The Danish Sign Language Corpus Project

Primary purposes of the corpus

The Danish Sign Language (DTS) corpus primarily aims to provide tools for the DTS Dictionary project, and will therefore be designed for:

- investigating the lexicon of DTS, e.g. as part of the lemma selection process for the DTS Dictionary, including lexical and phonological variation and mouthing movements.
- analysing the use and semantics of DTS signs as part of the editing process in the DTS Dictionary.
- providing DTS usage examples for use in the DTS Dictionary.
- providing collocational information of DTS signs as part of the editing process in the DTS Dictionary.

This means, that the basic annotation will probably be narrowed down to the following:

- Sign, preferably to a level of detail where phonological variants are told apart.
- Mouthing or mouth movement (preferably).
- Meaning in context.

This rather modest goal has been set partly to speed up the annotation process for the basic annotation level, partly in order to be able to engage several types of co-workers in the annotation process – primarily deaf consultants and hearing interpreter students – without the need of too much education in linguistic analysis; a good SL knowledge will suffice. Linguistic analysis will not be a part of the annotation task; it will be a part of the dictionary editing process.

Besides the dictionary-related purposes, the DTS corpus will be made accessible for SL researchers, including teachers and students at the DTS interpreter's education at UCC, Denmark. For this reason, it is important that the basic tier structure is designed in way that allows for expansion with project-specific tiers, so that parts of the corpus could be annotated in greater detail in connection with particular SL research projects.

Basic project info

The DTS corpus project is estimated for 6 years:

- 2014-2018: building a DTS corpus
- 2017-2020: expanding the DTS Dictionary on the basis of the DTS corpus

The DTS corpus project is in its first stage, where the methods and tools for building a DTS corpus are investigated. Furthermore, a prototype with basic annotation of a small number of DTS recordings will be built.

The primary aim of the project is to establish an annotated DTS corpus as a tool for the DTS Dictionary project. The expected outcome is 70 hours annotated on a basic level.
Because of limited funding, the first project stage will include only one camera recordings made in connection with the DTS Dictionary project 2001- 2008.

Basic annotation conventions of The Danish Sign Language Corpus Project compared to the BSL and NGT conventions described in the "Digging into Signs" project.

Yellow background signifies that a topic is still under consideration

Topic	**General practice**	***BSL guidelines section* /** **BSL exception/specification**	***NGT guidelines section* /** **NGT exception/specification**	**DTS - preliminary guidelines**
1. *Basic gloss*	All lexical signs are annotated using an identifying gloss (*ID-gloss*), written in upper case. This gloss corresponds to 'Annotation ID-gloss' in SignBank	*4.3* 'Annotation ID gloss' corresponds to lemma (citation form) only, not phonological variants	*5.3.2* 'Annotation ID gloss' may be lemma or phonological variant	ID-gloss may be lemma or phonological variant. Examples: SIGN, SIGN~a, SIGN~b ——— We are considering working with ID-glosses on two or three hierachical levels (cf. the DGS Corpus) partly in order to be able to assign uncertain variants to certain "super-types", partly in order to establish a level that facilitates one-to-one linking between the lexical sign base and our dictionary entry list. ——— ***Like NGT***
2. *Two-handed signs*	Two handed signs annotated on both right and left hand tiers.	*4* Start and end of two-handed signs follows dominant hand. Perseveration is not annotated. Meaningful/intentional perseveration on nondominant hand is annotated as buoy (see below).	*5.3.3.2* Start and end of two-handed signs is determined independently for each hand.	Start and end of two-handed signs is determined independently for each hand. Regular/irregular tokens can be found by combining the annotation with the basic info on the sign in the sign base. ——— We are considering the "double tokens" model used in the DGS Corpus. ——— ***Like NGT***
3. *Buoys*	Buoys get their own ID-gloss	*4.7* List, Pointer, Fragment and Theme buoys. Start and end of buoy follows start and end of co-occuring sign on dominant hand.	*5.3.7* Only List buoys (other buoys are not separately glossed, instead, non-dominant hand glosses are extended in duration; see also **2.** *Two-handed signs*, above)	Buoys are not specifically glossed, instead, non-dominant hand glosses (e.g. FIRST, SECOND, THIRD) are extended in duration. ——— ***Partially like NGT***

