Establishment of a corpus of Hong Kong Sign Language acquisition data: from ELAN to CLAN

Cat Fung H-M, Scholastica Lam, Joe Mak, Gladys Tang
Centre *for* Sign Linguistics and Deaf Studies
203, Academic Building #2, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
E-mail: cat_cslds@cuhk.edu.hk, schola_cslds@cuhk.edu.hk, joemakkl@yahoo.com.hk, gtang@arts.cuhk.edu.hk

Abstract

This paper introduces the Hong Kong Sign Language Child Language Corpus currently developed by the Centre for Sign Linguistics and Deaf Studies, the Chinese University of Hong Kong. When completed, the corpus will include both longitudinal and cross-sectional data of deaf children acquiring Hong Kong Sign Language. Our research team has decided to establish a meaning-based transcription system compatible with both the ELAN and CLAN programs in order to facilitate future linguistic analysis. The ELAN program, which allows multiple-tier data entries and synchronization of video data with glosses, is an ideal tool for transcribing and viewing sign language data. The CLAN program, on the other hand, has a wide range of well-developed functions such as auto-tagging and the 'kwal' function for data search and they are extremely useful for conducting quantitative analyses. With add-on programs developed by our research team and additional functions in CLAN developed by the CHILDES research team, the transcribed data are transferable from the ELAN format to CLAN format, thus allowing researchers to optimize the use of both programs in conducting different types of linguistic analysis on the acquisition data.

1. Introduction

The establishment of the Hong Kong Sign Language Child Language Corpus began in 2002 as one of the research outputs of two RGC-funded research projects entitled "Development of Hong Kong Sign Language by Deaf Children" and "Acquisition of Classifiers in Hong Kong Sign Language by Deaf Children". The major goal of this corpus is to collect, transcribe and tag acquisition data of Hong Kong Sign Language (hereafter HKSL) that would facilitate the long-term development of sign language acquisition research. When completed, the corpus will contain acquisition data collected both longitudinally and cross-sectionally. The transcription system of the corpus is based on the CHAT format with additional symbols for properties specific to sign languages, thanks to the assistance and advice from the research team of the Child Language Data Exchange System (CHILDES) headed by Brian MacWhinney. The finalized transcriptions are compatible with the CLAN program of CHILDES as well as the ELAN program developed by Max Planck Institute for Psycholinguistics Nijmegen, The Netherlands. A major strength of this transcription system is that researchers can have full access to the existing features or functions of both programs. On the other hand, researchers can compare signed and spoken acquisition data with ease using the CLAN interface. This paper describes the procedures we went through in transcribing the HKSL acquisition data: how the data were first transcribed in ELAN and then exported to a format compatible with CLAN. In Section 2 we will briefly introduce our transcription system. Section 3 describes the initial transcription procedure. Section 4 explains the technical steps involved in exporting the data from ELAN to CLAN. Section 5 discusses the difficulties we encountered in the process of transferring the data. Section 6 is the conclusion.

2. Transcription system developed by the Hong Kong Sign Language acquisition research team

Our research team aimed at achieving the following goals when developing the transcription system of the Hong Kong Sign Language Child Language Corpus:

(a) The transcription system must be transparent enough for easy viewing. That is, the transcribed data should be accompanied with an appropriate amount of linguistic information presented in an easy-to-read format.
(b) The transcription system should be compatible with other well-established computerized corpora so that researchers can make full use of the functions of these programs and solicit technical support from the developers of these programs when necessary.
(c) The transcription system should facilitate cross-linguistic and cross-modal comparative studies.

For the ease of data viewing, all lexical signs are glossed with English word(s) which bear the closest possible meanings, e.g. *BOOK*, *FATHER*, *DANGEROUS* (see Figure 1). If more than one English word is needed to stand for the meaning of a sign, an underscore is used to connect these English words, as in *NAME_SIGN* and *BIRTHDAY_CAKE*. [1] If there are several synonyms in

[1] In the sign language literature, English words that are used to gloss the meaning of a sign are usually connected by hyphens.

English that can match the meaning of a sign, only one is chosen to ensure the consistency and accuracy of data coding. Supplementary codings are adapted from the CHAT specification to mark grammatical properties specific to sign languages. For example, the gloss of a spatial verb is followed by a hyphen and a small letter that indicates the locative affixes, as in *PUT-a*, *PUT-b* and *PUT-c*. Note that at this initial stage of transcription the letters ‘a’, ‘b’ and ‘c’ are abstract in nature – they do not represent specific locations in the signing space. Rather, they simply show that locative marking is present with the glossed sign (see Figure 2 and 3).

