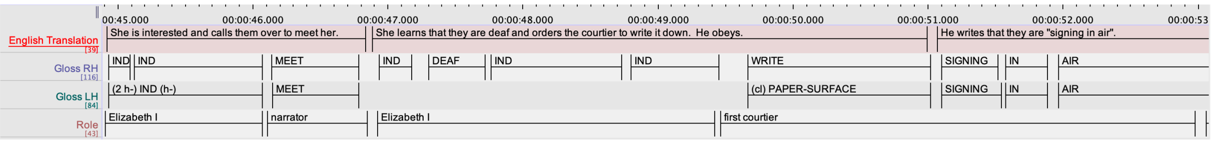
DGS Corpus In the DGS Corpus CA is annotated in the comment tier of the participants with the label 'CA'. The content of the comment tier is not shown in the Public DGS Corpus. *R. Konrad, personal communication, January 27, 2022*

Dicta-Sign-LSF-v2 Corpus No information available.

Digging into Signs Both approaches of BSL Corpus and Corpus NGT are described without further discussion. (Crasborn et al., 2015a, p. 8)

ECHO Corpus An extra tier **Role** is created to annotate the role the signer takes while doing a CA. (Nonhebel et al., 2004a, p. 9)

Example: In the transcript excerpt shown in Figure 4.36 the signer changes from the role of Elizabeth I to the narrator of the story, back to Elizabeth I and then the role of the first courtier. This is annotated in the tier **Role**.



00:45.000 | 00:00:46.000 | 00:00:47.000 | 00:00:48.000 | 00:00:49.000 | 00:00:50.000 | 00:00:51.000 | 00:00:52.000 | 00:00:53

English Translation [39]: She is interested and calls them over to meet her. | She learns that they are deaf and orders the courtier to write it down. He obeys. | He writes that they are "signing in air".

Gloss RH [116]: IND | IND | MEET | IND | DEAF | IND | IND | WRITE | SIGNING | IN | AIR

Gloss LH [84]: (2 h-) IND (h-) | MEET | (cl) PAPER-SURFACE | SIGNING | IN | AIR

Role [43]: Elizabeth I | narrator | Elizabeth I | first courtier

Figure 4.36: *Annotation of CA in the ECHO Corpus. Role changes are annotated in the **Role** tier. (Woll et al., 2004, BSL_PS_poem3.eaf, 00:00:44.920–00:00:46.840)*

HSL Corpus No information available.

LIS Corpus Occurrences of CA are suffixed with *-CL*. This means that CA and the use of classifers are labelled in the same way and therefore are not identifiable via the gloss only. (Santoro and Geraci, 2015, slide 24)

PJM Corpus No information available.

POLYTROPON Within the POLYTROPON no rules on constructed actions are needed, as they don't appear in the data. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus No information available.

Signs of Ireland CA is glossed as *G-CA* and can contain non-manual features too. (Matthews and Sheridan, 2015, slide 19)

SSL Corpus CA is glossed with a description of the shown behaviour followed by the marker *@ca*. (Wallin and Mesch, 2018, p. 21)

Example: From the data it comes apparent that conventions seem to have been changed to the marker *@ka*. One signer shows how a person is sitting with folded arms glossed as *SITTSTÄLLNING@ka* meaning 'sitting position'. (Mesch et al., 2012, Äventyr, `https://teckenspraakskorpus.su.se/#/video/sslc01_110.eaf?q=%2a%40ka&t=87.080`)

Funded by the Horizon 2020 Framework Programme of the European Union

4.15 SPECIAL GLOSSES

This section presents special glosses, most of them used in only one corpus.

Auslan Corpus No special glosses identified.

BSL Corpus No special glosses identified.

Corpus LSFB No special glosses identified.

Corpus NGT In NGT there is a sign usually accompanied with the mouthing 'op' and possible to mark directionally. This sign is glossed as *HOP*, meaning 'Hulpwerkwoord OP' or in English 'auxilary OP'. (Crasborn et al., 2020, p. 17)

Cases where the signer moves their hands without a fixed form or meaning, e. g. merely touching oneself for scratching or rubbing are glossed as *self-touch*. If the movement does have a communicative intention, as showing a part of the body to imitate a manual action of another person it is glossed as *imitation* (see also Section 4.14. (Crasborn et al., 2020, p. 29)

Example: An instance where the signer is scratching his nose is glossed as *self-touch*. (Crasborn et al., 2008b, CNGT0214.eaf, 00:01:17.280–00:01:17.480, https://hdl.handle.net/1839/00-0000-0000-0008-F0FF-1)

Corpus FinSL No special glosses identified.

Corpus VGT No special glosses identified.

DTS Corpus No special glosses identified.

DGS Corpus Signs not known to the team are glossed with the assumed meaning suffixed with *$CANDIDATE* to mark it as a lexical sign candidate, followed by a code for the region where the data was collected and a running number, e. g. *AUGUST-$CANDIDATE-MST05* for a sign for 'August' collected in the data collection region of Münster. With growing evidence these glosses will be changed to normal type glosses.

Multi-channel signs, also known as idiomatic signs or 'Spezialgebärden' in German are not glossed specially but as regular lexical signs in the Public DGS Corpus. In the DGS Corpus these types are prefixed with *$SPEZIAL-* in German and *$SPECIAL-* in English.

Extra linguistic manual activity like scratching or brushing of clothes is glossed as *$$EXTRA-LING-ACT^*. The segmentation and labelling of extra linguistic manual activity was not done from the beginning of the lemmatisation process, but was introduced in July 2016. The reason was that automatic recognition of manual activity will become a quality assurance step in finding lemmatisation gaps. Already identified manual activity will reduce the number of possible gaps. (Konrad et al., 2020b, pp. 10, 16; *R. Konrad, personal communication, January 24, 2022*)

Example: A sign for 'farmer' used by a signer from the region of Stuttgart glossed as: *FARMER-$CANDIDATE-STU57^* (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_stu_04: Experience of Deaf Individuals, 00:01:19:25–00:01:19:40, https://doi.org/10.25592/dgs.corpus-3.0-text-1211515)

Another signer starts a sequence of signing with a palm-up gesture, followed by scratching his face glossed as *$GEST-OFF $$EXTRA-LING-ACT^* (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_mue_05: Experience Report, 00:05:36:46–00:05:37:28, https://doi.org/10.25592/dgs.corpus-3.0-text-1210208)

Funded by the Horizon 2020
Framework Programme of the European Union

Dicta-Sign-LSF-v2 Corpus No special glosses identified.

Digging into Signs No special glosses identified.

ECHO Corpus No special glosses identified.

HSL Corpus No special glosses identified.

LIS Corpus No special glosses identified.

PJM Corpus No special glosses identified but it stands out that PJM Corpus has more blank spaces in the glosses than other corpora.

POLYTROPON Expressions specific to GSL without any direct translation equivalent are annotated as regular lexical signs but marked as 'GSL special expressions' in the lexical database. (Efthimiou et al., 2018, pp. 41–42)

SIGNOR Corpus No special glosses identified.

Signs of Ireland No special glosses identified.

SSL Corpus Object pronouns are glossed separately in the SSLC with the gloss *OBJPRO*. If the palm is oriented to the signer, it is glossed as *OBJPRO1*. Combinations with suffixes *.FL, >person* are possible. (Wallin and Mesch, 2018, p. 14)

Possessive expressions are glossed with *POSS* following the same rules as pointing signs and pronouns. (Wallin and Mesch, 2018, p. 14)

The gloss *AVGRÄNS* is mentioned extra in the conventions. It is a sign to define spaces and to explain if something is inside or outside. The findings regarding this gloss are preliminary and need further analysis. (Wallin and Mesch, 2018, p. 26)

Example: The signer tells a story where a person is leaving some charge which is exploding while the person is hiding around a corner. To indicate where the wall was the sign *AVGRÄNS* is used. In the same transcript *OBJPRO* and *POSS* are used several times. (Mesch et al., 2012, Från Örebrotiden, `https://teckenspraakskorpus.su.se//#/video/sslc01_007.eaf?q=avgr%C3%A4ns&t=478.866`)

4.16 GESTURES

Some corpora gloss gestures differently depending on the degree of conventionalisation of the gesture. If the same form is repeatedly used with the same meaning gestures get their own glosses in some corpora. Others use one gloss for all kinds of gestures. One gesture with a high lexical frequency is the so-called ‘palm-up gesture’. It is used in different contexts usually with an open flat hand with the palm showing upwards. This gesture is mentioned in most conventions and has an individual gloss in most corpora.

Auslan Corpus Prefixed with *G*, followed by a brief description of the meaning in context, e. g. *G:HOW-STUPID-OF-ME*. For the most common and reoccurring types of gestures the form of the gesture is added to the gloss in brackets, e. g. *G(5-UP):WELL* for the palm-up gesture. For important non-sign non-manual gesture activity an exception is made and they are transcribed in the **ID-gloss** tier. The glosses for non-manual gestures are labelled with *(NMS)* after the prefixed *G*. For non-manual gestures that involve the mouth the prefix *G* is not used, but *M* for mouthings and *MG* for mouth gestures, e. g. *G(NMS):LOOK-SURPRISED, M:BECAUSE, MG:BLOW*. (Johnston, 2019, pp. 41–45)

BSL Corpus Gestures are prefixed with *G* and followed by a description of the meaning, e. g. *G:HOW-STUPID-OF-ME*, as in the Auslan Corpus. Lexicalised emblems are glossed as lexical signs without the gesture prefix and collected in the Signbank, e. g. *GOOD*. It is mentioned that this is not consistent, e. g. *G:FUCK-OFF*.

The palm-up gesture is also seen as gesture and glossed as *G:WELL*, similar to the Auslan Corpus but without information on the handshape. As mentioned, CA is also glossed as gesture. (Cormier et al., 2017, p. 13)

Example: A gesture for getting attention in the BSL Corpus: *G:HEY* (Schembri et al., 2017, BF8n.eaf, 00:00:00.966–00:00:01.600)

Corpus LSFB All gestures are glossed as *GSIGN*, only the palm-up gesture has its own gloss, called *PALM-UP*. (Sinte et al., 2015, ‘22. Gestures’, ‘23. Palm up’)

Example: Examples for an uncategorised gesture, a palm-up gesture and a gesture that may be annotated as a CA later on. : *GSIGN, PALM-UP, GSIGN* (Meurant, 2015, CLSFBI0301.eaf, 00:11:09.038–00:11:09.325, 00:11:21.492–00:11:21.920 and 00:12:38.419–00:12:39.267).

Corpus NGT Gestures with a fixed meaning are glossed with their own ID-glosses, without further marking. Other gestures are glossed with a percentage character *%*. (Crasborn et al., 2020, p. 3)

The palm-up gesture is glossed as *PO*, independent of the exact appearance. If the palm is facing down the gloss *PV* is used, if it is facing downwards *PB* is used. (Crasborn et al., 2020, p. 27)

Example: In the data we could not find the use of the gloss *%*.

Corpus FinSL Two kind of gestures are distinguished in the Corpus FinSl: established and non-established gestures. Established gestures get their own ID-gloss suffixed with *_ele* and are added to the Signbank, e. g. *HEI-KUULE_ele, KÄMMEN-YLÖS_ele, KÄMMEN-ALAS_ele* being a gesture for attention, palm-up gesture and palm-down gesture. Non-established gestures are all glossed as *_ele* without further description. (Salonen et al., 2019, p. 20; Salonen et al., 2020, p. 199)

Funded by the Horizon 2020
Framework Programme of the European Union

Example: Examples for established gestures: *ETUSORMI-YLÖS_ele* for raising one's finger and *MENE-POIS_ele@sbb* to shoo away someone (Jantunen, 2018, CFINSL2014_011_03.eaf, 00:00:40.440–00:00:40.800, 00:03:42.000–00:03:42.200)

Corpus VGT Same approach as Auslan Corpus, but maybe with spaces in between the different information in the gloss, although this could be just a layout issue from the poster, e. g. *G:(5, palm up) EXACTLY, G: (1) AH!*. Common gestures are glossed with own ID-glosses prefixed with *G*. (Verstraete et al., 2015, 'Gestures')

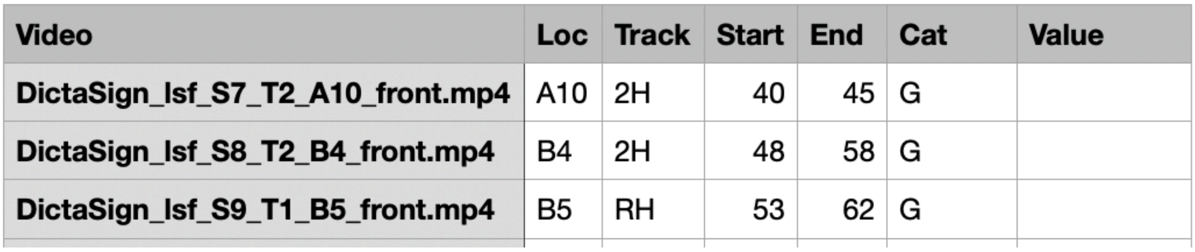
DTS Corpus No rules yet, except the gloss for the palm-up gesture: *PRESENTATION-GESTURE*. (Kristoffersen and Troelsgård, 2015, '22. Gestures', '23. Palm up')

DGS Corpus Gestures are glossed as *$GEST^* for manual gestures and *$GEST-NM^* for non-manual gestures. Beside these collective types, there are several gesture type entries specified by form and meaning similar to lexical signs, e. g. the palm-up gesture *$GEST-OFF^* and others like *$GEST-TO-PONDER1^*, *$GEST-DECLINE1^*, *$GEST-NM-NOD-HEAD1^*, *$GEST-NM-SHAKE-HEAD1^*. This differentiation is tentative and has not yet been reviewed. (Konrad et al., 2020b, p. 15; *R. Konrad, personal communication, January 27, 2022*)

Example: The signer is using a gesture without specified meaning followed by a palm up gesture glossed as: *$GEST^ $GEST-OFF^** (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_mue_05: Experience Report, 00:06:39:17–00:06:39:29, https://doi.org/10.25592/dgs.corpus-3.0-text-1210208)

Dicta-Sign-LSF-v2 Corpus Gestures are annotated with the category *G* without any further value. The annotation manual stresses the fact, that this category should be considered with caution, as some gestures can be considered as lexical signs as well. (Braffort, 2019, pp. 5–6)

Example: All gestures are categorised as *G* with no further value, as can be seen in Figure 4.37.



