SEDSU Deliverable 5

A video-linked Thai/Swedish child data corpus:
A tool for the study of comparative semiotic development

Jordan Zlatev, Mats Andrén and Soraya Osathanonda

Centre for Languages and Literature
Lund University

1. Introduction

The project Stages in the Evolution and Development of Sign Use (SEDSU) is *comparative*, in at least six different, though related senses:

1. It compares the cognitive and semiotic capacities of human beings across different ages (i.e. it is *developmental*)
2. It compares human cognitive and semiotic behavior in different cultures (i.e. it is *cross-cultural*)
3. It compares relevant aspects (such as spatial expressions) of different languages (i.e. it is *cross-linguistic*)
4. It compares different semiotic systems, such as language, gestures and pictures (i.e. it is *semiotic*)
5. It compares the cognitive and semiotic capacities of human beings and non-human primates (i.e. it is *comparative psychological*)
6. It aims to compare, on a macro-level, developmental/evolutionary processes in ontogeny and phylogeny, on the basis of their respective *stages*, defined by types of cognitive structures and learning processes.

It is both possible and desirable to perform several of these comparative studies in synchrony. The tool which we describe in this text, *a digitized bi-cultural data corpus*, is intended to be used in studies of the type (1)-(4), since:

- The corpus consists of 60 transcripts from interactions in everyday contexts between 6 children and their caregivers (10 transcripts per child), recorded longitudinally, for the period when the children are 18 to 27 months of age. As well known, this period typically coincides with the *vocabulary and grammar spurts* (Bates 2003), and is particularly interesting for the study of cognitive and semiotic *transitions*.
- Three of the children are growing up in typical Swedish middle class environments, while the other three are also part of middle class families, but in a rather different culture: that of urban Thailand (Bangkok).
- Correspondingly, three of the children are acquiring Swedish, a Germanic Indo-European language, while the three others have Thai as their first language; and the two languages have very different characteristics (Zlatev and Yangklang 2004; Zlatev and David 2003).

- The videos of the corpus are linked to the transcripts, on an utterance-by-utterance basis using the software CLAN (MacWhinney 2000), allowing the (preliminary) study of interaction between verbal and gestural communication.

Thus, the corpus affords studies of *cross-cultural linguistic/semiotic development*. Studies such as the following are planned as part of SEDSU:

- the social learning mechanisms that underlie the acquisition of communicative gestures in human infants can be investigated through a detailed analysis of the video recordings.
- longitudinal studies of the interaction between the adults and children could reveal both developmental changes in the capabilities of the individual child and changes in the interaction as a co-constructed phenomenon.
- by studying these factors in two different cultural and linguistic settings we hope to tease apart "universal" cognitive and mimetic capabilities of human children, from culture and language-specific properties of their gestures.

Furthermore, in conjunction with comparative-psychological studies (type 5, cf. above), we hope ultimately to be able to interrelate the processes of ontogeny and phylogeny (type 6, cf. above) and thus contribute to a novel theory of semiotic evolution, which is the main goal of the SEDSU project.

Comparing two very different languages and cultures such as Swedish and Thai (while maintaining other variables such as parental educational level and socio-economic class) is highly pertinent for SEDSU. Developmental psycholinguistics has shown that characteristics once thought to be universal, and therefore possibly a consequence of the human cognitive endowment, may vary to a considerable degree depending on the language acquired (cf. Choi and Bowerman 1991, Bowerman 1996). Even the established conception of a universal "noun spurt" has been questioned by demonstrating that Korean children, who are exposed to relatively more verbs than English speaking children, undergo a vocabulary spurt in which verbs participate at least at the same level as nouns (cf. Gopnik, Choi and Braumberger 1996). A self-evident conclusion is that it is necessary to broaden the database of languages studied developmentally if we are to understand which aspects of linguistic and cognitive development are truly universal, which vary depending on environmental factors, and what are the limits of this variation.

Thai is one of the many Non-Indoeuropean languages whose acquisition has hitherto not been studied in detail. Previous work includes Tuaycharoen (1977), who analyzed the phonetic and phonological development of one child, Tuaycharoen (1984) reporting on some strategies in acquiring classifiers by two children and Thanavisuth (1997) who studied prosodic and pragmatic characteristics of Thai infant directed speech (IDS). The pilot project *First language Acquisition of Thai,* which initiated the collection of the Thai data for the present corpus (see Section 2.1) produced some preliminary results concerning motion event constructions and their acquisition by Thai children (Zlatev and Yangklang 2003, 2004). Despite the value of this work, the ontogenesis of Thai grammar and semantics remains largely unknown. Relating this to general cognitive and semiotic development has not yet been attempted.

For all these various reasons, we believe that the corpus will be a very useful tool, not only for the studies carried out in Lund, but for SEDSU and for the wider community in general.

In **Section 2** of this report we present the background of this corpus, and the methodology we followed in constructing it. **Section 3** provides a specification of the corpus, including all the special symbols of the notation used, following the CHAT Transcription Format (MacWhinney 2000). **Section 4** outlines, in brief, a few empirical studies relevant for SEDSU that can be conducted using the corpus, pointing out ways in which the corpus can be extended and the data further analyzed.

Appendix A provides the codes of the CHAT format that have been used in the corpus. Appendix B lists the transcripts of the 60 data files. The electronic versions of these will be placed on a secure server and can be accessed by the members of the SEDSU consortium. After an initial trial period, the CHAT files will be contributed to the CHILDES data base (MacWhinney 2000), making them fully available to the general public. The video files, however, will be kept in Lund for the sake of personal data protection, but can be made available on DVD disks to any member of the SEDSU consortium after signing a confidentiality statement.