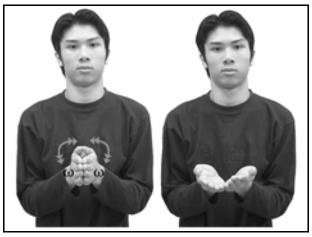


Figure 1: lexical sign for *BOOK* in HKSL

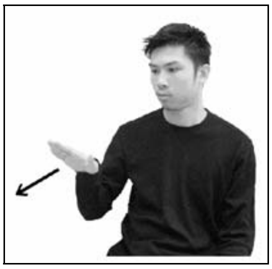


Figure 2: citation form of *PUT* in HKSL[2]

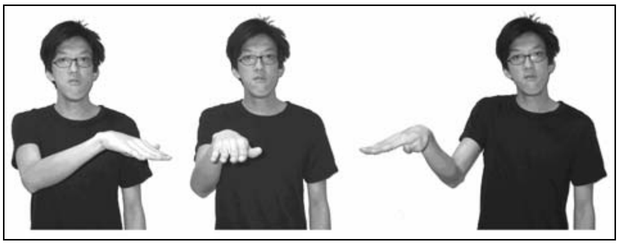


Figure 3: Spatial verb *PUT* with loci marked as *a*, *b* and *c*; glossed as *PUT-a*, *PUT-b* and *PUT-c*

In our acquisition corpus, lexical signs, gestures and simple classifier predicates are glossed on a single glossing tier (i.e. gloss 1), with the exceptions of simultaneous constructions involving independent morphological units produced separately by two manual articulators. In the latter case, the signs produced by the two hands would be glossed on the ‘gloss 1’ tier and ‘gloss 2’ tier (‘g1’ and ‘g2’ in short form) respectively. [3]

3. Transcription procedures

3.1 Initial Transcription in the ELAN program

Viewing of sign language transcription is relatively more convenient in the ELAN program than in the CLAN program because in the former multiple-tier annotations with time alignment are possible and the annotations are synchronized with the video images. It is therefore decided that the transcription of the HKSL acquisition data be done with the ELAN program first. The transcription is done by deaf researchers who are native signers of HKSL. Delimiters are also added by the deaf researchers at the last annotation of each sentence/utterance.

3.2 A table of glosses for consistency check and tagging

To check the consistency of the glosses, an add-on program is developed by our research team to examine the transcriptions on the ‘gloss 1’ and ‘gloss 2’ tier. Error messages are generated if the program notices any formatting typos in the annotations, such as a gloss with an open square bracket ‘[’ but not a close square bracket ‘]’. When all errors spotted by the program are corrected, a table containing all the glosses in the data will be generated for the purpose of consistency check, substitution and tagging. (See Figure 4)[4] The table consists of four columns: Glosses, Grammatical Category, Substitution and Files. Information of the first and the fourth column is generated by the add-on program. For the column of Glosses, the same English glossing items found in a selected set of files will only appear once. For instance, as shown in Figure 4, the sign *IX_1* appears in the ELAN file ‘CC02017.eaf’ and ‘CC030713.eaf’ respectively. The entry *IX_1* appears once only in the table, with the names of the files containing the sign listed in the fourth column. The researchers would need to go through this table with naked eyes to check the consistency of the English glosses. For example, it has been decided that in our transcription system the V-handshape sign should be glossed as *SEE* but sometimes it may be mistakenly glossed as *LOOK_AT*. This type of inconsistency is unavoidable because the data transcription has been done by more than one deaf researcher.[5] When this happens, the researchers can type in *SEE* in the Substitution column for the gloss entry

However, this convention is in conflict with the existing annotation convention of CHILDES. We therefore replaced underscores with hyphens.

[2] Photos in Figure 1 and 2 are taken from Tang (2006).

[3] Details of the glossing system for signs in general and simultaneous constructions involving two manual articulators will be given in another oral presentation from our colleagues.

[4] Since the add-on program is developed in an early stage of the establishment of the corpus, the program can only generate the glosses for the transcription using the internal transcription coding.

[5] When two or more English words match the meaning of a sign, the one with a more general meaning will be chosen, for example, we have chosen ‘MALE’ instead of ‘MAN’ or ‘BOY’. When two or more signs with the same meaning can only be translated with one English word, we use _1, _2, etc. to denote different signs, such as LIGHT_1 for brightness and LIGHT_2 for weight.

