MinJ–CHAT 0.9 Hebon (Preliminary Version)*

——CHILDES用日本語データ入力フォーマット——

宮　田　Susanne

MinJ–CHAT 0.9 Hebon (Preliminary Version)*: Entering Format for Japanese

Susanne Miyata

1．CHILDES，CHAT，J–CHATの概念

従来，言語獲得の研究において，研究者が独力で観察できる子どもの数に限りがあることが研究の大きな障害となっていたが，CHILDESという言語表現のデータ処理手法を用いれば，この壁を破ることができる。

CHILDESでは言語発話を共通のフォーマットによって処理することによって，苦労して集めたデータを多数の研究者の間で共有できるようになり，効率も大きく引き上げることができるようになる。また，CHILDESでは，ほかの言語との比較や，健常児と障害児の言語発達の比較も容易にできる。

CHILDESのパッケージは，データベース，分析プログラム集（CLAN），入力フォーマット（CHAT）の三つの部分が含まれている。これらはCD–ROMにまとめて入っているし，CLANとCHATのマニュアルは本にもなっている。

(1) データベース

データベースには，色々な言語の発達データ（縦断データが多いが，横断データ，障害児言語発達，第2言語自然獲得も含む），そして言語障害のデータ（失語症等）が集められている。

---

* MinJ–CHATの考え及びルールはJ–CHATプロジェクトメンバー（特にBrian Mac Whinney, Kazumi Matsuoka, Susanne Miyata, Hiromi Morikawa, Norio Naka, Yuriko Oshima–Takane）の電子メールの話し合い，及び第1回J–CHAT会議の際に承認されたものを，J–CHATマニュアル用にまとめたものである。

* 本章は問題点を検討するための臨時版であるので，最終版は内容的にも形式的にも変更される可能性がある。

* 使用希望者はJ–CHATプロジェクトに連絡すること。

* コピーライトはJ–CHATプロジェクトにある。

1993/6の CD–ROM には，英語のほかにアフリカーンス語，デンマーク語，オランダ語，フランス語，ヘブライ語，ハンガリー語，ポーランド語，スペイン語，タミル語，トルコ語などのデータのほかに，デンマーク語・日本語のバイリンガルデータも入力されていて，メンバーが自由にアクセスできるようになっている。

(2) CLAN

CLAN は分析や入力の際に役に立ちそうな Macintosh や IBM で動くプログラム集である。MLU 計算などが自動的に行われ，その結果がファイルにまとめられる。

(3) CHAT と MinCHAT

CHAT はデータを入力する時に使うフォーマットである。CLAN プログラムを利用するためには，少なくとも，CHAT のサブセットの MinCHAT のフォーマットに従う必要がある。MinCHAT より複雑なフォーマットに関しては，自分のデータと研究目的に合わせて，適宜利用すればよい。たとえば，発音記号 (PhonASCII)，イントネーション記号のほかに，誤用・言い間違い・関係ない発話・省略・方言・第２言語などが処理しやすいよう表記の仕方が工夫されている。

(4) J–CHAT と MinJ–CHAT

CHAT ファイルの必要条件の一つとして，ASCII 記号で作成されなければならないということがある（詳しいことは２．１．１に参考)。ASCII 記号には漢字や仮名が認められないので，日本語のデータでもローマ字で表記する必要がある。JCHAT はローマ字化によって生じる日本語特有の問題をどう扱うかまとめたものである。MinJ–CHAT は，日本語入力の際の最少必要条件をまとめたものである。

2．MinCHAT

MinCHAT フォーマットに従って作られたファイルの具体例：

@Begin
@Participants:　CHI Asuka Target_Child, MOT Kumiko Mother
*CHI:　basu atta !
%act:　omocha no basu o toriage, okaasan ni miseru
*MOT:　ja gareeji ni irete ne.
@End

CHAT フォーマットで作られたファイルは三つのタイプのラインからできている。

ヘッダーライン	@で始まる	場面や登場者に関する情報	（2．1．1に作照）
メインライン	*で始まる	実際の発話	（2．1．2に参照）
ディペンデント・ティヤー	%で始まる	前行の発話に関する情報	（2．1．3に参照）

