

A BRIEF INTRODUCTION TO THE *CHILDES* PROJECT

With Special Reference to Greek:

CHAT TRANSCRIPTION, LINKAGE, GRAMMATICAL CODING AND *CLAN* ANALYSIS¹

Ursula Stephany
University of Cologne

[December 23, 2010]

¹ I would like to thank Anastasia Christofidou, Dimitra Kati and Evangelia Thomadaki for their advice on the transcription of Greek and Anastasia Christopoulou for helping me with the application of Sonic Mode to Greek. Special thanks go to the more than 70 postgraduate students of the University of Athens who contributed to the coded Greek lexicon.

Table of Contents

1.	The CHILDES Project	3
2.	CHAT Transcription	3
2.1	MinCHAT	3
2.2	Form of Files	4
2.3	Files Headers	5
2.4	Main Line (Main Speaker Tier)	6
2.5	Some CHAT Conventions for the Main Line	7
2.6	Dependent Tiers	10
2.7	Transcription of Greek child data.....	10
3.	Linkage	10
3.1	Sonic Mode	11
4.	Coding	12
4.1	Lexicon-Based Automatic Morphological Coding of Transcripts ...	12
4.1.1	Introduction	12
4.1.2	How to Create a Unicode Version of Data Files and a Lexicon	14
4.1.3	How to Create a Language-Specific Lexicon	14
4.1.4	Generating the %mor Tier	16
4.1.5	How to Enlarge a Coded Lexicon and Code further Files ...	17
4.2	Coding Grammatical Errors and Self-Repairs	18
4.3	Syntactic Coding of Transcripts	18
5.	Overview of some CLAN Programs	19
6.	Analyzing Transcripts with the CLAN Programs	20
6.1	Introduction	20
6.2	COMBO	21
6.3	FREQ	22
6.4	KWAL	24
6.5	MLU	27
6.6	MODREP	27
6.7	How to Create Files to be Used in Search Strings	28
7.	Example of Transcribed and Coded Greek Data	28
	References	30
	Appendix I: Codes for Grammatical Morphemes	31
	Appendix II: Conventions for the Transcription of Greek	33

1. THE *CHILDES* PROJECT

The **CHILDES** [Child Language Data Exchange System] project consists of the following main parts:

CHILDES — | → **CHAT** [Codes for the Human Analysis of Transcripts]
 | → **CLAN** [Computerized Language Analysis]

CHILDES Database

CHILDES/BIB

CHAT (MacWhinney 2010, online manual)

CLAN (MacWhinney 2010, online manual)

For installing the CLAN programs on your computer go to the CHILDES site

<<http://childes.psy.cmu.edu/clang>>,

select the appropriate version of the programs (for MAC or PC) and install **clangwin.exe**.

The CLAN programs will only work properly (especially as far as Linkage and Automatic Coding are concerned), if you have a recent edition of the CLAN programs available on your computer. Before installing the most recent version of the CLAN programs on your computer, it is advisable to first **de-install** an eventual older version. In order to do this, open the Control Panel (Πίνακας ελέγχου), find Add/Remove programs, choose CLAN and remove it.

Addresses

Prof. Dr. Brian **MacWhinney**, Dept. of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA; email: <macw@cmu.edu>

The CHILDES Project: <<http://childes.psy.cmu.edu>>

Discussion of issues relating to child language learning:
 <info-childes@googlegroups.com>

Prof. Dr. Ursula **Stephany**, Institut für Linguistik, Allgemeine Sprachwissenschaft, Universität zu Köln, D 50923 Köln; e-mail <stephany@uni-koeln.de>

Working Papers of the Department of Linguistics of the University of Cologne:
<http://www.uni-koeln.de/phil-fak/ifl/asw/forschung/arbeitspapiere>

2. *CHAT* TRANSCRIPTION²

2.1. MinCHAT

1. You may write your file in the CED editor (Childes Editor) included in CLAN or in your common text editor. Files written in a common text editor must be saved as (unformatted) ASCII files (Text only...) and will carry the extension *.TXT. If you use

² This section is based on MacWhinney (2010, Part 1 referred to as 2010/I).

special characters, such as Greek γ δ θ , in your transcript, you must prepare a Unicode version of your file (see section 4.1.2 below). In order to be used with the CLAN programs, files must be saved in CED and should carry the extension *.cha (The lexicon file must carry the extension *.cut; see section 4.1.3).

2. In order for the CLAN programs to work properly, every file must begin with the @Begin line and end with the @End line. Files written in a common text editor (i.e. not in CED) must begin with the header @UTF8 (placed on the line before @Begin).
3. Every line must be ended by a carriage return (ENTER).
4. The line immediately following the @Begin line is @Languages (see section 2.3). The next line is @Participants: After a tab, this line lists three-letter codes for each participant (e.g. SPI) and the participant's name (e.g. Spiros) and role (e.g. Target_Child).
5. Further header lines indicate the participants' IDs (including the target child's age; e.g. 1;9.2), the media file (for oral data) (e.g. SPI-A-03), the target child's @Birth (e.g. 27-JUN-1972), the @Date of recording (e.g. 29-MAR-1974), the @Filename (e.g. SPI-A-03.cha), the @Situation, etc.³
6. The main tier contains what was actually said (eventually in a normalized form). Each main tier contains one utterance (or one proposition) and is introduced by the three-letter code for each speaker in capital letters preceded by '*' and followed by ':' and a tab (e.g. *SPI:→).
NB. In order to analyze several files by a common command it may be useful to code all target speakers by a common code (e.g. *CHI: or *NAR:).
7. Tiers dependent on the main tier are introduced by '%' followed by a three-letter code (in lower case), a colon, and a tab.
Examples: %mor: [morphological coding], %pho: [phonetic coding], %syn: [syntactic coding], %err: [errors], %com: [comments], %spa [speech acts].
Kinds and number of dependent tiers depend on the kind of data and on research questions.
(On the automatic insertion of the %mor: tier see section 4.1.4 below.)

Since transcription, coding and analysis by the CLAN programs are interdependent, users should learn the basics of CLAN and try out the results of a provisional transcription and coding sample before doing large-scale transcription and coding of their data.

2.2. Form of Files

The general form of files in CHAT format looks as follows:

```
@UTF8
@Begin
@Languages: ell
@Participants:      SPI Spiros Target_Child, MOT Mother of Target_Child
[ further headers ]
*SPI:  [spoken material]
%mor:  [morphosyntactic coding]
[ further dependent tiers, e.g. %com, %pho, %syn ]
```

³ For details see section 2.3 below and MacWhinney (2010/I:20-28).

*MOT: [spoken material]
 %mor: [morphosyntactic coding]
 [further dependent tiers]
 @End

Sample file coded in CHAT format:

```
@UTF84 [hidden header]
@Begin
@Languages: ell
@Participants: SPI Spiros Target_Child, MOT Mother of Target_Child,
               ULL Ursula Stephany Investigator
@Options:5
@ID:6 ell | stephany | SPI | 1;9.2 | male | | | Target_Child | |
@ID: ell | stephany | MOT | | | | lower middle | Target_Child's mother |
      primary school |
@ID: deu, ell | stephany | ULL | | | | Investigator | |
@Media: SPI-A-03
@Birth of SPI: 27-JUN-1972
@Date: 29-MAR-1974
@Interaction type: Looking at the picture book "Ich bin der kleine Bär"
                  (I am the little bear) and playing with toys at Spiros' home
@Location: Athens, Greece
@Time duration: 46 min.
@Filename: SPI-A-03.cha
@Transcriber: Ursula Stephany
@Warning: transcription has not yet been double-checked
*MOT: ti ine afto?
%com: referring to fairy tale book7
*SPI: mamisi [: paramiθi].
@End
```

2.3. File Headers⁸

Obligatory headers are @Begin, @Languages, @Participants, @ID and @End. @Begin and @End must be placed on the first and last line of the transcript, respectively, @Languages on the second line and @Participants on the third line. Eventually, the header @UTF8 precedes the @Begin header (see section 2.1 above).

Headers of the speaker tier (also called ‘main line’) as well as dependent tiers consist of three letters followed by a colon and one tab:

```
*MOT:      ela (e)ðo .
%mor:      V|erxome:IMP:PFV:2S ADV:LOC|eðo .
%com:      Mother addressing child
```

⁴ See MacWhinney (2010/I:20-21).

⁵ See MacWhinney (2010/I:23-24).

⁶ See section 2.3 below.

⁷ On the choice between @Comment: and %com: lines see section 2.5 and MacWhinney (2010/I:27).

⁸ See MacWhinney 2010/I:20-28).

The only headers not followed by a colon and a tab are @Begin and @End (and eventually @UTF8).

The **obligatory constant headers** @Languages, @Participants, @ID:

@Languages: The appropriate code for the language of the data must be chosen from the table presented in MacWhinney (2010/I:21). The present code for Greek is **ell** (MacWhinney 2010/I:21) (formerly, it was **gr** (MacWhinney 2009/I:25)).