LOOK_AT. By the same token, if typos are found in the glosses, e.g. *BOOK* is spelt as *BOO* by mistake, the researcher can enter the correct form in the Substitution column. The table with filled information on the substitution column will then be processed by the add-on program again and the substitutions will be performed automatically by the program in the selected ELAN files.

As for the column of Grammatical Category, the researchers will need to input the grammatical categories of all the gloss entries in the table manually. For example, *PUT* is a spatial verb and it is tagged as 'v:sp', whereas *IX_1* is tagged as 'n:pro' to show that it is a pronoun.[6] The completed table will become part of the source code for tagging in the future. The following figure shows the outlook of the table:

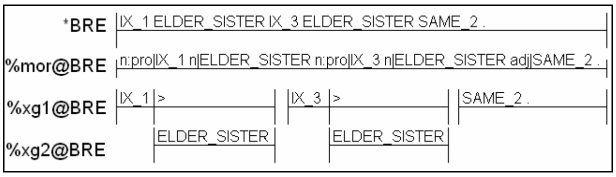
Glosses	Grammatical category	Substitution	ELAN files	
LOOK_AT	v:agr	SEE	CC040621.eaf	
BOO	n	BOOK	CC030713.eaf	
PUT	v:sp	<sub>	CC030523.eaf	CC040621.eaf
IX_1	n:pro	<sub>	CC020617.eaf	CC030713.eaf

Figure 4: Table generated by the add-on program for consistency check and tagging

4. Utterance and morphosyntactic tier – from ELAN to CLAN

4.1 Generation of utterance tier and morphosyntactic tier

When the consistency of the glosses is checked, the two tiers for glosses in the ELAN format and the glossing table will be processed by the add-on program to generate an utterance tier and a morphosyntactic tier for each signing participant. [7] The add-on program will automatically combine the glosses on the two glossing tiers to form an utterance tier. Sentence/utterance boundaries are detected on the basis of the delimiters added earlier on the two glossing tiers. The majoring of codings and symbols on the utterance tier are generated automatically by the add-on program, but a few more require manual input. The utterance tier becomes the main line of the transcription (_*BRE_ in Fig 5). At the same time, the information on the grammatical categories listed in the glossing table will be used to generate the morphosyntactic tier, in which each single gloss will be mapped with its corresponding tag. When the utterance and morphosyntactic tier are completed, the Elan files including all of the transcription tiers will then be exported to a CLAN readable format. The following figure shows the transcription of a sentence by a deaf adult in the ELAN program:



```
*BRE       IX_1 ELDER_SISTER IX_3 ELDER_SISTER SAME_2 .
%mor@BRE   n:pro|IX_1 n|ELDER_SISTER n:pro|IX_3 n|ELDER_SISTER adj|SAME_2 .
%xg1@BRE   IX_1 >                    IX_3 >                  SAME_2 .
%xg2@BRE        ELDER_SISTER              ELDER_SISTER
```

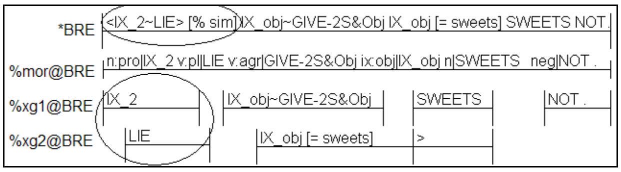
Figure 5: A sample of the transcription in the ELAN program

(meaning: "You address both me and her as 'elder-sister'.")

4.2 Coding for CHAT format

As the glosses correspond to individual signs only, certain utterance-level information, e.g. whether an utterance involves repetition or imitation, cannot be coded clearly on the two glossing tiers. In the process of generating the utterance tier, the add-on program can recognize certain set patterns of annotations, such as repetition of a sequence of signs. For example, if the signer produces the sign sequence 'A B, A B', additional symbols '< > [/]' matching the CHAT specification are added automatically by the add-on program to result in '<A B> [/] A B' on the utterance tier. Another auto-formatting generated by the add-on program is the switch '[+ imit]', which marks imitation of a whole utterance on the utterance tier of the deaf child. For example, the deaf adult produces a sequence of signs and the deaf child produces the same sequence of signs by imitation. Each of the imitated sign on the glossing tier of the deaf child is followed by '["]'. The add-on program, when generating the deaf child's utterance tier, will recognize these symbols and automatically add '[+ imit]' to the end of the imitated utterance. In the CLAN program, researchers can decide whether these utterances should be included in their analysis or not.