2．1 MinCHATファイルを作るための必要条件

2．1．1 ヘッダーライン（Obligatory Headers, @）

ヘッダーラインはファイルの内容に関する一般的な情報を含む。以下の三つのヘッダーが最少限必要である。

@Begin CLANプログラムにファイルの初めを知らせる。

@Participants: ファイルの登場者。ファイルの二番目の行になる。各登場者について XXX name role

のように登場者の3文字ID記号，その名前，そして役割を記録する。

@Participants: CHI Asuka Target_Child, MOT Mother, INV Tomoko Investigator

よく使われる役割は Target_Child, Mother, Father, Brother, Sister, Teacher, Playmate, Investigator

ID記号は登場者の名前あるいは役割の初めの3文字を使うことが多い。複数の子どものファイルの場合は，CHIやMOTのような役割をID記号にしたほうが，複数のファイルを同時に分析できるので便利である。

@End ファイルの終わりを知らせる。

2．1．2 発話の書き方：メインライン（Main Lines, *）

実際に発話されたことをメインラインに書く。メインラインは*で始まる。上の具体例を見ると

*CHI: basu atta!

*MOT: ja gareeji ni irete ne.

は子ども（CHI）と母親（MOT）の実際の発話を示している。*（asterisk）の後には登場者のID記号（大文字を用いた3文字）を配し，その後はコロン，タブ。その後はテキストが入り，最後には休止符（terminator: .?!）を付ける。

*XXX: text text text.

ワンポイント：CLANがスペースとタブを使い分けているので，スペースと間違えないよう気を付けたほうがいい。タブは通常8文字単位にする。

名前や地名は英語の場合と同様，大文字で始まる。

*CHI: kore Asuka no da yo!

"obasan", "okaasan", "neesan" などの親族名は，名前として使われた場合，大文字で始まる（一般的な名詞として使われた場合は小文字）。

*CHI: Mama wa Tookyoo.

*CHI: ojisan ga matteru.

%act: ningyoo o eki no mae ni narabete

聞き取れない部分は xxx として表わす。

*CHI: xxx.

聞き取れない単語の音声的な形を記録したい場合は，&guga のように & のあとに発音を書く。

*CHI: kore &hosamon da yo.

不完全な単語は，省略された部分を丸い括弧で付け加えられる。

*CHI: Obaachan n(o) (u)chi.

2．1．3 コメントの書き方：ディペンデント・ティヤー（Dependent Tiers, %）

話し手の動作，場面に関する情報，観察者のコメント，分析記号などはディペンデント・ティヤーに書く。オプションであり，使用する義務はない。上の具体例の中で %act: omocha no basu o toriage, okaasan ni miseru の行はディペンデント・ティヤーの例になる。

% (percentage) の後に 3 文字の小文字のティヤー・記号を配し，コロン，タブ。その後はコメントなどを記入する。文の最後にピリオドなどの終止符を付けてはいけない。

%xxx: text text text

上の例では %act（=activity）というティヤーを選んで，子どもの行動を記録した。ほかにもよく使われるものは %com（=comment）である。

日本語の場合はテキストの書き方は英語でも日本語でもよい。

*CHI: dame!

%act: taking away new book from Ree

or: %act: atarashii hon o Ree kara toriagete

ワンポイント：*CHI: [tab] とか %com: [tab] をいちいちタイプする必要はない。例えば，入力するときに，明日香ちゃんの発音の前の "a" を書き，それをファイルができたときに "CHI:" に置換することもできるし，マクロコマンドに登録し，マンタッチで呼び出す方法もある。

