

ANNOTATION CONVENTIONS IN CORPUS AND SIGNWIKI PROJECT

Sanna Nordlund, Leena Savolainen & Päivi Mäntylä
Corpus and SignWiki Project / The Finnish Association of the Deaf

1. INTRODUCTION

- The ID-glossing, i.e. lemmatization (Johnston 2008) and the creation of annotation conventions for ID-glossing Finnish Sign Language (FinSL) was started in the Corpus and SignWiki Project, at the Finnish Association of the Deaf (FAD) in Helsinki in 2014.
- Annotation conventions are mainly based on Auslan annotation conventions (Johnston 2014).
- Data: Language Policy Programme for the National Sign Languages in Finland (2010), the FinSL edition, 2 1/2 h of signed formal monologue
- Mission: to create the basis for a multi-purpose FinSL corpus annotation and its conventions that can be enriched and developed in the future
- Lexical database: we temporarily use the administrative site of the Suvu online dictionary (see Figure 1), but we plan to replace it by the Signbank!
- In the ongoing first phase, the focus is on ID-glossing the fully lexical signs, while other kinds of signs are given only a general tag, apart from some exceptions, so that they can be analyzed and annotated more specifically in the future.
- Cooperation with the Sign Language Centre at the University of Jyväskylä



Figure 1. Sign OLLA 'have', 'exist' in the administrative site of the Suvu online dictionary², our temporary lexical database. The sign has two phonological variants that differ in handshape. The handshape variants are described in the database.

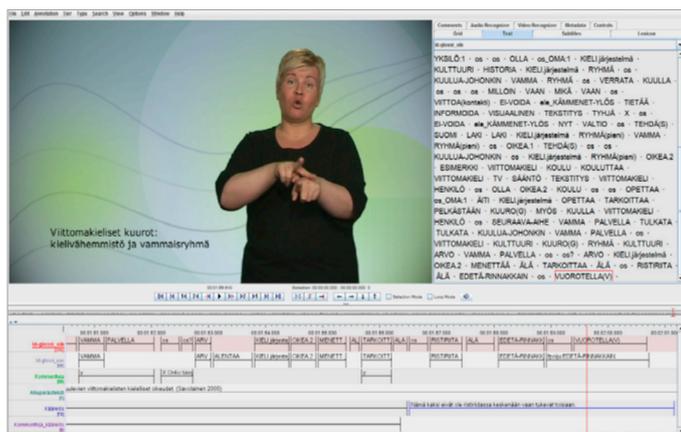
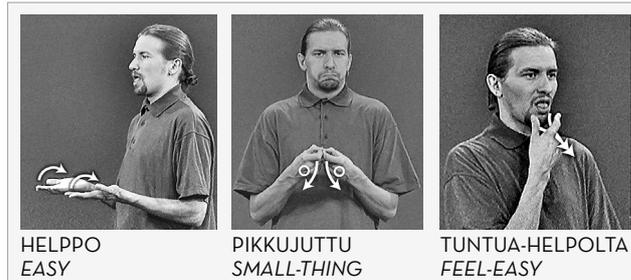


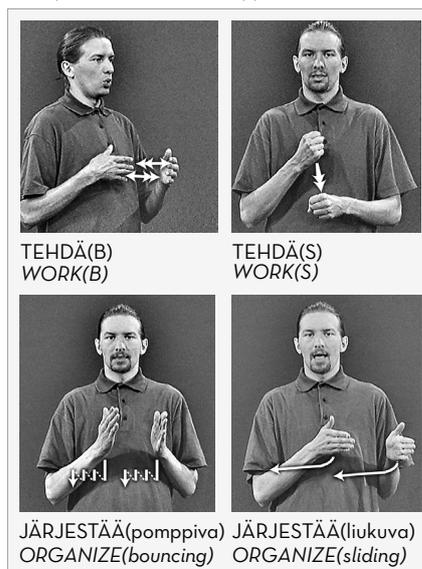
Figure 2. ELAN screenshot showing the tiers Id-glossi_oik (ID-gloss right hand), Id-glossi_vas (ID-gloss left hand), Kommentteja (comments), Alkuperäisteksti (original text), Käännös (translation) and Kommentteja_käännös (comments on translation). A tier Meaning for contextual meaning of each token is planned to be added to the basic template. The screenshot demonstrates also the annotation of two-handed signs on two tiers, the annotation of fragment buoys (fpouju) and the signs co-occurring with them as well as the length of the annotations. In addition to the annotation tiers, the text viewer shows many different ID-glosses.

2. ID-GLOSSING PRINCIPLES FOR LEXICAL SIGNS

- Each lexeme is given a unique ID-gloss.
- The citation form of the lexeme, the lemma, and its phonological variants are described in the lexical database (see figure 1).
- Choosing the ID-gloss:
 1. A unique Finnish word is used when possible, even for (near) synonyms. The word represents the core meaning of the sign.



2. If the same word needs to be used, the following specifications can be applied:



a. A symbol or a word referring to a structural feature (e.g. handshape or movement) of the sign is suffixed in parentheses.



b. A word that hints at the meaning is suffixed, separated by a dot.



c. A number suffix is added after a dot. (This is the very last option and it will probably be replaced with alternatives a and b, because numbers are opaque and hard to remember.)