@Participants are identified by a three-letter code each to be used on the main (speaker) line:

Example:

@Participants: SPI Spiros Target_Child, MOT Mother of Target_Child

There must be one **@ID**⁹ line for each participant. The general form of the @ID line is the following:

@ID: language | corpus | code | age | sex | group | SES | role | education |

Example:

@ID: ell | stephany | SPI | 1;9.2 | male | | Target_Child | |

@ID: ell | stephany | MOT | | | lower middle | Target_Child's mother |
primary school |

@ID: deu, ell | stephany | ULL | | | Investigator | |

As can be seen in the example, not all of the categories listed in the general form of the @ID line have to be specified for each participant. However, the respective positions in the general form must be indicated by bars separated by a space.

The target-child's age is computed with the help of the CLAN program **DATES** by indicating the child's birthday and the date of the recording.

Optional constant Headers,¹⁰ headers constant throughout the file, may contain useful information as shown in the example above. For **LINKAGE** (see section 3 below) to work properly the names of the media file and that of the transcript must exactly correspond to each other:

@Media: SPI-A-03.wav

@Filename: SPI-A-03.cha

2.4. Main Line (Main Speaker Tier)¹¹

The actual words spoken by the participants are transcribed on the speaker tier (or main line). Each main line begins with a three-letter code of the participant preceded by an asterisk and followed by a colon and one tab:

⁹ For details see MacWhinney (2010/I:23-24).

¹⁰ See MacWhinney (2010/I:25-26).

¹¹ See MacWhinney (2010/I:29-30).

*NAR: mia fora ke enan kero itan ena skilaki me onoma Balu .
 %eng: once upon a time there was a doggy called Balu .

For CHAT transcription lower case letters must be used, even for sentence-initial letters. Initial capital letters are only used for proper names. Phonetically deviant forms may be normalized on the main tier to a certain extent. Their phonetic form may be indicated on a %pho: dependent tier in cases of clearly recognizable strong deviance.¹² Another (probably preferable) possibility is to add the standard form in square brackets after the deviant form. In order to mark forms as deviant, “[*]” may be added after the corrected form so that nothing intervenes between the deviant and the standard form.¹³

Examples:

*SPI: musiki .
 %pho: mukiki
 or:
 SPI: mukiki [: musiki] [] .

Analyses with the CLAN programs give better results if main lines are kept as short as possible. They should therefore consist of a single utterance each. As is usual in typoscripts, words are separated by spaces. Every main line must end with a punctuation mark (see “Punctuation” in section 2.5).

2.5. Some CHAT Conventions for the Main Line (in alphabetical order)

Comments [% text]¹⁴

Comments consisting of a single word may be placed on the main line, longer comments should be placed on a separate %com tier. Comments referring to the entire transcript or to larger parts of it should be indicated by using the @Comment header placed in the appropriate position.

*NAR: eðo [% nu. 1] vlepo mia folia .
 or:
 *NAR: eðo vlepo mia folia .
 %com: narrator is looking at picture nu. 1

Deviant Forms¹⁵ [:] [*]

Deviant (ungrammatical) forms are followed by an asterisk in square brackets placed after the standard form given in square brackets. For the CLAN programs to function properly, the standard form must thus immediately follow the deviant form (see section 2.4 above).

In cases of (morpho)phonologically deviant forms and non-existing forms (e.g. *krionome*), the target form is added in square brackets after a colon and a blank. Lexical errors as well as errors concerning the choice of function words or the function of grammatical forms (e.g.

¹² On morphophonologically deviant forms see section 2.5 below.

¹³ Marking many forms by “[*]” will result in listing a huge number of heterogeneous forms by the CLAN programs. One way to avoid this is by distinguishing errors and marking them as e.g. “[p*]” (phonological-phonetic error) or “[g*]” (grammatical error). Another possibility is to retrieve deviant forms by specifying certain characteristics of the deviant or the standard form in a Combo command (see section 6.2).

¹⁴ See MacWhinney (2010/I:58).

¹⁵ See MacWhinney (2010/I:62-63).

wrong case use, tense errors, agreement errors) are marked by an asterisk but without indication of the target form(s). Errors such as these may be marked and corrected (by hand!) on the grammatical coding tier %mor (see section 4.2 below).

Examples:

NAR: o skilos irθe ke epiane [] tin ura tis γatas .
 NAR: krionome [: kriono] [].

NB. Indication of the target form in square brackets (e.g. [: kriono]) is important when it comes to automatic coding of the speaker's utterances, since the Clan program **MOR** will only code the (standard) forms given in brackets and not the deviant forms actually said. Thus, in the above examples, the forms *epiane* and *kriono*, but not *krionome*, will be grammatically coded.

In order to list forms marked by [*] by the search program **KWAL** (or **COMBO**) the option +s"[*]" must be used (see section 6.4 below).

Explanation¹⁶ [= text]

Brief explanations of the situation at hand may be given on the main line, longer ones on the %com tier.

*CHI: afto [= kivos] θelo .

Irrelevant Material www.¹⁷

Stretches of speech of adult interlocutors which are irrelevant to the dialogue need not be transcribed. This may be the case if at some point of the session other people enter the room and speak among themselves or if someone answers the phone.

*MOT: www.
 %exp: talks to neighbor on the telephone

NB. When the transcript is linked to the sound file, such stretches must be taken into consideration by an appropriate handling of bullets (see section 3 below).

Incomplete and omitted Words (...), 0PTL¹⁸

The missing parts of incomplete words are added in parentheses if the target form is clear.

*CHI: ela (e)δο .
 CHI: o (S)pi(r)o(s) [] to θeli .

Omitted words may only be added if they can be clearly guessed, such as grammatically obligatory forms. Omitted words should not be enclosed in parentheses but marked by "0" followed by an indication of their respective parts of speech.

*CHI: 0ART¹⁹ likos ine .

¹⁶ See MacWhinney (2010/I:57).

¹⁷ MacWhinney (2010/I:33).

¹⁸ MacWhinney (2010/I:34-35).

¹⁹ If necessary, the code may also be elaborated (e.g. 0ART:DEF:MASC:NOM:SG).

When coding files grammatically, the program MOR will code omitted words transcribed in this way by “?|0ART” so that they can be searched on the main tier as well as on the %mor tier by the option +s”*0*”. The CLAN programs FREQ or MLU disregard forms preceded by zero on the main tier by default.

Pauses²⁰ (.),

Short unfilled pauses are marked by (.) and longer ones by (..) or even (...). Unfilled pauses may also be marked by #, ##, or ###.

Filled Pauses (fp) may be transcribed by 'eh@fp' (or 'uh@fp', 'um@fp').

*NAR: afto ine ena (.) eh@fp psilo ðedro .

Phonological Fragment &²¹

The ampersand symbol '&' is used at the beginning of false starts or nonsense forms.

*CHI: &pro portofoli .

Proper Nouns and Titles

Proper names of persons and places as well as titles are transcribed with an initial capital letter. Multiple-word names should be joined by '_' in order to output them as single words in the analysis (*o Ajios_Nikolaos*). NB. Compounds are marked by '+' (e.g. *kokino+skufitsa*).

Punctuation²²

The default punctuation set for the CLAN programs consists of the following characters:

, . ; ? ! [] < >

The end of every speaker line has to be marked by a full stop, a question mark or an exclamation mark. The comma is reserved for syntactic boundaries between clauses. It is thus usually not placed at the end of the speaker line.

Repetition and Retracing [/], [//]²³

Repetition without correction is marked by [/], repetition with correction (retracing) is marked by [//]. If several words are repeated, they are placed in angle brackets:

*CHI: θelo portokali [/] portokali me zaxari .

*CHI: θelo ena milo [//] portokali .

*CHI: θelo <ena milo> [//] ena portokali .

Scoped Symbols < > []²⁴

When symbols placed in square brackets ('[]') refer to more than the immediately preceding word, the material they relate to must be surrounded by angle brackets ('< >'):

*CHI: θelo <ena milo> [//] ena portokali .

²⁰ MacWhinney (2010/I:51, 52, 83).

²¹ MacWhinney (2010/I:34, 71).

²² MacWhinney (2010/I:30, 49-50).

²³ MacWhinney (2010/I:53).

²⁴ MacWhinney (2010/I:58).

Unintelligible Speech xxx²⁵

Unintelligible words, parts of utterances or whole utterances are transcribed by 'xxx'.

2.6. Dependent Tiers²⁶

With the exception of the %mor tier, the lines of the dependent tiers (headed by %) do not have utterance delimiters (do not end in a punctuation mark). Useful dependent tiers are the following:

%com:

This is a general-purpose line for longer comments of all kinds.

%pho:²⁷

A phonetic-phonemic rendering of material, especially when deviating from standard pronunciation, may be given on this tier. NB. Once the transcript has been linked to the sound file, this line will be less important except for phonetic-phonological analyses.

%syn:²⁸

This line is useful for grammatical codings which are not part of the %mor line, e.g. functional categories such as subject, object and word order (see section 4.3 below).

2.7 Transcription of Greek Child Data

A sample transcript is presented in section 7 and conventions for transcribing Greek are given in Appendix II.

3. LINKAGE²⁹

It is nowadays possible “to link specific segments of the digitized audio or video to segments of the computerized transcript” (MacWhinney 2010/II:21). This can be done by **Sonic Mode** or **Transcriber Mode** within the Childes Project. Other possibilities are to use **ELAN** (Max Planck Institute for Psycholinguistics, Nijmegen, NL <<http://www.mpi.nl/elan>>) or the **Tool Box** software of the Summer Institute of Linguistics (<<http://www.sil.org/computing/toolbox/information.htm>>).³⁰

If the transcription (called “multimedia annotation” in ELAN) is done by using ELAN, the result will have to be converted to CHAT for computer-assisted analysis, since ELAN does not supply software for analyzing transcripts.³¹

In order to reach a decision on whether to use ELAN rather than the facilities of the CHILDES Project for transcription, it should be noted that in ELAN it is the sound or video

²⁵ MacWhinney (2010/I:16, 33).