2．1．4　ASCII ファイルである

ASCII ファイルを作るために，二つの条件がある。

(1) 文字は ASCII 記号（American Standard Codes for Information Exchange）に限る。

a b c d e f g h i j k l m n o p q r s t u v w x y z

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

1 2 3 4 5 6 7 8 9 0 - ^ [ ] ' ; / ¥ . ,

! @ # $ % ` & * ( ) – + " ~ : ? | > <

この ASCII 記号以外は，スペース，タブ，リターンしか使わない。

(2) ASCII ファイル（テキストファイル）として保存する。

エディター（ワープロ・ソフト）によってその保存の仕方が多少異なる。例えば IBM の Wordperfect の場合では DOS ファイルとして保存し，Macintosh の MS–Word の場合では「save as→text only」，NEC でよく使われる「一太郎」では「テキスト式」になり，WORDSTAR では最初から non–document として開き，MIFES やMcSink の場合はそのまま保存するとテキストファイルになる。

2．1．5　強引な転写を避ける

特定の転写法や分析記号を使用すると，転写記号の使い分けが難しい時もある。そういう場合に無理に使い分けないほうがよい。強引な使い方を避けるために CHAT にオプションがいくつか用意されている。また，%com ラインでコメントを書くこともできる。

2．2　ファイルをチェックする

ファイルに文法的な（フォーマット上の）間違えがあると，CLAN の分析プログラムが正確に働けない。よくある間違えは，終止符を打ち忘れたり，タブの変わりにスペースを打ってしまったなどである。CLAN のなかに入っている CHECK プログラムを使用するとフォーマット上の間違えを見つけ，その行番号や誤りの内容などを知らせてくれる。早めに気が付けば，誤解による誤りが避けられるので，頻繁に CHECK 使用をすることが望ましい。

2．3　ドキュメンタリーファイル（00readme. doc）

縦断研究のために収集された少数の子どもの複数のファイルの集合や横断研究のために収集されたいろいろな年齢のグループからなる多数の子どものファイルの集合をコーパス (corpus) と呼ぶ。各コーパスにドキュメンタリーファイル（00readme. doc）を付ける。データベースの利用者がデータを適切に解釈することができるように必要な情報を書き込む。以下に最少限必要な項目をまとめる。

(1) 謝　　辞 (Acknowledgements)

コーパスの利用者が引用文献としてなにを明記するべきか，などを指定する。そのほかにCHILDES 全体を利用したため MacWhinney and Snow 1990 を引用させる。

(2) 利 用 制 限 (Restrictions)

コーパスの利用を制限することができる。たとえば，コーパスをもとにした発表などのコピーを送るように指定することができる。しかし，多くの研究者が利用を限定しない。

(3) 警　　告 (Warnings)

データの限界を知らせる。たとえば発話の誤りを記述していない場合ではそれを知らせる。

(4) 偽　　名 (Pseudonyms)

登場者がデータの使用に同意しているかどうか，そして各登場者について偽名（通常）か本名（研究者の子どもの場合等）を使ったかどうかを記録する。

(5) 経　　緯 (History)

プロジェクトの経緯について，できるだけ詳しい情報を与える。プロジェクトに対する助成，そして目的や，資料収集法，転写法や転写の際に無視した部分，転写者，信頼性，記号の使用などを述べる。

(6) 記　　号 (Codes)

プロジェクトの固有な記号を使用した場合はここで説明する。

(7) 個人情報に関するデータ (Biographical Data)

各登場者について詳細な人口統計学的，言語的（方言など）な情報，心理検査によるデータなどをここでまとめる。項目として年齢，性別，兄弟，学歴，社会的階級，職業，以前の住所，宗教，趣味，友人などが含まれているとよい。親が育った場所か，家族が前にどこに住んでいたかなどの情報は社会言語学的な研究にとって非常に大切である。

(8) 目　　次 (Table of Contents)

日付と子どもの年齢そして，計算してあれば，MLU の値も含めたファイルの一覧表をここに入れる。

(9) 場 面 描 写 (Situational Descriptions)