4. *Lexical variants*	Lexical variants (same or similar/related meanings but differ in two parameters or more from each other) are suffixed.	*4.3* Use a numeral tag (note that the first lexical variant is not indicated with a number. This simplifies adding variants to existing glosses)	*5.3.9.1* Use a dash and a letter (lexical and phonological variants are not distinguished)	Lexical variants (same meanings but differ in two parameters or more from each other) are suffixed: SIGN~1, SIGN~2. If there is also phonological variation (same meaning but differ in only one parameter), suffixes ~a, ~b etc. are added: SIGN~1, SIGN~2~a, SIGN~2~b. —— ***Almost like BSL***
6. *Repetition*	Repeated signs are annotated seperately	*4.3*		Repeated signs are annotated separately, unless the repetition is a regular modification, e.g. the plural of a sign. In theese cases the modification is placed on a separate child tier (if annotated). —— ***Like BSL and NGT***
7. *Compounds*	One ID-gloss for lexicalised compounds in SignBank, or ^ between ID-glosses for possible compounds (e.g. GRAPHIC^ART with GRAPHIC and ART in SignBank but not GRAPHIC^ART)	*4.3*	*5.3.14* No ^ for possible compounds	No use of caret. One ID-gloss for each lexicalised compound in the sign base. Non-lexicalised (possible) compounds are glossed as two consecutive signs. —— ***Like NGT***
8. *Manual negative incorporation*	ID-Glossed using a negative suffix	*4.3*	*5.3.13*	These signs are glossed according to their meaning, typically (but not necessarily) including "-NOT" (which can also appear in glosses for other sign types)
9. *Directional verbs*	Directional verbs receive an ID-gloss	*4.3; 5* Grammatical/modification info on separate project-specific tiers.	*5.3.8* ID-gloss is pre- or suffixed with person number when it makes use of the 1SG location ‘near body’ or ‘on chest’	Directional verbs receive a normal ID-gloss. Grammatical/modification info is placed (if annotated) on separate tiers. —— ***Like BSL***

10. *Plurality*	Plural forms receive an ID-gloss	*4.3; 5* Grammatical info on separate project-specific tiers (see also **18.** *Points*).	*5.3.10* ID-gloss is suffixed .PL	Lexicalised plural forms receive an ID-gloss if they are not regular plural modifications of a corresponding singular form, or if the plural sign has meanings other than the mere plural of the base sign. CHILD / CHILDREN, TREE / FOREST. In other cases the plural is considered grammatical info, and is placed on separate tiers. ——— ***Partially like BSL***
11. *Numbers*	Numbers receive their own ID-glosses	*4.4* Numbers are written in words	*5.3.7* Numbers are written in digits	Numbers are written in words ——— ***Like BSL***
12. *Number sequences*	Number sequences receive one ID-gloss (see also **7.** *Compounds)*	*4.4* Carets specify the separate parts	*5.3.7* No specification of the separate parts	Separate ID-glosses, no carets. Example: NINETEEN-HUNDRED NINE EIGHTY.
13. *Number incorporation*	ID-gloss is suffixed with information about incoporated number	*4.4* ID-gloss is used for number	*5.3.7*	These signs are glossed according to their meaning, typically (but not necessarily) including glosses for the relevant incorporated number sign. Examples: TWO-HOURS / FOUR-HOURS / SIX-HOURS, TOMORROW / IN-TWO-DAYS / IN-THREE-DAYS, FIRST-FLOOR / SECOND-FLOOR.
14. *Ordinal numbers*		ID-glossed as lexical signs, some of which allow number incorporation (see **13**)	*5.3.7* Suffixed -ORD	Separate ID-glosses, no suffix. ——— ***Like BSL***