2. The construction of the corpus

2.1. Thai data

Work on the Thai part of the corpus began within the pilot project "First Language Acquisition of Thai" supported by the *Swedish Foundation for International Cooperation in Research and Higher Education* (STINT), and led by Jordan Zlatev at Chulalongkorn University in Bangkok during 2000-2001. Together with Peerapat Yangklang, who worked as a project assistant with the project and Soraya Osathanonda, who did large amounts of volunteer work, Dr. Zlatev collected data from three Thai children using an updated version of the by now classical methodology of Brown (1973). The children were followed for appr. 18 months, from 18 to 36 months of age, making appr. 20 minute long recordings every two weeks in the children's homes using a digital video camera.

This resulted in 36 recordings per child, and with rather primitive technology, Yangklang and Zlatev set about to make transcripts of as many of the recordings as possible. Until the end of the pilot project in December 2001, there were 42 such transcripts, of varying quality. They were not fully compatible with the CHAT format and were not linked to the video files, or checked for consistency. Dr. Zlatev was forced to return to Sweden to take up a position in Lund University in April 2001, and while the project produced other results, the construction of the longitudinal Thai corpus had only begun.

For the SEDSU project, we decided to focus on quality rather than quantity, and therefore selected a subset of the data, 10 of the provisional transcripts for each child, starting from the earliest one for each of the three children, with "code names" JAM, CHE and JOM, and choosing the others at the interval of approximately one a month, as shown in Table 1. A similar choice was performed for the Swedish part of the corpus (described in 2.2. below). Table 1 shows the ages of the 6 children, at the time of the respective recording (data point).

Table 1. The ages of the 6 children at the 60 data files

	Thai			Swedish		
Transcript	JAM	CHE	JOM	BEL	HAR	TEA
#1	1;6.21	1;6.08	1;7.28	1;6.09	1;7.09	1;6.10
#2	1;7.07	1;7.01	1;8.13	1;7.03	1;8.26	1;7.15
#3	1;8.08	1;7.16	1;9.03	1;7.28	1;9.15	1;8.26
#4	1;9.17	1;8.14	1;9.15	1;8.23	1;10.18	1;10.02
#5	1;10.07	1;9.24	1;10.20	1;10.04	1;11.18	1;11.07
#6	1;11.12	1;10.22	1;11.19	1;11.17	2;0.16	2;0.25
#7	2;0.10	1;11.26	2;0.24	2;1.03	2;1.10	2;1.17
#8	2;1.07	2;1.10	2;1.21	2;2.13	2;2.18	2;2.12
#9	2;1.27	2;2.16	2;2.20	2;3.23	2;3.09	2;3.27
#10	2;2.21	2;3.18	2;3.13	2;4.13	2;4.23	2;4.18

In the remaining part of this section we describe some particular features of working with the Thai data.

As in any transcription project, but especially one involving a non-European language with a non-Roman-based script, it was essential to develop a procedure guaranteeing consistency. This section describes such a procedure. The first 3 steps were took place during 2001-2002, while the remaining 5 were carried out within the first 12 months of the SEDSU project, April 2005-March 2006. Dr. Zlatev worked on the corpus within his research position at Lund University, and Soraya Osathanonda was payed for a total of 90 hours for research assistance (3 hours per data point) by a parallel project to SEDSU at Lund University, "Language, gestures and pictures in semiotic development".

Step 1. Transcription in Thai orthography

Between 15 and 20 minutes of each recording session was transcribed using standard Thai orthography by a native Thai research assistant (Yangklang or Osathanonda). Every line of the transcript corresponds to what the native coders perceived to be "a single utterance", on the basis of:

- intonational criteria: a single intonational unit
- pauses: no long pauses within an utterance
- turns: interruption by another speaker marking a new utterance

Each *utterance*, thus defined, could vary in length considerably.

Step 2. Transcription in phonemic notation

The Thai transcription was converted into a phonemic notation. The transliteration system shown in Table 2 for consonants and Table 3 for vowels was used. Long vowels were marked by doubling the vowel, rather than by using the symbol ":" which was used to mark phonetic vowel lengthening (cf. below). Tones were marked at the end of each syllable according to the scheme: Mid: 0, Low: 1, Falling: 2, High: 3, Rising: 4.

Table 2. The transliteration system for Thai consonants

	Labial	Alveolar	Palatal	Velar	Glottal
Stop, +voice –asp	b	d			
Stop, -voice –asp	p	t	c	k	?[1]
Stop, -voice +asp.	ph	th	ch	kh	
Fricative	f	s			h
Semivowel	w		j		
Nasal	m	n		N	
Lateral		l			
Trill		r			

Table 3. The transliteration system for Thai vowels

	Front	Central	Back
Close	i	U	u
Mid	e	q	o
Open	x	a	O

Each speaker was assigned a three letter code (in accordance with the CHAT format, see below), and the original Thai orthography was specified after the "dependent tier" %ort. So the outcome of the first 2 steps could be as shown in the exchange in (1), which is taken from the file CHE22b.

(1) *FAT: tham0pen0rU3plaaw1
%ort: ทำเป็นหรือเปล่า
*CHI: tham0pen0pen0
%ort: ทำเป็นทำเป็น

Step 3. Word segmentation

As shown in (1), Thai orthography does not place spaces between words. Spaces are not placed between clauses either, but rather between "sentences", loosely defined. To improve readability (for a non-Thai audience), and to allow the CLAN programs to perform automatic analyses (MLU, various frequency counts etc., as shown in Section 4) the phonemic transcription needed to be segmented into words. This was straightforward in most cases since the vast majority of Thai words, especially in the colloquial register, are monosyllabic. Thus example (1) could be simply converted to (2) by inserting spaces between all syllables.

