David Gil
Uri Tadmor
Department of Linguistics
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6, D-04103 Leipzig, Germany.

The MPI-EVA Jakarta Child Language Project was a joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, and the Center for Language and Culture Studies, Atma Jaya Catholic University. The project was officially started in January 1999, and recordings and data processing began in April 1999. Funding was provided by the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. The goal of the project was to record, transcribe, and enter into a computerized database a corpus of naturalistic data from a large sample of Jakarta Indonesian child language. A total of eight children were studied longitudinally over the course of five years. The children's ages at their first recordings ranged from 1:7 to 4:6, and each child was recorded at average intervals of 7-10 days over a period of 2-4 years. In addition, data relating to each age group was used for latitudinal studies. Recordings were made at various settings, mostly indoors but also outdoors. Specific situational descriptions are contained in each file.

The target children were chosen for practical reasons. We chose families with which our research assistants were already familiar and where we were reasonably confident that we would be able to maintain regular recording sessions for several years. However, the families are fairly representative of Jakarta's population, belonging to different socioeconomic strata and different ethnic groups. Our *sine qua non* condition was that the target children were acquiring Jakarta Indonesian as their first language, and that the major home language of all families (used between parents and children as well as among children) was Jakarta Indonesian. However, other members of the household (mostly grandparents) sometimes spoke languages other than Indonesian, as is the case in most Indonesian families.

Data collection and processing was carried out by graduates of language-related departments of Indonesian universities who underwent a stringent selection process. The successful candidates received training in field methods, phonetic transcription, morphological analysis, and data entry.

The data collection and processing followed a regular routine. On a weekly basis our research assistants took a camcorder to the field and recorded the target children in their (the children's) homes. The aim was to record natural language in a natural setting. Other than the research assistants and the target children, participants sometimes included parents, siblings, grandparents, friends, and others. The assistants then returned to the Field Station and captured the video recordings to digital video files that were then burnt to CDs. The digital video files were made in PAL format (MPEG-1, 352 x 288 pixels, 25 fps). This allowed us to fit about one hour of video onto a regular 650MB data CD. The CDs were then viewed and coded by the research assistants, each assistant working on

the sessions that he or she recorded. Coding was done directly into our customized FileMaker database software.

Each utterance comprises a single record in the database. Each record consists of five fields: transcription using conventional orthography (of any recorded utterance, uttered by anyone); phonetic transcription; interlinear glossing; English translation; and comments specific to the particular utterance regarding linguistic matters as well as the nonlinguistic context.

Our research supervisors checked a large random sample of the coded files, consisting of about 20% of the total, to ensure data integrity and consistency of the data processing methods.

The personal names used in the corpus were not replaced by pseudonyms. Names and nicknames are frequently used in argument positions in Indonesian (where speakers of English, for example, would use pronouns), and in fact comprise about 10% of the total data in the corpus. Altering them would have significantly distorted the data. Moreover, personal names are subject to special morphological processes (e.g. various types of truncation and nickname derivations), and this important linguistic information would have been lost had the names been replaced by pseudonyms. It should also be noted that the names of participants, when mentioned, comprise single names and nicknames, not complete names. Moreover, using personal names in the context of linguistic data citation does not violate Indonesian legal, academic, or cultural norms. However, when quoting from the database users who so wish may substitute names with codes or pseudonyms, as long as this is clearly noted.

Transcription conventions

Jakarta Indonesian is not commonly used in print, Standard Indonesian being used instead. However, it is often used in advertisements, billboards, short text messages, email chats, and personal letters. Some newspapers also use it for writing headlines. Although the orthography is far from standardized, it is roughly based on the spelling of Standard Indonesian, with a few additional conventions. Most characters are used with their IPA values, with a few exceptions:

Graph/digraph	*Description*	*IPA*
ny	palatal nasal	ɲ
ng	velar nasal	ŋ
j	voiced palatal stop	ɟ
sy	palatal fricative	ç
e	mid central vowel	ə
e	front central vowel	e, ɛ

Note that in the conventional spelling of Jakarta Indonesian (as well as in Standard Indonesian), the mid central vowel and the front central vowel are not distinguished, even though they constitute separate morphemes. Moreover, in Jakarta Indonesian glottals are not spelled consistently. The glottal stop is sometimes unwritten, sometimes it is represented by an apostrophe, sometimes by *q*, and sometimes (in final word position) by *k*. The glottal fricative can be represented by h or (rarely) by *kh*; sometimes it is not written, and sometimes an *h* is written even though no glottal phoneme is present (based on the orthography of the cognate in Standard Indonesian).