²⁶ For further details see MacWhinney (2010/I:65-70).

²⁷ MacWhinney (2010/I:33, 68).

²⁸ For further details see MacWhinney (2009/I:82). The %syn dependent tier is no longer mentioned in the 2010 edition of the manual.

²⁹ On linkage see MacWhinney (2010/II: chapter 4).

³⁰ In contrast to ELAN, Toolbox also incorporates programs for computer-assisted linguistic analysis.

³¹ For details on converting transcripts done with ELAN to CHAT see MacWhinney (2010/II).

data which forms the basis of the program so that the annotation is secondary and is aligned to the oscillogram of the sound. It follows from this that it is hardly possible (or at least very cumbersome and time-consuming) to use ELAN for elaborating existing transcriptions, e.g. those originally done in CHILDES. Furthermore, ELAN does not show the full contributions of two or more speakers participating in a dialogue on the screen. Rather, the consecutive contributions of a single speaker at a time (e.g. Child or Mother, but not both) are presented on the screen, something which is fine for narrations but rather inadequate for dialogues (especially those between young children and their mothers). Although the contributions of both (or more) speakers participating in a dialogue can be seen on the lower part of the screen (below the oscillogram), their stretches of speech may not be fully visible, since the transcription of a piece of text usually takes up more space than the corresponding part of the oscillogram. ELAN may, however, be advantageous when transcribing videos and gestures rather than audio files, although the CHILDES Project also provides a **Video Mode** (see MacWhinney 2010/II).

Within the Childe Project, linking can be done in two ways: either by **Sonic Mode** (MacWhinney 2010/II:22) or by **Transcriber Mode** (MacWhinney 2010/II:22-24). The first of these guarantees a more accurate alignment of the transcript with the sound tier, while the second one is faster, but often less precise (MacWhinney 2010/II:22). In order to use either of these modes **QuickTime 7** (or above) must be installed on the computer (<<http://www.apple.com/quicktime/download>>).

3.1. Sonic Mode³²

Sonic Mode accepts audio files in either **.wav** or **.mp3 format** (MacWhinney 2010/II:21). The sound file to be worked on must be available on the hard disk rather than simply on CD. Furthermore, the .wav file (or .mp3 file) and the .cha file must carry the same name and will differ only by the extension .wav vs. .cha.

Open CLAN and begin to establish a cha file by typing in the headers (or open a cha file which has been prepared previously and which shall be linked to the sound tier presently). Go to **Mode** in the task line on top of your screen and choose **Sonic Mode**. The wav file will open automatically (if the @Media Header containing the name of the wav file and the cha file containing the @Filename correspond to each other!).³³

Drag the cursor over the first segment of the wave form in order to highlight it. When you release the mouse, the segment will play. Roughly transcribe what you have just heard. In order to listen to the same segment of the sound tier again (as often as needed), hold down the Shift key and click the left key of your mouse (**Shift+click**). In order for this to work, the corresponding section of the sound tier has to stay highlighted.

Once you have corrected your original transcription of an utterance place a **bullet** at the end of the utterance by clicking on the “s” button to the left of the waveform. The corresponding section of the oscillogram has to stay highlighted for this to be possible. NB. Instead of clicking on the “s” button, you may also use **Esc+I** (insert time code).

The “bullet contains information regarding the exact onset and offset of the highlighted section” (MacWhinney 2010/II:22). This information is normally hidden. In order to expand

³² See MacWhinney (2010/II:22).

³³ If Sonic Mode should ask for a CD in spite of the fact that the sound file has been copied to the hard disk, simply ignore this.

the bullets, type *ESC-A*. Retyping this will again hide the information. A bullet wrongly set may be removed by using the Back-Space key.

Proceed by highlighting the next segment of the oscillogram and listen to it in order to produce a first rough transcription. Continue as indicated above.

You can move the oscillogram by using the arrow on the right-hand bottom side of the screen rather than the block on the scroll-bar, since the latter may move the oscillogram too quickly. MacWhinney (2010/II:22) notes that “scrolling in the sound file can take some time as the sound files for long recordings are very large and take up processing capacity.”

There are two ways³⁴ of listening to a certain utterance in a transcript in which bullets have been placed or of continuing to work on a file in which bullets have been set up to a certain point and further bullets shall be added (e.g. the next day):

- (1) - Place the cursor to the right of the bullet in question.
 - Press F5. This will mark the corresponding section of the oscillogram and the sound will play.
 - In order to stop playing the sound, click the left mouse key.

or

- (2) - Place the cursor immediately to the left of the bullet in question and triple-click in order to mark the corresponding section of the oscillogram.
 - Press F5 and the sound will play.
 - In order to stop playing the sound, click the left mouse key.

4. CODING

4.1. Lexicon-Based Automatic Morphological Coding of Transcripts³⁵

4.1.1. Introduction

The automatic coding system presented here is based on the CLAN program MOR (see MacWhinney 2010/II) as it has been extended to languages with richer morphologies by Steven Gillis (see **MinMOR** in CHILDES). While automatic coding systems such as those devised for English by MacWhinney and for Dutch by Gillis derive morphologically complex forms from their base by rules, the system created for languages with richer morphologies like Greek by Gillis only have a rudimentary rule component and mainly rely on a lexicon in which both inflectional forms and derivations are listed.

MinMOR contains 3 basic files with the extension **.cut** (**ar.cut**, **cr.cut**, and **lex.cut**) which can be used for coding data from any language. These files are required by MOR in order for it to work properly. The lexicon file **lex.cut** must be enlarged according to the data you want to analyze (see section 4.1.3 below).

³⁴ Both of these work with the version of the Clan programs issued on March 17, 2010 as well as more recent ones.

³⁵ The lexicon-based automatic coding system based on the CLAN program MOR was originally devised by Steven Gillis (in collaboration with Gert Durieux) at the Department of Germanic Languages of the University of Antwerp, Belgium, for the morphological coding of languages with a richer morphology than English, such as Greek, for which no rule-based automatic coding system is yet available.

The files **ar.cut** and **cr.cut** must be placed in the **lib** subdirectory. The lexicon file (e.g. **GREEKLEXStephany.cut**) must be placed in a **lex** subdirectory to be created within the **lib** subdirectory. The Childes directory with its subdirectories may look like this:³⁶

```

CHILDES
  CLAN
    Data
      German
        GermanL2
      Greek
        GreekL1
        GreekL2
        GreekNarratives
    lib
      [preinstalled folders/files]
      ar.cut
      cr.cut
      lex

```

This is what you lib directories may look like when working with different languages, e.g. German and Greek:

```

lib-deu
  [preinstalled folders/files]
  ar.cut
  cr.cut
  lex
  DaZAF_LEX_2000.cut
lib-ell
  [preinstalled folders/files]
  ar.cut
  cr.cut
  lex
  GREEKLEXStephany.cut

```

When using the Commands Window and operating the CLAN programs, make sure that the different subdirectories in the Commands Window are correctly set. In order to set the directories, press the button 'working' (or 'lib'), locate and mark the desired directory and press the 'Select directory' button. If the Output subdirectory is left unspecified, the output of the Clan commands will be placed in the Working subdirectory. For working with Greek data, the subdirectories in the Commands Window may be set like this:

working	C:\childes\clan\data\GreekL1
output	
lib	C:\childes\clan\lib-ell
mor lib	C:\childes\clan\lib-ell

³⁶ When working with different types of data or languages for which different coded lexicons are needed (e.g. Greek, German etc.), a separate **lib** directory with a different **lex** subdirectory may be created for each of these. These directories may be called **lib-deu** for German or **lib-ell** for Greek.

4.1.2. How to Create a Unicode Version of Data Files and a Lexicon

Since the CLAN programs, including MOR, only run on Unicode files, you must use an unformatted version of your transcripts (carrying the extension .cha) for automatic coding as well as data analysis.

If you use the CHILDES editor CED to create your files or extend the lexicon, a Unicode version will result automatically. Otherwise, add the initial header **@UTF8** to your files.

If you use the word processing program WORD to establish your files or extend (or establish) a coded lexicon, it will be more comfortable to save these as word documents while you go along. However, these files must in the end be saved as **text only** files. But if you have used non-ASCII characters (e.g. Greek characters) in these files, WORD will warn you that these will not be represented properly in the text-only version of the file. In order to avoid this and preserve the foreign characters correctly, proceed as follows:

Save as...	text only
Converting the file to	check Other (rather than Windows or MS-DOS)
Select and mark the line	Unicode (UTF-8) (in the list on the right-hand side of your screen)
Save your file	

How to convert a .txt file into a .cut file: Incorporate the resulting text file of e.g. the lexicon **GREEKLEXStephany.txt** into the **lex** subdirectory within **lib-ell** (or just **lib**, if you only work with Greek). Open the file **GREEKLEXStephany.txt** within **CLAN** and **Save as... GREEKLEXStephany.cut** typing in the extension “cut” by hand. Remove the original **GREEKLEXStephany.txt** from the **lib-ell** subdirectory since only the cut version will be needed for coding your data.