一般的な場面の情報，たとえば家の間取り，つまりコーパス中に変わらない情報をここで

記録する。ファイルによって変わる場面情報は各ファイルに記録する。

3．MinJ—CHAT

3．1　文字化について

CLAN のプログラムを使いたいとき，そして，CHILDES のデータベースにデータを提供したいときには，ASCII 記号（つまりローマ字）を使うことが現在前提である。日本語をローマ字化するにあたっていくつかの問題が生じる。

まずヘボン式か訓令式にするかという問題がある。ヘボン式は発音に近くて外来語に強い。訓令式では動詞の変化などが説明しやすい。研究者の要求に合わせて選ぶことができるが，ヘッダーでどちらを使用したかを明記しなければならない。

ヘボン式と訓令式の間にできるだけの互換性を保つために，コーパス内では一貫して同じローマ字方式を使用する必要がある。たとえば地名や固有名詞を別の方式で文字化したり，外来語を別扱いにすることは避けなければいけない。

CLAN の CHSTRING プログラムとともにヘボン ‹–› 訓例 change. txt（中則夫　93/11）を使えば，お互いの書き方に換えることができる。change. txt にはヘボン ‹–› 訓令の変換だけでなく，仮名との交換も含まれている。入力の際に仮名のほうが便利な場合もある。

ワンポイント：MinJ–CHAT はヘボン式を基にして説明を進める。訓令の使用の場合に多少の差が生じるので，［4．訓令式］［作成中］にも参照。

3．2　J—CHAT 用ローマ字つづり方表（ヘボン式に基づく）

あ a	い i	う u	え e	お o
か ka	き ki	く ku	け ke	こ ko
さ sa	し si	す su	せ se	そ so
た ta	ち chi	つ tsu	て te	と to
な na	に ni	ぬ nu	ね ne	の no
は ha	ひ hi	ふ fu	へ he	ほ ho
ま ma	み mi	む mu	め me	も mo
や ya		ゆ yu		よ yo
ら ra	り ri	る ru	れ re	ろ ro
わ wa				を o

ん n

が ga	ぎ gi	ぐ gu	げ ge	ご go
ざ za	じ ji	ず zu	ぜ ze	ぞ zo
だ da	ぢ ji	づ zu	で de	ど do
ば ba	び bi	ぶ bu	べ be	ぼ bo
ぱ pa	ぴ pi	ぷ pu	ぺ pe	ぽ po

きゃ kya	きゅ kyu	きぇ kye	きょ kyo
しゃ sha	しゅ shu	しぇ she	しょ sho
ちゃ cha	ちゅ chu	ちぇ che	ちょ cho
にゃ nya	にゅ nyu	にぇ nye	にょ nyo
ひゃ hya	ひゅ hyu		ひょ hyo
みゃ mya	みゅ myu		みょ myo
りゃ rya	りゅ ryu		りょ ryo
ぎゃ gya	ぎゅ gyu		ぎょ gyo
じゃ ja	じゅ ju		じょ jo
ぢゃ ja			ぢょ jo
びゃ bya	びゅ byu		びょ byo
ぴゃ pya	ぴゅ pyu		ぴょ pyo

ア a	イ i	ウ u	エ e	オ o
カ ka	キ ki	ク ku	ケ ke	コ ko
サ sa	シ shi	ス su	セ se	ソ so
タ ta	チ chi	ツ tsu	テ te	ト to
ナ na	ニ ni	ヌ nu	ネ ne	ノ no
ハ ha	ヒ hi	フ fu	ヘ he	ホ ho
マ ma	ミ mi	ム mu	メ me	モ mo
ヤ ya		ユ yu		ヨ yo
ラ ra	リ ri	ル ru	レ re	ロ ro
ワ wa				
ン n				

