東邦学誌
第33巻第２号
2004年12月
論　文

Making of the Yuu Corpus：Longitudinal Spontaneous Speech Data of a Japanese Boy

Mayumi Nishibu

Contents:

1. Introduction

1.1 Outline

This paper describes the main features and the creation process of a longitudinal speech 'corpus' compiled by the author. The following sections show how the real language data was collected (in section 2), how the collected utterances were transcribed (in section 3), and how the transcription has been modified and coded in accordance with the Japanese-specific format designed by the CHILDES project (in section 4).[1] A description of and example text from the Yuu Corpus will be given in section 5.

1.2 Corpus and Longitudinal Speech Data of Japanese Children

The term 'corpus' in linguistics means a database in which electronically coded texts of natural languages are stored. The development of language corpora has enabled linguists to examine a large number of language samples quickly on analytical software. Accordingly, an increasing number of studies on lexicography, grammar, historical and stylistic investigation, language acquisition, and teaching have been conducted empirically and quantitatively based on corpora. The corpora frequently used by linguists are the following: corpora of spoken and written English from various genres (e.g., British

National Corpus, the Bank of English, and American National Corpus), corpora of historical English (e.g., Helsinki Corpus), corpora of children's written and spoken language (e.g., the CHILDES Database), and corpora of second language Learner English (e.g., Longman Learner's Corpus and International Corpus of Learner English).[2] Although the largest number of corpora deal with English, there are many corpora dealing with other languages as well.

The corpus introduced in this paper is a corpus of first language acquisition. It consists of spontaneous speech of a Japanese boy (from the age of 1;09 to 2;08), which was collected longitudinally in conversations between him and his mother (and sometimes with other participants, such as his grandparents). The files of the Yuu Corpus are constructed in accordance with the CHAT format, which is a principled transcription system used in the CHILDES project (MacWhinney, 2000). At present, only a few CHAT-formatted corpora of Japanese longitudinal speech data are open to the public.[3]

The longitudinal speech data of Japanese children that are available to researchers are quite limited. Researchers hesitate to make their data open, probably because the longitudinal observation of young children's speech is such a laborious task that they treasure their data, or because the publication of conversations between child and mother may infringe on the privacy of the participants. Thus, many previous studies of Japanese language acquisition have been conducted based on a restricted number of published longitudinal data, such as Noji's diary-style data (Noji, 1977) and Okubo's data (Okubo, 1967); otherwise, studies are based on researchers' own private data. Therefore, the accumulation of longitudinal corpora of children's speech is the most promising way to correct the paucity of objective, accessible data for studies of Japanese language acquisition.

2. Data Collection

2.1 Subject's Background

The subject boy (Yuu, a shortened form of his real name) was born on 26th March in 1995 as the first child of a Japanese couple, and he is growing up in Tokyo. His younger sister was born in October of 1997. Her mother usually speaks standard Japanese, but sometimes speaks her native Osaka dialect. Yuu and his mother moved into his grandparents' house in Osaka in preparation for his mother's delivery in September 1997, and stayed there until the end of the recording period. Yuu was a talkative and active boy. He liked to play with '*Shimajiroo*,' a tiger puppet, 'Thomas' toy trains, and dolls and various toys related to the TV superheroes '*Ultraman*' and 'Mega-ranger.'

2.2 Method

Conversational interactions and monologues of Yuu were tape-recorded by his mother during indoor free play for 30 minutes per week. The participants were Yuu and his mother in most of the sessions, and his grandparents in several sessions after September 1997 (when the child was 2;05 in age). The date and time for the recording were not fixed, and the mother started recording whenever she thought the child was in a good mood. Notes on the contextual information of every recording were provided by the mother. The mother tried to make the child speak during the recording. Several recording sessions were suspended or canceled when the child was extremely quiet (roughly, no words for more than 5 minutes). The recordings were canceled and rescheduled when their quality was poor, with a lot of noise or with faint voices for long durations. The period of recording ran from August 1996 to April 1998 (from when the child was 1;05 in age to 2;09), but the first and last few months of recorded speech have not yet been transcribed.

3. Transcription

The transcription of the recorded utterances was made by the author, using the CED editor.[4] The transcription was done in *Romaji* (Hebon) instead of Japanese *Hiragana*, since analytical software runs faster and more accurately with ASCII.

At first, Morikawa's format was employed for the transcription of the tape. In Morikawa's format, the annotation of morphological analyses of inflected words was parenthesized and inserted after each word on the main line of the utterances. Later in 1999, the author changed all of the transcriptions into the standard JCHAT format (Oshima-Takane and MacWhinney, 1995).[5] This revision was necessary in order to use the CLAN program effectively and to maintain uniformity with the Aki Corpus (Miyata, 1995), the first longitudinal corpus of L1 Japanese included in the CHILDES database.[6] In the standard version of the JCHAT format, morpho-syntactic analyses of every word are placed in a new line called a 'morpho-syntax tier,' which simplifies the main line.

4. Coding and Modification

Since the orthography of the Japanese language does not employ a space between words, we need to split all phrases or clauses into words and insert a space between words on the transcript. This format of Japanese texts is called '*Wakachi-gaki*.' To make a *wakachi-gaki*-formatted transcription, WAKACHI 98 (Miyata and Naka, 1998), a standardized version of *Wakachi-gaki* proposed for analyzing child language, was employed.

The transcription was modified and coded in accord with the CHAT format for Japanese (i.e., JCHAT 1.0). The 'headers,' 'dependent tiers,' and 'special form markers' were inserted in the transcript.

The headers, (which begin with the '@' sign,) were coded to give basic information for the file.

The dependent tiers, (which begin with a '%' sign) used in the files were %tim (to indicate the time duration of the tape), %act (to describe the actions of the speaker that are necessary to understand the transcription), %err (to indicate the use of erroneous forms), and %mor (to indicate grammatical categories and morphological analyses). Regarding %mor, morpho-syntactic analysis was conducted with the JMOR program (Naka, 1998), and every utterance on the main line was followed by a %mor line. However, %mor makes a file three times (or more) as heavy, and the JMOR program is still under development (or revision). The results of the morpho-syntactic analysis showed about 60% accuracy. Therefore, the data including %mor tiers have been compiled into another version of the files.

Special form markers, (which consist of the symbol '@' in conjunction with one or two additional letters, placed at the end of a word,) were inserted in the transcript. They included the @fp marker for sentence final particles(*shuujoshi*), @i for interjections(*kantooshi*), @o for onomatopoeic forms, @l for letters, @c for caregiver language (or motherese/baby-talk) (*yoojigo*), and my original markers such as @k for the names of animation characters, @u for unclearly pronounced words, and @n for phonetically clear but semantically unclear words.

An example of a text in the Yuu Corpus is shown in Example 1.

Example 1. Extraction from a File of the Yuu Corpus

@Begin
@Participants: CHI Yuuya Target_Child, MOT Tsuruyo Mother,GMO grandmother
@Filename: yuuya39.cha
@Age of Yuuya: 2;6.06
@Date: 2-OCT-1997
@Location: in the living room at Mother's parents' house
@Time start: 11:00
@Situation: playing with toys (the Megaranger jigsaw puzzle, the Ultraman picture cards, animation character picture books)
*CHI: mama [/] mama [/] mama , dekita yo „ [>] kore.
%act: playing with the Megaranger jigsaw puzzle

```
*MOT:	dekita [〈] ?
*MOT:	waa@i , honto da .
*MOT:	sugoi nee@fp , yuu-kun .
*CHI:	kore wa koko ?
*MOT:	sore koko kana@fp ?
*CHI:	hora@i .
*MOT:	a@i , honto da .
*MOT:	megaburuu@k soko da nee@fp .
*CHI:	kore doko ?
*CHI:	kore koko ?
*CHI:	hikooki da „ kore .
*CHI:	kore hikooki .
*MOT:	hikooki da nee@fp .
*MOT:	hikooki mitaina katachi shiteru naa@fp .
*CHI:	un .
*CHI:	kore wa ?
*MOT:	n ?
*CHI:	kore [/] kore ?
*MOT:	soo [/] soo [/] soo [/] soo .
*MOT:	beruto ga mieteta mon nee@fp .
*CHI:	kore +. . .
*MOT:	moo yuu-kun zembu oboeteru nee@fp .
*CHI:	un .
*MOT:	kore wa nande wakatta kana@fp , yuu-kun ?
*CHI:	〈kore wa nee〉[〉] [/] kore wa nee chotto dame machigaeta nda yo@fp.
*MOT:	un [〈] .
*MOT:	machigaeta no@fp ?
```

5. Contents of the Yuu Corpus

The Yuu Corpus currently contains 51 files covering the ages between 1;09 and 2;08. Each file includes approximately 30-minute-long weekly-collected interactions between the child and his mother (and his grandparents). The details of the files in the Yuu Corpus are listed in Table 1.

Mean Length of Utterance (MLU, henceforth) was calculated with the CLAN program.[7] Interjections, utterances consisting of singing a song and unclearly pronounced words were excluded from the calculation. The MLU of Yuu was relatively high (1.766 at 1;09 and 2.443 at 2;08), which means that Yuu's language development was progressing relatively quickly and steadily.