Proceed in a similar way with any data file (transcript) you may write in WORD rather than the CED editor in order to convert a .txt file into a **.cha file**.

4.1.3. How to Create a Language-Specific Lexicon

In order to run MOR³⁷ on your first transcript you need a rudimentary Lexicon, such as **greeklex.cut** containing at least one complete coded entrance. For Greek, this entry could read as follows:

Xristos {[scat N:PROP]} "Xristos:MASC:NOM:SG"

Nowadays you do not have to start from scratch for Greek, since the lexicon **GREEKLEXStephany** (as of Sept. 2010) comprising nearly 9,500 entries will be provided. Still, your transcript will most likely contain forms not represented in this lexicon so that you have to extend it in order to fully code your data.

Use the following command to create a lexicon of all (as yet uncoded) word forms found on all speakers' tiers of the file you want to code morphologically (e.g. the file SPI-A-03.cha):

MOR +xl @³⁸

³⁷ See section 4.1.4 below.

The above command will result in a file called **SPI-A-03.ulx.cex**. It contains all word forms occurring on all speakers' tiers in the left-hand column and {[scat ?]} in the second column; e.g.

```
fevji  {[scat ?]}
γata  {[scat ?]}
pulakia {[scat ?]}
puli   {[scat ?]}
```

All entries found in this file have to be coded by hand using your usual text editor (or the CED editor) and providing for all possible grammatical interpretations of each form:

1. Add the appropriate s[yntactic]cat[egory] replacing the question mark by the major part of speech of the grammatical word form found in the left-hand column (e.g. {[scat N]}). Stick to the grammatical codes provided in Appendix I, which are based on MacWhinney (1995:113ff., 2000/I:167ff., part 2) as far as possible. You may add subclasses to the major parts of speech separating them by a colon (e.g. {[scat N:PROP]}).
2. Place a tabulator after the right-hand brace and enter the grammatical coding of the specific word form of the first column enclosed in quotation marks.¹ If a word form is grammatically ambiguous, add a new line for each grammatical interpretation of the form or use slashes (e.g. NOM/ACC).

¹ The quotation marks used in the lexicon must not be the ones usually provided by WORD, thus not "xxx", but "xxx".

After all forms have been coded, your file will contain three columns:

```
fevji  {[scat V]}      "fevγo:IPFV:NONPAST:3S"
γata   {[scat N]}      "γata:FEM:NOM/ACC:SG"
pulakia {[scat N]}      "puli:DIM:NEUT:NOM/ACC:PL"
puli   {[scat N]}      "puli:NEUT:NOM/ACC:SG"
```

NB. Unknown or undecidable parts of speech may be coded as "unknown": {[scat unknown]}, but it is preferable not to enter such non-standard forms into the coded lexicon, except if there are special reasons for doing this.

Such a small coded file based on a small amount of data of a language you want to study (e.g. Kabiye, a Gur language spoken in Northern Togo, West Africa), may represent the beginning of a coded lexicon for the language in question.³⁹

In the case of Greek, integrate the coded .ulx.cex file into the Greek lexicon file GREEKLEXStephany.cut.

³⁸ If you only want to code the child's (e.g. Spiros') utterances, add +t*SPI to this command.

³⁹ If the transcription of a language such as Kabiye requires special symbols, you may select these from Lucida Sans Unicode and set the Font and Commands Font in Clan accordingly so that these symbols will be supported by the Clan programs. In order to set the fonts select "Set Font" and afterwards "Set Commands Font" in the pull-down menu View in Clan and change the Arial Unicode MS to Lucida Sans Unicode. In order to change the default status to the original one, choose "Arial Unicode MS" (not "@Arial Unicode MS").

Here are some examples taken from the lexicon **GREEKLEXStephany.cut**:

```

afini  {[scat V]}      "afino:IPFV:NONPAST:3S"
afisi  {[scat V]}      "afino:PFV:NONPAST:3S"
afti   {[scat PRO:DEM]} "aftos:FEM:NOM/ACC:SG"
afti   {[scat PRO:DEM]} "aftos:MASC:NOM:PL"
afto   {[scat PRO:DEM]} "aftos:NEUT:NOM/ACC:SG"
akomi  {[scat ADV]}    "akomi"
akrivos {[scat ADV]}   "akrivos"
alo     {[scat PRO:INDEF]} "alos:NEUT:NOM/ACC:SG"
ala     {[scat CONJ:COO]} "ala"
aneveni {[scat V]}     "aneveno:IPFV:NONPAST:3S"
anevi   {[scat V]}     "aneveno:PFV:NONPAST:3S"
anevike {[scat V]}     "aneveno:PFV:PAST:3S"
apo     {[scat PREP]}   "apo"
arpaksi {[scat V]}     "arpazo:PFV:NONPAST:3S"
arpazi  {[scat V]}     "arpazo:IPFV:NONPAST:3S"
arxi    {[scat N]}     "arxi:FEM:NOM/ACC:SG"
arxizi  {[scat V]}     "arxizo:IPFV:NONPAST:3S"
arxizun {[scat V]}     "arxizo:IPFV:NONPAST:3P"
as       {[scat PTL]}   "as"
avya    {[scat N]}     "avyo:NEUT:NOM/ACC:PL"
berðevome {[scat V]}   "berðevo:MP:IPFV:NONPAST:1S"
birdaki {[scat N]}     "bird@e:DIM:NEUT:NOM/ACC:SG"
bori    {[scat V:MDL]} "boro:IPFV:NONPAST:3S"
ðagoni  {[scat V]}     "ðagono:IPFV:NONPAST:3S"
ðagose  {[scat V]}     "ðagono:PFV:PAST:3S"

```

4.1.4. Generating the %mor Tier

Once all (standard) forms contained in the .cha file(s) you want to code (e.g. SPI-A-03.cha) occur in your current Lexicon, you are ready to code your transcript morphologically. The following command will generate the %mor: tier and add this line to each main line (speaker tier) in your transcript:

MOR @⁴⁰

The result of this command will be a file with the extension .mor.cex (e.g. SPI-A-03.mor.cex).⁴¹

For the sentence "o skilos iðe tin γata" (in a file taken from Hickmann's picture story *The cat*) the %mor tier will look like this:

```

*NAR:      o skilos iðe tin γata .
%mor:      ART:DEF|MASC:NOM:SG N|skilos:MASC:NOM:SG V|vlepo:PFV:PAST:3S
           ART:DEF|FEM:ACC:SG N|γata:FEM:NOM/ACC:SG .

```

⁴⁰ If you only want to code the child's (e.g. Spiros') utterances, add +t*SPI to this command.

⁴¹ If you run MOR on the same .cha file a second time, the preceding .mor.cex file will be overwritten and disappear. In case you want to preserve it, rename it before running MOR a second time.

Note that the codings placed within braces and square brackets in the lexicon appear in front of the vertical bar on the %mor tier, whereas the given lexeme together with the coding of its grammatical form occurring in a specific context are placed after the vertical bar.

If a word form is associated with two or more codings in the Lexicon (e.g. *afti*), all alternatives will be provided in the coded transcript separated by "^". The file has to be disambiguated by hand. A convenient way to disambiguate multiple codings is to use the **Disambiguator Mode** of the Childes editor CED. After opening the .mor.cex file you want to disambiguate, select the **Disambiguator tier** in the **Mode** pulldown menu.

The program will automatically mark the first instance of multiple codings in your file and display the alternatives at the bottom of the screen. Double-clicking on the adequate alternative will erase the inadequate coding(s) on the %mor tier and make the program go to the next instance of an ambiguous form.⁴²

4.1.5. How to Enlarge a Coded Lexicon and Code further Files

In order to check new files for forms not yet included in your coded lexicon, use the same command as in section 4.1.3 above, but be sure to work with the elaborated version of the lexicon into which the new codings of the first file you have worked with have been integrated:

MOR +xl @

This command will create a file containing only those word forms which are not yet contained in the elaborated version your lexicon.

In order to enlarge your coded lexicon, it is not necessary to proceed file by file, but you can run the above command on several files at once (by placing all of them into the FILE IN window).

Unite the output file of the above command (.ulx.cex) with your Lexicon, order all entries alphabetically, and code the as yet uncoded new entries.

A convenient way to incorporate new entries into your lexicon is the following: Open your lexicon in WORD and place the file with the new entries after the last entry of the existing lexicon. Mark all new entries in yellow. Order all entries of the file (lexicon plus new entries) alphabetically. Now the new entries will appear at the correct places and will stick out because they are marked in yellow. In order to code the new entries copy as much as possible of neighboring entries in order to go more quickly and avoid typing errors. Example:

```
afini  {[scat V]}    "afino:IPFV:NONPAST:3S"
afisi  {[scat V]}    "afino:PFV:NONPAST:3S"
afiso
afta
afti   {[scat PRO:DEM]} "aftos:FEM:NOM/ACC:SG"
afto   {[scat PRO:DEM]} "aftos:NEUT:NOM/ACC:SG"
```

For coding *afiso* copy {[scat V]} "afino:PFV:NONPAST:3S" and replace "3" by "1".

⁴² See MacWhinney (2010/II:127, 138-139).

For coding *afta* copy {[scat PRO:DEM]} "aftos:NEUT:NOM/ACC:SG" and replace "SG" by "PL".