3. START AND END OF A SIGN

- Theoretically, we accept the view according to which a sign is a relatively long unit, and so-called transitional movements are part of the sign (Jantunen 2015). Therefore, only little space is left between the annotations, often just one frame.
- Defining exact criteria for the beginning and end of a sign has proven to be difficult: for example, the order in which the sign's structural features appear varies, and the beginning and the end of two signs often merge with each other.
- For two-handed signs, the start and end of the sign follows the dominant hand, as in BSL conventions.
- When the two hands produce two different signs, the tiers remain independent, as in NGT conventions.
- For buoys and signs co-occurring with them, the tiers also remain independent (see Figure 2).

4. OTHER PRACTICES COMPARED TO THE DIGGING PROJECT

TOPIC	THE PRACTICE IN CORPUS AND SIGNWIKI PROJECT	EXAMPLE	
TWO-HANDED SIGNS	The ID-gloss is annotated on both hand tiers (on the length of the annotations, see section 3.)	TEHDÄ(B) TEHDÄ(B)	WORK(B) WORK(B)
BUOYS	<ul style="list-style-type: none"> • List buoys and fragment buoys are annotated as in BSL, except that the codes are in Finnish and in lower case. The handshape is not specified. • Only intentional usage of buoys is annotated, as in BSL. • On fragment buoys, also the ID-gloss of the sign of which it is a fragment is given after the colon (see Figure 2), as in Auslan (Johnston 2014). 	os:lpouju lpouju os:fpouju fpouju:LAKI	pt:lbouy lbouy pt:fbouy fbouy:LAW
REPETITION	Repeated signs are annotated separately as in BSL and NGT corpora, when articulated clearly separately. Grammatical repetition is not annotated at this stage.	os os os	pt pt pt
MANUAL NEGATIVE INCORPORATION	A negative suffix NEG separated by a plus. Note: The word EI 'not' is used in signs having inherent negative meaning.	TIETÄÄ+NEG	KNOW+NEG
NUMBER INCORPORATION	The incorporated number is added in parentheses at the moment. The parentheses might be replaced by a plus in the future to indicate the morpheme status of the number handshape.	TUNTI(4) TUNTI+4	HOOR(4) HOOR+4
SIGN NAMES, FINGERSPELLING, DEPICTING SIGNS AND GESTURES	A transparent code (prefix) in lower case letters. More specific annotation of these sign types has not been decided yet.	nv (nimiviittoma) sa (sormiaakkoset) kv (kuvailleva viittoma) ele (ele)	sn fs ds gest
POINTING SIGNS	<ul style="list-style-type: none"> • At the first stage, only a code, apart from some exceptions. • Pointing towards signer's chest: seems to be lexicalized to mean the 1st person singular. No evidence on the lexicalization of the other pointings as 2nd or 3rd person pronouns (see, e.g. Meier 1990, Johnston 2013). • Meaning of the pointing is planned to be annotated on the meaning tier. 	os (osoitus) os:l	pt pt:l

5. CONSISTENCY WITH THE PRACTICE OF GLOSSING SPOKEN LANGUAGES?

The Leipzig Glossing Rules³ have become a worldwide standard in spoken language research⁴. In our opinion, it would be important to discuss the consistency between the standard suggested by the Digging into Signs project and the Leipzig Glossing Rules. A consistent practice would facilitate the participation in the typological discussion. However, if differing conventions are chosen, they should be justified in the guidelines.

6. CONCLUSIONS

- Since the FinSL and BSL conventions are both based on Auslan annotation conventions, they have quite a lot in common. However, in some points we agree more with the conventions for NGT.
- According to our view, the basic principle should be that the ID-gloss tier is reserved only for the ID-glosses (lemmas), and all the other (grammatical & semantic) information is annotated on separate tiers.
- We prefer using a different word for each ID-gloss, if possible, and transparent suffixes when needed (see section 2). As well, we prefer using abbreviations for different sign types (e.g. depicting signs and fingerspelling), since they are easier for the annotator and the user of the corpus to remember than meaningless codes.

REFERENCES

- Crasborn, Onno, Richard Bank & Kearsey Cormier 2015.** *Digging into Signs: Towards a gloss annotation standard for sign language corpora. Draft, February 2015.*
- Jantunen, Tommi 2015.** *How long is the sign?* *Mouton de Gruyter, Linguistics 2015; 53(1): 93-124.*
- Johnston, Trevor 2008.** *Corpus linguistics and signed languages: no lemmata, no corpus. Proceedings of the Third Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (LREC), Marrakech, Morocco, 82-87.*
- Johnston, Trevor 2013.** *Formational and functional characteristics of pointing signs in a corpus of Auslan (Australian sign language): are the data sufficient to posit a grammatical class of 'pronouns' in Auslan?* *Corpus linguistics and linguistic theory, Vol. 9, No. 1, p.109-159.*
- Johnston, Trevor 2014.** *Auslan Corpus Annotation Guidelines, Sydney, Australia: Macquarie University.*
- Meier, Richard P. 1990.** *Person deixis in American Sign Language. Susan D. Fischer & Patricia Siple (eds.), Theoretical Issues in Sign Language Research. Volume 1: Linguistics, 175-190. Chicago & London: The University of Chicago Press.*



FINNISH
ASSOCIATION
OF THE DEAF

KONEEN SÄÄTIÖ



1 <http://www.auslan.org.au>

2 <http://suvi.viittomat.net>

3 <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>

4 *The Generic Style Rules for Linguistics*, <http://www.eva.mpg.de/linguistics/staff/martin-haspelmath/generic-style-rules.html>