Video	Loc	Track	Start	End	Cat	Value
DictaSign_lsf_S7_T2_A10_front.mp4	A10	2H	40	45	G	
DictaSign_lsf_S8_T2_B4_front.mp4	B4	2H	48	58	G	
DictaSign_lsf_S9_T1_B5_front.mp4	B5	RH	53	62	G	

Figure 4.37: *Gestures are all categories as G with no further value in the the Dicta-Sign-LSF-v2 Corpus(LIMSI, 2020, Dicta-Sign-LSF_Annotation.csv)*

Digging into Signs The conventions of BSL Corpus and Corpus NGT are described without further discussion. It is noted that the palm-up gesture receives its own gloss in the Corpus NGT called *PO*. (Crasborn et al., 2015a, p. 8)

ECHO Corpus Gestures are prefixed with *g-* in brackets followed by the meaning of the gesture in small caps between quotes, e. g. *(g-)"well"*. The palm up gesture is glossed as *(g-)pu*. (Nonhebel et al., 2004a, p. 3)

Example: The signer is ending one stretch of signing with a gesture glossed as *(g-)"c'est la vie"* or in English *(g-)"that's life"*. (Crasborn et al., 2004, NGT_WE_poems.eaf, 00:04:10.950–00:04:13.000)

HSL Corpus There are no information on the annotation of gestures within the HSL Corpus, but the multimodal analysis from the *HuComTech's gesture research project* by Abuczki (2013) is mentioned. (Bartha et al., 2016, p. 4)

LIS Corpus Gestures are glossed as *gesture*. (Santoro and Geraci, 2015, slide 24)

PJM Corpus Gestures are prefixed with *G:* followed by the handshape and an approximate meaning in brackets, e. g. *G: 1 (AUTOSTOP), G: A+A (SMUTNY)* the latter being a sad gesture. This is a similar approacht to the Auslan Corpus but with spaces in between the information and different use of brackets. Three fixed gestures have their own glosses: *@* for dragging attention, *&* to indicate 'never mind' and^ for pauses. Additionally the palm-up gesture is glossed as *%*. (Rutkowski et al., 2015, slide 37–38)

POLYTROPON Gestures are glossed as *$GEST* without further information. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus For gestures a distinction is made between gestures and fillers, both of which the annotators describe intuitively. The frequency of gestures is 500 signs in the corpus. The most frequent fillers are listed on the poster. The palm up gesture is treated separately to mark the end of utterances (Jerko and Vintar, 2015, 'Gestures', 'Top ten fillers with frequencies', 'Palm up')

Signs of Ireland Gestures are prefixed with *gesture* in brackets followed by a hyphen and an explanation, e. g. *(gesture)-WELL*. It becomes not clear how the palm-up gesture is glossed. (Matthews and Sheridan, 2015, slide 19)

SSL Corpus Gestures are preliminary annotated with names and the suffix *@g*. The palm-up gesture is glossed accordingly as *PU@g* and the gesture to call someone's attention as *PÅKALLA-UPPMÄRKSAMHET@g*.

Example: In the data we find different gestures one of them seems to incorporate a negative meaning glossed as *PU-NEG@g*. (Mesch et al., 2012, Var Är du, grodan?, https://teckenssprakskorpus.su.se/#/video/sslc02_308.eaf?q=%2a%40g&t=64.584)

Some other gestures meaning 'hello', 'calm', 'wait' and a palm-up gesture meaning 'what more' in the corpus are glossed as *HEJ@g, LUGNA@g, VÄNTA@g, PU-VAD-MERA@g*. (Mesch et al., 2012)

Funded by the Horizon 2020
Framework Programme of the European Union

4.17 OTHER LANGUAGES

While signing some signers use signs from other SLs or words from spoken languages. Most corpora mark these instances, some only in the lexical database others directly in the gloss. A special kind of language mix are cued-speech systems. Most cued-speech systems were developed to teach articulation of phonemes to deaf children. The use of these cued-speech systems is often labelled separately.

Auslan Corpus Information on the language used (if not Auslan) is appended after a period, e. g. *COOL.ASL*.

The use of cued speech systems are annotated by appending information on the system used after a period, e. g. *GAVE.SE* for Signed English. (Johnston, 2019, p. 24)

BSL Corpus Signs from other SLs, as well as the use of cued speech systems, are glossed with ID glosses and taken into the Signbank where they are tagged as doubtful, to indicate that they belong to another language. Additionally, a note is made that it may be a borrowing. (Cormier et al., 2017, p. 7)

Corpus LSFB No information available.

Corpus NGT Possible uses of other SLs are not glossed separately but as an ID-gloss from NGT. The Corpus NGT team didn't want to make a judgement as to whether a sign is from NGT or not. Some signs are coded for also occurring in another SL. The use of spoken languages, i. e. in fingerspellings or mouthings, is not specifically tagged. (*O. Crasborn, personal communication, January 29, 2022*)

Corpus FinSL Signs from other SLs are glossed with ID glosses in the source language and marked as such in the associated daughter tiers with the code *@lv*. Information on the source language is also added to the glossary, e. g. the *POWER* is from American Sign Language (ASL). (Salonen et al., 2019, pp. 32–33; Salonen et al., 2020, p. 199)

Example: No example found in the data.

Corpus VGT No information available.

DTS Corpus No rules for other languages, except for the mouth-hand-system. The target word rendered through the mouth-hand-system is written in conventional spelling suffixed with *(M)*, e. g. *Sahara(M)*. (Kristoffersen and Troelsgård, 2015, '17.b Mouth-Hand-System')

DGS Corpus Signs from other SLs are glossed with gloss names from the surrounding vocal language and information on the source language as a suffix, e. g. *NO-ASL1* where the 'NO' stays in English also for the German gloss version or *GERMANY-INTS1* for a sign from International Sign. (Konrad et al., 2020b, p. 10)

Occurrences of cued speech are all glossed *$CUED-SPEECH^*. If the cued speech sign has been lexicalised it receives a separate child gloss connected to the parent gloss for cued speech, e. g. the gloss *WHITE11* is a child type of the parent gloss *$CUED-SPEECH^*. (Konrad et al., 2020b, p. 16)

If German is used in the form of a (voiceless) oral articulation without an accompanying sign the gloss *$ORAL^* is used in the tier for signs and the articulated word is added in the tier for mouthings and mouth gestures. (Konrad et al., 2020b, p. 15)

Funded by the Horizon 2020 Framework Programme of the European Union

Example: The signer using an ASL sign for 'understand' glossed as: *UNDERSTAND-ASL** in the English and German version of the transcript. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_fra_05: Experience of Deaf Individuals, 00:02:50:24–00:02:50:49, https://doi.org/10.25592/dgs.corpus-3.0-text-1212176)

In the same transcript, the two signers use the letter 'H' from the cued-speech system to refer to a city, the cued-speech is glossed as *$CUED-SPEECH** with the mouthing *husum* which is the name of the city. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_fra_05: Experience of Deaf Individuals, 00:12:49:27–00:12:50:40 and 00:12:53:06–00:12:53:30 https://doi.org/10.25592/dgs.corpus-3.0-text-1212176)

Figure 4.38 shows how an instance of a signer articulating the word 'Slovakia' only orally instead of using a name sign is glossed.

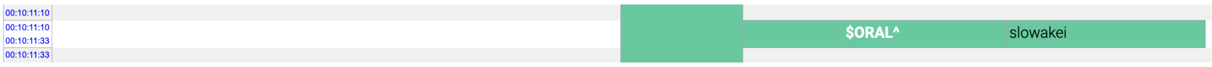

00:10:11:10
00:10:11:10
00:10:11:33
00:10:11:33
$ORAL^
slowakei

Figure 4.38: *The word 'Slovakia' articulated orally in the Public DGS Corpus. (Konrad et al., 2020a, Public DGS Corpus, dgskorpus_ber_02: Free Conversation, 00:10:11:10–00:10:11:33, https://doi.org/10.25592/dgs.corpus-3.0-text-1413703)*

Dicta-Sign-LSF-v2 Corpus No information available.

Digging into Signs No information available.

ECHO Corpus No information available.

HSL Corpus No information available.

LIS Corpus No information available.

PJM Corpus No information available.

POLYTROPON Within the POLYTROPON no rules on other used languages are needed, as they don't appear in the data. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus No information available.

Signs of Ireland No information available.

SSL Corpus No information available.

Funded by the Horizon 2020 Framework Programme of the European Union

4.18 FORM DESCRIPTION/PHONOLOGY

Some corpora annotate phonological information on signs (handshape, orientation, location, movement) directly in the transcripts other collect this information in the lexical database. Additionally to the citation form some corpora also annotate deviations of a token from the type directly in the gloss by adding labels or in separate tiers. While some corpora use a phonetic writing system like HamNoSys, others use a free description or controlled vocabulary.

Auslan Corpus Phonological or phonetic information can be transcribed in the tier **LH/RH-Transcript** and its daughter tiers. HamNoSys, other dedicated notation system or – as with pointing signs – codes can be used for this. (Johnston, 2019, pp. 64–65)

BSL Corpus In principle, both tiers for ID-glosses (left and right hand) are accompanied by daughter tiers for handshape, location, movement, other phonology, and orientation, but these tiers are not publicly available. In the case of depicting signs the handshape is part of the gloss (see Section 4.13). (Cormier et al., 2017, pp. 15–16; *K. Cormier, personal communication, January 31, 2022*)

In the BSL Signbank further information on phonology can be stored and seen by registered users.

Corpus LSFB No information available.

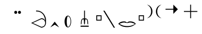
Corpus NGT To describe the phonetics of manual signs seven tiers are provided, although they are not systematically used yet, nor do they have controlled vocabularies, with exception of the tier for phonetic reduction. (Crasborn et al., 2020, p. 31)

Corpus FinSL As already mentioned in Section 4.2, are the variants labelled with information on their handshape, movement, location, and orientation. But as only one parameter is described this is not a full annotation of the form of a sign.

Corpus VGT No information available.

DTS Corpus Due to limited resources signs are only represented with glosses without any form description. (Troelsgård and Kristoffersen, 2018, p. 195)

DGS Corpus Each type or subtype is stored in the lexical database with a HamNoSys notation of its citation form. Form deviation of a token is annotated either by using HamNoSys symbols to spot the relevant form aspect or by marking the token as deviant (in German: 'a' for 'abweichend'). As already mentioned HamNoSys is also used in the DGS corpus to indicate the handshape(s) of productive signs. In the types list of the Public DGS Corpus a video of the citation form of each type is given in combination with the HamNoSys notation. (Konrad et al., 2020b, p. 7; *R. Konrad, personal communication, January 27, 2022*)

Example: The type *KATZE1A* for 'cat' is stored in iLex with the English gloss *CAT* and the HamNoSys  describing its exact form. (Konrad et al., 2020a, type 'CAT1A', https://doi.org/10.25592/dgs.corpus-3.0-type-16495)

Dicta-Sign-LSF-v2 Corpus No information available.

Digging into Signs No information available.

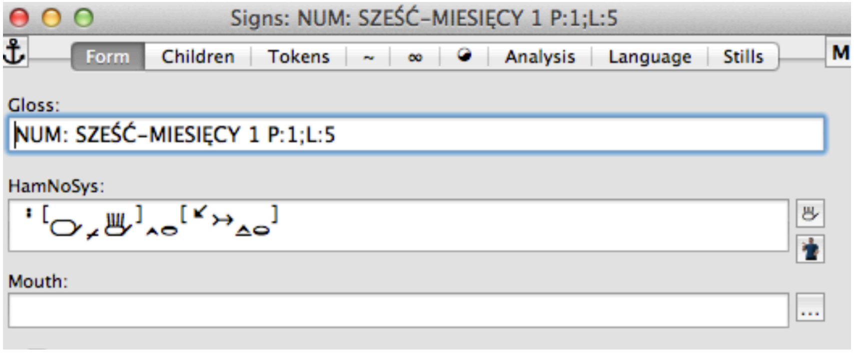
ECHO Corpus Location is annotated in an extra tier, see Section 4.12.

HSL Corpus As already mentioned in Section 4.1 is the handshape annotated in more detail by using pictograms for the dictionary work. (Bartha et al., 2016, p. 5)

LIS Corpus Information on the phonology is added on extra tiers. (Santoro and Geraci, 2015, slide 16)

PJM Corpus Information on the form of the sign is given for each type and subtype via a HamNoSys transcript. (Rutkowski et al., 2015, slide 43)

Example: Figure 4.39 shows an example of a HamNoSys transcript.