After completing the coding of your extended lexicon make sure to convert it into a **cut** file (see 4.1.2). Use your enlarged Lexicon to generate a %mor: tier in some file(s) by the command from section 4.1.4 repeated here for convenience: **MOR @**

4.2. Coding Grammatical Errors and Self-Repairs

In order to do a detailed analysis of errors occurring in child language or learner languages, it may be useful to distinguish error types in the transcript. Here are some suggestions for coding errors and self-repairs on the %mor tier (by hand!):

Morphophonemic errors (#)

CHI: su dono [: dino] [] ena peynidi .
%mor: ... V|dino:IPFV#:NONPAST:1S ...

Wrong use of grammatical categories ()*

CHI: krionome [: kriono] [].
%mor: ... V|kriono:IPFV:PASS*=ACT:NONPAST:1S ...

Successful self-repair (\$)

CHI: su dono [: dino] [] [/] dino ena peynidi .
%mor: ... V|dino:IPFV\$:NONPAST:1S ...

or

CHI: su dono [: dino] [] [/] dino ena peynidi .
%mor: ... V|dino:IPFV#\$:NONPAST:1S ...

Unsuccessful self-repair (%)

CHI: etsi lene [/] lenun [: lene] [].
%mor: ... V|leo:IPFV:NONPAST:3P% ...

Another way of distinguishing between different types of errors is to add specifications to [*] on the Main line, e.g. [mp*] for morphophonemic errors, [gr*] for wrong use of grammatical categories. Warning: Try such markings out on a small file and use the Clan programs Kwal or Combo to list them!

4.3. Syntactic Coding of Transcripts⁴³

For syntactic coding a dependent tier %syn may be added to the transcript below the %mor line. This has to be done by hand.

Here are some codes which may be used on the %syn line:

ADJ	adjective
ADV	adverbial
ADV:LOC	locative adverbial

⁴³ On syntactic analysis with CHILDES also see MacWhinney (2008) and MacWhinney (2010/II: ch. 11) on GRASP.

ADV:TEMP	temporal adverbial
C:TEMP	temporal clause
CONJ:COND	conditional conjunction
CONJ:COO	coordinating conjunction
CONJ:SUBOR	subordinating conjunction
CONJ:TEMP	temporal conjunction
CONJ:TEMP*	wrongly used temporal conjunction
DO	direct object
IO	indirect object
LOC	locative adverbial
NEG	negative
PP	prepositional phrase
PP*	wrongly used prepositional phrase
PRED:ADJ	predicative adjective
PRED:N	predicative noun
Q	question word
S	subject
S0*	wrongly omitted subject (German)
TEMP	temporal adverbial
V2	verb second rule observed in main clause (German)
V2*	verb second rule not observed (German)
OV0	verb wrongly omitted in main clause
V2:AUX	auxiliary correctly placed in second position (German)
V:PP	past participle of verb
VF	verb final position observed in subordinate clause (German)
VF*	verb final position not observed (German)
VF*:AUX	auxiliary not in final position in subordinate clause (German)
VF*:INF	infinitive not in final position (German)
OVF	verb omitted in final position (German)
MC	main clause

NB. For languages with a rich inflectional morphology, such as Greek, syntactic coding can most often be avoided by making a clever use of Clan commands operating on the %mor tier (see section 6 below). Also, some of these codings may be integrated into the Maine line (e.g. “0S” for a wrongly omitted subject).

5. OVERVIEW OF SOME CLAN PROGRAMS

DATES	Takes two time values and computes the third (e.g. computes age of child/learner on the basis of date of birth and date of the interview).
FREQ	Frequency analysis and type/token ratio. ⁴⁴ Examples of application: alphabetical list of words (or morphemes) indicating frequency of each word form (morpheme); frequency of grammatical categories. Reverse dictionary.
COMBO	Finds combinations of keywords and lists all examples comprising a given keyword or combination of keywords. Examples of application: inflectional and derivational morphology, word order, discontinuous morphemes, questions, negations; picture stories, experimental data.

⁴⁴ Warning: Since the type/token ratio is dependent on corpus size, the ratios indicated by FREQ are unreliable.

KWAL	Finds words (grammatical forms, lexemes, grammatical categories) and lists all examples comprising a given keyword. Examples of application: morphological and lexical analysis.
MOR	Provides automatic morphological coding by generating a %mor: tier for all (or selected) speaker tiers.
MLU	Computes mean length of utterance and number of utterances. Examples of application: assessment in first and second language acquisition, language impairment, aphasia.
MODREP	The Model-and-Replica Analysis matches words on a "model" tier with words on a "replica" tier. Examples of application: phonetic (or graphic) variation in language acquisition, language impairment, aphasia, unimpaired speech.
CHECK	Verifies if transcripts correspond to CHAT conventions. ⁴⁵

6. ANALYZING TRANSCRIPTS WITH THE CLAN PROGRAMS

6.1. Introduction

Before starting the analysis, the directory in which the files to be analyzed are located should be set as the Working directory. 'Lib' must be set to the directory in which the files for automatic morphological coding are located (see section 4.1.1 above).

After setting the directories proceed as follows:

1. Press the button 'CLAN' in the Commands window and select the appropriate program (or type its name in the Command window).
2. Type in the desired options.
3. Select the cha file (or files) to be analyzed by pressing the button 'FILE IN'. Locate the file, mark it and double-click so that it will appear in the right-hand window.
4. Press 'Done' in the FILE IN window and subsequently 'Run' in the Commands window.

The Options available with each CLAN program appear on the screen if you type the name of the program (e.g. **FREQ**, **KWAL**) in the Commands Window and then press Enter.

The following sections list some useful commands concerning the CLAN programs **COMBO**, **FREQ**, **KWAL**, **MLU**, and **MODREP**.⁴⁶ Since **Combo** is a very complex program, the reader should start with **Freq**, proceed to **Kwal** and only then try **Combo**.

If you are not sure in which way a certain command operates on your data and how to interpret its results, a good method is to create a very small test file and operate the command on this file so that you can exactly see what the respective CLAN program lists or calculates. An example of such a file is presented at the end of section 6.4 below.

⁴⁵ For details see MacWhinney (2010/II:48-51).

⁴⁶ For further details on the Clan programs see MacWhinney (2010/II).

6.2. COMBO

combo +s"*o^[*:s]" +k @

Lists all utterances containing a word ending in /o/ as well as the corrected form ending in /s/, which immediately follows it and is placed in square brackets (+k treats upper and lower case letters as different.)

combo +t*SPI +t%mor +s"V:COP*^ADJ*" +k @

Lists all utterances containing a copula immediately followed by an adjective in Spiros' data. If omitted copulas have been transcribed by "0V:COP" on the main tier and are therefore coded by "?|0V:COP" on the %mor tier, this command will only find non-omitted copulas. In order to retrieve both omitted and non-omitted ones, the search string should be +s"*V:COP*^ADJ*".

combo +t*SPI +t%mor +s"V:COP*^ADJ*" +x +k @

Lists all utterances containing a copula immediately preceding or following an adjective (+x option).

combo +t*SPI +t%mor +s"V:COP*^^ADJ*" +k @

Lists all of Spiros' utterances containing a copula immediately or eventually (^*^) followed by an adjective (e.g. "ine pio meyaló" as well as "ine meyaló").

combo +t*SPI +t%mor +s"ADJ*^N|*" +k @

Searches adjectives immediately followed by a common noun in Spiros' data.

combo +t%mor +s"ADJ*^^N|*" +k @

Searches adjectives immediately or eventually followed by a common noun in Spiros' data.

combo +t%mor +t*SPI +s"!V*" +k @

Lists all of Spiros' verbless utterances (i.e. those in which no Verb code appears on the %mor tier).

combo +t*SPI +t%mor +s"V*^(PRO*+N*)" +k +x @

Lists all of Spiros' utterances in which a verb is immediately preceded or followed (+x option) by a pronoun or noun (PRO*+N*).

NB. Parentheses must not immediately be preceded by "*".

combo +t*SPI +t%mor +s"V*^^^(PRO*+N*)" +k @

Lists all of Spiros' utterances in which a verb is immediately or eventually (^*^) followed by a pronoun or noun (PRO*+N*).

combo +t*SPI +t%mor +s"(PRO*NOM*+N*NOM*)^^^V*" +k @

Lists all of Spiros' utterances in which a pronoun or noun in the nominative immediately or eventually (^^^) precedes a verb. NB. For this search to give the desired results the %mor tier must have been disambiguated.

combo +t*SPI +t%mor +s"V^(PRO*NOM*N|*ACC*)" +x +k @

Lists all of Spiros' utterances containing a verb, a pronoun in the nominative case, and a noun in the accusative case in any order.

combo +t*SPI +t%mor +s"PRO^(V*^^^N|*)" +k @

Lists all of Spiros' utterances containing a pronoun directly followed by a verb, and a noun eventually following the verb. For this command to work properly, the %mor tier must have been disambiguated.