(2) *FAT: tham0 pen0 rU3 plaaw1
%ort: ทำเป็นหรือเปล่า
*CHI: tham0 pen0 pen0
%ort: ทำเป็นทำเป็น

However, it is not always clear if certain multi-syllabic expressions should be treated as (a) mono-morphemic words, (b) multi-morphemic words including lexical compounds or (c)

[1] Due to requirements of the CHAT notation, the glottal stop symbol "?" was not included in the transcription. Its presence is nevertheless derivable from the data since Thai syllables can not begin with a vowel or end with a short vowel. Whenever that seems to be the case in the data, there is an "invisible" glottal stop before the initial vowel or after the final short vowel, e.g. (?)aw0 ('want'), lx(?)3 ('and').

phrases consisting of one or more words. In deciding how to analyze particular examples, we used the following criteria, following Zlatev and Yangklang (2004):

1. Mono-morphemic word IFF at least one of the syllables in the expression does not have a transparent separate meaning, e.g. **naa2taaN1** ('window'). Even though this expression is probably a compound diachronically, the compounding is not transparent for present-day speakers.
2. Multi-morphemic word ("^" between the syllables) IFF all the syllables have transparent separate meanings, but the meaning of the whole is not derivable by combining that of the parts, e.g. phuu2^jaj1 ('person'+'big' = 'adult'). Lexical compounds are one subclass of this category. Derivations such as **khwaam0^suk1** (PROPERTY + 'happy' = 'happiness') are also included in this category, even though their derivation is semantically regular.
3. Phrase (SPACE between the syllables) IFF the syllables have separate meaning, and combine systematically to give the meaning of the whole, e.g. **maa4** ('dog') **noj4** ('little') = 'little dog'.

In the same group as the second category, and thus marked in the same way (with a "^" connecting the parts) were expressions that appeared to be formulaic, e.g. may0^pen0^raj0 ('never mind'), as well as reduplications, e.g. dek1^dek1 ('children').[2]

Step 4. Linking the transcripts to video

The sections of the video tapes corresponding to the transcripts where converted first to DV format using the iMovie, and these large data files, appr. 4 GB per 15 minute fragment were further compressed by Media Cleaner to files of appr. 150 MB (Sorenson V3). The latter were used for linking to the transcripts, while the DV files were archived. Since the time coding for the DV and the compressed files are identical, the high quality video files can be used for further analysis, once the linking is performed. Linking was done on an utterance by utterance basis, using the appropriate facilities of the CLAN program.

Step 5. Changes to the transcripts

5.1. Corrections to the transcriptions

Working with linked media files and quality headphones allowed a more precise transcription, so Osathanonda could make a number of corrections in the %ort lines, and Zlatev made corresponding changes to the main tiers using the transliteration system described above.

5.2. Marking deviations from standard pronunciation

Certain adaptations were made to bring the transcription closer to the actual speech produced. Deviations from the normative pronunciation were represented using the CHAT convention of placing the citation form in square brackets after the transcription of the sub-standard form, e.g. **lUU3 [: rUU4]** ('or').[3]

[2] Using this notation compounds and other multi-morphemic words, formulaic expressions and reduplications can be treated as *single lexical items* (which is intuitively correct) by the CLAN programs, at the same time as analysis can easily be performed on their parts if required. For example, by adding the switch +b^ in the command line of the program MLU the constituent morphemes will be counted separately.

[3] Most of the CLAN programs make an automatic substitution of the second form for the first, which gives greater reliability in e.g. counting word types. This substitution can be easily cancelled and analysis performed on the first forms by using the switch "+r5".

5.3. Pauses, vowel lengthening and repetition/retracing

All pauses were marked in the transcriptions as short (#) or long (##). Sometimes instead of a pause there seemed to be a vowel lengthening, which was marked using the column, e.g. **maa:4** ('dog'). Repetitions and re-tracings were marked using the CHAT conventions, i.e. the repeated or re-traced material was surrounded by angle brackets "< >" (if consisting of more than one words) and followed by [/], or [//].The first indicates a repetition, the second a retracing with some change, usually a self-correction.[4]

5.4. Utterance delimiters

In the preliminary transcripts all utterances ended with a full stop (".") . Since it is possible to use other utterance delimiters, and this allows a rough division into speech acts, at the same time as it improves readability of the transcripts, Zlatev and Osathanonda went though the 30 data files, and on the basis of word order, question words and intonation, specified

- *questions* with the delimiter "?", including monosyllabic expressions such as **hU3?**
- *requests* and *exclamations* with the delimiter "!", such as **paj!** ("go")
- *statements* (and any speech acts which were not questions or requests) with the delimiter "."

We point out, however, that this is a very rough speech act analysis, and that appropriate categories and codes should be constructed for detailed pragmatic studies of the corpus.

Step 6. Clause segmentation

As pointed out above, each line of the preliminary transcripts correspond to an *utterance*: "Each main line should code one and only one utterance." (MacWhinney 2000: 16), following the spontaneous intuitions of the native speakers rather than grammatical criteria. At the same time, it is desirable to be able to accommodate units such as clauses, which give a clearer picture of grammatical development.