Gloss: NUM: SZEŚĆ–MIESIĘCY 1 P:1;L:5

HamNoSys:

Mouth:

Figure 4.39: *Incorporated number in the PJM Corpus with information on the form of the sign specified in HamNoSys (Rutkowski et al., 2015, slide 29)*

POLYTROPON Form descriptions of the sign are not provided in the transcripts but in the POLYTROPON lexical database in the form of a HamNoSys notation for the manual activity and a notation of the non-manual activity with the SiS-Builder notation tool. (Efthimiou et al., 2018, p. 41)

SIGNOR Corpus No information available.

Signs of Ireland No information available.

SSL Corpus No information available.

4.19 OTHER PROPERTIES

This section contains all further topics that are annotated in individual corpora.

Auslan Corpus Indecipherable signs are glossed as *INDECIPHERABLE*. (Johnston, 2019, p. 46) False starts are indicated as a suffix in brackets, e. g. *BOY(FALSE-START)*. (Johnston, 2019, p. 50)

BSL Corpus If uncertainties exist on which ID-gloss is the right one for a single token, two or more glosses can be used separated by a forward slash, e. g. *LOOK/THINK*. If the identity of the sign is uncertain a question mark can be prefixed to the gloss, e. g. *?HOME*. Unidentifiable signs are glossed as *UNDECIPHERABLE*, but also *ADD-TO-SIGNBANK(UNKNOWN)* is used in some of these cases. This inconsistency is known to the corpus creators and should be fixed in the future. Signs that are not finished by the signers are glossed with *(FALSE-START)* suffixed, e. g. *DOG(FALSE-START)*. (Cormier et al., 2017, pp. 14–15)

Example: Annotation of an uncertain gloss for a gesture and a lexical sign and a case of a sign that was not finished in the BSL Corpus: *?G:WELL, ?TRUE, MOTHER(FALSE-START)* (Schembri et al., 2017, BF6n.eaf)

Corpus LSFB Signs that are indecipherable are glossed with *INDECIPHERABLE*, false starts with the gloss *SIGN(initiate)*. (Sinte et al., 2015, ‘Comparison’)

Example: An indecipherable name sign: *NS:INDECIPHERABLE(CALIN)* (Meurant, 2015, CLSFBI0103.eaf, 00:06:07.894–00:06:08.664)

Corpus NGT Signs with a false start are glossed with a prefixed tilde, e. g. *˜GLOSS*. (Crasborn et al., 2020, p. 28)

Signs that are not visible or recognisable, e. g. if another body part covers the sign or it is outside the video window the gloss *!* is used. If the sign is partly visible the exclamation mark is prefixed, e. g. *!GEBAREN-A*. (Crasborn et al., 2020, p. 28)

Uncertainties are glossed with question marks. One question mark is marking an educated guess, two question marks are used when the annotator does not know the sign and three question marks are used if the sign stays unknown after discussing the sign between several annotators. If the annotator is unsure if a movement is really a sign, *+-* is used. (Crasborn et al., 2020, p. 28)

Proposals for new glosses are prefixed with a dollar sign, e. g. *$GLOSS*. If there is no proposal for discussion the dollar sign is used by itself. (Crasborn et al., 2020, p. 3)

Example: Examples for other properties in the Corpus NGT: *˜ORGANISEREN, ZELFVERZEKERD?, !* (Crasborn et al., 2008b, CNGT0128.eaf, 00:03:45.520–00:03:45.880, 00:05:58.520–00:05:58.840, 00:01:19.520–00:01:20.760, `https://hdl.handle.net/1839/00-0000-0000-0009-2D6F-3`)

Corpus FinSL Idioms are glossed as lexical signs but marked on the translation equivalents in the Finnish Signbank. (Salonen et al., 2019, p. 13)

Corpus VGT Proposals for new ID-glosses are glossed as *SIGN !GLOSS*. The proposals are discussed in the team meeting and, if agreed upon, added to the lexical database. Unclear or unknown signs are glossed as *??* or *GLOSS??*. (Verstraete et al., 2015, ‘Proposals for a new ID-gloss’, ‘Uncertainties’)

Funded by the Horizon 2020
Framework Programme of the European Union

DTS Corpus No further properties.

DGS Corpus Indecipherable manual activity is labelled as *$UNCLEAR^*. Every annotator in the DGS Corpus has a gloss with their own name following the system *$$NAME-$SAM*, e. g. *$$MARIA-$SAM* to gloss tokens which should be discussed with the advisers. Unclear mouthings are glossed as *?* and as *??* if no team member can identify it. (Konrad et al., 2020b, p. 16; *R. Konrad, personal communication, January 27, 2022*)

Dicta-Sign-LSF-v2 Corpus No further properties.

Digging into Signs The different approaches on annotating uncertainties and corpus organisation are described but not further discussed. (Crasborn et al., 2015a, pp. 8–9)

ECHO Corpus An extra tier **Comments/notes** is added for comments. (Nonhebel et al., 2004a, p. 9)

Example: In the BSL poem a pun is explained in the tier **Comments/notes** with the annotation: *This sign also means 'SAY-NO-TO-SOMEONE' so it is a pun that by speaking to the deaf person they say no to the deaf person's needs*. (Woll et al., 2004, BSL_PS_poem3.eaf, 00:01:19.985–00:01:20.495)

HSL Corpus No further properties.

LIS Corpus No further properties.

PJM Corpus Indecipherable, invisible, unclear or doubtful signs are glossed with *###*. If a new gloss is needed, but there is no proposal yet the gloss *#???* is used. (Rutkowski et al., 2015, slides 39–40)

POLYTROPON The POLYTROPON Corpus contains only short sentences and no longer narrative signing, therefore some linguistic phenomena do not appear within the data. *(E. Efthimiou, personal communication, January 24, 2022)*

SIGNOR Corpus Indecipherable signs are glossed with *NEJASNA KRETNJA* (unclear sign). There are no special glosses for other issues, as only two annotators work on the data.

From the picture on the poster it comes apparent that there are further conventions, that are not described. For example are some glosses suffixed with information in brackets, others use different words separated by slashes: *OSNOVNA (ŠOLA)* meaning 'basic(school)', *MED(VMES)* meaning 'during(in between)', *ZRAVEN/PRI/OB* meaning 'next to/with/at'. (Jerko and Vintar, 2015, 'Uncertancies')

Signs of Ireland If there is doubt if a movement is a sign or not it is glossed as *Mov?*. If there is doubt, if the correct gloss has been chosen it is glossed as *GLOSS? ?)* followed by an explanation in brackets. If a sign is not known to the annotator and should be double-checked it is glossed as *??* followed by an explanation in brackets. If the sign is unknown to all annotators it is glossed as *GLOSS:???*. Suggestions for new signs which have to be discussed are glossed as *NEW GLOSS!*. If a new gloss is needed, but there is no proposal yet the sign is glossed as *NEW GLOSS?*. Recognised signs with a false start are glossed as *GLOSS(FALSE-START)*, if it is not clear which sign it is, due to a false start, this is glossed as *SIGN? (FALSE-START)*. (Matthews and Sheridan, 2015, slides 20–22)

The pictures in Matthews and Sheridan (2015, slide 11) suggest that there are further conventions not mentioned in the text, e. g. *#FLAT, DEAF*, 'what's that...', INDEX+me*.

SSL Corpus In the ELAN transcripts an extra tier called *Glosa_DH_extra* is used to merge the dominant and non-dominant hand annotation. This is needed for analysis like the concordance view also available via the corpus page. (Wallin and Mesch, 2018, pp. 28–31)

Several conventions exist for unfinished signs, unsure annotations and other special cases. If the annotator is in doubt if the hand movement is a sign or not or the sign is invisible the gloss *ZZZ@z* is used. If no existing gloss fits the sign the base *GLOSS* is used with different suffixes: *:glosa* if the sign is in the SSL Dictionary but not in the ELAN vocabulary, the same gloss is used if the sign is not in the Dictionary, but then also a proposal must be made, if the sign resembles another sign *GLOSS(glosa)* is used and if none of these options can be used *GLOSS:(?)* is used. False starts are suffixed with *@&*. If the sign is not identified *zzz@&* is used. If the signer stops signing in the middle of a sign to think about what signing next this is glossed as *tp@&*. (Wallin and Mesch, 2018, pp. 19–21)

So called 'homesigns', or 'idiosyncratic signs', are glossed with the suffix *@hg*, e. g. *TÅG@hg, KÖPA@hg*. (Wallin and Mesch, 2018, p. 20)

Example: A signer using an unknown sign is glossed as *GLOSA:(?)*. The same signer uses a lot of CA, some of which are glossed as *GLOSA:(?)@ka*; another sign is not finished, this is glossed as *zzz@&*. (Mesch et al., 2012, Äventyr, https://teckenspraakskorpus.su.se/#/video/sslc01_110.eaf)

5 NON-MANUALS

In comparison to manual signs, non-manuals have received less attention. Only some of the annotation conventions and manuals analysed in the previous chapters cover the topic of how to annotate movement happening on body and face. Even when there are conventions, we see in the actual data that a lot of non-manual activity is not annotated. As this is perfectly understandable in the light of the huge amount of resources that flow into a detailed annotation, the following listing is not to be understood as a criticism. The comparison of conventions and their use in published data helps us to differentiate between annotation schemes that are solely theoretical and such that have verified through application to an annotation process. This evaluation helps us to find an efficient and useful common format for non-manual activity within EASIER.

With the exception of annotation conventions for mouth movements, only four resources listed in deliverable D6.1. describe how to annotate non-manual movement, namely Crasborn et al. (2020), Nonhebel et al. (2004a), Bartha et al. (2016) and Wallin and Mesch (2018), as well as Johnston (2019). As Bartha et al. (2016) does not specify how they annotate different non-manual activity other than by using controlled vocabularies, they are not listed hereafter.

Efthimiou et al. (2018) mention the use of the SiS-Builder embedded non-manuals notation tool within their Polytropon lexical database (See Goulas et al., 2010). From Rutkowski et al. (2015, slide 8) it comes apparent that the PJM Corpus annotates non-manuals, at least for body movements and mouthings.

Section 5.1 to Section 5.6 describe the methodology suggested by different annotation conventions for different non-manual phenomena. Section 5.7 collects examples of the annotations of non-manual action. Three of the ten corpora provide transcripts with annotations of non-manuals in their downloadable data.

5.1 BODY

Three of the four annotation conventions describe how they annotate body movements. All use an extra tier for the body, Corpus NGT uses two separate tiers: one for body movement, the other for the body position. The movements are coded with respect to the neutral body position and encompass leaning, bending, rotating, shifting etc. . The SSLC uses eight different codes, the Corpus NGT has no controlled vocabulary for this tier and the Auslan Corpus does not elaborate on the used vocabulary. (Johnston, 2019, p. 51; Crasborn et al., 2020, p. 41; Wallin and Mesch, 2018, pp. 39–40)

5.2 HEAD

The four conventions all suggest annotating head movements on a separate tier. The Corpus NGT uses two different tiers, one for head movements, the other for head position. (Crasborn et al., 2020, p. 41; Johnston, 2019, p. 10; Nonhebel et al., 2004a, p. 5; Wallin and Mesch, 2018, p. 37)

In the SSLC and Corpus NGT extensive controlled vocabularies are used to annotate the different movements. SSLC uses 12 different codes (Wallin and Mesch, 2018, pp. 37–39), Corpus NGT includes 17 codes (Crasborn et al., 2020, p. 42). In the ECHO Corpus a code system with three entries is used: 'n' for nodding, 's' for shaking and 't' for tilts (Nonhebel et al., 2004a, p. 5). The Auslan Corpus does not specify how the movement is annotated, in the provided pictures some glosses are visible: *RIGHTWARDS, RAPID-LITTLE-SHAKES, TILT FORWARD, TILT RIGHT, NOD* (Johnston, 2019, pp. 51, 62, 94).

5.3 MOUTH

The Auslan corpus distinguishes between mouth action that is aligned with manual glosses and stand alone mouth gestures. Aligned mouth actions are further separated into mouthings and mouth gestures. Both have extra daughter tiers to annotate the grammatical class of the word that is mouthed for mouthings and the form and meaning of mouth gestures. Stand alone non-manual gestures are prefixed with *M* for mouthings and *MG* for mouth gesture followed by the meaning in context, e. g. *M:BECAUSE*. Mouthings are further categorised into nine different types, mouth gestures into six different types, these categories are adapted from Sutton-Spence and Day (2001) (Johnston, 2019, pp. 52–55).

The SSLC uses the categorisation of mouth actions presented in Crasborn et al. (2008a). The five categories are: 'M-type' for mouthings, 'A-type' for adverbial mouth gestures, 'E-type' for semantically empty mouth gestures, '4-type' for enacting mouth gestures (mouth 4 mouth) and 'W-type' for whole-face activities. Three extra categories were added by the team of the SSLC: 'B-type' for backchanneling, 'No mouth action' for non-existing mouth movenments and 'obestämd' for movement that does not fit any of the above categories (Wallin and Mesch, 2018, pp. 32–33).

The NGT Corpus uses six tiers to describe mouth actions in detail, the tiers comprise a transcription of the mouthing in Dutch orthography, the corresponding Dutch lemma, a categorisation of different mouthing types, information on spreading, the number of syllables and additional meaning of the mouth action. There are no controlled vocabularies, but conventions on how to annotate mouthings, mouth gestures and other actions. The classification of mouth types also follows Crasborn et al. (2008a) and adds five subtypes for the category 'Mouthing': 'M' for regular mouthing, 'M-back' for backchanneling, 'M-add' for mouthings that are not related to a manual signs but overlap with it, 'M-solo' for non-overlapping mouthings and 'M-spec' for mouthings that specify the meaning of the manual sign. (Crasborn et al., 2020, pp. 38–40).