NB. The +x option is invalid with commands comprising the string "^^^".

combo +t*SPI +t%mor +s"V*3*^^N|*NOM*" +x +k @

Lists all of Spiros' utterances containing a third person verb form and a noun in the nominative in any order. NB. For this command to work properly, the %mor tier must have been disambiguated.

combo +t*MAI +t%mor +s"ena^^^%mor:^^^PRO*" +k +r2 @

Lists all of Mairi's utterances in which *ena* is used as a pronoun and not as the homophonous numeral. NB. For this command to work properly, the %mor tier must have been disambiguated.

combo +t*SPI +s@pos1^@pos2 @

Lists all of Spiros' utterances in which a lexical item included in the Include file pos[ition]1 (pos1.cut) immediately precedes a lexical item included in the Include file pos2 (pos2.cut) (see Thomas 1994:283).⁴⁷ This command may be used for listing examples with two clitics in a row.

combo +t*MOT +s"mikr*^^^@dimin @

Lists all of Mother's utterances in which a form of the adjective *mikros* is eventually followed by a diminutive included in the Include file dimin.cut containing **aki**, **its** and **ul** (e.g. "ta mikra ta arkuðakia").

6.3. FREQ

freq +t*SPI +r2 +k @

+t*SPI limits the output to Spiros's data ignoring the other speakers

⁴⁷ Include files are files to be used in search strings. On creating such files see section 6.7 below.

+r2 lists elements in parentheses as these are marked in the transcript
 +k treats upper and lower case as different⁴⁸

Lists all of Spiros' word forms in alphabetical order indicating their frequencies. Mother's and Investigator's utterances are not listed. Material enclosed in parentheses is represented as found in the transcript.

freq +t*SPI +r2 +k +o @

+o sorts Spiros' output by descending frequency

This command shows, among other things, that Spiros very often does not use the definite article where required since the number of "0ART:DEF" exceeds the number of tokens of "o", "i" or "ta".

freq +t*SPI +r2 +k +d @

+d outputs the line numbers of each word form in the file, its frequencies and the corresponding examples

freq +t*SPI +r2 +d +k +s"[*]" @

+s"..." searches particular strings within word boundaries
 +s"[*]" finds all utterances containing an error marked by '*' indicating the frequency of errors⁴⁹

freq +t*SPI +r2 +d +k +s"*aki" @

+s"*aki" finds all diminutive forms ending in *-aki* used by SPI and their frequencies indicating line numbers and the respective utterances in which the diminutives occur

freq +t%mor +t*SPI +k @

Produces an alphabetical list of lexemes according to their parts of speech and grammatical coding indicating frequencies. NB. For this command to work properly the %mor tier must have been disambiguated.⁵⁰

freq +t*SPI +t%mor +k +s"N|*GEN*" @

Finds all common nouns in the genitive singular or plural in SPI's speech indicating their frequencies.

freq +t*SPI +t%mor +s"ART:DEF*" +k @

⁴⁸ This option is valid in the version of the CLAN of September 2010. In the CLAN version of March 2010, upper and lower case are treated as different by default and the option +k will treat them as the same.

⁴⁹ If the asterisk is not quoted by the back slash, Freq will take it to indicate 'any string' occurring in square brackets.

⁵⁰ When disambiguating the %mor tier pay attention to the space occurring in front of the punctuation mark terminating the line. Forms preceding the punctuation mark without a space will be distinguished from those with a space between the form in question and the punctuation mark (e.g. "mesa ." and "mesa." are treated as two different forms by the CLAN programs).

Lists all tokens of the definite article occurring in Spiros' data. The search string for omitted definite articles is +s"*0ART:DEF*" (if such tokens are transcribed by "0ART:DEF" on the main tier). (See also section 6.4 below.)

freq +t*SPI +t%mor +s"ART:DEF%" +k @

Indicates the frequencies of definite articles occurring in Spiros' data without any grammatical subdivisions, such as gender, case or number.

freq +t*SPI +t%mor +s"N|%" +s"V|%" +k @

Indicates the respective frequencies of common nouns and main verbs in Spiros' data.

freq +t*SPI +s"0*" @

Lists the frequencies of omitted words coded by an initial zero on SPI's speaker tier (e.g. "0V:COP", "0DEF:ART").

freq +t*SPI +s"(*)" +r2 @

Outputs the frequencies of word forms with omitted endings. NB. Depending on the context, such forms are not necessarily ungrammatical (e.g. "s(e) ena trapezi").⁵¹

freq +t*SPI +s"babas" @

Finds all forms in Spiros' speech in which the final /s/ of the form is either present or missing (i.e. both "babas" and "baba(s)").

freq +t*SPI +s"baba(s)" +r2 @

Finds all forms in Spiros' speech in which the final /s/ of the form is missing (i.e. forms transcribed as "baba(s)").

freq +t*SPI +s"babas" +r3 @

Finds all forms in Spiros' speech in which the final /s/ of the form is present (i.e. forms transcribed as "babas").

6.4. KWAL

kwal +t*SPI +s"o" @

Lists all examples in which the definite article form "o" occurs in Spiros' data. (In order to list instances of omitted articles search "0ART*").

kwal +t*SPI +s"o" +r2 @

⁵¹ Filling in vowels in the transcript even if these are normally omitted in colloquial speech will enable MOR to recognize the respective complete forms contained in the lexicon, since parentheses are disregarded by MOR (e.g. "s(e) ena" will be coded as if it was written "se ena").

Lists all examples in which “o” is used by Spiros, e.g. “o” or “u [: o]”.

kwal +s"[*:s]" @

Lists all utterances containing a form ending in /s/ in Standard Greek and the corresponding deviant form in the child’s speech (e.g. forms in which the final /s/ may have been omitted) which has been corrected by a form in square brackets following the deviant form.

kwal +t*SPI +s"*?" @

Lists all of Spiros’ interrogative clauses.

kwal +t*SPI +t%mor +s"PTL:NEG*" @

Lists all of Spiros’ utterances containing a negative particle.

kwal +t*SPI +t%mor +s"*0PTL:NEG*" @

Lists all of Spiros’ utterances in which a negative particle has been omitted.⁵²

kwal +t*SPI +t%mor +s"*PTL:NEG*" @

Lists all of Spiros’ utterances containing a negative particle as well as those in which it has been omitted.

kwal +t*SPI +s"[*]" @

Lists all utterances containing incorrect forms indicated by "[*]" on the main tier.

kwal +t*SPI +t%mor +s"****" @

Lists all of Spiros’ utterances containing incorrect forms marked by an asterisk on the %mor tier (e.g. V|kriono:IPFV:PASS*=ACT:NONPAST:1S).

kwal +t*SPI +s"*(*)*" +r2 @

Lists all of Spiros’ utterances containing forms with omitted parts.

kwal +t*SPI +s"aft*" +r2 -w2 +w2 @

Lists all of Spiros’ utterances containing a form of *aftos*, such as *aftos*, *afto*, *afta*, including the two preceding utterances (-w2) and the two following ones (+w2).

kwal +t*SPI +s"%aki" +s"%ula" +r2 @

Lists all of Spiros’ utterances containing diminutives ending in *-aki* or *-ula*. In contrast to the options +s"*aki" and +s"*ula", keywords only consist of the suffix rather than the individual lexemes used.

⁵² The omission is marked by "0PTL:NEG" on the main tier and is coded by "?|0PTL:NEG" on the %mor tier.

`kwal +t*SPI +t%mor +s"PRO:PRS*" +k +r2 @`

Lists all of Spiros' utterances containing a personal pronoun ignoring the %mor lines associated with the other speaker tiers.

`kwal +t*SPI +s@diminut.cut +r2 @`

Lists all of Spiros' utterances containing one of the diminutive suffixes contained in the include file dimin.cut. In this way, all diminutives can be retrieved in grammatically uncoded transcripts.⁵³

`kwal +t*SPI +s@locprep.cut +r2 @`

Lists all of Spiros' utterances containing a locative preposition contained in the include file locprep.cut.

`kwal +o@ -t% +k +r2 +d +f @`

- +d outputs the file in legal CHAT format without line numbers so that it can be used as input for other CLAN programs
- +o@ preserves the header tiers
- t% drops out all dependent tiers
- +f saves the result in a file

This command is useful for printing transcripts without dependent tiers or for recoding transcripts in a different way.

`kwal +t*SPI +t%mor +s"PTL:NEG*" +k +d @ | mlu`

Calculates the MLU of Spiros' utterances containing a negative particle by **Piping** KWAL and MLU.

`kwal +t*SPI +t%mor -s"PTL:NEG*" +k +d @ | mlu`

Calculates the MLU of Spiros' utterances not containing a negative particle by **Piping** KWAL and MLU.

The latter two commands may be tested on a small file such as the following:

```
@UTF8
@Begin
@Languages: ell
@Participants: SPI Spiros
*SPI: to θelo .
%mor: PRO|to V|θelo .
*SPI: δen to θelo .
%mor: PTL:NEG|δen PRO|to V|θelo .
*SPI: ela .
%mor: V|ela .
*SPI: oxi .
```

⁵³ On creating files to be used in searches see section 6.7 below.

```
%mor: PTL:NEG|oxi .
@End
```

The result of the first command (+s"PTL:NEG*") will be a ratio of 1.75 morphemes per utterance (4/7), whereas the second command (-s"PTL:NEG*") will result in a ratio of 1.5 (2/3). A larger speech sample of a linguistically more advanced child might allow to test the hypothesis that negative utterances tend to be less complex overall than non-negated ones.

6.5. MLU

```
mlu -t%mor @
```

Outputs a word/utterance ratio of each speaker. If morphemes have been hyphenated, the output will be a morpheme/utterance ratio.

