Third International Conference on Language Resources and Evaluation

A Way of Documenting Persian Data

Abstract

This paper investigates the documentation of Persian data by a computer program called CHILDES (Child Language Data Exchange System) designed by MacWhinney and Snow in 1985. The data collected from three Iranian children aged between 1;6 to 3;6 were transcribed. The transcribed data were analyzed syntactically and morphologically. The result indicated although a new format were applied for Persian data to meet the CHILDES standards, Persian like English can be analyazed by the above database program successfully.

Introduction:

The method for the analysis of the data collected from three Iranian children aged between 1;6 to 3;6 was a CHILDES format. The CHILDES (Child Language Data Exchange System) project, designed by MacWhinney and Snow in 1985, initially aimed to collect a nonstandardised database of computerized corpora of face to face conversational interactions. The system is designed for use with both normal and disordered populations. ‘Now’, as MacWhinny (1995:154) points out, ‘researchers have access to the results of nearly a hundred major research projects in over a dozen languages across the last 25 years’. This computational tool which will facilitate the sharing of transcript data allows researchers to enter the transcript data into computer files and analyze it by standard data-processing techniques. It has three major components: the database, the CHAT (codes for the Human Analysis of Transcripts) transcription systems and the CLAN (Computerized Language Analysis)programs (see MacWhinney, 1991). The CLAN programs are designed to perform a large number of automatic analyses on the transcript data that have been placed into the CHAT format. In this study the CHAT format for morphological analysis and syntactic analysis was employed. The transcriptions, which will be explained later, were analyzed in two tiers: %mor for morphological analysis and %syn for syntactic analysis. The main tiers are identified by symbol* while symbol% is used for the analytical tiers. This format is called a CHILDES analysis in this project since from time to time it was necessary to include a third tier for the phrasal analysis which is not anticipated in the CHAT transcription systems and furthermore, many new conventions and codes needed to be added to this system for Persian.

Transcription

The videotaped data from three Iranian children aged between 1;6-3;6 were collected for two years. The collected data were transcribed immediately after each session. The data was transcribed orthographically and occasionally phonetically. The vowels were transcribed phonetically since the vowel system of Persian is simple and may be summarized as follows: front high i, front mid-high e, front-mid a, back-high u, back mid-high o, mid back a: . The transcription for long ‘a’ throughout the study was decided to be a: and this convention was followed for transcribing the data into the CHILDES database. Winfuhr (1979) gives the following table for Persian consonants

Stops	fortis	p t ch k
	Lenis	b d j g
Fricatives	fortis	f s sh x
	Lenis	v z sh q
Nasals		m n
Liquids		l r
Glides		y h ?

The transcription for all consonants, except q (fricative, lenis), ? (glide) and x (fricative, fortis), was done orthographically. In addition, the data was transcribed in a way to meet the minimum set of standards for a CHAT (codes for the Human Analysis of Transcript) profile. In order that the CLAN (Computerized Language Analysis) programs run successfully on the transcribed data Macwhinney (1991:8,() established the following guidelines:

1. Every character in the file must be in the basic ACII character set.
2. Every line must end with a carriage return.
3. The first line in the file must be an Begin header line.
4. The last line in the file must be an End header line.
5. There must be an Participants header line listing three-letter codes for each participant, the participant’s name, and the participant’s role.
6. Lines beginning with * indicate what was actually said. These are called ”main lines”. Each main line should code one and only one utterance: When a speaker produces several utterances in a row, code each with a new main line.

7. After the asterisk on the main line comes a three-letter code in upper case letters for the participant who was the speaker of the utterance being coded. After the three letter code comes a colon and then a tab.
8. What was actually said is entered starting in the ninth column.
9. Lines beginning with the %symbol can contain anything. Typically, these lines include codes and commentary on what was said. They are called “dependent tier” lines.
10. Dependent tier lines begin with the %symbol. Then comes a three-letter code in lower case letters for the dependent tier type, such as “mor” for morphology and then a tab. The text of the dependent tier begins in the ninth column.
11. Continuations of main lines and dependent tier lines begin with a tab.

*a*Begin

*a*Participants: FAA Faeze Child, DAD father

*a*Date: 22-Jun-93

*a*Age of FAA: 2;8

*a*Filename: FAEZE, CHA

*a*Situation: free talk

*FAA: uno beza:r dige

%mor: pron|un-omarker|o be|vimp|za:r adv|dige

%syn: <XVY>.< XY+O: NP>.