The DGS Corpus uses one tier for mouth actions and distinguishes between mouthings, mouth gesture, the case of a signer indicating that someone is speaking and the imitation of sounds. Mouthings are annotated as open text entries, written in lower case and according to the intended word. Incomplete mouthings are supplemented in curly brackets. Mouth gestures are not further classified and all annotated as *[MG]*. The indication of someone speaking is glossed as # and the imitation of sound with the prefix *LM:*, e. g. *[LM:miau]* for the sound a cat makes. Uncertainties are marked with two question marks. Start and end of the mouth action is not separately annotated but aligned with the glosses for the manual signs, but it can spread over several token tags. (Konrad et al., 2020b, pp. 16–17)

The ECHO Corpus follows an approach of describing mouth actions as meaning-independent sequentially ordered combinations of a set of open and closed segments. The annotation is based on the visible contrast and done with features, e. g. the feature open can be '-open' for not open and '+open' for open, depending on this feature the label for annotation is chosen, e. g. */BILABIAL/* or

/CHEEKS/ for not open. The BSL and NGT datasets use slightly different labels and codes than the SSL dataset. Detailed descriptions of this approach can be found in Nonhebel et al. (2004b) and Nonhebel et al. (2004c) (Nonhebel et al., 2004a, pp. 6–9).

5.4 BROWS

All four manuals suggest annotating eyebrow movement on a separate tier and in respect to the neutral position of relaxed eyebrows. Two positions are named: raised and lowered (or furrowed) which are annotated with labels – 'höjda' for raised and 'sänkta' for lowered in the SSLC (Wallin and Mesch, 2018, p. 36), 'UP' in the Auslan Corpus (Johnston, 2019, p. 93) – or codes – 'r' for raised and 'f' for furrowed in the ECHO Corpus (Nonhebel et al., 2004a, p. 5).

5.5 EYES

All four conventions describe that they divide the annotation of the gaze from the annotation of the eyes. The latter being directed towards opening or closing degrees, squints, and blinks. The Auslan corpus does annotate the eye movement together with the brow movements in one single tier and until 2010 the gaze was only annotated when co-occurring with pointing signs. (Crasborn et al., 2020, p. 41; Johnston, 2019, p. 10; Nonhebel et al., 2004a, pp. 5–6; Wallin and Mesch, 2018, p. 33)

The ECHO corpus uses an elaborate system to annotate the eyes and gaze. For eye aperture four codes are used: 'w' for widened eyes, 's' for squints, 'c' for closed eyes and 'b' for blinks which are annotated in more detail by adding a number for the amount of repetition, e. g. a blink with four closing motions would be glossed as *b-4*. The eye gaze is annotated with a coding system with 10 entries, separating between left and right in different angles, down- and upward, to each or both hands and toward present persons or the camera (Nonhebel et al., 2004a, pp. 5–6).

The SSLC has two distinct coding systems for gaze and degree of opening. The first containing five codes: 'adr' for a gaze towards the addressee, 'hö' towards the right, 'va' towards the left, 'ner' for a gaze going down and 'up' for an upwards gaze, the latter four: 'bl' for blinks, 'sl' for closing, 'vid' for widened eyes and 'kis' for squints (Wallin and Mesch, 2018, pp. 33–35).

Auslan Corpus has a coding system for the gaze comprising four codes: 'a' for a gaze towards the addressee, 't' towards a target, 'o' for other directions and 'z' if it cannot be coded (Johnston, 2019, p. 52).

Corpus NGT annotates both phenomena without a controlled vocabulary (Crasborn et al., 2020, p. 41).

5.6 OTHER

The Auslan Corpus additionally provides a tier to annotate a global description of the face (Johnston, 2019, p. 51).

Corpus NGT also annotates the face as a whole and additionally the nose, whereby no controlled vocabulary is used (Crasborn et al., 2020, p. 41)

The Echo Corpus has an extra tier to annotate the movement of the cheeks. Two codes are used: ‘i’ for cheeks drawn in and ‘p’ for puffed cheeks (Nonhebel et al., 2004a, p. 9).

The SSLC also annotates the cheeks but in one tier with the nose. The codes used are: ‘upp’ for raised cheeks raised and ‘rynk’ for wrinkles on the nose (Wallin and Mesch, 2018, pp. 36–37).

In the Public DGS Corpus non-manual gestures without an aligned manual sign/gesture are glossed as *$GEST-NM^* in the tier for manual signs (Konrad et al., 2020b, p. 15). In the DGS Corpus these non-manual gestures are specified in more detail, e. g. *$GEST-NM-KOPFNICKEN1-$SAM*, *$GEST-SCHULTERZUCKEN1-$SAM*, *$GEST-ZUNGE-RAUSSTRECKEN-$SAM*, for head nodding, shrugging and the extrusion of the tongue.

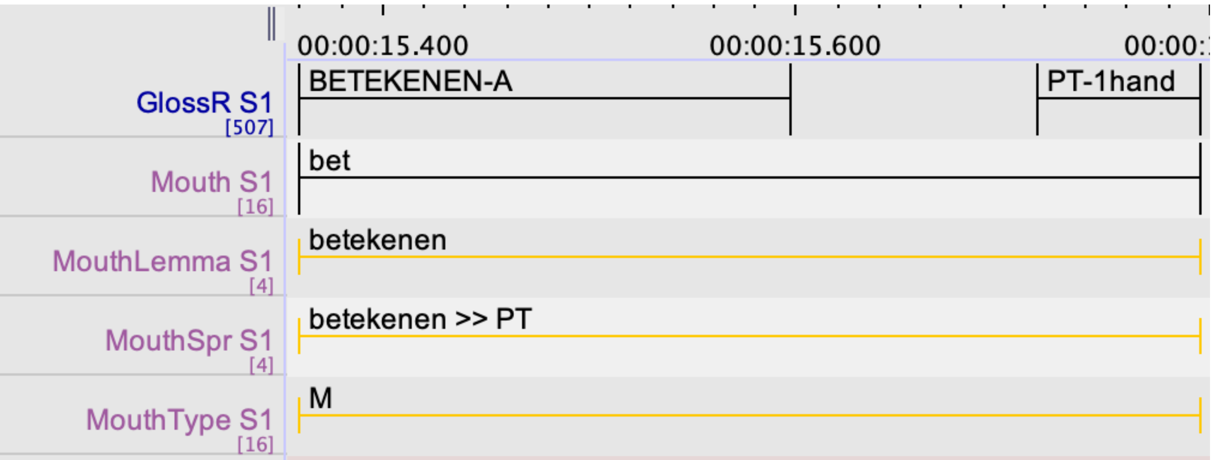
5.7 DATA EXAMPLES

Figures 5.1 to 5.4 show extracts of data with annotation of non-manual activity. For the annotation of eyebrow movements in the ECHO Corpus two different coding systems are used in the datasets:

BSL dataset Labels ‘u’ and ‘f’ are used (e. g. Woll et al., 2004, BSL_PS_poem3.eaf)

NGT dataset Labels ‘r’ and ‘f’ are used (e. g. Crasborn et al., 2004, NGT_WE_poems.eaf)

SSL dataset Labels ‘r’ and ‘f’ are used (e. g. Bergman and Mesch, 2004, SSL_JI_fab1.eaf)



00:00:15.400 00:00:15.600 00:00:

GlossR S1 [507]	BETEKENEN-A	PT-1hand
Mouth S1 [16]	bet	
MouthLemma S1 [4]	betekenen	
MouthSpr S1 [4]	betekenen >> PT	
MouthType S1 [16]	M	

Figure 5.1: *Mouthing ‘bet’ in the Corpus NGT categorised as simple mouthing and marked to be spreading over to the next gloss PT. (Crasborn et al., 2008b, CNGT0128.eaf, 00:00:15.360–00:00:15.800, `https://hdl.handle.net/1839/00-0000-0000-0009-2D6F-3`)*

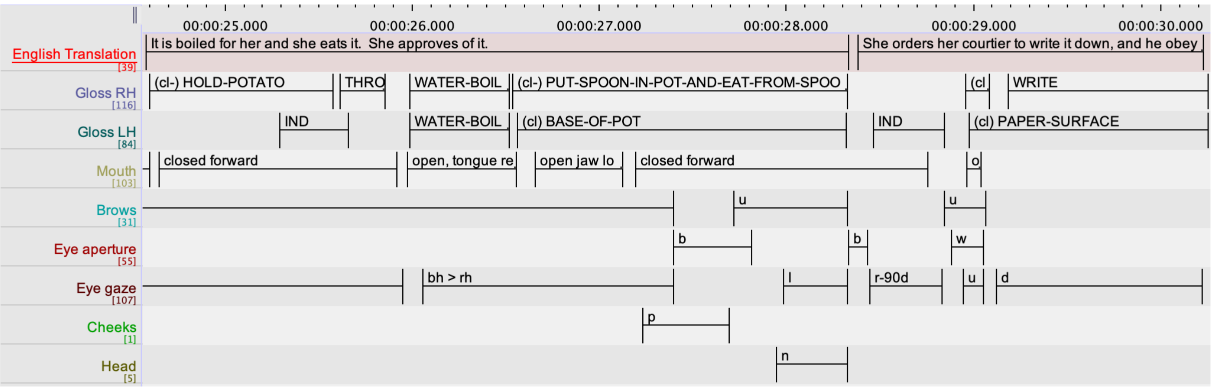

00:00:25.000 00:00:26.000 00:00:27.000 00:00:28.000 00:00:29.000 00:00:30.000
English Translation [39]: It is boiled for her and she eats it. She approves of it. | She orders her courtier to write it down, and he obey
Gloss RH [116]: (cl-) HOLD-POTATO | THRO | WATER-BOIL | (cl-) PUT-SPOON-IN-POT-AND-EAT-FROM-SPOO | (cl | WRITE
Gloss LH [84]: IND | WATER-BOIL | (cl) BASE-OF-POT | IND | (cl) PAPER-SURFACE
Mouth [103]: closed forward | open, tongue re | open jaw lo | closed forward | o
Brows [31]: u | u
Eye aperture [55]: b | b | w
Eye gaze [107]: bh > rh | l | r-90d | u | d
Cheeks [1]: p
Head [5]: n

Figure 5.2: *Annotation of several mouth actions, brow movements, eye apertures, eye gaze, cheeks and head movements. (Woll et al., 2004, BSL_PS_poem3.eaf, 00:00:24.578–00:00:30.262)*

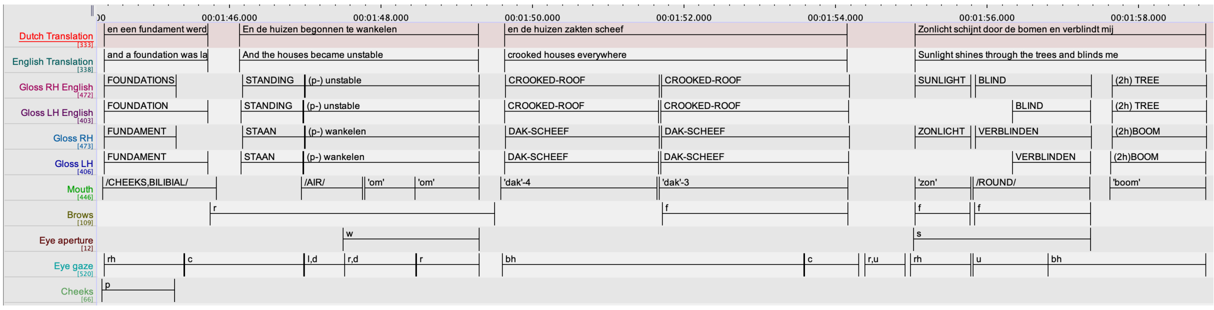

00:01:46.000 00:01:48.000 00:01:50.000 00:01:52.000 00:01:54.000 00:01:56.000 00:01:58.000
Dutch Translation [333]: en een fundament werd | En de huizen begonnen te wankelen | en de huizen zakten scheef | Zonlicht schijnt door de bomen en verblindt mij
English Translation [338]: and a foundation was la | And the houses became unstable | crooked houses everywhere | Sunlight shines through the trees and blinds me
Gloss RH English [472]: FOUNDATIONS | STANDING | (p-) unstable | CROOKED-ROOF | CROOKED-ROOF | SUNLIGHT | BLIND | (2h) TREE
Gloss LH English [403]: FOUNDATION | STANDING | (p-) unstable | CROOKED-ROOF | CROOKED-ROOF | BLIND | (2h) TREE
Gloss RH [473]: FUNDAMENT | STAAN | (p-) wankelen | DAK-SCHEEF | DAK-SCHEEF | ZONLICHT | VERBLINDEN | (2h)BOOM
Gloss LH [406]: FUNDAMENT | STAAN | (p-) wankelen | DAK-SCHEEF | DAK-SCHEEF | VERBLINDEN | (2h)BOOM
Mouth [446]: /CHEEKS,BILIBIAL/ | /AIR/ | 'om' | 'om' | 'dak'-4 | 'dak'-3 | 'zon' | /ROUND/ | 'boom'
Brows [109]: r | f | f | f
Eye aperture [12]: w | s
Eye gaze [520]: rh | c | l,d | r,d | r | bh | c | r,u | rh | u | bh
Cheeks [66]: p