15. *Sign names*	Prefixed	*4.5* Prefixed SN: followed by first name only, unless fingerspelled then fingerspelling conventions are followed (see **17.** *Fingerspelling*). If sign name is homonym with lexical sign, ID gloss for homonym is added in brackets	*5.3.17* Prefixed * followed by both first and last name	Separate ID-glosses, no prefix/suffix, and no regard to phonologically identical lexical (non-name) signs. Examples: Known, unique individual: WILLIAM-STOKOE Known, unique building, institution etc.: THE-PARLIAMENT Lexicalised surname or first name: SOPHIE, PETER, RASMUSSEN At the moment we have no rules for unknown (or partially unknown) sign names.
17. *Finger-spelling*	Prefixed	*4.8.4* Prefixed FS: followed by word if fingerspelled fully, or by intended word followed by actual letters used if not fingerspelled fully	*5.3.16* Prefixed # followed by *perceived* letters. The presumed intended word is on the *Meaning* tier.	A word rendered through fingerspelling is written in conventional spelling followed by (H). Example: Maria(H) At the moment we have no rules for words that are rendered incorrectly (or only partially). ——— ***Almost like BSL and NGT***
18. *Pointing signs*	Pointing signs receive the gloss PT. with suffixes.	*4.8.1* *Function* of points including pronominal, locative or determiner functions is annotated. Points to buoys are annotated as PT: followed by type of buoy with particular fingers unspecified.	*5.3.11* *Direction* of points is annotated for spatial directions like up and down. Location of points is annotated for pointing to specific fingers of the weak hand. The referent of a pointing sign is annotated on the *Reference* child tier(s). Grammatical class distinctions may be specified on the *GrammClass* child tier.	Pronominal points to the 1SG location ‘near body’ or ‘on chest’ are glossed: I All other points are glossed: POINT Direction and location of POINT are (if annotated) placed in a separate tier called "locus". Function and reference of POINT are not described at the moment, but could be annotated on separate tiers.

19. *Classifier/ depicting signs*	One of four elements for movement (MOVE, PIVOT, AT, BE), followed by classifier handshape	*4.8.2* Additionally add prefix for type of depicting sign: whole entity (DSEW), part entity (DSEP), or handling (DSH). Handshape list includes same handshapes as for NGT but some additional handshapes based on Brennan (1992) and pilot annotations.	*5.3.19* Meaning to be annotated on *Meaning* tier. More restricted handshape list than BSL.	At the moment 50 classifiers (classificatory verb stems) are identified, and given ID-glosses, all starting with PF-. Classifier constructions are annotated with these glosses, and a free text description of the movement/meaning is added on a separate child tier. We will consider adding a formal movement description like MOVE, PIVOT, AT, BE
20. *Shape constructions*		*4.8.2* Annotated as type of classifier/depicting sign (DSS) but without movement	*5.3.20* Glossed as SHAPE + handshape. Meaning to be annotated on *Meaning* tier	At the moment appr. 10 basic shape signs are identified, and given ID-glosses. We will consider adding a prefix to these glosses (and a sign level meaning tier, where the meaning in the current context can be specified).
21. *Type-like Classifier / depciting signs*		*4.8.2* For common whole entity signs mainly depicting humans, animals and vehicles, specify handshape as well as orientation, and the referent (human, animal, entity, vehicle)	Annotate as classifier/depicting signs.	Annotated as classifiers/depicting signs.
22. *Gestures*		*4.8.3* Prefixed G	*5.3.18* Generic character % for non-emblematic gesticulation. Emblematic gestures are included as lexical items in Signbank without a % prefix.	At the moment no rules.
23. *Palm up*		*4.8.3*	*5.3.12*	ID-gloss: PRESENTATION-GESTURE —— ***Like NGT***

24. *Manual constructed action*		*4.8.3* Prefixed G:CA	Marked with %, meaning described on *Meaning* tier	At the moment no rules.