However, a number of additional codings for different types of simultaneous constructions need to be added manually by the researchers. For example, in certain simultaneous constructions, the two manual articulators produce signs that do not combine syntactically to form a phrasal category (e.g. the co-articulation of *IX_2* and *LIE* as in Figure 6). On the utterance tier additional symbols '<A~B> [% sim]' are added to indicate that the sequence of signs enclosed by the angle brackets does not reflect the actual order of appearance, i.e. the two signs are produced simultaneously rather than sequentially.



```
*BRE       <IX_2~LIE> [% sim] IX_obj~GIVE-2S&Obj IX_obj [= sweets] SWEETS NOT .
%mor@BRE   n:pro|IX_2 v:pl|LIE v:agr|GIVE-2S&Obj ix:obj|IX_obj n|SWEETS  neg|NOT .
%xg1@BRE   IX_2        IX_obj~GIVE-2S&Obj     SWEETS         NOT .
%xg2@BRE      LIE          IX_obj [= sweets]  >
```

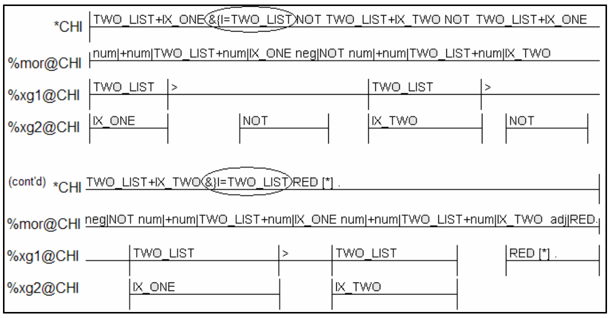
Figure 6: Representation of simultaneous signing in ELAN interface

(meaning: "You lie then I won't give you any sweets.")

[6] Pronouns in Hong Kong Sign Language are indexical signs represented by 'IX' in the corpus. '1', '2', '3' represent 1st, 2nd and 3rd person respectively.

[7] Some of our earlier files were transcribed with a glossing system incompatible with the CHAT specifications. Another function in the add-on program was designed to convert these glosses into forms compatible with the CHAT format.

In some cases, a sign is first held in the signing space for a prosodic function and is then re-activated again to form a larger morphosyntactic complex with the co-occurring signs. In Figure 7 below, the *TWO_LIST* is first held by the weak hand and is reactivated again later and combines with *IX_TWO* to form a noun phrase. Two sets of symbols, namely, '&{l=SIGN' and '&}l=SIGN', are added on the utterance tier to indicate the duration for which the sign *TWO_LIST* is held in the signing space.[8]



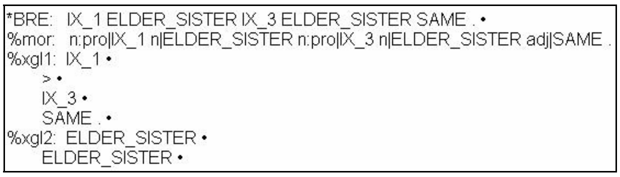
```
*CHI       TWO_LIST+IX_ONE&{l=TWO_LIST NOT TWO_LIST+IX_TWO NOT  TWO_LIST+IX_ONE
%mor@CHI   num|+num|TWO_LIST+num|IX_ONE neg|NOT num|+num|TWO_LIST+num|IX_TWO
%xg1@CHI   TWO_LIST  >                       TWO_LIST  >
%xg2@CHI   IX_ONE          NOT               IX_TWO      NOT

(cont'd) *CHI  TWO_LIST+IX_TWO&}l=TWO_LIST RED [*] .
%mor@CHI   neg|NOT num|+num|TWO_LIST+num|IX_ONE num|+num|TWO_LIST+num|IX_TWO  adj|RED.
%xg1@CHI       TWO_LIST        >    TWO_LIST            RED [*] .
%xg2@CHI       IX_ONE               IX_TWO
```

Figure 7: Representation of simultaneous reactivation in sign holding in ELAN interface – CC 3;5;23[9]

(meaning: "There are two: this one is not, that one is not; this and that are red.")