Figure 5.3: *Annotation of several mouth actions, brow movements, eye apertures, eye gaze and cheeks movements. (Crasborn et al., 2004, NGT_WE_poems.eaf, 00:01:44.350–00:01:58.900)*

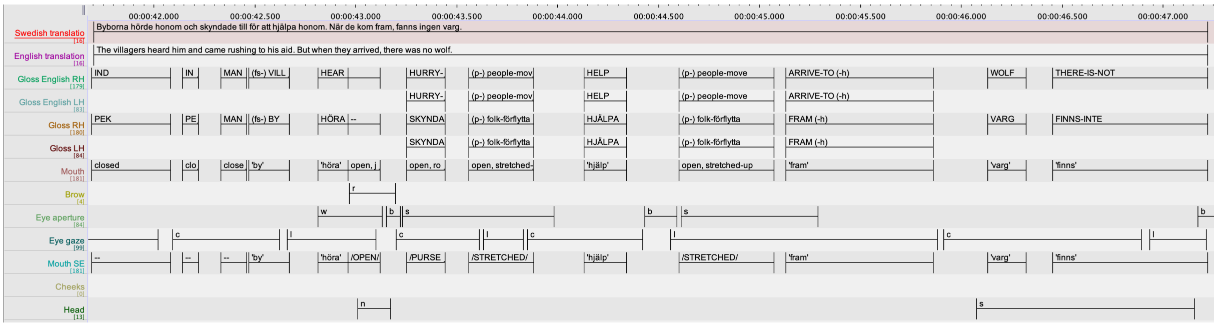
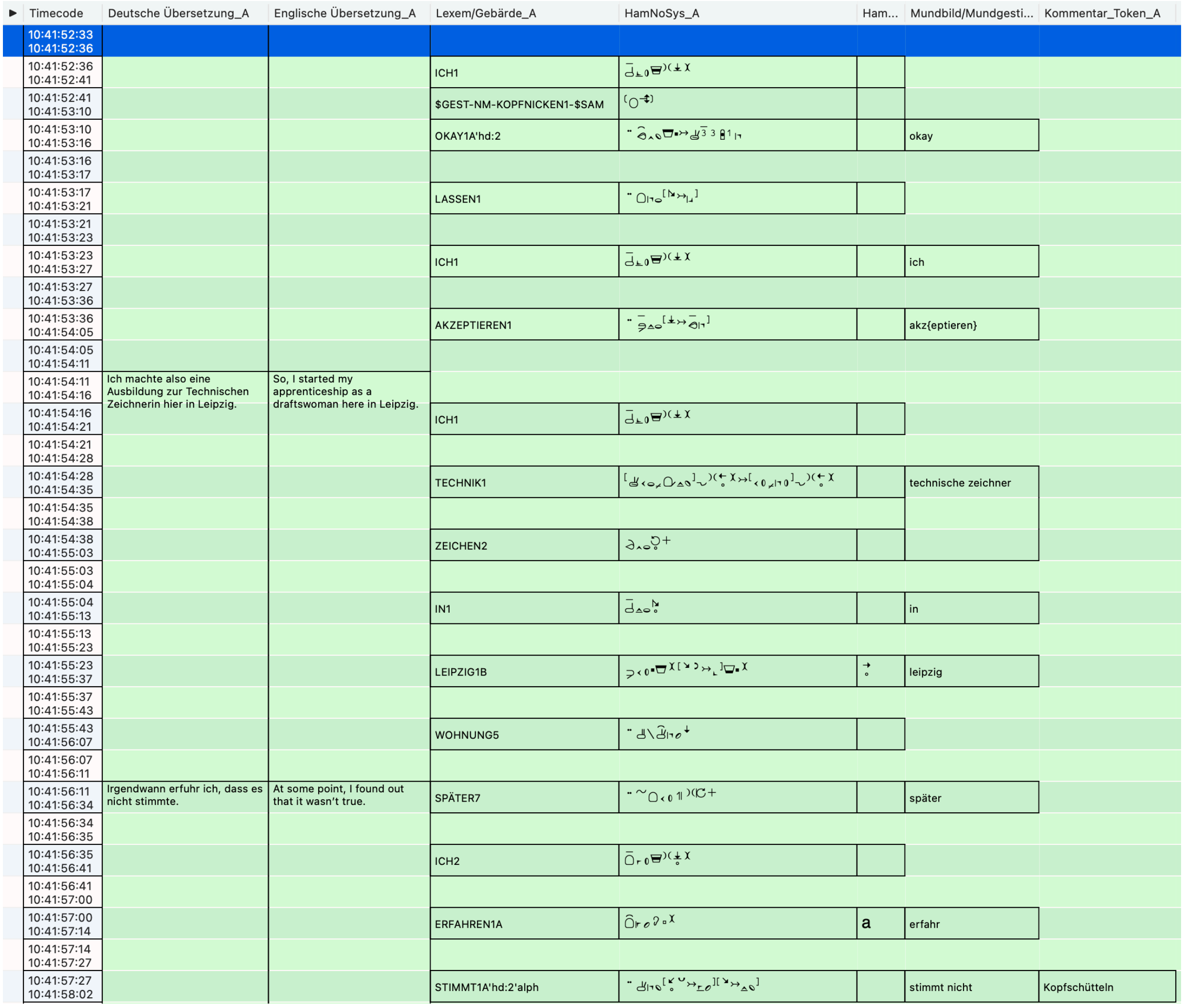

00:00:42.000 00:00:42.500 00:00:43.000 00:00:43.500 00:00:44.000 00:00:44.500 00:00:45.000 00:00:45.500 00:00:46.000 00:00:46.500 00:00:47.000
Swedish translatio [16]: Byborna hörde honom och skyndade till för att hjälpa honom. När de kom fram, fanns ingen varg.
English translation [16]: The villagers heard him and came rushing to his aid. But when they arrived, there was no wolf.
Gloss English RH [179]: IND | IN | MAN | (fs-) VILL | HEAR | HURRY- | (p-) people-mov | HELP | (p-) people-move | ARRIVE-TO (-h) | WOLF | THERE-IS-NOT
Gloss English LH [83]: HURRY- | (p-) people-mov | HELP | (p-) people-move | ARRIVE-TO (-h)
Gloss RH [180]: PEK | PE | MAN | (fs-) BY | HÖRA | -- | SKYNDA | (p-) folk-förflytta | HJÄLPA | (p-) folk-förflytta | FRAM (-h) | VARG | FINNS-INTE
Gloss LH [84]: SKYNDA | (p-) folk-förflytta | HJÄLPA | (p-) folk-förflytta | FRAM (-h)
Mouth [181]: closed | clo | close | 'by' | 'höra' | open, j | open, ro | open, stretched- | 'hjälp' | open, stretched-up | 'fram' | 'varg' | 'finns'
Brow [4]: r
Eye aperture [84]: w | b | s | b | s | b
Eye gaze [99]: c | l | c | l | c | l | c | l
Mouth SE [181]: -- | -- | -- | 'by' | 'höra' | /OPEN/ | /PURSE | /STRETCHED/ | 'hjälp' | /STRETCHED/ | 'fram' | 'varg' | 'finns'
Cheeks [0]
Head [13]: n | s

Figure 5.4: *Annotation of several mouth actions, brow movements, eye apertures, eye gaze, mouthings in Swedish, and head movements. (Bergman and Mesch, 2004, SSL_JI_fab1.eaf, 00:00:41.700–00:00:47.220)*

Funded by the Horizon 2020
Framework Programme of the European Union



▶ Timecode	Deutsche Übersetzung_A	Englische Übersetzung_A	Lexem/Gebärde_A	HamNoSys_A	Ham...	Mundbild/Mundgesti...	Kommentar_Token_A
10:41:52:33 10:41:52:36							
10:41:52:36 10:41:52:41			ICH1	[illegible]			
10:41:52:41 10:41:53:10			$GEST-NM-KOPFNICKEN1-$SAM	[illegible]			
10:41:53:10 10:41:53:16			OKAY1A'hd:2	[illegible]		okay	
10:41:53:16 10:41:53:17							
10:41:53:17 10:41:53:21			LASSEN1	[illegible]			
10:41:53:21 10:41:53:23							
10:41:53:23 10:41:53:27			ICH1	[illegible]		ich	
10:41:53:27 10:41:53:36							
10:41:53:36 10:41:54:05			AKZEPTIEREN1	[illegible]		akz{eptieren}	
10:41:54:05 10:41:54:11							
10:41:54:11 10:41:54:16	Ich machte also eine Ausbildung zur Technischen Zeichnerin hier in Leipzig.	So, I started my apprenticeship as a draftswoman here in Leipzig.					
10:41:54:16 10:41:54:21			ICH1	[illegible]			
10:41:54:21 10:41:54:28							
10:41:54:28 10:41:54:35			TECHNIK1	[illegible]		technische zeichner	
10:41:54:35 10:41:54:38							
10:41:54:38 10:41:55:03			ZEICHEN2	[illegible]			
10:41:55:03 10:41:55:04							
10:41:55:04 10:41:55:13			IN1	[illegible]		in	
10:41:55:13 10:41:55:23							
10:41:55:23 10:41:55:37			LEIPZIG1B	[illegible]	[illegible]	leipzig	
10:41:55:37 10:41:55:43							
10:41:55:43 10:41:56:07			WOHNUNG5	[illegible]			
10:41:56:07 10:41:56:11							
10:41:56:11 10:41:56:34	Irgendwann erfuhr ich, dass es nicht stimmte.	At some point, I found out that it wasn't true.	SPÄTER7	[illegible]		später	
10:41:56:34 10:41:56:35							
10:41:56:35 10:41:56:41			ICH2	[illegible]			
10:41:56:41 10:41:57:00							
10:41:57:00 10:41:57:14			ERFAHREN1A	[illegible]	a	erfahr	
10:41:57:14 10:41:57:27							
10:41:57:27 10:41:58:02			STIMMT1A'hd:2'alph	[illegible]		stimmt nicht	Kopfschütteln

Figure 5.5: *Annotation of a head nodding unaligned with a manual gloss $GEST-NM-KOPFNICKEN1-$SAM, a supplemented mouthing akz{eptieren} and a simultaneous head shake in the DGS Corpus. (Konrad et al., 2020a, DGS Corpus, dgskorpus_lei_12: Experience of Deaf Individuals, 10:41:52:33–10:41:58:02, https://doi.org/10.25592/dgs.corpus-3.0-text-1584617)*

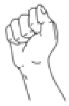
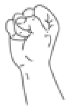
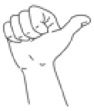
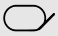
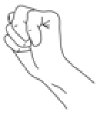
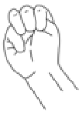
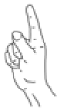
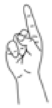
6 CODING OF HANDSHAPES

Handshape codes are used for different purposes within corpora, as shown in Chapter 4. The following table compares different handshape codes and shows the wide variety used within the different corpora.

PJM Corpus handshape codes are based on the PJM finger alphabet, which is not provided within the conventions available to us (Rutkowski et al., 2015, slide 13). DGS Corpus uses HamNoSys, but in the case of fingerspelling, handshape codes are based on the DGS manual alphabet (Konrad et al., 2020b, p. 24).

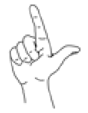
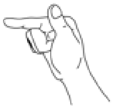
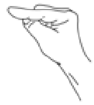
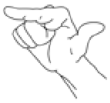
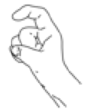
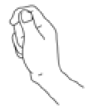
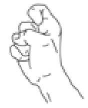
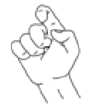
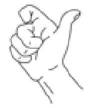
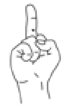
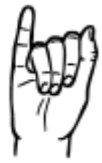
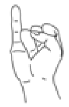
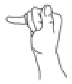
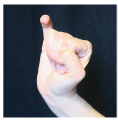
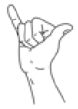
All information in the following table is taken from the according transcription conventions or manual, Johnston (2019, p. 105) for Auslan corpus, Cormier et al. (2017, p. 19) for BSL Corpus, Crasborn et al. (2020, p. 46) for Corpus NGT, Salonen et al. (2019, pp. 22–24) for CFinSL and pictures marked with *, and Wallin and Mesch (2018, pp. 45–46) for SSLC. The handshape codes used in the Corpus LSFB and the pictures marked with † are published on the Corpus LSFB website[29]. Corpus LSFB uses almost the same codes as Auslan Corpus, but deviates in some cases, so extra attention must be paid. The handshape pictures, if not marked differently, are taken from the HamNoSys handshapes table in iLex.