For this purpose we includes the CHAT symbol [^c] at (what we interpreted as) clause boundaries. The relationship between utterances and clauses were in many cases not one to one. For example interaction given in (3) between GRM and GAR (from JOM27a.cha) show how the first utterance contains two clauses while the second, strictly speaking none.

(3) *GRM: neen0 thUU4 dii0^dii0 na3 [^c] jaa1 tok1 saj1 huua4 dek1 na3 [^c]!
[Neen carry well particle]c [FUT fall on head child particle]c
'Neen, be careful! It will fall on the child's head!'
*GAR: khrap3.
Polite-MASC
'Yes, madam'

Deciding clause boundaries in spoken language, and particularly in a serial verb language with a high degree of implicitness such as Thai is far from trivial. For more than one clause per utterance, we used the criteria defined by Zlatev and Yangklang (2004). Utterances which

[4] All CLAN programs except MLU and MODREP include the repeated/retraced material by default, and in order to exclude it, the switch "+r6" needs to be used.

were “less than” clauses and were not marked with the [^c] symbol were utterances which did not express *predications*. This involved:

- single word utterances which were not verbs (action or property terms)
- multiple word utterances which were enumerations

Step 7. Final conversion into CHAT format

Steps 1-6 had the consequence that they incrementally introduced more and more elements of the CHAT notation, but in order for the files to be analyzed by the CLAN programs, they should be fully consistent with the CHAT format. Every file was for that purpose given the following “headers ties” (cf. MacWhinney 2000), here illustrated for the file JOM19(1).

```
@Begin
@Languages:             th
@Participants:          CHI Target_Child, FAT Father, GRM Grandmother, AUN
                        Aunt
@ID:                    th|Zlatev|CHI|1;7.28||||Target_Child||
@Exceptions:            *N* *O* *U*
@Coder:                 Peerapat Yangklang, Soraya Osathanonda, Jordan Zlatev
@Birth of CHI:          28-DEC-1998
@Date:                  26-AUG-2000
@Age of CHI:            1;7.28
@Location:              Bangkok, Thailand
@Situation:             CHI is visiting relatives on a Saturday evening and
            is walking around in the house's garden with his father. GRM and
            AUT participate in the interaction. Taping was done from about
            5:30 to 6:20, but the first half hour was lost because of a
            technical error. CHI is a bit tired during the last 20 minutes
            and it is getting dark.
@End
```

The first specifies the beginning of the transcript, and the last one, on the last line of the transcript, its end. The second states that the main language of the transcript is Thai. The third states the participants, giving each a three-letter code, which is later used for each participant’s *main tier*.
The other header ties give:

- the “id” of the target child for this session,
- three letter symbols which are spelled with capital letters to signify specific Thai consonants and vowels (cf. Table 2 and 3)
- the names of the people responsible for the coding, in the order of their involvement
- the date of birth of the child
- the data of recording
- the age of the child at the time of recording
- the location of the recording
- a short description of the circumstances of the recording

The interaction is transcribed on main tiers, using the three letter codes of each participants, as shown in (3) above. Following each main tier can be a number of different *dependent tiers*, specifies as three level codes, following a percept sign. The following were used in Thai data:

- %ort Thai orthography
- %exp explanations
- %sit situational information
- %act non verbal actions

- %gpx gestures

A number of other standard CHAT notations were used, listed in Appendix A.

Step 8. Consistency checking

In order to guarantee consistency in the transcription and the notation, Zlatev and Osathanonda carefully read through the 30 transcribed files, line by line, correcting any found mistakes and inconsistencies. Finally the files were automatically checked for notational consistency using the CLAN program CHECK.

2.2. Swedish data

The Swedish part of this corpus is based on earlier work by Ulla Richthoff and Sven Strömqvist in the nineties in what is called the Strömqvist-Richtoff corpus. Richthoff (2005) presents a rather comprehensive specification of this corpus. Their corpus consists of transcribed data from six Swedish children between 18 months and four years. For the purposes of the SEDSU project three of these children were chosen, for whom analogue video recordings were available. These three children are referred to as BEL, HAR and TEA in this text. A subset of ten transcriptions for each child were selected with the ambition to match as closely as possible the criteria used in the Thai part of the corpus; 15 minutes of transcribed video data for each child, every month, between 18 and 27 months. All in all this makes up 30 transcriptions and 7.5 hour of video data in the Swedish part of the Thai/Swedish corpus.

All work on the Swedish data within the SEDSU project have been made by Mats Andrén, PhD student at SoL-centre at Lund University.

Step 1. Transcriptions

The transcriptions on which this corpus is based are inherited from the Strömqvist-Richthoff corpus. This means that criteria for determining what counts as an utterance and similar issues of relevance have not been determined by us. Again, we refer to the specification within the work of Richthoff (2005) for more information about choices made in the Strömqvist-Richthoff corpus. However, this means that there are some small differences in the notation used in the Swedish part and the Thai part of the corpus. For example, the Swedish part does not use the exclamation mark ("!") at all as an utterance delimiter. The symbols "+" and "_" are used as morpheme boundaries (derivational and inflectional respectively). For technical reasons, it was not possible to use these in the Thai data, which instead uses "^" as a morpheme boundary.

Step 2. Linking of transcripts to video

The procedure for linking the transcripts to video is almost identical to the procedure used for the Thai part of the corpus, including the quite extensive work of converting analogue videos into compressed digital video files. 15 minutes from each video file were linked to the transcripts and there are a total of 13924 utterances linked to video, by specifying the start and endpoint of the utterances. However, both transcripts and the converted video files generally cover more than 15 minutes, so it is possible to extend the corpus in the future if we would desire to do so. The number of data-points could also be extended, since there are transcriptions and analogue videos available for the period until the children are about 36 months old.