4.3 Exporting the data from ELAN to CLAN

After the utterance tier is generated and the additional codings are included manually, the transcribed ELAN data will be exported to a CLAN readable format by using the function 'ELAN2CHAT' in the CLAN program. Using the 'CHAT2ELAN' function, data from CLAN files can also be transferred back to the ELAN program. Any changes in the ELAN/CLAN file can be converted back to the CLAN/ELAN interface using these two functions. The following table shows the outlook of the exported files in the CLAN format.



```
*BRE:   IX_1 ELDER_SISTER IX_3 ELDER_SISTER SAME . •
%mor:   n:pro|IX_1 n|ELDER_SISTER n:pro|IX_3 n|ELDER_SISTER adj|SAME .
%xgl1:  IX_1 •
    > •
    IX_3 •
    SAME . •
%xgl2:  ELDER_SISTER •
    ELDER_SISTER •
```

Figure 8: Representation of the tiers corresponding to one signer in the HKSL acquisition corpus in CLAN interface

(meaning: "You address both me and her as 'elder-sister'.")

Note that in Figure 8 the bullet at the end of each line corresponds to the video clip linked to the utterance or the sign on the same line. The video clip will be played in the CLAN video-player when the button is clicked. .

One major advantage of our transcription system and add-on program is that the functions/features of both ELAN and CLAN are made accessible to the researchers. As ELAN allows multiple-tier entries and synchronization of video data with glosses, it is an ideal tool for transcribing and viewing sign language data. The CLAN program, on the other hand, has a wide range of functions, like auto-tagging and 'kwal' function for searching data, which can facilitate the linguistic analysis of sign language data. Exporting the sign language data to a readable format in CLAN also allows researchers to compare the acquisition data between spoken language and sign language.

5. Problems encountered in the course of setting up the corpus

We encountered a number of problems in the course of establishing the current acquisition corpus. Generating morphosyntactic tiers with the add-on program requires the tagging table as mentioned in Section 3.2. The grammatical categories are input manually for the whole batch of data and the process is repeated when a new batch of data is transcribed. To facilitate the tagging process, the research team is now switching to CLAN using the auto-tagging function and the establishment of the HKSL lexicon is now in progress.

On the other hand, the add-on program can only generate the tagging table for the transcription data following the internal transcription system which is used in an earlier stage. Despite of the transferring function of the add-on program, the research team is now switching to the CHAT transcription system on 'gloss 1' and 'gloss 2' tiers. Further development of the add-on program is required in order to support the existing 'substitution' function.

6. Conclusion

At present, the transcriptions in the Hong Kong Sign Language Child Language Corpus consist of glosses for the manual articulators and the data are convertible between CLAN and ELAN. The development of such a transcription system and the add-on program makes the functions/features of both ELAN and CLAN accessible to the researchers. On the other hand, as the data are readable in the CLAN format, researchers can make use of the functions and other child language data in the CHILDES to conduct cross-linguistic and cross-modal comparisons.

7. Acknowledgement

The development of the Hong Kong Sign Language Child Language Corpus was initially supported by the Research Grant Council (RGC Grant #CUHK4278/01H; Direct Grant #2081006). Our corpus has been supported by Mr. Alex K. Yasumoto's private donation since 2006. In

[8] Note that in our corpus, classifier predicates are glossed according to the adjective/verb root of the predicates and the handshape morphemes only. Other morphemic units, such as locatives, are not yet included in the glosses at this stage. Below is the transcription of an example that involves a two-handed classifier predicate meaning "a cup on the table":

utt: put+CL_hand:cup+be_located+CL_sass:table [= a cup on the table]
g1: put+CL_hand:cup [= a cup on the table]
g2: be_located+CL_sass:table

[9] CC is short for the name of a longitudinal subject in the Hong Kong Child Language Corpus. This data is taken from the corpus in which 'CHI' stands for the subject 'child' in the data.

addition, we would also like to thank the four deaf research assistants, Kenny Chu, Pippen Wong, Anita Yu and Brenda Yu, for data collection and transcription. Our thanks also go to Brian MacWhinney for his valuable advice in helping us devise a transcription system compatible with the CHAT format of the CHILDES corpus.

8. References

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Tang, G. (2006) *A Linguistic Dictionary of Hong Kong Sign Language*. Hong Kong: Chinese University Press.

Vermeerbergen, M., Demey, E. (2007). Comparing aspects of simultaneity in Flemish Sign Language to instances of concurrent speech and gesture. In M. Vermeerbergen, L. Leeson & O. Crasborn (Eds) *Simultaneity in Sign Languages: Form and Function*. Amsterdam/Philadelphia: John Benjamins, pp.257--282.

MacLaughlin, D., Neidle, C., Greenfield, D. (2000). *Sign Stream TM User's Guide (Version 2.0)*. University of Boston.