Table 6.1: ***Coding of handshapes in different corpus resources***

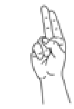
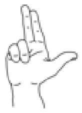
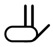
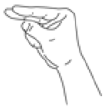
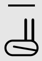
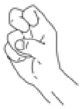
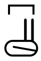
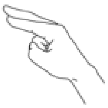
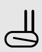
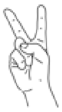
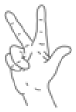
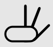
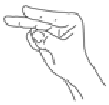
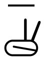
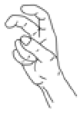
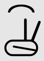
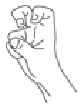
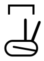
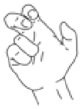
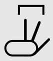
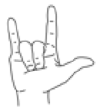
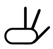
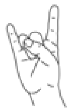
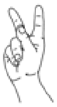
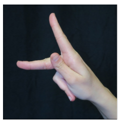
Handshape	HamNoSys	Auslan Corpus	BSL Corpus	Corpus NGT	Corpus LSFB	CFinSL	SSLC
		-	FIST	-	-	A	-
		S	FIST	S	S	S	G
		6	A-DOT	-	6	A1	B
	3 \ 4	N	-	-	N	-	-
	4 \ 5	M1	-	-	M1	-	-
		-	1	1	-	-	-
		1	-	1	1	G	L

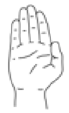
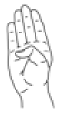
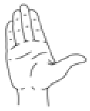
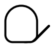
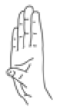
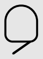
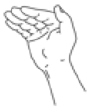
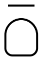
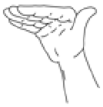
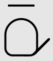
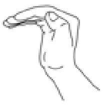
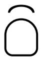
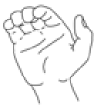
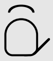
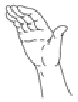
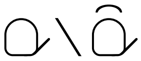
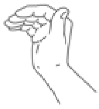
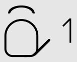
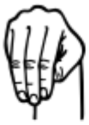
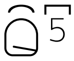
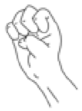
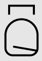
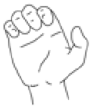
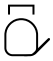
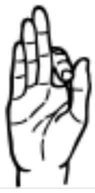
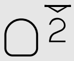
[29] https://www.corpus-lsfb.be/signesCode.php

Funded by the Horizon 2020
Framework Programme of the European Union

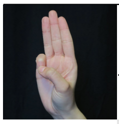
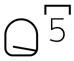
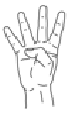
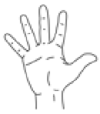
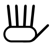
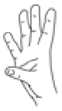
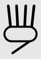
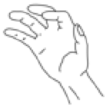
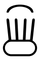
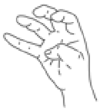
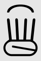
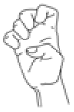
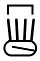
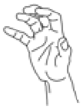
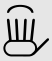
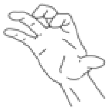
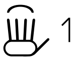
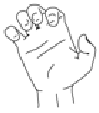
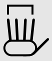
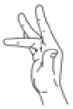
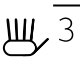
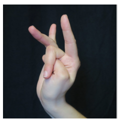
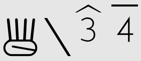
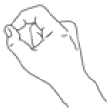
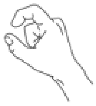
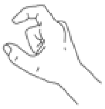
Handshape	HamNoSys	Auslan Corpus	BSL Corpus	Corpus NGT	Corpus LSFB	CFinSL	SSLC
		7	GUN	-	7	L	Lt
		-	1	-	-	T	-
		-	1	-	-	-	Lv
		-	-	-	-	T	-
		-	-	1_curved	-	G	Lb
		IRISHT	GRIP	money	IRISHT	Ax	Q
		X	HOOK	-	X	X	Lbt
		-	-	-	BENT7	-	-
	1	BENT7	HOOK	-	-	-	-
	3	!	RUDE	-	!	-	R
	* 5	-	-	-	-	I	-
	5	I	PINKY	-	I	I	I
	5	-	PINKY	-	-	-	-
	† 5	BENTI	-	-	BENTI		-
	5	Y	Y	Y	Y	Y	lt -

Funded by the Horizon 2020
Framework Programme of the European Union

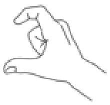
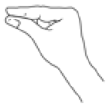
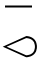
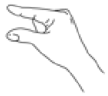
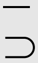
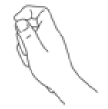
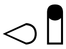
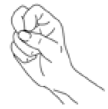
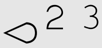
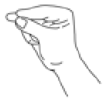
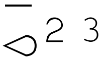
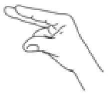
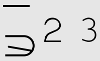
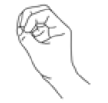
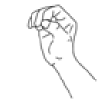
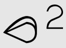
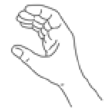
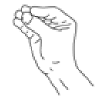
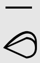
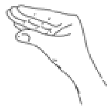
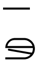
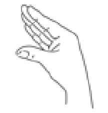
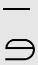
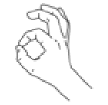
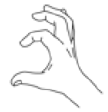
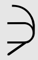
Handshape	HamNoSys	Auslan Corpus	BSL Corpus	Corpus NGT	Corpus LSFB	CFinSL	SSLC
		H	SPOON	-	H	H	N -
		HTHUMB	-	-	HTHUMB	-	-
		-	SPOON	-	-	-	-
		-	2	-	-	-	-
	 2 3	R	WISH	-	R	R	-
		2	2	V	2	V	V
		8	ASL-VEH	-	8	Vl	Vt
		-	2	-	-	-	-
		-	2	V	-	V	Vb
		BENT2	-	-	BENT2	Vc	Vbt
		BENT8	SPHERE	-	BENT8	-	-
	 2 5	ILY	-	-	ILY	-	-
	2 5	IRISHH	HORNS	-	-	Gi	-
	$\overline{3}$ 1	-	2	-	-	K	-
 †	$\overline{3}$ $\hat{4}$ $\hat{5}$	P	-	-	P	-	-

Handshape	HamNoSys	Auslan Corpus	BSL Corpus	Corpus NGT	Corpus LSFB	CFinSL	SSLC
		-	FLAT	B	-	B	-
		BB	FLAT	B	B	B	D
		B	FLAT	B	BA	B	J
		-	-	-	-	-	Jt
		-	FLAT	B	-	-	Jv
		BENTB	-	-	BENTB	-	
		-	CYL	B_curved	-	Bc	-
		CURVEDB	-	-	-	-	-
		-	-	-	-	-	Jb
		-	-	-	CURVEDB	Bc	-
	* 	-	-	-	-	M	-
		E	O	-	E	E	-
		-	CYL	-	-	-	-
	* 	-	-	-	-	F	-

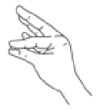
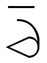
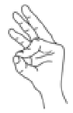
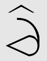
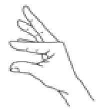
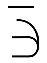
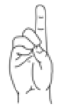
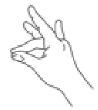
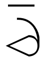
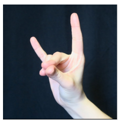
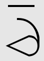
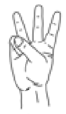
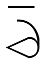
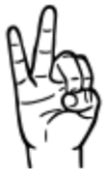
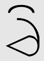
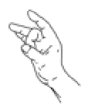
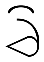
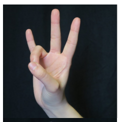
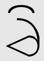
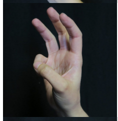
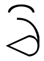
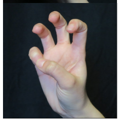
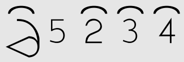
Funded by the Horizon 2020
Framework Programme of the European Union

Handshape	HamNoSys	Auslan Corpus	BSL Corpus	Corpus NGT	Corpus LSFB	CFinSL	SSLC
 †		M	3	-	M	-	-
		4	4	4	4	4	4
		5	5	5	5	5	Y
		-	-	-	-	-	Yt
		-	5	-	-	5c	-
		-	-	-	BENT4	-	-
		-	-	-	-	E	-
		BENT4/BENT5	SPHERE	C_spread	BENT5	5c	Ybt
		-	5	-	-	-	Yb
		BENT4/BENT5	SPHERE	-	-	-	-
		MID	OPEN-8	-	MID	2	Lbs
 †		-	-	-	MIDO	-	-
		GO	SMALL_CLOSED	Baby_O	GO	Lo	-
		-	-	-	GC	-	-
		-	SMALL_OPEN	Baby_C	-	-	-

Funded by the Horizon 2020
Framework Programme of the European Union

Handshape	HamNoSys	Auslan Corpus	BSL Corpus	Corpus NGT	Corpus LSFB	CFinSL	SSLC
		GC	-	-	-	Lc	-
		FLATGO	SMALL_CLOSED	Baby_beak	FLATGO	Lq	Lvs
		GCFLAT	SMALL_OPEN	Baby_beak_open	GCFLAT	-	Lvt
		-	SMALL_CLOSED	-	-	-	-
		-	-	-	-	-	Vbs
		12	-	-	12	Lq	-
		HCFLAT	SMALL_OPEN	-	HCFLAT	-	-
		O	O	O	O	O	O
		-	-	-	-	O	-
		BC	CYL	C	BC	C	S
		-	CLOSED	Beak	FLATO	Bq	A
		FLATBC	OPEN	-	FLATBC	-	Ao
		-	OPEN	Beak_open	-	-	-
		F	SMALL_CLOSED	T	-	F	H -
		FC	-	-	FC	-	-

Funded by the Horizon 2020
Framework Programme of the European Union

Handshape	HamNoSys	Auslan Corpus	BSL Corpus	Corpus NGT	Corpus LSFB	CFinSL	SSLC
		FLATF	-	-	FLATF	-	-
		-	SMALL_CLOSED	3/T	-	M	-
		FLATFC	-	-	FLATFC	-	-
	3 4 5	D	-	-	D	D	-
	 3	IRISHK	-	-	IRISHK	-	-
 †	 3 4	-	-	-	IRISHH	-	-
	 5	3	3	-	3	W	W
 *	 2 3	-	-	-	-	7	-
	 3	-	-	-	-	2o	Z
 †	 4	-	-	-	VI	-	-
 †	 5	BENT3	-	-	BENT3	-	-
 †		CLAW3	-	-	CLAW3	-	-

*: Handshape image taken from Salonen et al. (2019).
†: Handshape image taken from https://www.corpus-lsfb.be/signesCode.php
All other handshape images taken from the HamNoSys handshapes table in iLex. (Hanke and Storz, 2008)

Funded by the Horizon 2020 Framework Programme of the European Union

7 HARMONIZATION OF GLOSS STANDARDS FOR EASIER

To process corpus data in the EASIER translation pipeline, the annotations of different corpora need to be converted into a single format that can encode all information that is relevant to the pipeline. As a first step towards this, the annotation formats of the different corpora need to be converted into a single unified interchange format. In this chapter we provide a first glimpse at a potential structure for this interchange format, based on the observations gathered in this report.

7.1 INTERCHANGE FORMAT

In sign language corpus annotations, glosses function as a text-based aid to humans. As such they need to encode information in a succinct way that allows the perusal of sequences of glosses. The EASIER interchange format, on the other hand, is primarily intended for the communication between components of a software pipeline and therefore has different constraints and priorities. Information does not need to be encoded as succinctly, but needs to be easily and unambiguously parseable. So instead of representing a sign through a single gloss word, we propose using a JSON container structure that explicitly identifies individual components of the gloss, such as its name and variant, plus relevant meta-information like its source language and dataset.

The exact structure of a container depends on the kind of gloss that it represents. Entries for basic lexical types provide different information than special types, such as those for buoys, fingerspelling, etc. Token entries can refer to their associated type entry for general lexical information, but must specify start and end times and whether they deviate from the base form.

The following sections provide some examples for what different kinds of containers might look like.

7.1.1 Type Entry for a Regular Sign

Type entries represent the context-independent base form of a sign. The most common type entry is that of a regular lexical sign, i. e. one whose gloss centres around a spoken language word approximating the sign meaning (see Section 4.1), as opposed to specially glossed types like buoys or fingerspelling.

The example shown in Listing 7.1 encodes the type gloss *WHIPPED-CREAM1*, taken from the Public DGS Corpus.[30] It is structured as follows:

id: The unique ID of the container in the EASIER system.

parent: This gloss is a subtype of *TO-STIR1^*. This field specifies the ID of its parent type.

language: The language of the sign, identified by its ISO639-3 language code.

kind: Specifies that the container represents a type entry (as opposed to a token instance) and that the type is a regular sign (as opposed to a specially glossed type like a buoy or fingerspelling).

[30] https://doi.org/10.25592/dgs.corpus-3.0-type-13156

```
{
    "id": 123456,
    "parent": 123440,
    "language": "gsg",
    "kind": ["type", "regular"],
    "name": {
        "deu": ["Schlagsahne"],
        "eng": ["whipped", "cream"]
    },
    "lex variant": 1,
    "phon variant": null,
    "hands": 1,
    "hamnosys": "hamfist,hamindexfinger,hamfingerhookmod,
       hamextfingerdo,hampalmdr,hamcircled,hamsmallmod,
       hamrepeatfromstartseveral",
    "wordnet": ["07621388-n"],
    "source": {
        "gloss": {
            "deu": "SCHLAGSAHNE1",
            "eng": "WHIPPED-CREAM1"
        },
        "id": "4670",
        "parent": "13156",
        "dataset": "Public DGS Corpus"
    }
}
```

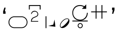
Listing 7.1: *Example of a container for the corpus-independent representation of a type entry for a regular lexical sign.*

name: The component of a regular sign that is based on a spoken language word. The spoken language used is identified using its ISO639-3 language code. In the case of the Public DGS Corpus both German and English versions of each gloss name are provided. The language-specific gloss name is represented as a list of individual words. This avoids the ambiguity of whether the dash in a gloss name marks a multi-word expressions like 'whipped cream' or a hyphenated word like 't-shirt'. The language-specific capitalisation of each word is restored.

lex variant: An integer disambiguating lexical variants (see Section 4.2). Glossing conventions that use other values are converted, e. g. 'A' to 1, etc. For conventions that encode additional information for variants (words, handshape codes), integers are assigned and the additional information encoded in additional fields.

phon variant: An integer disambiguating phonetic variants. Value conversion is handled like for `lex variant`. As no phonetic variants are given for this specific gloss, but the source corpus does specify them in general, the value is set to *null*. For corpora that never specify phonetic variants, the field would be omitted.

hands: Integer specifying whether the base form of the sign is performed in a one-handed (1) or two-handed (2) manner. For token entries, a similar field will specify whether a sign was executed with the dominant, non-dominant or both hands.

hamnosys: The phonetic transcription of the sign type in HamNoSys format. As HamNoSys symbols require a special font to display them, they are instead provided as the ASCII names of each symbol. The sequence in this example equates to ‘’.

wordnet: Sense identifiers for the EASIER interlingual index (based on Open Multilingual Wordnet). Lists the different possible meanings the sign can have.

source: A sub-container providing information on the original corpus data on which the type entry is based. These fields are included mainly to allow developers to verify whether the gloss was parsed correctly and to refer back to the data.

gloss: The original gloss string(s) of the type entry as they were used in the corpus. Like the `name` field, it allows for multiple language variants.

id: The unique identifier of the type entry that is used in the original corpus. Unlike the EASIER ID this might not be an integer, e. g. in some corpora it is identical to the gloss string.

parent: The corpus ID of the supertype (see EASIER field of same name).

dataset: Identifier of the corpus from which the entry originates.