Additional topics

7.b Affixes				A small number of sign prefixes and suffixes have been identified, typically calques from spoken Danish. These signs have ID-glosses starting or ending with ^. Examples: UN^, ^S-GENITIVE
17.b *Mouth-Hand-System*				A word rendered through the mouth-hand-system is written in conventional spelling followed by (M). Example: Sahara(M). At the moment no rules for words that are rendered incorrectly (or only partially).
17.c Initialised signs				29 glosses (one for each letter in the Danish alphabet) are used for the annotation of initialised signs which are not considered lexicalised, and therefore not given individual ID-glosses. The glosses of these signs typically begins with INITIALISED- Examples: INITIALISED-C, INITIALISED-F We will consider adding a sign level meaning tier, where the meaning in the current context can be specified.

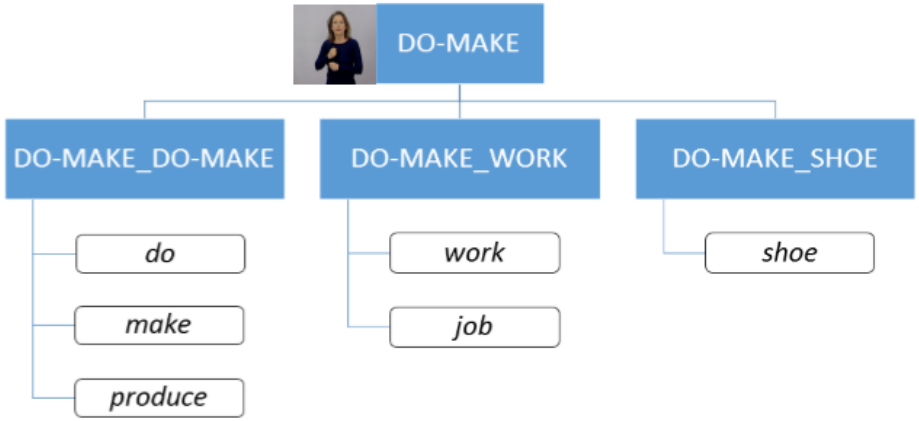
Reuse of dictionary ID-glosses

The DTS corpus project is closely related to the DTS Dictionary, and the ID-glosses used in the dictionary project constitute the base of the type inventory that will be used for the corpus annotation.

Thus, the 2.900 glosses that currently represent lemmas or lemma variants in the dictionary are ”ready to use”, and another 5.000 glosses can be chosen from a raw base with signs that are not (yet) selected as dictionary lemmas. At the moment, however, this database includes some messy data (duplicates etc.), but we hope to be able to combine the clean-up (and merger with the main sign base) with the identification of new or unclear signs encountered during the annotation work.

Multi-level glosses

For the basic corpus annotation, tokens can be identified at different levels of detail, the most detailed being a level that resembles the variant level in the DTS Dictionary. On top of this level we consider following the model of the DGS Corpus by adding one or two levels of “super types”. This approach enables the annotator to identify a base sign, if the token does not exactly match one of the more detailed sub-ordinate types. It also facilitates linking between dictionary and corpus not at “super type” level, but at a lower level, allowing the lexicographer to move meanings, partially or entirely, from one entry to another, without affecting existing corpus annotations.



Example of a sign with three sub-types, linked to three possible different dictionary entries.

Uncertainties

At the moment, the test annotations of the DTS Corpus project has only been of adapted, "nice" SL texts from the DTS dictionary, i.e. text based on natural utterances, but not 100% natural language. For this reason, we haven't yet defined rules for uncertainty-related problems such as unknown signs, partial or erroneous fingerspelling, invisible elements, false starts etc.

Jette H. Kristoffersen jehk@ucc.dk
Thomas Troelsgård ttro@ucc.dk

The Danish Sign Language Courpus
The Danish Sign Language Dictionary

Centre for Sign Language, UCC, Denmark

info@tegnsprog.dk

OVER ORDBOG DANSK TEGNSPROG