ガ ga	ギ gi	グ gu	ゲ ge	ゴ go
ザ za	ジ ji	ズ zu	ゼ ze	ゾ zo

ダ da　ヂ ji　ヅ zu　デ de　ド do

バ ba　ビ bi　ブ bu　ベ be　ボ bo

パ pa　ピ pi　プ pu　ペ pe　ポ po

キャ kya		キュ kyu	キェ kye	キョ kyo
シャ sha		シュ shu	シェ she	ショ sho
チャ cha		チュ chu	チェ che	チョ cho
ニャ nya		ニュ nyu	ニェ nye	ニョ nyo
ヒャ hya		ヒュ hyu		ヒョ hyo
ミャ mya		ミュ myu		ミョ myo
リャ rya		リュ ryu		リョ ryo
ギャ gya		ギュ gyu		ギョ gyo
ジャ ja		ジュ ju		ジョ jo
ヂャ ja				ヂョ jo
ビャ bya		ビュ byu		ビョ byo
ピュ pya		ピュ pyu		ピョ pyo
	ウィ wi		ウェ we	ウォ wo
クァ kwa	クィ kwi		クェ kwe	クォ kwo
	スィ si			
ツァ tsa	ツィ tsi		ツェ tse	ツォ tso
		トゥ tu		
	ティ ti	テュ tyu		
ファ fa	フィ fi	フュ fyu	フェ fe	フォ fo
	ウィ wi		ウェ we	ウォ wo
				ジョ jo
	ズィ zi			
	ディ di	デュ dyu		
		ドゥ du		
	ヴィ vi			
			イェ ye	

長音：

かあ　カー　kaa

けい　ケー　kee

きい　キー　kii

こう　コー　koo

くう　クー　kuu

促音：

っか kka	っき kki	っく kku	っけ kke	っこ kko
っさ ssa	っし sshi	っす ssu	っせ sse	っそ sso
った tta	っち tchi	っつ ttsu	って tte	っと tto
っは hha	っひ hhi	っふ ffu	っへ hhe	っほ hho
っら rra				
っが gga	っぎ ggi	っぐ ggu		
	っじ jji	っず zzu	っぜ zze	
っだ dda			っで dde	っど ddo
	っび bbi	っぶ bbu	っべ bbe	
っぱ ppa	っぴ ppi	っぷ ppu	っぺ ppe	っぽ ppo

っきゃ kkya	っきゅ kkyu		っきょ kkyo
っしゃ ssha	っしゅ sshu	っしぇ sshe	っしょ ssho
っちゃ tcha	っちゅ tchu	っちぇ tche	っちょ tcho
っじゃ jja	っじゅ jju	っじぇ jje	
っぴゃ ppya			っぴょ ppyo

ッカ kka	ッキ kki	ック kku	ッケ kke	ッコ kko
ッサ ssa	ッシ sshi	ッス ssu	ッセ sse	ッソ sso
ッタ tta	ッチ tchi	ッツ ttsu	ッテ tte	ット tto
ッハ hha	ッヒ hhi	ッフ ffu	ッヘ hhe	ッホ hho
ッラ rra				
ッガ gga	ッギ ggi	ッグ ggu		
	ッジ jji	ッズ zzu	ッゼ zze	
ッダ dda			ッデ dde	ッド ddo
	ッビ bbi	ッブ bbu	ッベ bbe	
ッパ ppa	ッピ ppi	ップ ppu	ッペ ppe	ッポ ppo

ッキャ kkya	ッキュ kkyu		ッキョ kkyo
ッシャ ssha	ッシュ sshu	ッシェ sshe	ッショ ssho
ッチャ tcha	ッチュ tchu	ッチェ tche	ッチョ tcho
ッジャ jja	ッジュ jju		ッジョ jjo
	ッディ ddi		
ッピャ ppya			ッピョ ppyo

特別記号：

1．んン　母音音節の前は n'，例えば「まんいん」man'in

2．っッ　音節の終には Q，例えば「こらっ」koraQ

3．3　分かち書きのルール

ローマ字で入力する際，単語（自立語）をスペースで分ける。

*MOT: ja gareeji ni irete ne.