7.1.2 Token Entry for a Regular Sign

While type entries represent an abstract base form, token entries represent the specific production of a sign in the context of an utterance. As such they must specify which recording transcript they originate from, locate the token in the time frame of the recording, connect it to a type entry, and specify whether and how it deviates from its base form.

Listing 7.2 shows the container for a token occurrence of the *WHIPPED-CREAM1* sign type that was described in Section 7.1.1. It is structured as follows:

id: The unique ID of the container in the EASIER system.

type: The ID of the type entry that represents the base form of the token.

kind: Specifies that the container represents a token instance of a regular sign.

is modified: Specifies whether the way the sign is executed differs from how it was defined in the type entry. If the corpus also provides information on how it differs (e. g. different handshape or a one-handed sign being performed with both hands), this would be specified in an additional field `modification`.

hand: Which hand produces the sign. May be ‘right’ or ‘left’ for one-handed and asymmetric two-handed signs and ‘both’ for symmetric two-handed signs (cf. Section 4.3).

timecode start: Specifies where in the transcript recording the production of the sign begins. The timecode is given as a list of integers representing hour, minute, second, and a numerator/denominator pair to specify a sub-second value. The sub-second denominator can match the frame rate of the video recording (50fps in the example) or be set to 1000 to represent milliseconds.

timecode end: Timecode for where the production of the sign ends.

```
{
    "id": 678453,
    "type": 123456,
    "kind": ["token", "regular"],
    "is modified": true,
    "hand": "right",
    "timecode start": [0, 8, 18, 11, 50],
    "timecode end": [0, 8, 18, 36, 50],
    "source": {
        "id": "1001435",
        "type": "4670",
        "transcript": "1250646",
        "dataset": "Public DGS Corpus",
        "modification": {
            "hamnosys": "*"
        }
    }
}
```

Listing 7.2: *Example of a container for the corpus-independent representation of a token entry of a regular lexical sign.*

source: A sub-container providing information on the original corpus data on which the token entry is based.

> ***id:*** The unique identifier of the token entry in the original corpus.
>
> ***type:*** The corpus ID of the type (see EASIER field of same name).
>
> ***transcript:*** The identifier of the annotated transcript in which the token occurs. The transcript is used to connect token entries to the video recording(s) to which they refer.
>
> ***dataset:*** Identifier of the corpus from which the entry originates.
>
> ***modification:*** A sub-container that specifies on what data the values of the EASIER fields `is modified` and `modification` are based. In this case, the Public DGS Corpus uses the ‘hamnosys’ field in its token entry to indicate with a ‘*’ that a modification occurred, but does not specify what exactly the modification is.

7.1.3 Token Entry for Fingerspelling

The previous examples represented the type and token format for regular signs. As glosses for other kinds of signs encode different information, their containers provide different fields.

Listing 7.3 shows an example of a fingerspelling token. The example is taken from the Corpus LSFB and is a fingerspelling of the family name ‘Descornet’, glossed as *FS:DESCORNTET(DESCORNET)*. The gloss specifies both the letters that were spelled out by the signer and the spelling that the annotators presume was intended. In this particular case, the difference is that the signer accidentally included a second ‘T’ in the name.

The fingerspelling token container uses the following fields that did not occur in previous examples:

```
{
    "id": 6553348,
    "kind": ["token", "fingerspelling"],
    "spelling": ["D", "E", "S", "C", "O", "R", "N", "T", "E", "T"],
    "word": {
        "fra": ["Descornet"]
    },
    "alphabet": "lsfb1",
    "hand": "right",
    "timecode start": [0, 0, 56, 957, 1000],
    "timecode end": [0, 1, 0, 384, 1000],
    "source": {
        "gloss": "FS:DESCORNTET(DESCORNET)",
        "id": "456789",
        "transcript": "CLSFBI0301",
        "dataset": "Corpus LSFB"
    }
}
```

Listing 7.3: *Example of a container for the corpus-independent representation of a token entry of fingerspelling.*

spelling: The sequence of fingerspelled letters. Represented as a list of strings to allow for fingerspelling signs that represent multiple letters, e. g. ‘ch’ or ‘sch’. Correct marking of such multi-letter signs depends on whether they can be identified reliably in the gloss format or through other methods.

word: The word(s) spelled out by the fingerspelling sequence. Like for gloss names, the source language is identified, capitalisation is corrected, and multi-word expressions are applied. Spelling errors and incomplete spellings are corrected for when possible.

alphabet: Identifier for the fingerspelling alphabet that is used. The identifier will refer to a set of known fingerspelling alphabets. As some signed languages have more than one alphabet (e. g. a one-handed and a two-handed alphabet), identifiers will consist of the language name and additional disambiguating markers.

7.2 CHALLENGES

As we have seen in the previous chapters of this report, the considered corpora differ both in what information they encode and how they encode information, including the degree of granularity and the definition and boundaries of specific phenomena.

Where possible we differentiate between whether a particular bit of information is not specified because it does not apply to the gloss in question or because it is generally not annotated in the dataset. This is important to judge which inferences can be drawn from an annotation. For example, in Listing 7.1 the gloss *WHIPPED-CREAM1* has no phonological variants, but the Public DGS Corpus does specify them in general, so the `phon variant` field is set to `null`. In a corpus that does not mark phonological variants, the same gloss would omit that field entirely. Similarly, the `parent` field, indicating that *WHIPPED-CREAM1* is a subtype of the type *TO-STIR1^*, would be `null` in *TO-STIR1^* and would be omitted in datasets without double glossing (cf. Section 4.1).

Certain kinds of implicit information might turn out to be difficult to extract through automatic processes. For example, some corpora annotate two-handed signs as individual glosses on separate left and right hand tiers with different start and end (see Section 4.3). A conversion script will not only have to identify these glosses as being connected, but also have to decide whether their overlapping occurrence indicates a single two-handed sign or two separate performances of the same sign type, e. g. when describing the independent movement of two persons.

Different annotation tools encode the sub-second unit of timestamps in different ways. ELAN uses milliseconds (Hellwig et al., 2021) while iLex bases it on the frame rate of the given video recording. Both approaches have advantages and disadvantages. Milliseconds are a constant measure independent from the source recording, but introduce rounding errors when used for recordings with frame rates that can not be represented in steps of $^1/_{1000}$, e. g. for a video with 30 frames per second, each frame is a step of $33.\overline{3}$. Using frame counts avoids the rounding issue, but makes comparing data with different base frame rates slightly more complex and introduces questions about handling multi-camera recordings which might involve multiple frame rates or have offsets involving a fraction of a frame. The interchange format is designed to allow for both approaches, but this does not fully resolve the aforementioned issues. Comparisons between different data sources will also require the definition of a tolerance value for what is considered 'the same moment'.

The interchange format will be developed further as converters for individual corpora are written. While not all of the information that can be extracted from the various corpora will end up being used in the EASIER translation pipeline, all of it must be parsed and understood to some degree to correctly simplify and filter it. We also believe that providing as much information as possible will help to provide flexibility in later phases of development and encourage experimentation.

Funded by the Horizon 2020
Framework Programme of the European Union

ACKNOWLEDGEMENTS

We would like to thank the following people for their most valuable input and advice: Annelies Braffort, Kearsy Cormier, Onno Crasborn, Eleni Efthimiou, and Reiner Konrad.

REFERENCES

Abuczki, Ágnes (2013). 'A *mondjuk* nem konceptuális használatának vizsgálata multimodális kontextusban'. In: *Alknyelvdok7*. Ed. by Tamás Váradi. Budapest, Hungary: Hungarian Academy of Sciences, Institute of Linguistics, pp. 3–17. ISBN: 978-963-9074-59-0. URL: http://www.nytud.hu/alknyelvdok13/proceedings13/ANyD7-Abuczki-Agnes.pdf (visited on 21/01/2022).

Bartha, Csilla, Margit Holecz and Szabolcs Varjasi (2016). 'The SIGNificant Chance Project and the Building of the First Hungarian Sign Language Corpus'. In: *10th International Conference on Language Resources and Evaluation (LREC 2016). Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* (Portorož, Slovenia). Ed. by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Hedegaard Kristoffersen and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 1–6. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/16021.pdf.

Bergman, Brita and Johanna Mesch (2004). *ECHO data set for Swedish Sign Language (SSL)*. Language Resource. Department of Linguistics, University of Stockholm. URL: http://sign-lang.ruhosting.nl/echo/ (visited on 30/11/2021).

Braffort, Annelies (2019). *Dicta-Sign-LSF Corpus Annotation manual*. Annotation Convention. Paris, France: LIMSI, CNRS, Université Paris-Saclay. 8 pp. URL: https://www.ortolang.fr/market/corpora/dicta-sign-lsf-v2?path=/data/annotations (visited on 14/09/2021).

Brentari, Diane (1998). *A prosodic model of sign language phonology*. Language, speech, and communication. Cambridge, Mass: MIT Press. 376 pp. ISBN: 978-0-262-02445-7.

Cormier, Kearsy, Jordan Fenlon, Sannah Gulamani and Sandra Smith (2017). *BSL Corpus Annotation Conventions*. Annotation Convention. Version 3.0. London, United Kingdom: Deafness Cognition and Language (DCAL) Research Center, University College London. 19 pp. URL: https://bslcorpusproject.org/wp-content/uploads/BSLCorpus_AnnotationConventions_v3.0_-March2017.pdf (visited on 21/07/2021).

Cormier, Kearsy, Jordan Fenlon and Adam Schembri (2015). 'Indicating verbs in British Sign Language favour motivated use of space'. In: *Open Linguistics* 1.1, pp. 684–707. ISSN: 2300-9969. DOI: 10.1515/opli-2015-0025.

Crasborn, Onno A. and Richard Bank (2015). *Corpus NGT Anonymisation Protocol*. Version: 1. URL: https://www.academia.edu/40438732/Corpus_NGT_Anonymisation_Protocol (visited on 31/01/2022).

Crasborn, Onno A., Richard Bank and Kearsy Cormier (2015a). *Digging into Signs: Towards a gloss annotation standard for sign language corpora*. Project report. Radboud University & University College London, p. 11. 11 pp. URL: https://www.ru.nl/publish/pages/973124/dis_annotation_guidelines_4may2015.pdf (visited on 14/09/2021).

Crasborn, Onno A., Richard Bank, Inge Zwitserlood, Els van der Kooij, Anne Meijer, Anna Sáfár and Ellen Ormel (2015b). *Annotation conventions for the Corpus NGT*. Annotation Convention. Version 3. Nijmegen, Netherlands: Centre for Language Studies & Department of Linguistics, Radboud University. 39 pp. DOI: 10.13140/RG.2.1.1779.4649.

Crasborn, Onno A., Els van der Kooij, Dafydd Waters, Bencie Woll and Johanna Mesch (2008a). 'Frequency distribution and spreading behavior of different types of mouth actions in three sign languages'. In: *Sign Language & Linguistics* 11.1, pp. 45–67. ISSN: 1387-9316, 1569-996X. DOI: 10.1075/sll.11.1.04cra.

Crasborn, Onno A. and Anna Sáfár (2016). 'An annotation scheme to investigate the form and function of hand dominance in the Corpus NGT'. In: *A Matter of Complexity: Subordination* *in Sign Languages*. Ed. by Roland Pfau, Markus Steinbach and Annika Herrmann. De Gruyter Mouton, pp. 231–251. DOI: doi:10.1515/9781501503238-010.

Crasborn, Onno A. and Han Sloetjes (2008). 'Enhanced ELAN functionality for sign language corpora'. In: *6th International Conference on Language Resources and Evaluation (LREC 2008). Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* (Marrakech, Morocco). Ed. by Onno A. Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst D. Thoutenhoofd and Inge Zwitserlood. Paris, France: European Language Resources Association (ELRA), pp. 39–43. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/08022.pdf.

Crasborn, Onno A., Els van der Kooij, Annika Nonhebel and Wim Emmerik (2004). *ECHO data set for Sign Language of the Netherlands (NGT)*. Language Resource. Department of Linguistics, Radboud University Nijmegen. URL: http://sign-lang.ruhosting.nl/echo/ (visited on 30/11/2021).

Crasborn, Onno A., Inge Zwitserlood and Johan Ros (2008b). *The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands.* Language Resource. Centre for Language Studies, Radboud University Nijmegen. URL: http://hdl.handle.net/hdl:1839/00-0000-0000-0004-DF8E-6 (visited on 29/11/2021).

