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 to be used 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

1. Basic gloss	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	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 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 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 separately, unless the repetition is a regular modification, e.g. the plural of a sign. In these cases the modification is placed on a separate child tier (if annotated). *Like BSL and NGT*
7. 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	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 a normal ID-gloss. Grammatical/ modification info is placed (if annotated) on separate tiers. *Like BSL*
10. Plurality	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 are written in words *Like BSL*
12. Number sequences	Separate ID-glosses, no carets. Example: NINETEEN-HUNDRED NINE EIGHTY.
13. Number incorporation	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	Separate ID-glosses, no suffix. *Like BSL*

15. Sign names	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	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	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	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	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 / depicting signs	Annotated as classifiers/depicting signs.
22. Gestures	At the moment no rules.
23. Palm up	ID-gloss: PRESENTATION-GESTURE *Like NGT*
24. Manual constructed action	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

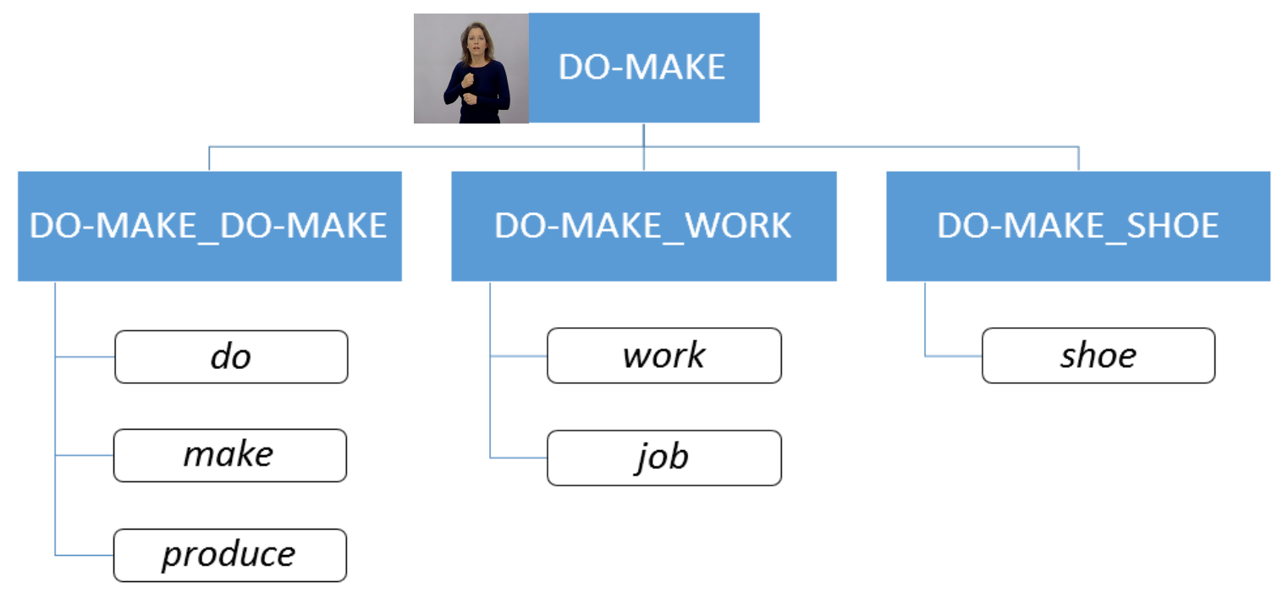
The DTS corpus project is closely related to the DTS Dictionary, and the ID-glosses used in the dictionary project constitute the base of the type inventory that will be used for the corpus annotation.

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 annotaion 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.



DO-MAKE

DO-MAKE_DO-MAKE

do

make

produce

DO-MAKE_WORK

work

job

DO-MAKE_SHOE

shoe

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.

OVER ORDBOG DANSK TEGNSPROG

The Danish Sign Language Courpus
The Danish Sign Language Dictionary

Jette H. Kristoffersen
Thomas Troelsgård

Centre for Sign Language, UCC, Denmark

info@tegnsprog.dk