Step 3. Modifications of the transcriptions

In our versions of the transcriptions, a number of changes have been made in respect to the original ones in addition to the video linking mentioned before.

3.1. CHAT syntax checking - Substitution of some symbols

Since the original transcriptions used some notation conventions not included in the CHAT standard, it was not possible to do syntax checks on the files in the Strömqvist-Richthoff corpus. Therefore some symbols were substituted for others and the transcribed files now all pass through the syntax check in CLAN. See Appendix A for details on symbols used in the transcriptions, including the minor differences between the Swedish and Thai transcription.

3.2. Coding of gestures

Even though coding of gestures is not part of corpus itself, but rather a matter for further analysis, we have already begun to do some initial marking of pointing gestures for two of the children (BEL and TEA). Pointing gestures are marked in the transcriptions with the %gpx tier, which is part of the CHAT standard. In sum, there are so far 1181 pointing gestures marked in the Swedish part of the corpus (see Section 4.2). In addition to this, cases where similar gestures (in form or function) occur, or borderline cases, have been marked too, since these are often theoretically interesting. There are so far 217 such cases marked in the transcriptions.

3.3. Modifications of file headers

The "@Age of CHI"-field was added in the file headers for compliance with the Thai part of the corpus. Also, the "@Coder" field in the files was updated to include credits for work on the transcriptions made by Mats Andrén. The file headers still differ from the Thai part of the corpus in that they do not include neither of the "@Exceptions:"-, "@Location:"- nor "@Birth of CHI:"-fields, since these fields were judged unnecessary. Concerning locations, all the Swedish transcriptions are made from interactions in middle class families in the western parts of Sweden.

3.4. Translation

The "@Situation" descriptions in the transcriptions were translated from Swedish to English.

3.5. Removed tiers

In the original transcripts there were a number of so-called *dependent tiers*, similar to the ones in the Thai corpus. Three of them, which are not in the Thai part of the corpus, were removed. These two are:

- %tim - Time codes which is an artefact of the transcription process in the original transcriptions in the Strömqvist-Richthoff corpus which were mainly done by using audio recordings on tape.
- %cod - This tier included codes that were of specific interest only to the studies conducted by Richthoff (2005), and were removed here to make the transcriptions more readable .
- %pho – This tier included phonetic information. It is removed for the same reason as %cod.

The rest of the dependent tiers (%sit, %act, %exp) were left in the Transcripts, even though their specifications are written in Swedish. There is also some redundancy since the original transcript used the %act-tier to code some instances of pointing gestures (but not all!). These

may, or may not, disappear in future versions of the corpus, but are left for the moment since extra information can always be helpful when doing analysis.

3.6. Corrections of typos and added symbols

Minor changes and corrections were made in the few cases were typos and missing notation were encountered. Typical examples of "missing notation" are unmarked instances of overlapping talk.

3. Specification of the Thai/Swedish corpus

The following 6 tables provide an overview of the contents of all 60 transcripts of the Thai/Swedish corpus.

Table 4.1. Data files for JAM (Thai, girl)

File name	Age	Other participants	Length of video file (min:sec)	Activities	Utterances
JAM18.cha	1;6.21	MOT, FRE	15:30	Book-looking, playing with a baby	302
JAM19(1).cha JAM19(2).cha	1;7.07	MOT, BRO	14:21	Eating, free play	275
JAM20(1).cha JAM20(2).cha	1;8.08	MOT, BRO, SIT, INV	15:41	role play, book-looking, TV-watching, pretend play	373
JAM21(1).cha JAM21(2).cha	1;9.17	SIT, INV, GRA, AUN	15:22	Free play, role play	576
JAM22(1).cha JAM22(2).cha	1;10.07	MOT, SIT, BRO, INV	15:32	Playing with dolls, TV-watching, give-and-take	622
JAM23(1).cha JAM23(2).cha	1;11.12	MOT, SIT, INV, NEI, FRI	20:45	Eating, pretend play, playing with dolls	672
JAM24(1).cha JAM24(2).cha	2;0.10	MOT, AUN, INV, BRO	16:30	Free play, pretend plays with telephone	396
JAM25a(1).cha JAM25a(2).cha	2;1.07	MOT, BRO, AUN, GRM UNC, VIS, INV	16:02	Running around, eating, pretend play, role play	534
JAM25b.cha	2;1.27	MOT, BRO, FRE, VIS, INV	15:48	Drawing, role play, makeup making	513
JAM26.cha	2;2.21	MOT, SIT, INV, BRO	15:37	Role play, playing with dolls, eating	570
Total					4833

Table 4.2. Data files for CHE (Thai, girl)

File name	Age	Other participants	Length of video file (min:sec)	Activities	Utterances
CHE18(1).cha CHE18(2).cha	1;6.08	MOT, VIS, BRO, MAI	15:28	Naming games, picture book looking	418
CHE19a(1).cha CHE19a(2).cha	1;7.01	MOT, FAT, BRO	18:19	Free play, book looking, role play	140
CHE19b(1).cha CHE19b(2).cha	1;7.16	FAT, BRO, GRM, MOT, NEB	15:01	Walk to the temple, running and chacing, bike riding	371
CHE20a.cha	1;8.14	GRM, MAI, BRO	12:56	Book-looking, naming	429
CHE21b.cha	1;9.24	FAT, BRO, INV	17:10	Running, teasing, pretend play, video watching	429
CHE22b.cha	1;10.22	FAT, MOT, BRO, GRF	16:04	Dressing, preparing for a trip, riding in the car	368
CHE23b.cha	1;11.26	MOT, INV	16:17	Free play with toys, TV-watching,	360
CHE25a.cha	2;1.10	MOT, BRO, INV	20:02	Free play, singing, dancing, karaokee	475
CHE26a.cha	2;2.16	MOT, BRO, INV	15:56	Free play, singing, fighting (with BRO), TV-watching	455
CHE28a.cha	2;3.18	MOT, BRO, INV	15:33	Karaoke, picture drawing	346
Total					3791

Table 4.3. Data files for JOM (Thai, boy)

File name	Age	Other participants	Length of video file (min:sec)	Activities	Utterances
JOM19(1).cha JOM19(1).cha	1;7:28	FAT, GRM, AUN	15:15	Walking in the yard, exploring	355
JOM20(1).cha JOM20(2).cha	1;8:13	MOT, FAT	15.13	Playing in the playground, riding a bike	400
JOM21a.cha	1;9:03	MOT, FAT, VIS	15:47	Playing with visitors, throwing around clothes, eating	430
JOM21b.cha	1;9.15	MOT, FAT, GRM	15:39	Playing with the dog, walking in the yard	474
JOM22b.cha	1;10.20	MOT, FAT, VIS	15:20	Riding play-car, helping Mom, talking to visitors	346
JOM23b.cha	1;11.19	FAT, MAI, MOT	16:51	Playing with toys, TV-watching, feeding the fish, watching the caterpillars in the bushes	339
JOM24b.cha	2;0.24	MOT, MAI, FAT	15:50	Book-looking, naming letters, playing with caterpillars, making orange juice	568
JOM25b.cha	2;1.21	MOT, MAI, AUN, INV	15:32	Eating, walking around, free play	461
JOM26b.cha	2;2.20	MOT, FAT, VIS, MAI	16:56	Playing with baby, at a restaurant, chasing flies	527
JOM27a.cha	2;3.13	MOT, GRM, INV, GAR	18:43	Picking mangoes, playing with the dog, observing new surroundings, riding a bike	365
Total					4265

Table 4.4. Data files for BEL (Swedish, girl)

File name	Age	Other participants	Length of video file (min:sec)	Activities	Utterances
bel18_09.cha	1;6.09	MOT, FAT	14:16	book-looking, eating, block-playing	401
bel19_03.cha	1;7.03	MOT	15:00	lego-playing, puzzle-doing, book-looking	461
bel19_28.cha	1;7.28	MOT	15:00	book-looking, puzzle-doing, card-playing	385
bel20_23.cha	1;8.23	FAT, MOT	15:00	book-looking	551
bel22_04.cha	1;10.04	MOT	15:00	book-looking, puzzle-doing	480
bel23_17.cha	1;11.17	FAT	15:00	toy-jar-playing, book-looking, eating	475
bel25_03.cha	2;1.03	MOT, FAT	15:00	book-looking, toy-figure-playing	503
bel26_13.cha	2;2.13	MOT	15:00	card-playing, book-looking	535
bel27_23.cha	2;3.23	MOT	15:00	pretending to eat, book-looking	458
bel28_13.cha	2;4.13	MOT	15:00	doll-playing, book-looking	424
Total					4673

Table 4.5. Data files for HAR (Swedish, boy)

File name	Age	Other participants	Length of video file (min:sec)	Activities	Utterances
har19_09.cha	1;7.09	MOT	15:00	lego-playing, puzzle-doing, book-looking	299
har20_26.cha	1;8.26	MOT	15:00	puzzle-doing, book-looking	477
har21_15.cha	1;9.15	MOT, GMM	15:00	puzzle-doing, lego-playing, book-looking	391
har22_18.cha	1;10.18	MOT	15:00	book-looking, lego-playing, car-playing	412
har23_18.cha	1;11.18	MOT	15:00	lego-playing, car-playing, drawing	434
har24_16.cha	2;0.16	MOT	15:00	talking about family, lego-playing, book-looking	433
har25_10.cha	2;1.10	MOT, GMM	15:00	book-looking, toy-car-talk	485
har26_18.cha	2;2.18	MOT	15:00	lego-playing, book-looking	482
har27_09.cha	2;3.09	MOT	15:00	eating, puzzle-doing, book-looking	439
har28_23.cha	2;4.23	MOT	15:00	puzzle-doing, book-looking	409
Total					4261

Table 4.6. Data files for TEA (Swedish, girl)

File name	Age	Other participants	Length of video file (min:sec)	Activities	Utterances
tea18_10.cha	1;6.10	MOT	15:00	Toy-playing, pretending	359
tea19_15.cha	1;7.15	MOT	15:00	book-looking, puzzle-doing, toy-playing	382
tea20_26.cha	1;8.26	MOT	15:00	doll-playing, book-looking	484
tea22_01.cha	1;10.02	MOT	15:00	doll-playing, book-looking	528
tea23_07.cha	1;11.07	MOT	15:00	book-looking, eating	566
tea24_25.cha	2;0.25	MOT	15:00	eating, doll-play, puzzle-doing, book-looking	617
tea25_17.cha	2;1.17	MOT	15:00	doll-playing, book-looking	556
tea26_12.cha	2;2.12	MOT	15:00	doll-playing	564
tea27_26.cha	2;3.27	MOT	15:00	doll-playing, book-looking	508
tea28_19.cha	2;4.18	FAT	15:00	book-looking, toy-playing	426
Total					4990

4. Preliminary analyses and future extensions

In this section we provide examples of the kind of studies that can be efficiently performed on the corpus, using some of the CLAN programs. The analyses are not meant to be in any way comprehensive, but simply illustrative.

4.1. MLU

Given the segmentation of the corpus in sentence-length units, as defined in Section 2, a convenient first impression of the development of linguistic proficiency in the 6 children may be provided by using the program MLU, measuring "medium length of utterance". The CLAN command in (4) calculates the number of utterances, the number of words (including multimorphemic words and formulas) and divides the two in order to produce a standard measure of children's grammatical complexity: medium length of utterance, mlu (Brown 1973).

(4) mlu *.cha +t*CHI

The results, shown in Table 5, show that all 6 children display rather rapid development according to this measure, with JAM, CHE and TEA showing rather spurt-like patterns, i.e. a relative rapid increase at a particular data point (marked in bold face in Table 5), while the other three children show a rather more continuous progression.

Table 5. MLU estimates for the three Thai and three Swedish children

	Thai			Swedish		
Transcript	JAM	JOM	CHE	BEL	HAR	TEA
#1	1.267	1.438	1.015	1.700	1.023	1.050
#2	1.096	1.462	1.100	1.603	1.079	1.227
#3	1.250	1.467	1.205	1.759	1.096	1.096
#4	1.166	1.602	1.389	1.837	1.201	1.115
#5	1.129	1.767	**2.028**	1.349	1.132	1.086
#6	1.229	2.045	2.167	1.764	1.225	1.147
#7	**1.635**	2.135	1.935	2.015	1.326	1.233
#8	1.612	2.321	1.920	2.251	1.443	1.326
#9	1.605	2.245	1.888	2.495	1.747	**1.959**
#10	1.916	2.022	2.753	2.114	2.413	2.333

Since the question of "continuity vs. discontinuity" in development is of crucial importance for the SEDSU project, these differences will be further investigated. It is possible to run the MLU program counting "morphemes" rather than words, in order to see if the pattern found in Table 5 would be maintained. This is done by adding the parameter +b with value "^" for Thai and "+" and "_" for Swedish, making the MLU program regard these symbols as word (or rather morpheme) delimiters, as in (5).

(5) mlu *.cha +t*CHI +b^

Another similar measure is *medium length of clause*. This can be achieved for the Thai data, where clauses where coded by simply adding the parameter +c[^c] to the commands in (4) or (5).

Since we have available more unlinked transcripts and video files around the particular points of interest, linking and adaptation of this extra data, following the procedures of Section 2, can be performed in the future if deemed necessary.

4.2. Gestures

Of course, for the purposes of SEDSU, and in particular for Work Package 5 "Imitation and mimesis" analysis of the children's linguistic development will not be sufficient. As pointed out in Section 1, our intention is for the Thai/Swedish corpus to be used for studies of comparative semiotic development, and in particular for investigating the relationship between language and gestures on ontogeny.

For the purpose we have so far coded a number of gestures, most of them acts of pointing in the transcriptions, using the %gpx dependent tier. Table 6 shows this work in progress, which is not a proper part of this deliverable, but rather paves the way for those to come.

Table 6. Number of gestures so far coded (%gpx codings)

	Thai			Swedish		
Transcript	JAM	JOM	CHE	BEL	HAR	TEA
#1	10	17	3	52		18
#2	15	11	17	60		77
#3	9	3	13	22		85
#4	9	23	6	61		146
#5	12	7	5	57		97
#6	5	5	1	42		104
#7	1	7	10	25		84
#8	14	8	13	78		21
#9	5	8	20	25		30
#10	13	10	2	5		90

Furthermore, since CHAT files can be automatically converted to the program ELAN developed by the Max Planck Institute for Psycholinguistics (www.mpi.nl/tools/elan.html), which allows much greater precision in the analysis of gestures, we plan to use this possibility at a further stage of the development of the Thai/Swedish child data corpus.

Acknowledgements

First, we wish to thank Brian MacWhinney, for his inspirational work in developing CHILDES and setting a paradigm example of data sharing in the community, as well as for his ever-ready assistance.

For the Swedish data we are very much in debt to Sven Strömqvist and Ulla Richthoff for allowing us to use their corpus. Only after working on a subset of the data do we fully realize how much effort they have put into collecting the audio and video data, and then on producing very careful transcripts.

We would also like to thank the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) for generously sponsoring the pilot project "First Language Acquisition of Thai" from January 2000 to March 2001. Without the hospitality and support of the Department of Linguistics and the Center for Research in Speech and Language Processing (CRSLP) at Chulalongkorn University, this research would have been impossible. We therefore wish to express our gratitude, especially to director of CRSLP, Dr. Sudaporn Luksaneeyanawin. We would also like to thank Peerapat Yangklang who carried out most of the collection and transcription of the Thai data.

Finally, we thank the Faculty of Humanities and Theology at Lund University for funding the project *Language, gestures and pictures in semiotic development* which made it possible for Soraya Osathanonda to do the linking of the Thai video files to the transcripts, and correct the latter.

References

Bates, E. (2003) On the nature and nurture of language. In R. Levi-Montalcini, D. Baltimore, R. Dulbecco, & F. Jacob (Series Eds.) & E. Bizzi, P. Calissano, & V. Volterra (Vol. Eds.), *Frontiere della biologia* [Frontiers of biology]. *Il cervello di Homo sapiens* [The

brain of homo sapiens]. Rome: Istituto della Enciclopedia Italiana fondata da Giovanni Trecanni S.p.A., pp. 241-265.