Crasborn, Onno A., Inge Zwitserlood, Els van der Kooij and Richard Bank (2020). *Annotation conventions for the Corpus NGT*. Annotation Convention. Version 4. Nijmegen, Netherlands: Centre for Language Studies & Department of Linguistics, Radboud University. 46 pp. URL: https://www.ru.nl/publish/pages/1013556/corpusngt_annotationconventions_v4_1.pdf (visited on 22/07/2021).

Ebling, Sarah, Reiner Konrad, Penny Boyes Braem and Gabriele Langer (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, pp. 30–56. ISSN: 1533-6263. DOI: 10.1353/sls.2015.0024.

Efthimiou, Eleni, Kiki Vasilaki, Stavroula-Evita Fotinea, Anna Vacalopoulou, Theodoros Goulas and Athanasia-Lida Dimou (2018). 'The POLYTROPON Parallel Corpus'. In: *11th International Conference on Language Resources and Evaluation (LREC 2018). Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* (Miyazaki, Japan). Ed. by Mayumi Bono, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Hedegaard Kristoffersen, Johanna Mesch and Yutaka Osugi. Paris, France: European Language Resources Association (ELRA), pp. 39–44. ISBN: 979-10-95546-01-6. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18043.pdf.

Fenlon, Jordan, Adam Schembri and Kearsy Cormier (2018). 'Modification of indicating verbs in British Sign Language: A corpus-based study'. In: *Language* 94.1, pp. 84–118. ISSN: 1535-0665. DOI: 10.1353/lan.2018.0002.

Gabarro-Lopez, Silvia and Laurence Meurant (2014). 'The Use of Buoys across Genres in French Belgian Sign Language (LSFB)'. In: *Actes du IXème colloque de linguistique des doctorands et jeunes chercheurs du Laboratoire MoDyCo. La question des genres à l'écrit et à l'oral.* COLDOC 2013. Paris, France, pp. 43–54.

Goulas, Theodoros, Stavroula-Evita Fotinea, Eleni Efthimiou and Michalis Pissaris (2010). 'SiS-Builder: A Sign Synthesis Support Tool'. In: *7th International Conference on Language Resources and Evaluation (LREC 2010). Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* (Valletta, Malta). Ed. by Philippe Dreuw, Eleni Efthimiou, Thomas Hanke, Trevor Johnston, Gregorio Martínez Ruiz and Adam Schembri. Paris, France: European Language Resources Association (ELRA), pp. 102–105. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/10008.pdf.

Funded by the Horizon 2020 Framework Programme of the European Union

Hanke, Thomas, Sung-Eun Hong, Susanne König, Reiner Konrad, Gabriele Langer, Silke Matthes, Rie Nishio and Anja Regen (2019). *Segmentierung / Segmentation*. Project Note AP03-2010-01. Hamburg, Germany: DGS-Korpus project, IDGS, Universität Hamburg. DOI: 10.25592/uhhfdm.817.

Hanke, Thomas and Jakob Storz (2008). 'iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography'. In: *6th International Conference on Language Resources and Evaluation (LREC 2008). Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* (Marrakech, Morocco). Ed. by Onno A. Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst D. Thoutenhoofd and Inge Zwitserlood. Paris, France: European Language Resources Association (ELRA), pp. 64–67. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/08011.pdf.

Hellwig, Birgit, Dieter Van Uytvanck, Micha Hulsbosch, Aarthy Somasundaram, Maddalena Tacchetti, Jeroen Geerts and Han Sloetjes (2021). *ELAN - Linguistic Annotator*. Manual. Version 6.2. Nijmegen, Netherlands: The Language Archive, MPI for Psycholinguistics. URL: https://www.mpi.nl/corpus/manuals/manual-elan.pdf (visited on 26/01/2022).

Institute for Language and Speech Processing - Athena Research Center (2018). *Polytropon Parallel Corpus*. Version 1.0.0. URL: http://hdl.handle.net/11500/ATHENA-0000-0000-4C77-6 (visited on 30/11/2021).

Jantunen, Tommi (2013). 'Signs and Transitions: Do They Differ Phonetically and Does It Matter?' In: *Sign Language Studies* 13.2, pp. 211–237. ISSN: 1533-6263. DOI: 10.1353/sls.2013.0004.

— (2018). *Corpus of Finnish Sign Language (CFinSL)*. Language Resource. Jyväskylän yliopisto. URL: http://urn.fi/urn:nbn:fi:lb-2019012321 (visited on 30/11/2021).

Jerko, Boštjan and Špela Vintar (2015). 'SIGNOR. Annotating for Slovene Sign Language Corpus'. Poster. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/Jerko_Vintar-SSL-poster1.pdf (visited on 30/09/2021).

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, pp. 106–131. ISSN: 1569-9811. DOI: 10.1075/ijcl.15.1.05joh.

— (2013). *Auslan Corpus Annotation Guidelines*. Annotation Convention. Version 2013-02-22. Sydney, Australia: University of Sydney. 83 pp. URL: https://media.auslan.org.au/attachments/AuslanCorpusAnnotationGuidelines_Johnston.pdf.

— (2019). *Auslan Corpus Annotation Guidelines*. Annotation Convention. Version 2019-08. Sydney, Australia: University of Sydney. 105 pp. URL: https://www.academia.edu/40088269/Auslan_Corpus_Annotation_Guidelines_August_2019_version_.

Konrad, Reiner, Thomas Hanke, Susanne König, Gabriele Langer, Silke Matthes, Rie Nishio and Anja Regen (2012). 'From form to function. A database approach to handle lexicon building and spotting token forms in sign languages'. In: *8th International Conference on Language Resources and Evaluation (LREC 2012). Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon* (Istanbul, Turkey). Ed. by Onno A. Crasborn, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Jette Hedegaard Kristoffersen and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 87–94. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/12023.pdf.

Konrad, Reiner, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn and Marc Schulder (2020a). *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public corpus of German Sign Language, 3rd release*. Language Resource. Version 3. Universität Hamburg. DOI: 10.25592/dgs.corpus-3.0.

Konrad, Reiner, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio and Anja Regen (2020b). *Public DGS Corpus: Annotation Conventions*. Project Note AP03-2018-01. Version 3. Hamburg, Germany: DGS-Korpus project, IDGS, Universität Hamburg. DOI: 10.25592/uhhfdm.1860.

Kopf, Maria, Marc Schulder and Thomas Hanke (2021). *Overview of Datasets for the Sign Languages of Europe*. Project deliverable D6.1. Version 1.0. EASIER Consortium. DOI: 10.25592/UHHFDM.9560.

Kristoffersen, Jette Hedegaard and Thomas Troelsgård (2015). 'The Danish Sign Language Corpus Project. Basic annotation conventions of The Danish Sign Language Corpus Project compared to the BSL and NGT conventions described in the "Digging into Signs" project.' Poster. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/Kristoffersen_Troelsgaard_DTS_poster.pdf (visited on 30/09/2021).

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) (2020). *Dicta-sign-LSF-v2*. Language Resource. URL: https://hdl.handle.net/11403/dicta-sign-lsf-v2/v1 (visited on 29/11/2021).

Liddell, Scott K. (2003). *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press. ISBN: 978-0-521-81620-5. DOI: 10.1017/CBO9780511615054.

Matthews, Patrick and Sarah Sheridan (2015). 'Digging into Signs Workshop'. Presentation. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/ISL-Matthews-Sheridan-Digging-into-Signs-Workshop.pdf (visited on 30/09/2021).

Mesch, Johanna, Lars Wallin, Anna-Lena Nilsson and Brita Bergman (2012). *Datamängd. Projektet Korpus för det svenska teckenspråket 2009-2011*. Version 1. Sign Language Section, Department of Linguistics, Stockholm University. URL: https://ling33.ling.su.se/sslc/video/ (visited on 30/11/2021).

Meurant, Laurence (2015). *Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB)*. LSFB-Lab, University of Namur. URL: http://www.corpus-lsfb.be (visited on 30/11/2021).

Nonhebel, Annika, Onno A. Crasborn and Els van der Kooij (2004a). *Sign language transcription conventions for the ECHO Project*. Annotation Convention. Version 9. Nijmegen, Netherlands: University of Nijmegen. 11 pp. URL: http://hdl.handle.net/2066/57889 (visited on 21/07/2021).

— (2004b). *Sign language transcription conventions for the ECHO Project. BSL and NGT mouth annotations*. Annotation Convention. Version 2. ECHO project, University of Nijmegen. URL: http://hdl.handle.net/2066/57889 (visited on 20/01/2004).

— (2004c). *Sign language transcription conventions for the ECHO Project. SSL mouth annotations*. Annotation Convention. Version 2. ECHO project, University of Nijmegen. URL: http://hdl.handle.net/2066/57889 (visited on 20/01/2004).

Rutkowski, Paweł, Joanna Filipczak and Anna Kuder (2015). 'PJM Corpus Annotation Guidelines'. Presentation. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/RutkowskiFilipczakKuder2015_Final.pdf (visited on 30/09/2021).

Rutkowski, Paweł, Sylwia Łozińska, Joanna Filipczak, Joanna Łacheta and Piotr Mostowski (2013). 'Jak powstaje korpus polskiego języka migowego (PJM)?' In: Polonica 33, pp. 297–308. URL: https://www.plm.uw.edu.pl/wp-content/uploads/2012/07/Rutkowski-et-al-Jak-powstaje-korpus-polskiego-jezyka-migowego.pdf (visited on 25/11/2021).

Salonen, Juhana, Antti Kronqvist and Tommi Jantunen (2020). 'The Corpus of Finnish Sign Language'. In: *12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Funded by the Horizon 2020 Framework Programme of the European Union

Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (Marseille, France). Ed. by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Hedegaard Kristoffersen and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 197–202. ISBN: 979-10-95546-54-2. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/20004.pdf.

Salonen, Juhana, Tuija Wainio, Antti Kronqvist and Jarkko Keränen (2019). *Suomen Viittomakielten Korpusprojektin (CFINSL) Annotointiohjeet*. Annotation Convention. Jyväskylä, Finland: Department of Linguistics and Communication Sciences, Sign Language Center, University of Jyväskylä. 40 pp. URL: https://www.jyu.fi/hytk/fi/laitokset/kivi/opiskelu/tutkinto-ohjelmat-ja-oppiaineet/viittomakieli/tutkimus-2/suomen-viittomakielten-korpusprojekti/cfinsl_annotointiohjeet_2019_2versio.pdf (visited on 28/07/2021).

Sandler, Wendy and Diane Lillo-Martin (2006). *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press. ISBN: 978-1-139-16391-0. DOI: 10.1017/CBO9781139163910.

Santoro, Mirko and Carlo Geraci (2015). 'Italian Sign Language (LIS) Corpus'. Presentation. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/Digging-LIS-corpus-final.pdf (visited on 30/09/2021).

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, Kearsy Cormier, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn and Marc Schulder (2017). *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition)*. Language Resource. London: University College London. URL: http://www.bslcorpusproject.org/ (visited on 26/11/2021).

Sinte, Aurélie, Christophe De Clerck, Sibylle Fonzé, Susana Sanchez, Gauthier Raes and Laurence Meurant (2015). 'Corpus LSFB (French Belgian Sign Language). Current annotation conventions compared to the "Digging into signs" suggestions'. Poster. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/LSFB_Sinte-et-al._Poster.pdf (visited on 30/09/2021).

Sutton-Spence, Rachel and Linda Day (2001). 'Mouthings and Mouth Gestures in British Sign Language'. In: *The Hands are the Head of the Mouth: the Mouth as Articulator in Sign Languages (International Studies on Sign Language and the Communication of the Deaf, vol 39.)* Ed. by Penny Boyes Braem and Rachel Sutton-Spence. Hamburg: Signum Press, pp. 69–87. ISBN: 978-3-927731-83-7.

Troelsgård, Thomas and Jette Hedegaard Kristoffersen (2018). 'Improving Lemmatisation Consistency without a Phonological Description. The Danish Sign Language Corpus and Dictionary Project.' In: *11th International Conference on Language Resources and Evaluation (LREC 2018). Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* (Miyazaki, Japan). Ed. by Mayumi Bono, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Hedegaard Kristoffersen, Johanna Mesch and Yutaka Osugi. Paris, France: European Language Resources Association (ELRA), pp. 195–198. ISBN: 979-10-95546-01-6. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18009.pdf.

Verstraete, Sam, Hannes De Durpel, Hilde Nyffels, Eline Demey, Myriam Vermeerbergen and Mieke Van Herreweghe (2015). 'The Flemish Sign Language Corpus (Corpus VGT): Annotation Practice'. Poster. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/Verstraete_etal2015.jpg (visited on 30/09/2021).

Wallin, Lars and Johanna Mesch (2015). ‘Swedish Sign Language Corpus’. Presentation. Digging into Signs Workshop: Developing annotation standards for Sign Language Corpora (University College London). URL: https://bslcorpusproject.org/wp-content/uploads/MeschWallin_DiggingIntoSigns_London2015.pdf (visited on 30/09/2021).

— (2018). *Annoteringskonventioner för teckenspråkstexter*. Annotation Convention. Version 7. Stockholm, Sweden: Avdelningen för teckenspråk, Stockholms universitet. 49 pp. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-152163 (visited on 21/07/2021).

Woll, Bencie, Rachel Sutton-Spence and Dafydd Waters (2004). *ECHO data set for British Sign Language (BSL)*. Language Resource. Department of Language and Communication Science, City University (London). URL: http://sign-lang.ruhosting.nl/echo/ (visited on 30/11/2021).

Funded by the Horizon 2020
Framework Programme of the European Union