[Pron Omarker]

*DAD: Xob bolandesh kon az un zir daresh beya:r

*FAA: ekast.

%mor: v|shekast&past_3s.

%syn: V

*a*End.

If the main line indicated the child’s actual speech, the target utterance was given orthographically on the morphological tier. For example, in the last main line of the above transcription the child said ekast instead of shekast so the target language was used and analysed on the morphological line. The conventions which are used on %mor and %syn lines are explained below.

Morphological and syntactic Coding and Analysis

The morphological and syntactic analysis of Persian data give a systemtatic and overall picture of the children's grammatical development. Moreover, many researchers of child language are interested in examining the role of universals in language acquisition through examining the role of universals in language acquisition through examining the syntactic development in children's corpora from different languages. MacWhinney (1991:95) suggested a system for morphological and syntactic coding for the corpora which is extremely detailed and will be employed fully in the future. MacWhinney (1991) suggested two ways of morphological coding: a) superficial morphological analysis can be done on the main line. B) %mor line should be used for a deeper morphological analysis. This study favoured approach b). However, some of the conventions that MacWhinney suggested were not included as they were not necessary for this study, e.g. the errors and omitted categories. The following conventions were employed for morphological coding and analysis:

1. Each word on the %mor line is separated by spaces to correspond to a space delimited word on the main line. However, the minor and vocative utterances and some categories, e.g. present perfect or reflex pronoun, on the morphological line did not correspond to a space limited word o the main line, e.g.

   *SHA: ba:ba: beya: 'daddy come'

   %mor: be-vimp|ya:

2. The coding on the %mor line ends in a full stop or a question mark.
3. The symbol | on the %mor line separates a morpheme from its grammatical definition, for example:

   FAA: ino.

   %mor: pron|in-omarker|o.

4. –hyphen is uses to indicate the attachement of an affix or an inflection to a stem, e.g.

   *SHA: nada:ri 'you don't have'

   %mor: neg|na-v|dar&pres-INF|i&2s.

5. The symbols (&), (-) are used to indicate the combined categories in a single morpheme, e.g.

   *FAA: koume? 'which one is it'

   %mor: q|kodum-cop|e&pres_3s.

The following morphological codings were used:

q	question
cop	copula
pres	present tense
past	past tense
pres perf	present perfect
past part	past participle
omarker	object marker
vimp/Vimp	imperative verb
n	noun
adj	adjective
adv	adverb
det	determiner
1s	first person singular
2s	second person singular
3s	third person singular
1pl	first person plural
2s	second person singular
3s	third person singular
1pl	first peron plural
2pl	second person plural
3pl	third person plural
neg	negative
prep	preposition
poss	possessive
reflex pron	refelexive pronoun
aux	auxiliary
INF	inflection
PP	prepositional phrase

The syntactic coding was done on the %syn tier. Clauses are either enclosed in single brackets followed by full stop or only ended in a full stop. The phrase structures are indicated in square brackets on the same line or the following line. Capital letters were used for the syntactic coding. The example below illustrates this:

*FAA: naqashi adam tush mikeshe

%mor: n|naqashi n|adam PP|tu&pron|sh&3s mi- v|kesh&pres-3s|e.

%syn: <CompVI>

[N VI]

The following grammatical conventions were employed:

I	Inflection
V	Verb
CompV	Compound verb
S	Subject
N	Noun
C	Complement
WHQ	WH Questions
X	any grammatical elements
A	Adverb
ADJ	Adjective
AUX	Auxiliary
Pr	Preposition
Pron	Pronoun
(E)	Contracted copula after complement
D	Determiner

Conclusion:

The brief explanation above shows that the transcribing and analysis of the collected data was extremely time consuming and labour intensive. However, This computer tool which will facilitate the sharing of transcript data allows researchers to enter the transcript data into computer files and analyze it by standard data-processing techniques. This program is not only useful in giving information about English data but also it can be successfully adapted to other languages, especially Persian.

References:

MacWhinney, B. and Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, **12**,271-96.

Mac Whinney.B. (1991) The CHILDES Project: Computatioan Tools for Analyzing Talk. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1995). Computational Analysis of Interaction. In P. Fletcher and B. MacWhinney (Eds). The handbook of Child Language, Oxford: Blackwell.

Windfuhr (1979). Persian Grammar History and State of the Study. The Hague: Mouton.