Punctuation marks and codes used in the transcription line include:

.	End of utterance, statement
?	End of utterance, question
!	End of utterance, imperative (but not exclamation)
,	Comma
...	Interruption (in the middle or at the end of an utterance)
xx	Unanalyzable, treated as a word
xxx	Unanalyzable, not treated as a word
	www Unintelligible speech in another language, including prayers in Arabic
'...'	Quotations, role play (i.e. when speaker pretends to be someone else)
"..."	Titles (of songs, books, movies, etc.)
0	Action with no utterance (action is described in the comment field)

Interlinear glosses

Each Indonesian word has a single gloss equivalent in the glossing line. The gloss contains as many morphemes as are analyzed in the Indonesian form, separated by hyphens. Lexical morphemes are generally translated into English. If a single Indonesian morpheme has an equivalent consisting of more than one English word, the words are separated by a period; for example Indonesian *adik* is glossed as 'younger.sibling'. For glossing grammatical morphemes (affixes, function words, and reduplication patterns) three approaches were used. If there was an unambiguous English equivalent it was used as the gloss. For example, Indonesian *ke* was glossed as 'to', and Indonesian *ini* was glossed as 'this'. If the morpheme could be easily described by a grammatical term, an abbreviation of that term was used. For example the negator *tidak* was glossed as NEG, and the relativizer *yang* was glossed as REL.

The following grammatical abbreviations were used:

Abbreviation	*Meaning*
1	first person
2	second person
3	third person
AGT	agent, agentive
COMP	complementizer

CONTR	contrastive
EPIT	epithet
EXCL	exclamation
FILL	filler
FUT	future
IMIT	imitative (inc. nonlexicalized onomatopoeia and interjections)
LOC	locative
MUT.RED	mutative reduplication (a special type of full reduplication where some of the phonemes of the second element undergo mutation)
NEG	negator
OATH	oath
PERS	person marker
PFCT	perfect
PL	plural
POSS	possessive
RED	reduplication
REL	relativizer
SG	singular
TOP	topic marker
TRU	truncation

However, for a number of function morphemes (basically affixes, clitics, and particles), we were unable to settle on a single uncontroversial and agreed-upon gloss that would provide a clear indication of its function. In some cases this was because the form in question has a variety of seemingly different functions, in other cases because the form has been analyzed in different ways by different scholars, and in yet other cases for both of the above reasons. These forms are accordingly glossed with an upper-case replication of the forms' conventional spelling.

For the benefit of users unfamiliar with Jakarta Indonesian, some further information on these forms is provided in the following five tables, covering separate words, prefixes, suffixes, circumfixes, and complex discontinuous morphemes respectively. In the tables below, the first column shows the form as it is appears in the interlinear gloss, the second column provides one or more recommended glosses representing some aspects of its function(s), while the third column presents a very concise description of its function(s). These descriptions are just rough and ready suggestions as to the nature of the forms in question, invitations to the user to come up with more explicit analyses, based on the data in this corpus. Please note that most exclamations are simply glossed as EXCL.

Separate Words

Gloss	*Suggested abbreviation(s)*	*Description*
AH	HORT	Hortative particle, typically used to express speaker's intention to perform activity

