Using CHILDES and CLAN for Student Projects

Sonja Eisenbeiss (University of Essex)

seisen@essex.ac.uk

Information about CHILDES

- Child Language Data Exchange System
- http://childes.psy.cmu.edu/
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

CHILDES: Types of Data

- transcript data from child learners
- part of the data from clinical populations
- a broad range of languages, but focus on English
- individual recordings for groups of learners
- recordings over longer periods for some learners
- spontaneous interactions
- picture book descriptions (e.g. “frog story”)

CHILDES: Data Format

- some corresponding video/audio-files
- some time-linked video/audio-files
- encoding of linguistic information for some files (word class, syntax).

Transcription in CHAT

- The most common format for transcribing child language data is the CHAT-format, developed for the largest child language data deposit: CHILDES (http://childes.psy.cmu.edu/).
- Digital CHAT-files can be searched using the CLAN tools provided by CHILDES.
- Files from various transcription programmes (ELAN) can be exported in CHAT-format.

Transcription: CHAT-Format

- Transcripts are written in a text editor and stored as unformatted ASCII files (*text only* or *plain text*).
- All lines are ended by a carriage return (ENTER).
- Every transcript must begin with the line: @BEGIN and end with the line: @End.
- Between @BEGIN and @End:
  - headers with information about the transcript (obligatory: @Participants)
  - main tier for transcription
  - dependent tiers for further annotations

CHAT-Format: Basic Structure

- @BEGIN
- @Participants
- [other headers]
- *JOE: [spoken material]
- %mor: [morpho-syntactic coding]
- *INT: [spoken material]
- %mor: [morpho-syntactic coding]
- @End

CHAT-Format: Headers

- three letters followed by a colon and a tab
- obligatory: @ Participants, on the second line of the transcript; e.g.:
  @Participants: JOE Joe child, INT Interviewer
- optional; e.g.:
  - @Birth of Learner: ...
  - @Age of Learner: ...
  - @Date: ...
  - @Language: ...
  - @Transcriber: ...

CHAT-Format: Main Tiers

- what was actually said, one utterance per tier
- introduced by "*", the three-letter code for the participant and a tab; e.g.:
  *JOE:	the boy put the leash on the cat.
- orthographical transcription in lower case Latin letters; except for proper nouns (e.g. *John*) and "I"
- numbers spelled out (*ten*, not *10*)
- normalisation of phonetically deviant forms (phonetic information about forms can be presented on a %pho dependent tier)

Main Tiers: Markers

- unfilled pauses: #
- filled pauses: eh@fp
- interruption: +/.
- self-interruption: +//.
- repetition w/o correction: [/]
- repetition with correction: [//]
- unintelligible speech: xxx
- material coded on phonol. tier: yyy
- doubtful material: [?] or [=? text]
- omitted parts of words: ( )
- to refer to more than one word: < >