名詞，動詞などのほかに格助詞（wa ga no ni o mo），終助詞（yo，ka，zo etc）なども自立語として扱われる。

3．4　格助詞の書き方

格助詞は発音通りに書く：

は→wa

を→ o

へ→ e

3．5　同音意義語の表記

ローマ字化によって同音意義語の区別が付かなくなる場合，括弧を用いて同音意義語の表記をすることができる。

ame(sweets)

ame(rain)

同音意義語の表記は通常，テキストに対応する英語の単語を使う。単語と括弧の間にスペースを入れない。

助詞（よ，が，か，から，ね，に，の，を，と，は，よ）と混同する可能性のある同音意義語に関しては，必ず同音意義語の表記をする。この場合，助詞には同音意義語の表記をしない。

*ASU: yoochien e itte takusan e(picture) o kaita yo.

助詞と同音意義語の義務的表記の一覧表

e	へ
e(handle)	柄
e(picture)	絵
ga	が
ga(moth)	蛾
ka	か

ka(mosquito) 蚊
kara から
kara(empty) 空
kara(shell) 殻
ne ね
ne(price) 値
ne(root) 根
ni に
ni(two) 二
no の
no(field) 野
o を
o(tail) 尾
to と
to(door) 戸
wa は
wa(circle) 輪
yo よ
yo(night) 夜
yo(world) 世

ワンポイント：入力しているときは，同音意義語であることが意外に気が付きにくい。入力が終わってからまとめて探したほうが確実で効率がよい。上の助詞を replace コマンドで探し，必要なときだけに括弧の付いた形に書き換える。多数のファイルを同時に開くと便利。

3．6 コンマの使い方

呼びかけ，スクランブル，デイスロケーションの場合はコンマを使用する。

, 呼びかけ。コンマによって呼びかけを主語などの使用と区別する。

*ASU: Mama, kore mite ! 呼びかけ
*ASU: kore mite, Mama ! 呼びかけ
*ASU: Mama mita ? 主語

ただし，呼びかけだけの発音は %com 行を付ける。

*ASU: Mama !

%com: yobikake

また，呼びかけかどうかはっきりしない場合も %com を付ける。

*ASU: Mama yatte!

%com: yobikake ka shugo ka hakkiri shinai

„スクランブル（"scrambling" Saito 1985, "afterthought" Martin 1975）やデイスロケーションの場合はダブルコンマを使用する。

*ASU: mita„ kore ?

*ASU: kore tabeta„ kore ?

*ASU: kita„ Mama ga.

4．訓 令 式

作成中(未定)

5．文献，連絡先

文　　献：

B. MacWhinney, C. Snow 1990
The Child Language Data Exchange System: An Update
Journal of Child Language 17, 457–472.

寺尾康 1989
発話資料のデータベース化
言語1989/6, 122−124

マニュアル：

B. MacWhinney 1991
The CHILDES Project: Tools for Analyzing Talk.
Lawrence Erlbaum Assoc. ISBN 0–8058–1006–4 $29.95

CD–ROM 注文先：

Dr. Brian MacWhinney
Dept of Psychology, Carnegie Mellon University
Pittsburgh, 15213 U.S.A.
childes@andrew.cmu.edu or childes@andrew.bitnet

CHILDES プロジェクト連絡先：

Child Language Data Exchange System
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
e–mail: childes@andrew.cmu.edu
childes@andrew.bitnet
Brian MacWhinney: brian@andrew.cmu.edu

J–CHAT プロジェクト連絡先：

日本国外連絡先：Yuriko Oshima-Takane
Department of Psychology
McGill University
1205 Dr. Penfield Avenue
Montreal, PQ H3A 1B1 Canada
e-mail: yuriko@hebb.psych.mcgill.ca
日本国内連絡先：Susanne Miyata
Department of Communication
Aichi Shukutoku Junior College
23 Sakuragaoka, Chikusa-ku,
Nagoya, 464 Japan
e-mail: h44816g@nucc.cc.nagoya-u.ac.jp
PBC02454 (NIFTY)
JCHAT LIST 連絡先：Hiromi Morikawa-Paul
Department of Human Development and Family Life
University of Kansas
Lawrence, Kansas 66045 U.S.A.
e-mail: hiromi@kuhub.cc.ukans.edu