Table 1. Construction of the Yuu Corpus

age	file name	date			length of session (h:m:s)	total length of session (h:m:s)	number of utterances	MLU of the child
1;09	Yuu1	29	Dec.	96	00:26:11	1:59:11	Yuu 1469 Mother 1313	1.766
	Yuu2	4	Jan.	97	00:13:00			
	Yuu3	8	Jan.	97	00:20:00			
	Yuu4	15	Jan.	97	00:36:16			
	Yuu5	22	Jan.	97	00:23:44			
1;10	Yuu6	30	Jan.	97	00:36:46	1:48:55	Yuu 1455 Mother 1259	1.770
	Yuu7	5,6	Feb.	97	00:23:14			
	Yuu8	12,13	Feb.	97	00:28:12			
	Yuu9	19,20	Feb.	97	00:20:43			
1;11	Yuu10	26	Feb.	97	00:29:53	1:21:35	Yuu 903 Mother 868	1.772
	Yuu11	5	Mar.	97	00:30:07			
	Yuu12	22	Mar.	97	00:21:35			
2;00	Yuu13	27	Mar.	97	00:32:16	2:10:20	Yuu 1698 Mother 1427	1.925
	Yuu14	9	Apr.	97	00:35:12			
	Yuu15	16	Apr.	97	00:24:39			
	Yuu16	23	Apr.	97	00:38:04			
2;01	Yuu17	3,4	May.	97	00:21:56	1:56:20	Yuu 1489 Mother 1310	1.903
	Yuu18	7	May.	97	00:31:19			
	Yuu19	14	May.	97	00:28:41			
	Yuu20	22	May.	97	00:34:24			
2;02	Yuu21	28	May.	97	00:25:36	1:41:49	Yuu 1315 Mother 1343	2.225
	Yuu22	4,5	Jun.	97	00:16:12			
	Yuu23	12	Jun.	97	00:25:37			
	Yuu24	18	Jun.	97	00:38:28			
	Yuu25	25	Jun.	97	00:21:32			
2;03	Yuu26	3	Jul.	97	00:33:30	2:00:00	Yuu 1272 Mother 1430	2.200
	Yuu27	10,11	Jul.	97	00:26:30			
	Yuu28	17	Jul.	97	00:35:05			
	Yuu29	24	Jul.	97	00:24:55			
2;04	Yuu30	31	Jul.	97	00:37:29	2:00:00	Yuu 1180 Mother 1076	2.406
	Yuu31	6	Aug.	97	00:22:41			
	Yuu32	16	Aug.	97	00:32:45			
	Yuu33	21	Aug.	97	00:27:15			
2;05	Yuu34	27	Aug.	97	00:38:30	2:09:05	Yuu 1174 Mother 1261	2.511
	Yuu35	4	Sep.	97	00:10:00			
	Yuu36	10	Sep.	97	00:37:02			
	Yuu37	19	Sep.	97	00:09:17			
	Yuu38	25	Sep.	97	00:34:26			
2;06	Yuu39	2	Oct.	97	00:22:18	1:40:20	Yuu 1388 Mother 1383	2.427
	Yuu40	12	Oct.	97	00:17:56			
	Yuu41	18	Oct.	97	00:20:19			
	Yuu42	24	Oct.	97	00:39:47			
2;07	Yuu43	30	Oct.	97	00:14:33	1:42:09	Yuu 1250 Mother 1097	2.454
	Yuu45	6	Nov.	97	00:40:00			
	Yuu46	13	Nov.	97	00:20:00			
	Yuu47	22	Nov.	97	00:27:36			
2;08	Yuu48	29	Nov.	97	00:26:29	2:02:29	Yuu 1689 Mother 1280	2.443
	Yuu49	6	Dec.	97	00:38:25			
	Yuu50	13	Dec.	97	00:21:35			
	Yuu51	20	Dec.	97	00:36:00			

6. Concluding Remarks

Several years have been spent compiling this longitudinal speech data. It took two years to observe and record Yuu's language, another two years to transcribe the tape, and another two years to code the data in the JCHAT format. Nonetheless, a few more years will be required to refine the corpus (in morpho-syntactic and phonological aspects, in particular), though general linguistic studies can be conducted with the current version of the Yuu Corpus.

In retrospect, there were ways in which the process of creating the corpus could have been improved. Those can be summarized as follows.

First, it is recommended that the main line of each utterance should be kept as simple as possible, and that the other features, such as morpho-syntax analyses, errors and actual pronunciation, should be coded in a new line. This would help us to cope with various revisions and with the evolution of the transcription and coding formats.

Second, I recommend that the researcher him/herself should transcribe the whole recording from the beginning to the end. This makes the transcription more consistent and accurate. In addition, this task forces the researcher to peruse every utterance in the recording, and this thorough acquaintance with the recorded utterances may give him/her correct inferences and insights in the linguistic analysis.

Third, leaving the time-consuming manual coding for the time being may turn out to be advisable, because the more labor a coding procedure requires, the more likely it is that software or a program that completes the coding quickly will be developed. The software or program will eventually be distributed to researchers.[8)]

The use of the Yuu Corpus is currently limited to the author, but the corpus will be open to other researchers in the near future.

Notes

* I am indebted to Dr. Susanne Miyata and Norio Naka for their generous assistance in compiling this corpus.

1) CHILDES (Child Language Data Exchange System) was established mainly by Catherine Snow and Brian MacWhinney in 1984. The goals of this system are to 'automate the process of data analysis,' 'obtain better data in a consistent, fully-documented transcription system' and 'provide more data for more children from more ages, speaking more languages' (MacWhinney, 2000, 1: 4). Three separated but integrated tools were developed: the CHAT transcription and coding format, the CLAN analysis program and the database.

2) Useful lists of corpus resources can be found in the appendix of Biber, et al. (1998) or Meyer (2002).

3) At the time the author started the transcription in 1999, the Aki Corpus (Miyata, 1995) was the only

JCHAT-formatted longitudinal speech data available to the public. At present, the Ryo Corpus by Miyata and the Jun Corpus by Ishii (although the data for the important stage from 2;00 to 3;04 are not given by Ishii) are available in the CD-ROM of MacWhinney (2000). These two as well as two other new corpora (by Miyata and by Hamasaki) are available at the following webpage (http://xml.talkbank.org:8888/talkbank/file/CHILDES/East Asian /Japanese/).

4) The CED editor is a program designed specially to transcribe data in CHAT format.

5) The JCHAT format is a transcription and coding format specially designed for Japanese. See Oshima-Takane and MacWhinney (1995).

6) The CLAN (Computerized Language Analysis) program is designed specially to analyze data transcribed in the format of CHILDES.

7) MLU (Mean Length of Utterance) is the average length of an utterance, normally calculated in morphemes. This parameter provides a rough guide to the stage of linguistic development of a child. See Brown (1973) for details on MLU. The MLU for Japanese used here was *Jiritsu-go Fuzoku-go* MLU. *Jiritsu-go* means autonomous words, equivalents to content words, which include nouns, verbs, adjectives, adverbs, adjectival nouns, interjections, and conjunctions. *Fuzoku-go* means affiliating words, which include auxiliary verbs and particles.

8) In the course of compiling the Yuu Corpus, a program that automatically changes texts into WAKACHI format was developed, and the version of JMOR was revised. The standard version of the CHAT format has been evolved from texts into 'sonic' transcripts of utterances.

References

Aitchison, Jean (2003) *A Glossary of Language and Mind.* Oxford: Oxford University Press.

Biber, Douglas, Susan Conrad, and Randi Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Brown, Roger (1973) *A First Language: The Early Stages.* Cambridge, Mass: Harvard University Press.

Oshima-Takane, Yuriko and Brain MacWhinney (1995) *CHILDES Manual for Japanese,* McGill University / Chukyo University.

MacWhinney, Brian (2000) *The CHILDES Project: Tools for Analyzing Talk. 3rd ed., vol.1: Transcription Format and Programs, vol.2: The Database.* Mahwah, NJ: Erlbaum.

Meyer, Charles F. (2002) *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.

Miyata, Susanne (1995) 'The Aki Corpus: Longitudinal Speech Data of a Japanese Boy Aged 1;5.7－3;0.0,' *The Bulletin of Aichi Shukutoku Junior College, 34,* 183-191.

Miyata, Susanne and Norio Naka (1998) 'Wakachigaki Gaidorain WAKACHI98 v.1.11,' *Educational Psychology Forum Report,* 98-003, The Japanese Association of Educational Psychology.

Naka, Norio (1998) 'JMOR.' *JCHAT' 98 CD-ROM.* JCHAT Project.

Noji, Junya (1977) *Yoojiki no Gengo Seikatsu no Jittai, 4 vols.* Hiroshima: Bunka Hyooron Shuppan.

Okubo, Ai (1967) *Yooji Gengo no Hattatsu.* Tokyo: Tokyodoo Shuppan.