Berman, R. and D. Slobin. (1994) *Relating events in narrative. A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum.

Bowerman, M. (1996) Learning how to structure space for language–A cross-linguistic perspective. In P. Bloom, M. Peterson, L. Nadel & M. Garret (eds.), *Language and space*. Cambridge, Mass.: MIT Press.

Brown, R. (1973) *A first language: The early stages.* Cambridge, Mass: Harvard University Press.

Choi, S.and M. Bowerman (1991) Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition* 41: 83-121.

Gopnik. A., S. Choi and T. Baumberger (1996), 'Cross-linguistic Differences in Early Semantic and Cognitive Development'. *Cognitive Development* 11/2: 197-227.

MacWhinney, B. (2000) *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum.

Richthoff, U. (2005). *En svensk barnspråkskorpus – Uppbyggnad och analyser.* Department of Linguistics, Göteborg University.

Thanavisuth, C. (1997) *Phonetic and pragmatic characteristics if infant directed speech in Thai*. PhD Dissertation, Department of Linguistics, Chulalongkorn University.

Tuaycharoen, P. (1977) *The phonetic and phonological development of a Thai baby: From early communicative interaction to speech.* Ph.D. Thesis. University of London.

Tuaycharoen, P. (1984) Developmental stages in the acquisition of classifiers in Thai. *Selected papers from the International Symposium on Language and Linguistics*, Chiang Mai, Jan 11-14 1984, Bangkok: Thepmongkol Karnpim.

Zlatev, J. and C. David (2003) Motion event constructions in Swedish, French and Thai: Three language types? *Manusya, Journal of Humanities*, 6: 18-42.

Zlatev, J. and P. Yangklang (2003) The acquisition of motion constructions in Thai. In Lars-Olof Delsing, Cecilia Falk, Gunlög Josefsson and Halldór Sigurdsson (eds.) *Grammar in Foucs. Vol II, Festskrift till Christer Platzack*, 383-394. Lund: Wallin & Dalholm.

Zlatev, J. and P. Yangklang (2004). A third way to travel: The place of Thai in motion event typology, In Sven Stromqvist & Ludo Verhoeven (eds.) *Relating Events in Narrative: Cross-linguistic and Cross-contextual Perspectives*, 159-190. Mahwath, N.J. Earlbaum

Appendix A

The following set of codes has been used within the Thai/Swedish corpus. All are standard for the format of the data within CHILDES, known as CHAT format (cf. MacWhinney 2000), with a few noted exceptions. The table also describes notation differences between the Thai and Swedish part of the corpus.

Code	Meaning	Example
*CHI:	Thai only: Main tier header for Target child	*CHI: kaan0la3 khraN3 nUN1
*BEL, *HAR, *TEA	Swedish only: Main tier header for Target child	*BEL: han e gä .
%ort:	Thai only: Header for transcription in Thai	%ort: เขาก็ตกใจ
%com:	Comment line	%com: child appears confused
%exp:	Explanation	%exp: mistaken referenc
%act:	Visible act.	%act: reaching for the toy
%gpx:	Gestures	%gpx: point
%sit:	Description of an event or state in the current Situation	%sit: Mother enters the room
word^word	Thai only: Compounds, derivations as well as formulaic expressions. This is not CHAT standard, but was used for technical reasons.	thOON3^faa3, kwaam0^s maj0^pen0^raj0
word_word	Swedish only: Inflectional morpheme boundary. Not CHAT standard but used for technical reasons.	fisk_ar_na
word+word	Swedish only: Derivational morpheme boundary.	flyg+plan
$word_1$ [: $word_2$]	Citation forms: $word_2$ is the citation form of $word_1$	liip3 [: riip3]
#	Short pause or hesitation	dek1 kO2 # mOON0 hen4
##	Long pause	su1nak3 kO2 wiN2 ##
:	Extra-long vowel	lxx:w3
[>] and [<]	Swedish only: Indicates overlapping speech. In addition to [>] and [<] marking the first and the second overlapping utterance respectively, the exact scope of the overlapping speech is marked by using < and > around the overlapping parts.	*FAT: vänta <lite> [>] . *BEL: <ä tjå> [<] bad_e .
+<	Thai only: Indicating overlapp	
<string> [/] string	Repetition	kop1 [/] kop1
<string1> [//] string2	Retracing with change	<kop1 pay0> [//] kop1 ma
[^c]	Thai only: Clause boundary	mii0 kop1 maj3 [c] maj2
.	Utterance delimiter: declarative	*CHI: mii kop1.
?	Utterance delimiter: question	Uh?
!	Thai only: Utterance delimiter: request or exclamation (The Swedish part of the corpus uses "." in this case.)	paj!
+…	Trailing off	*CHI: lxxw3 kO2 ## +…
word@o	Thai only: Onomatopoetic expression	eN4eN4eN4eN4@o
word@i	Thai only: Interjection	aaw3@i
word@f	Thai only: Family-specific form	kok1kik3@f
word@c	Thai only: Child-invented pause	aap3a1@c
word@s	Thai only: Second language form	wan0@s
word@wp	Thai only: Word play	ooj0uj2aa2@wp
word@si	Thai only: Singing	la0lala@si

[=! description]	Swedish only: Paralinguistic information.	*CHI: aa [=! whispers] . *MOT: 0 [=! laughing] .
xxx	Unintelligible utterance	
xx	Unintelligible word within an utterance	
www	Untranscribed speech	
0	Non-verbal act	

Appendix B

Transcripts of the 60 data files.