CHAT: Dependent Tiers

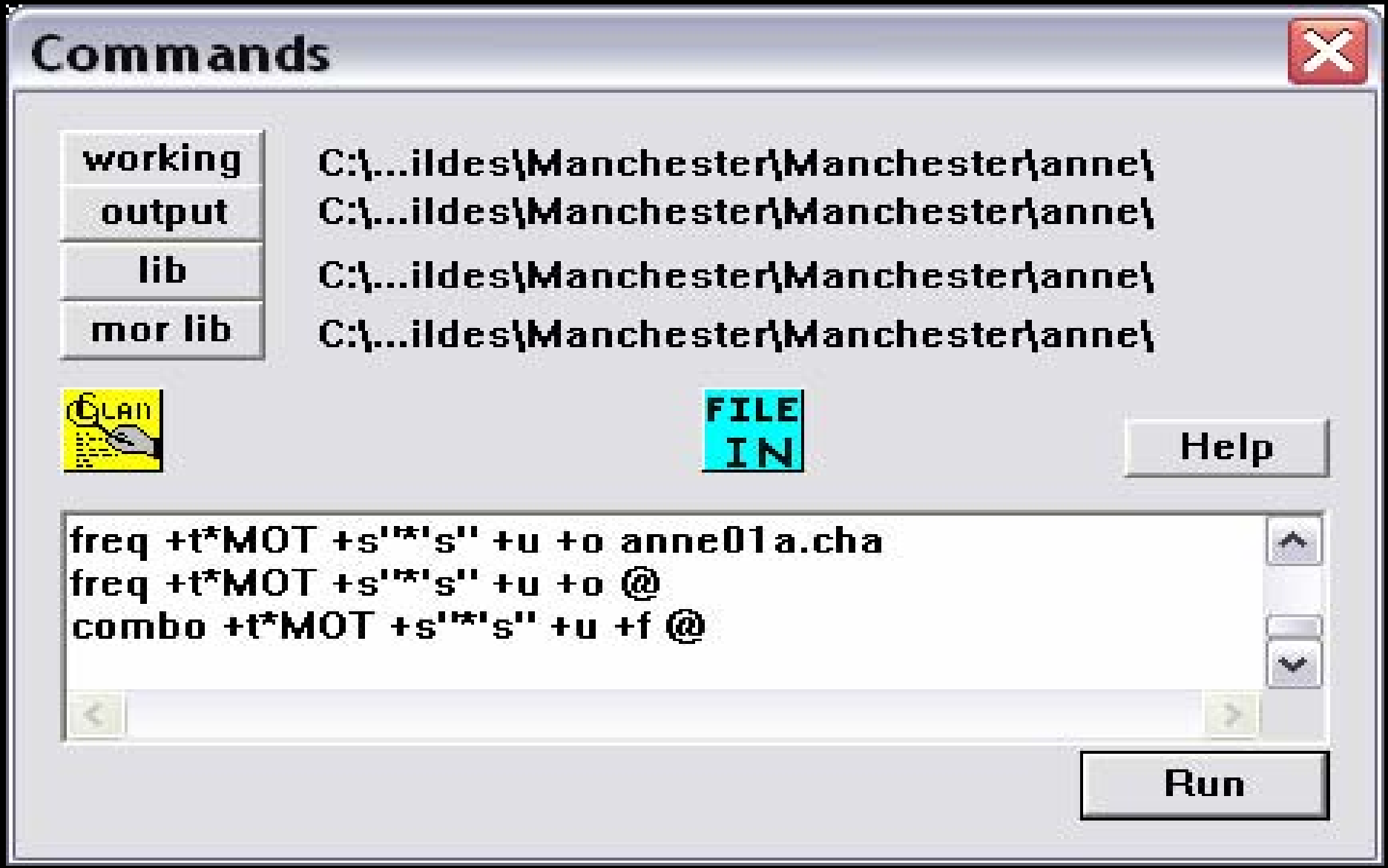
further annotations, e.g.

- %mor [morphosyntactic coding]
- %pho [phonological coding]
- %syn [syntactic coding]
- %err [errors]
- %com [comments]
- %spa [speech acts]

CLAN: Windows

- the commands window where you specify the folders, files, and commands you want to use
- the CLAN output window, where you will see the results of your searches. If you have not specified an output file, your results will be displayed in this window. If you have saved your outputs into a file (as you will be asked to do for this exercise), you will not be able to see it in the output window, but the name and location of the output file will be displayed in the output window.

CLAN: Command Window



CLAN: Steps

- specify your WORKING DIRECTORY, where the files you will be working with are stored
- specify your output directory, where any output files will be stored
- select a command (type of select from CLAN)
- select one or more transcription files for analysis (type name or select from FILE IN)
- optionally use some so-called switches to modify the commands.

CLAN: Core Commands

- FREQ: will provide you with type and token frequency information
- COMBO: will find utterances matching a given set of criteria
- MLU: will calculate the MLU (mean length of utterance)

CLAN: Useful Switches

+f saves output to file. For each transcription that you have chosen to analyse, an output file will be generated. By default, this output file will have the name of the transcription file and an extension that will show you which command was used to create the output (e.g. frq, mlu or cmb).

+s searches for a string in a file.

+t restricts the search to a particular tier – e.g. the tier of a particular speaker.

+u treats all files together.

+o orders FREQ lists according to token frequency

+w –w1 and +w1 provide one preceeding/following line, -w2 and +w2 will provide two preceeding/following lines, etc.

CLAN: Search Strings

^	immediately followed by
+	inclusive OR
!	logical NOT
*	“joker”
“”	strings including blanks, etc. should be put in quotes

CLAN: Search Strings

^	immediately followed by
+	inclusive OR
!	logical NOT
*	“joker”
“”	strings including blanks, etc. should be put in quotes

Some examples

freq sarah009.cha (frequency list of all words)

combo +sthank* +t*CHI sarah134.cha
(all child utterances with “thank”)

combo +s"*'s*" +t*MOT sarah134.cha
(all maternal utterances with “’s”: John**’s** hat/here)

combo +s"aux|**" +t*CHI +t%MOR sarah134.cha
combo +s"aux|**" +t%MOR sarah134.cha
(all auxiliaries (produced by child))

Some Project Ideas

- polite routines and formulas (*thank you, please*, etc.) in the child's language and in the child's input
- forms of auxiliaries in the child's input and in the child's own speech
- corrections and recasts in child-directed speech
- abstract and concrete nouns in the child's speech
- building up a lexicon of motion verbs