AYO	COHRT	Cohortative particle, inviting interlocutor(s) to join in performing an activity. Sometimes also used as an exhortative particle, urging others to perform activity
DAH	PFV, PFCT	Aspect marker, ranging in usage from perfective to perfect
DEH	IMP, CONC	Pragmatic particle with a variety of functions, including completion, command, and concession.
DENG	CORR	Particle expressing self-correction; variant of DING
DIH	IMP	Imperative particle.
DING	CORR	Particle expressing self-correction
DOGE	VCBL	Vocable (meaningless expression used in singing) [Toba Batak]
DONG	IMP, EMPH	Pragmatic particle with a variety of functions including softening (in imperatives) and 'of course' (in declaratives)
EH	CORR	Exclamation expressing self-correction
GIH	IMP	Strong imperative particle
HAYO	EXHRT	Variant of AYO used to challenge interlocutor, e.g. to perform a risky activity or to provide an answer to a riddle.
KAH	Q	Question particle, used primarily to express polar interrogatives [Standard Indonesian]
KAN	Q, EMPH	Reduced form of the negative marker *bukan*, used as a question particle, usually to form tag questions, and as an emphatic particle (‘you know’) preceding emphasized phrase
KEK	ASSOC:DISJ, INDF	Particle with several functions, including associative disjunction ('or things like that'), and indefinites (in construction with content interrogatives)
KOK	FOC, CONTR, 'how come'	Contrastive focus particle, which in initial position acquires interrogative force to mean 'how come'
LAH	IMPER, CONC, FOC	Pragmatic particle expressing a variety of meanings, including imperative, concessive, and contrastive focus
MAH	TOP	Pragmatic particle typically occurring between a contrastive topic and a following comment
MARI	COHORT	Cohortative particle, inviting interlocutor to join in performing an activity. Also used as a polite imperative and in ritualistic leave-taking (‘goodbye’).
NAH	PRES, EXCL	Presentative particle, often used to introduce the main point of an argument. Also an exclamation to express satisfaction, especially at interlocutor’s understanding (‘That’s it!’, ‘You got it!’) and after completing a task (‘There we/you go!’).
PAN	Q, EMPH	Forms tag questions or emphasizes following phrase, much like KAN [Betawi]

PUN	FOC	Focus particle, with functions similar to 'also', 'even' and others, also used to form indefinites (in construction with content interrogatives)
SIH	FOC, EXPL	Pragmatic particle with a variety of functions. In declaratives occurs after the topic to denote contrastive focus and clause finally to mark explanations. In interrogatives, requests clarification or repetition of previously provided information.
SOK	'presumptuously'	Usually precedes adjectives and means 'presuming to be something (conveyed by that adjective) that one is actually not'
TEH	FOC	Focus particle with uses similar to SIH [Sundanese]
TO	Q, EMPH	Tag ('right?') or emphatic, used much like KAN by speakers of Javanese background.
TOH	Q, EMPH	Variant of TO

Prefixes

Gloss	*Suggested abbreviation(s)*	*Description*
BA-	DEPAT-, MED-	Voice marker, sometimes analyzed as a depatientive or middle voice marker [Papuan Malay]
BE-	DEPAT-, MED-	Voice marker, sometimes analyzed as a depatientive or middle voice marker
BER-	DEPAT-, MED-	Voice marker, sometimes analyzed as a depatientive or middle voice marker
DI-	PAT-, PASS-	Patient-oriented voice marker, sometimes analyzed as a passive voice marker
KE-	DEAG-, INVOL-	Voice marker, sometimes analyzed as a deagentive or passive voice marker; depending on stem, it can also mark involuntary activity
MA-	AG-, ACT-	Actor-oriented voice marker, sometimes analyzed as an active voice marker
MEN-	AG-, ACT-	Actor-oriented voice marker, sometimes analyzed as an active voice marker [Standard Indonesian]
N-	AG-, ACT-	Actor-oriented voice marker, sometimes analyzed as an active voice marker
PE-	AG-, HAB-, INSTR	Derives agentive, habituative, or instrumental nouns from intransitive verbs
PEN-	AG-, HAB-, INSTR-	Derives agentive, habituative, or instrumental nouns from transitive verbs
SE-	one-, same-, as-	Basic meanings are 'one' ,'same', and 'as', also used to derive words with a variety of functions
TA-	DEAG-, INVOL-, SUPERL-	Voice marker, sometimes analyzed as a deagentive or passive voice marker; depending on stem, it can also

		mark involuntary activity [Papuan Malay]
TE-	DEAG-, INVOL-, SUPERL-	Voice marker, sometimes analyzed as a deagentive or passive voice marker; depending on stem, it can also mark involuntary activity or the superlative
TER-	DEAG-, INVOL-, SUPERL-	Voice marker, sometimes analyzed as a deagentive or passive voice marker; depending on stem, it can also mark involuntary activity or the superlative