NB. MLU works on the %mor tier by default. The option “-t%mor” will make it work on the speaker tier.

Since MLU indicates the number of utterances of a file (or a group of files) for each speaker separately, it may be used to determine and compare the length of files as far as the number of utterances of the target child is concerned.

```
mlu -t%mor -s"*@i" @
```

Outputs a word/utterance ratio disregarding interjections coded by “@i” (e.g. “vre@i”).

```
mlu -t%mor -b- @
```

Outputs a word/utterance ratio in spite of the fact that morphemes have been hyphenated.

6.6. MODREP⁵⁴

```
modrep +b*SPI +c%pho +k @
```

Compares the child's variable renderings on the %pho tier to the forms indicated on the speaker tier *SPI.

```
modrep +b%mod +c%pho +k @
```

Compares the child's variable renderings of forms indicated on the %pho tier to the standard forms given on the %mod tier.

```
modrep +b%mod +c*SPI +k @
```

⁵⁴ For sophisticated searches by Piping ModRep with Combo see MacWhinney (2010/II).

Compares the child's variable renderings of forms indicated on the speaker tier to standard forms given on the %mod tier.

Example:	result of the command: modrep +b%mod +c*SPI +k @
*SPI: papa	3 paraθiro
%mod:paraθiro	1 papa
*SPI: soso	1 soso
%mod:paraθiro	1 papasoso
*SPI: papasoso	
%mod:paraθiro	

NB. For ModRep to work properly the lines compared must contain the same number of words so that a one-to-one alignment is possible.

6.7. How to Create Files to be Used in Search Strings⁵⁵

The first line of each Search File must be @UTF8 if you want to use Greek characters. Use one item per line only (e.g. δe* (for *den*, *dem* and *de*)). End each line (including the last one) by a carriage return (ENTER). Save the search file as a .cut file (e.g. a file called neg.cut including all negative particles). Incorporate the search file into the **lib** (or **lib-ell**) directory.

Search file neg.cut for Greek:

δe*
mi*
oxi

7. EXAMPLE OF TRANSCRIBED AND CODED GREEK DATA

```
@UTF8
@Begin
@Languages: ell
@Participants: SPI Spiros Target_Child, MOT Mother of Target_Child,
               ULL Ursula Stephany Investigator
@ID: ell | stephany | SPI | 1;9.2 | male | || Target_Child ||
@ID: ell | stephany | MOT | || lower middle | Target_Child's mother |
      primary school |
@ID: deu/ell | stephany | ULL | || Investigator ||
@Media: SPI-A-03.wav
@Birth of SPI: 27-JUN-1972
@Date: 29-MAR-1974
@Interaction type: Looking a the picture book "Ich bin der kleine Bär"
                  (I am the little bear) and playing with toys at Spiros' home
@Location: Athens, Greece
@Tape location: CD1
@Time duration: 46 min.
@Filename: SPI-A-03.cha
@Transcriber: Ursula Stephany
*MOT: xxx arkuðes .
%mor: ?|xxx N|arkuða:FEM:NOM/ACC:PL .
*SPI: ales .
%mor: PRO:INDEF|alos:FEM:NOM/ACC:PL .
```

⁵⁵ See MacWhinney (2010/II:59).

*MOT: ales .
 %mor: PRO:INDEF|alos:FEM:NOM/ACC:PL .
 %com: confirmingly
 *SPI: ales .
 %mor: PRO:INDEF|alos:FEM:NOM/ACC:PL .
 *MOT: ales i arkuðes .
 %mor: PRO:INDEF|alos:FEM:NOM/ACC:PL ART:DEF|MASC/FEM:NOM:PL
 N|arkuða:FEM:NOM/ACC:PL .
 *SPI: niau .
 %mor: ONOM|niau .
 %com: in a high voice
 *ULL: niau .
 %mor: ONOM|niau .
 *SPI: mu@o .
 %mor: ?|mu@o .
 *SPI: jajaki [: psaraki] .
 %mor: N|psari:DIM:NEUT:NOM/ACC:SG .
 %com: referring to a fish in book
 *MOT: psaraki .
 %mor: N|psari:DIM:NEUT:NOM/ACC:SG .
 *MOT: ti ine ?
 %mor: PRO:INT|ti V:COP|ime:IPFV:NONPAST:3S/P
 *MOT: podikaki .
 %mor: N|podikos:DIM:NEUT:NOM/ACC:SG .
 *MOT: o vatraxos .
 %mor: ART:DEF|MASC:NOM:SG N|vatraxos:MASC:NOM:SG .
 *MOT: ti in(e) afto ?
 %mor: PRO:INT|ti V:COP|ime:IPFV:NONPAST:3S/P PRO:DEM|aftos:NEUT:NOM/ACC:SG ?
 %act: pointing to frog
 *SPI: vakotos [: vatraxos] .
 %mor: N|vatraxos:MASC:NOM:SG .
 *MOT: u@o !
 %mor: ?|u@o !
 *MOT: ta arkuðakia !
 %mor: ART:DEF|NEUT:NOM/ACC:PL N|arkuði:DIM:NEUT:NOM/ACC:PL !
 *MOT: posa ine ?
 %mor: PRO:INT|posos:NEUT:NOM/ACC:PL V:COP|ime:IPFV:NONPAST:3S/P ?
 *SPI: e(na) zio [: ðio] .
 %mor: NUM|ena:NEUT:NOM/ACC:SG NUM|ðio .
 %com: whiningly
 *SPI: t(r)ia zio [: ðio] .
 %mor: NUM|tris:NEUT:NOM/ACC NUM|ðio .
 *SPI: oo@o .
 %mor: ?|oo@o .
 *MOT: posa pola !
 %mor: PRO:INT|posos:NEUT:NOM/ACC:PL QUANT|polis:NEUT:NOM/ACC:PL !
 *ULL: pali i manula fonazi to arkuðaki .
 %mor: ADV|pali ART:DEF|FEM:NOM:SG N|mana:DIM:FEM:NOM/ACC:SG V|fonazo:IPFV:NONPAST:3S
 ART:DEF|NEUT:NOM/ACC:SG N|arkuða:DIM:NEUT:NOM/ACC:SG .
 %com: pages 23 to 24 in picturebook
 *ULL: lei <ela (e)ðo ! > [""]
 %mor: V|leo:IPFV:NONPAST:3S n|quote
 *ULL: <mi pas makria apo ti manula > [""]
 %mor: n|quote
 *MOT: ti lei , ayapi mu ?
 %mor: PRO:INT|ti V|leo:IPFV:NONPAST:3S N|ayapi:FEM:NOM/ACC:SG PRO:PERS|eyo:CLIT:GEN:1S ?
 *SPI: matia [: makria] # 0PREP 0ART manula !
 %mor: ADV:LOC|makria ?|0PREP ?|0ART N|mana:DIM:FEM:NOM/ACC:SG !
 %com: in a low male voice
 *MOT: <makria ap(o) ti manula> [""]
 %mor: n|quote
 @End

References

- MacWhinney, Brian (1994). New horizons for CHILDES research. In Sokolov & Snow (eds.) 1994: 408-452.
- MacWhinney, Brian (2000). *The CHILDES Project. Tools for Analyzing Talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Ass.
- MacWhinney, Brian (2008). Enriching CHILDES for morphosyntactic analysis. In Heike Behrens (ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives*, 165-197. Amsterdam/Philadelphia: John Benjamins.
- MacWhinney, Brian (2010). *The CHILDES Project: Tools for Analyzing Talk*. Electronic Edition. Part 1: The CHAT Transcription Format. Part 2: The CLAN Programs. <<http://childes.psy.cmu.edu/manuals/>> (Sept. 2010)
- MacWhinney, Brian & Catherine E. Snow (1985). The child language data exchange system. *Journal of Child Language* 12: 271-295.
- MacWhinney, Brian & Catherine E. Snow (1990). The child language data exchange system: An update. *Journal of Child Language* 17: 457-472.
- Sánchez-Martínez, Juan Carlos (1994). Untersuchungen zu Tempus und Aspekt bei spanisch-deutsch bilingualen Kindern. Romanisches Seminar, Universität zu Köln. Ms.
- Sokolov, Jeffrey L. & Brian MacWhinney (1990). The CHIP framework: Automatic coding and analysis of parent-child conversational interaction. *Behavioral Research Methods, Instruments, and Computers* 22: 151-161.
- Sokolov, Jeffrey L. & Catherine E. Snow 1994. Transcript analysis using the Child Language Data Exchange System. In Sokolov & Snow (eds.) 1994: 1-25.
- Sokolov, Jeffrey L. & Catherine E. Snow (eds.) 1994. *Handbook of Research in Language Development Using CHILDES*. Hillsdale, NJ: Lawrence Erlbaum Ass.
- Stephany, Ursula & Conny Bast (2001). Working with the CHILDES Tools: Transcription, Coding and Analysis. In Ursula Stephany, Conny Bast and Katrin Lehmann, *Computer-Assisted Transcription and Analysis of Speech*. Arbeitspapier No. 41 (N.F.), Institut für Sprachwissenschaft, Universität zu Köln, Nov. 2001.
- <http://www.uni-koeln.de/phil-fak/ifl/asw/forschung/arbeitspapiere> [select PDF 2.Teil]

Appendix I

Codes for Grammatical Morphemes

MACWHINNEY, B. (2000), *The Childes Project*. Tools for analyzing talk, 3rd ed. Vol. 1: Transcription format and programs (pp. 167-169). Mahwah, N.J.: Erlbaum; MACWHINNEY, B. (1995), *The Childes Project*. 2nd ed. (pp. 113-115). These codes are originally based on LEHMANN, C. (1982). Directions for interlinear morphemic translations. *Folia Linguistica* 16: 199-224. Revised and enlarged by U. Stephany, Dept. of Linguistics, University of Cologne, Germany. – Codes referring to parts of speech, rather than grammatical categories, are noted with asterisks.

1	first person
1P	first person plural
1PE	1st Plural exclusive
1PI	1st Plural inclusive
1S	1st person singular
2	2nd person
2P	2nd person plural
2S	2nd person singular
3	3rd person
3P	3rd person plural
3S	3rd person singular
ABESS	abessive ('without x')
ABL	ablative ('from x')
ABS	absolutive
ABST	abstract
ACC	accusative
ACH	achieve ('manage to')
ACT	active
ADESS	adessive ('toward x')
ADJ	adjective, adjectival*
ADJR	adjectivalizer
ADP	adposition*
ADV	adverb(ial)*
ADVERS	adversative
ADVN	adverbial noun*
ADVR	adverbializer
AFF	affirmative
AFFECT	affective
AG	agent
AGR	agreement
AGTV	agentive
AL	alienable
ALL	allative
ALLOC	allocutive
ANA	anaphoric
ANI	animate
ANT	antipassive
AORIST	aorist
APP	apposition
APPL	applicative
ART	article*
ASP	aspect
ASS	assertive
AT	attributor
ATTEN	attenuative
AUG	augmentative
AUX	auxiliary*
BEN	benefactive
CARD	cardinal number
CAT	catenative

CAUS	causative
CESS	cessive "stop"
CGN	conjugalional marker
CIRC	circumstantial
CLFR	classifier
CLIT	clitic*
CMN	common
CMPLR	complementizer
CMPLX	complex (morpho- logically)
COLL	collective
COM	comitative ('together')
COMP	comparative
COMPL	completive
CONC	concessive
COND	conditional
CONJ	conjunction*
CONN	connective
CONSEC	consecutive
CONT	continuous, continua- tive
COO	coordinating
COP	copula*
CORR	correlative
COU	count
CP	comparative
DAT	dative
DCLN	declensional marker
DECL	declarative
DEF	definite
DEICT	deictic
DEM	demonstrative
DESID	desiderative
DET	determiner*
DIM	diminutive
DIREC	directional
DIST	distal
DISTR	distributive
DO	direct object
DU	dual
DUB	dubitative
DUR	durative
DYN	dynamic (nonstative)
ELAT	elative ('out of X')
EMPH	emphatic
EMPTY	empty
EPIT	epithet
ERG	ergative
ESS	essive ('as x')
EV	evidential

EVE	event
EXCL	exclusive
EXIST	existential
FACT	factive, factitive
FEM	feminine
FIN	finite
FNL	final (goal)
FOC	focus
FREQ	frequentative
FUT	future
GEN	genitive ('of x')
GENER	generic
GER	gerund
HAB	habitual
HE	head
HON	honorific
HORT	hortative
HUM	human
ILL	illative ('into x')
IMM	imminent
IMP	imperative
IMPRS	impersonal
INAL	inalienable
INANI	inanimate
INCH	inchoative
INCL	inclusive
INCPT	inceptive
INDEF	indefinite
INESS	inessive ('in X')
INF	infinitive*
INFER	inferential
INJ	injunctive
INSTR	instrumental
INT	interrogative
INTENT	intensive
INTERJ	interjection*
INTNS	intensifier
INTRANS	intransitive
INVIS	invisible
IO	indirect object
IPFV	imperfective
IRR	irrealis
ITER	iterative
JUSS	jussive
LAT	lative ('moving to')
LOC	locative
MAIN	main
MAN	manner
MASC	masculine
MASS	mass

MDL	modal
MEAS	measure
MOD	modifier
MP	mediopassive
N	noun*
NARR	narrative
NEG	negative
NEUT	neuter
NEUTRAL	neutral
NH	nonhuman
NOM	nominative
NOML	nominal
NONPAST	nonpast
NONVIR	nonvirile
NR	nominalizer
NUM	numeral, numeric
OBJ	object
OBL	oblique
OBLIG	obligatory
OPT	optative
ORD	ordinal numeral (‘first’)
OTHER	other
PART	participle*
PARTIT	partitive
PASS	passive
PAST	past
PAT	patient
PEJ	pejorative
PERF	perfect
PERM	permissive (‘may’)
PFV	perfective
PL	plural
PLACE	place
PLPF	pluperfect
POL	polite
POSS	possessive (X’s)
POST	postposition*
POT	potential
PP	past participle
PRDV	predicative
PRE	prefix
PREP	preposition*
PRES	present
PRESPT	present participle
PRESUM	presumptive
PRET	preterite
PRH	prohibitive
PRO	pronoun*
PROG	progressive
PROL	prolative (‘along X’)
PROP	proper
PROS	prospective (‘by tomorrow’)
PROT	protracted (‘keep on’)
PROX	proximal
PRS	personal (pronoun)
PSBL	possible
PTL	particle*
PURP	purposive

QUANT	quantifier*
QUE	question
QUOT	quotative
REAL	realized, nonfuture
RECENT	recent
RECIP	reciprocal
REFL	reflexive
REL	relative*
REM	remote
REPET	repetition
REPORT	reportative
RES	resultative
RETRO	retrospective
SEQ	sequential
SG	singular
SIMUL	simultaneous
SP	superlative
SPEC	specific
SS	same subject
STAT	stative
SUBJ	subject
SUBJV	subjunctive
SUBL	sublative (‘onto X’)
SUBOR	subordinating
SUFF	suffix
SUG	suggestive
SUPER	superessive (‘on X’)
TANG	tangible
TEMP	temporal, time
TERM	terminative
TNS	tense
TOP	topic
TRANS	transitive
TRANSL	translative (‘becoming X’)
TRY	try or strive to achieve
USIT	usitative
V	verb*
VAL	validator
VIR	virile
VIS	visible
VOC	vocative
VOL	volitional
VR	verbalizer
WH	wh-question word
YN	yes-no question word

Appendix II

Conventions for the Transcription of Greek

Ursula Stephany

(Revised in collaboration with E. Thomadaki and A. Christofidou [19 March 2010])⁵⁶

Greek letters	Transcription	Examples
α	a	kala
αι, αη	ai	xaiðevo, xaiðema, aiðoni, kelaiðima
ε, αι	e	pede, kerðos, keo, keros
ι, η, υ, ει, οι, υι	i	pino, lima, kima, ime, ikos, ios (υιός)
ια, εια, οια	ia	ðiamoni, voiθia, aletria, adria, enia, peðia, fiðia, peðakia
ιε, ειε	ie	adimetrieme, aγapieme
ιο, ειω, ιω	io	sxolio, peðio, teliono
ο, ω	o	boro, loγos, omos
οι	oi	voiðamaksa, koroiðo
ου	u	puli, kupa
β, ββ, (α, ε)υ, (ε)υβ	v	voli, savato, avli, evoikos
γ [ᾱ]	gh	γata, maγos
γ [j], ι (etc.)	j	jirizo, jeros, ajios, jos
γκ, γγ	g	agizo, pagos, egonos, agrafa, garaz, ginja, gemi, egrafo, sigrafeas, sigenis, egenis
γχ	nx	enxisi, vronxos
δ	dh	ðen, ðino, peði
ζ, σ	z	zoi, mazi
θ	th	θelo, eθnos
κ, κκ	k	kano, kita, kenos, eklisia
λ, λλ	l	kalos, leo, pali, malon
λι, λλι + V	li	liakaða, elia, teliono, maliaros, ilios
μ, μμ	m	mama, monos, omos, amos
μπ	b	babas, bubuki, kubi
ν, νν	n	ne, eno, jeneos, nona
νι, ννι	ni	niata, kunia, enja (εννιά), enia (έννοια)
ντ	d	dada, adras
ντζ, τζ	dz	padzuri, dzami, pidzama, kodzam
ξ, κς	ks	ekso, ksero, ekstratia
π, ππ	p	peto, apo, ipos
ρ, ρρ	r	ora, rixno, arostos
σ, σσ, ς	s	soma, γlosa, mesos, kozmos
τ, ττ	t	tote, prato
τσ, τς	ts	tsakizo, katsaros, mats
φ, (α, ε)υ, (ε)υφ	f	fos, afti, efxi, eforos
χ	x	exo, xano, loxos, exis, xino, maxi, xeri
ψ	ps	psixolojia, psari, psino
πι, τι, ρι, κι, ðι + V	pi, ti, ri, ki, dhi	papia, fotia, xorio, sikia, karðia

Notes

- Homophonous forms or identical transcriptions of two different Greek words will be disambiguated on the %mor tier. The same is true of segmentally identical forms bearing contrastive stress.
- The use of capital letters is reserved for proper nouns.

⁵⁶ Based on a proposal prepared in collaboration with E. Thomadaki, D. Katis and A. Christofidou, 10 April 2006.