Suffixes

Gloss	*Suggested abbreviation(s)*	*Description*
-AN	-NOUN, -COMPR, -RECP	Derivational suffix with a variety of seemingly unrelated meanings, including deriving nouns, comparative (for adjectives), reciprocal (for some verbs), and many others
-E	-3, -POSS, -3:POSS, -ASSOC, -DEF	Marker of a range of functions from possessive (usually third person) through to definiteness; may be analyzed as expressing a generalized relationship of association [Javanese]
-IN	-END.POINT, -APPL, -TRANS, -BEN, -CAUS	Voice marker, sometimes analyzed as an end-point or applicative voice marker; has a range of functions, including causative, benefactive, and transitivizer
-I	-TR	Forms transitive verbs
-KAN	-END.POINT, -APPL, -TRANS, -BEN, -CAUS	Voice marker, sometimes analyzed as an end-point or applicative voice marker; has a range of functions, including causative, benefactive, and transitivizer
-NO	-APPL.IMP	Forms imperatives of applicative verbs [Javanese]
-NYA	-3, -POSS, -3:POSS, -ASSOC, -DEF	Marker of a range of functions from possessive (usually third person) through to definiteness; may be analyzed as expressing a generalized relationship of association
-NYE	-3, -POSS, -3:POSS, -ASSOC, -DEF	Marker of a range of functions from possessive (usually third person) through to definiteness; may be analyzed as expressing a generalized relationship of association
-O	-IMP	Imperative suffix [Javanese]
-WAN	-AG:M	Suffix forming masculine agentive nouns
-WATI	-AG:F	Suffix forming feminine agentive nouns

Circumfixes

Gloss	*Suggested abbreviation(s)*	*Description*
KE.AN	ABST-[root]-CIRC, ADV-[root]-CIRC	Derives abstract nouns and adversative passives

PENG.AN	VN-[root]-CIRC	Derives verbal nouns
PER.AN	NOUN-[root]-CIRC	Derives collective and other nouns
PE.AN	NOUN-[root]-CIRC	Derives collective and other nouns (rare)
SE.NYA	ADV-[root]-CIRC	Derives adverbs
SE.RED.NYA	ADV-RED-[root]-CIRC	Derives 'superlative' adverbs ('as x as possible')
RED.AN	SIMIL-[root]-AN	Derives similitudinals (conveying a concept similar but not identical to the meaning of the root)

8. Biographical data of the eight principal target children

Code	*Sex*	*Age at start*	*Age at end*	*Socioeconomic Status*	*Ethnic background*	*Other languages spoken at home*
RIS	Female	1;8	6;1	lower	Father Sundanese, Mother Betawi	Traditional Betawi
PIT	Female	4;4	8;9	middle	Father Javanese, Mother Javanese	Javanese
IDO	Male	3;4	6;5	middle	Father Javanese, Mother Sundanese	Javanese
HIZ	Male	1;7	5;11	upper middle	Father Javanese, Mother Manado	Javanese
PRI	Female	2;7	6;1	upper middle	Father Javanese, Mother Chinese	None
MIC	Male	2;0	3;10	upper middle	Father Chinese, Mother Chinese	None
LAR	Female	2;10	6;4	middle	Father: Chinese-Betawi-Javanese, Mother Chinese	Javanese (nanny)
TIM	Male	1;6	5;0	middle	Father Papuan-Dutch, Mother Batak	Toba Batak

There are 8 files I the PIT-OPI folder. These files were recorded as part of the PIT project, but PIT herself was not present during the recording. She was unavailable, but the researcher decided to proceed with the recording since her cousin OPI was present and we were also tracing his linguistic development.

Articles based on the use of this corpus should cite the following source:

Gil, David, & Tadmor, Uri, 2007. The MPI-EVA Jakarta Child Language Database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.

1. The data may not be used for any commercial purposes without prior written consent from the contributors.
2. Any citation of the data must be properly attributed to the authors according to the relevant academic norms.
3. The contributors request that a copy of any published work citing the data be sent to them at the following address: Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany.