CHILDES workshop

Universität Mannheim

Vivienne Rogers

Wednesday 15th April 2015

This document is based on an earlier workshop handout by Prof Florence Myles from Centre for Research in Linguistics and Language Sciences, Newcastle University.

1 Aims of this course:

- Introduce you to the CHILDES website
- Overview of transcription system
- Show you how to find and download existing corpora
- Perform common analyses on existing corpora

2 Introduction to CHILDES website

2.1 What is CHILDES?

a) Definitions
- CHILDES: Child Language Data Exchange System
- CLAN: Computerized Language Analysis
- CHAT: Codes for the Human Analysis of Transcripts

b) Data available
- over 130 corpora
- from Spanish to Cantonese and Farsi to Tamil
- monolingual and bilingual L1 acquisition, L2 acquisition, adult aphasics, etc.
- all is available freely at http://childes.psy.cmu.edu/. The adult data (L1 and L2) is at http://talkbank.org/.

c) Tools provided
- transcripts database
- programs transcripts analysis
- methods for linguistic coding
- systems for audio and video linking

2.2 Overview of the website

d) Manuals
- CHAT transcription manual: provides a standardised format for producing computerised transcripts of face-to-face conversational interactions. http://childes.psy.cmu.edu/manuals/chat.pdf
- CLAN programs manual: describes the use of the CLAN program which is designed specifically to analyze data transcribed in the format of the Child Language Data Exchange System (CHILDES). http://childes.psy.cmu.edu/manuals/clan.pdf
- Database manuals: 8 documents which describe the CHILDES data. http://childes.psy.cmu.edu/manuals/

e) Database
- Browsable transcripts: view the corpus by JAVA based viewer, so you can browse the data over the web without CLAN program. You cannot view these with Internet Explorer.
- Downloadable transcripts: you can download the corpus and run the transcript in your local machine. To do this, you need to: unzip the corpus into folders and then use CLAN program to open the *.CHA files.

3 Introduction to transcription conventions

Each file has a set of headers so that the computer can recognise certain features of each file. Some file headers are obligatory (language, list of participants, ID headers). Others depend on the research question and factors you may think could influence your results (e.g. length of exposure to L2, school, age). Headers always start with @.

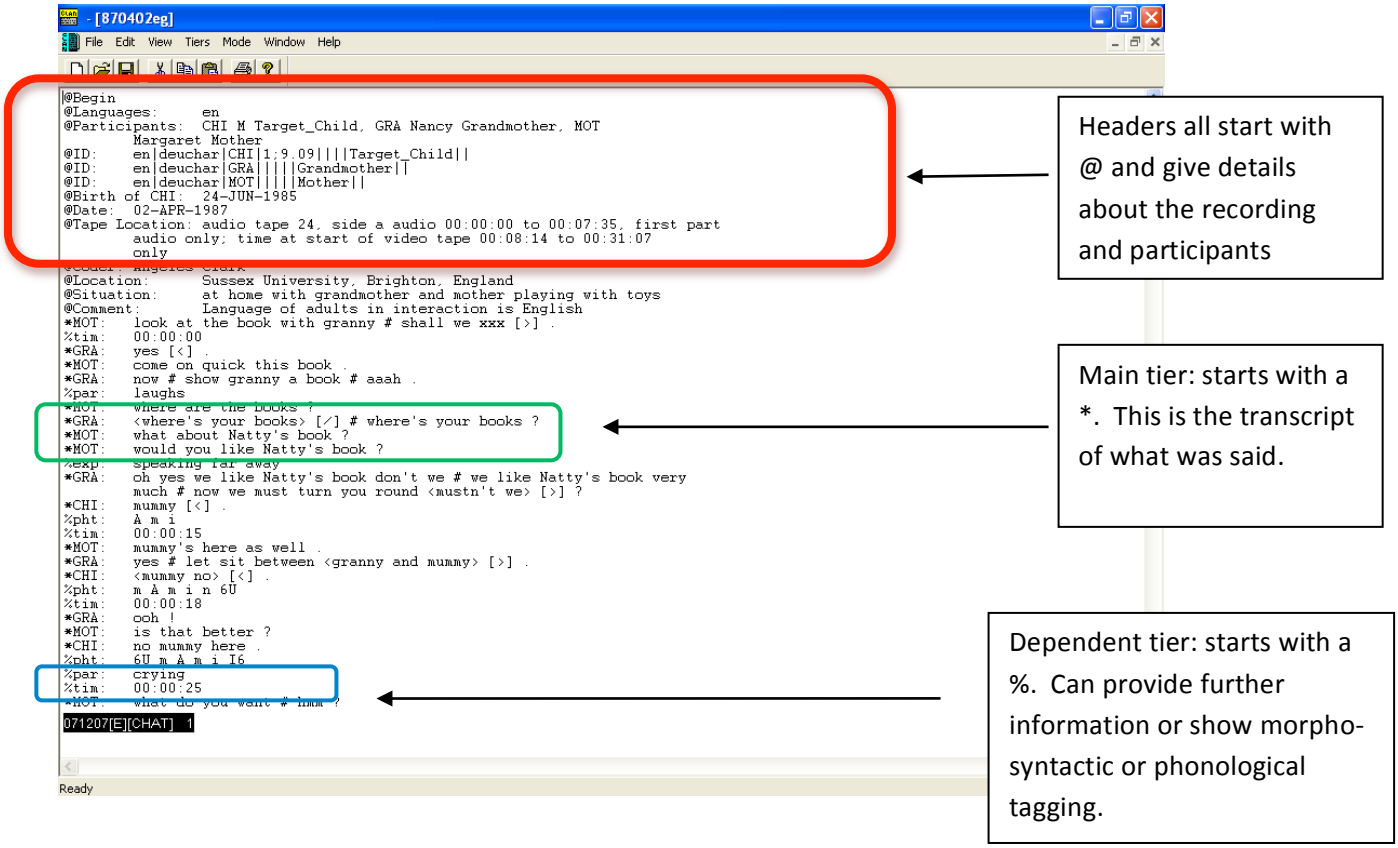
Each file always begins with @Begin and ends with @End.

a) Main tiers

This line gives the basic transcription of what the speaker said. The structure of the main lines in CHAT is fairly simple. Each main tier line begins with an asterisk (*). After the asterisk, there is a three-letter speaker ID, a colon and a tab. The transcription of what was said begins after these codes.

b) Dependent tiers

Dependent tiers are lines typed below the main line that contain codes, comments, events, and descriptions of interest to the researcher. These lines start with %.



Some common codes:

- +/. Interruption
- +... trailing off
- xxx unintelligible speech not treated as a word
- xx unintelligible speech treated as a word
- [?] best guess

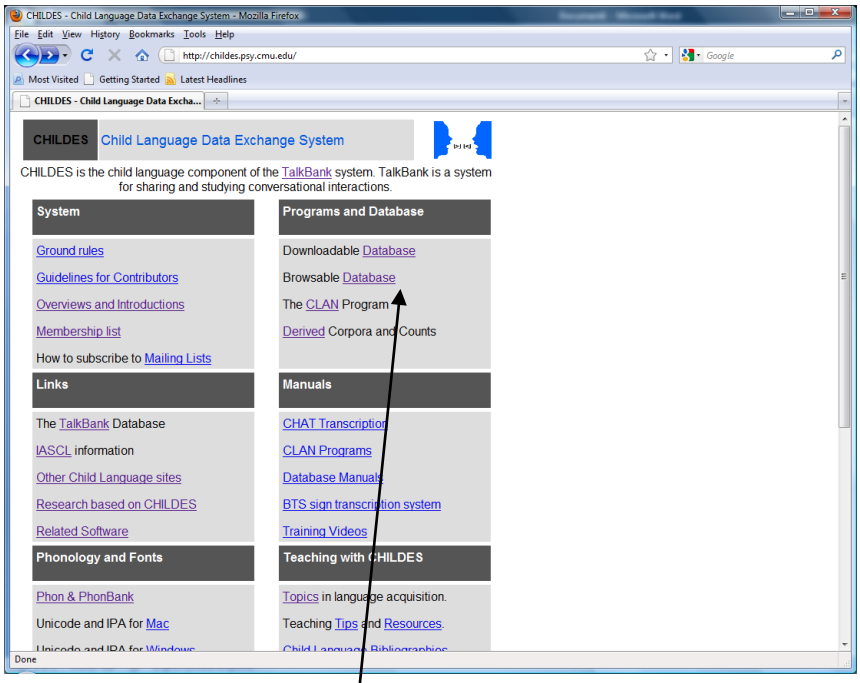
- @g imitation
- (.) or # pause
- [=r] reported speech
- [/] repetition

4 Finding and downloading files

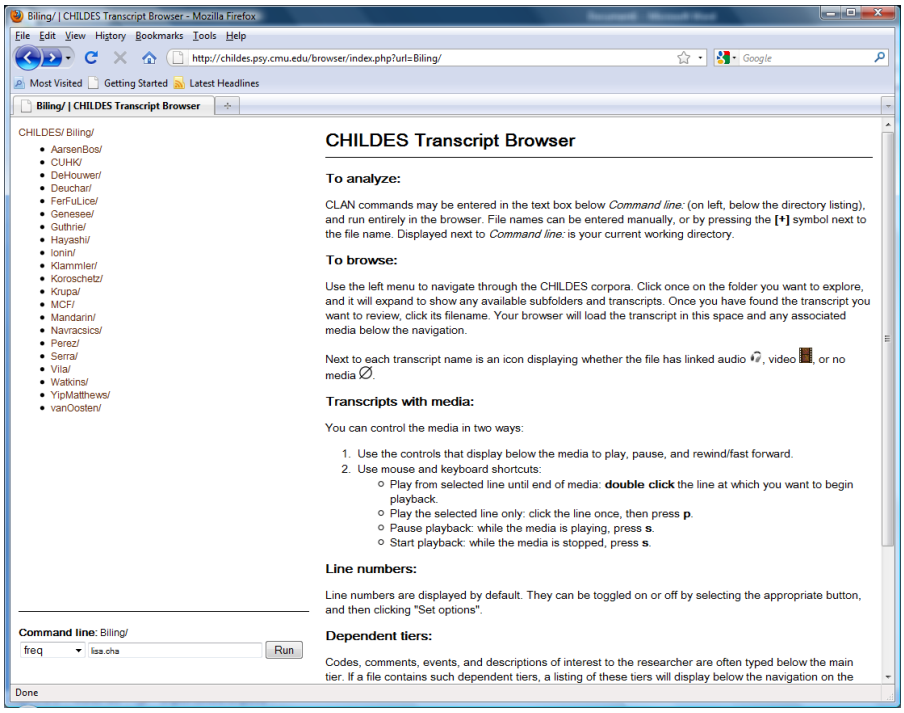
There are two options within CHILDES – you can either browse the databases or you can download databases to work with when you are not connected to the internet. You will need to have installed CLAN on your computer to open these files (it's already on all the computers in seminar room J).

4.1 How to browse the child L1 database on the internet

1) Go to childes website: http://childes.psy.cmu.edu



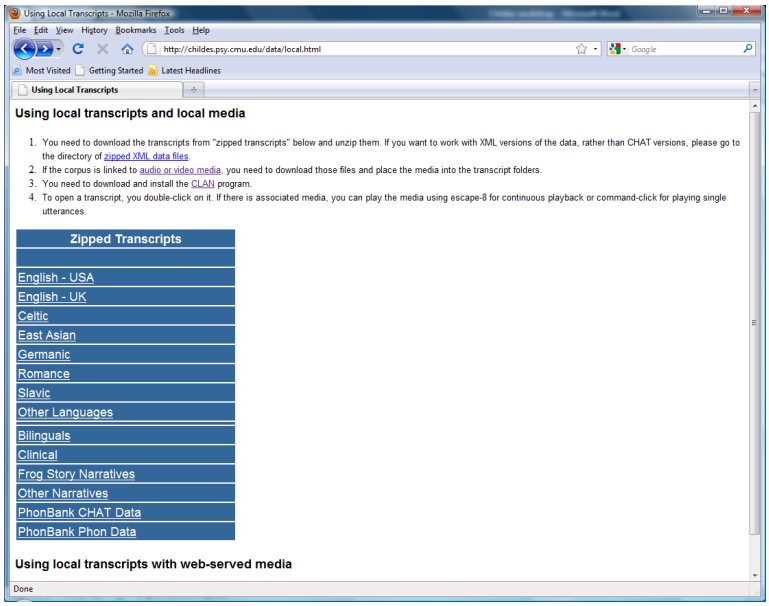
2) Click on browsable database.
3) Select the group of databases you would like to view, e.g. "biling" for all the databases working with bilingual children.



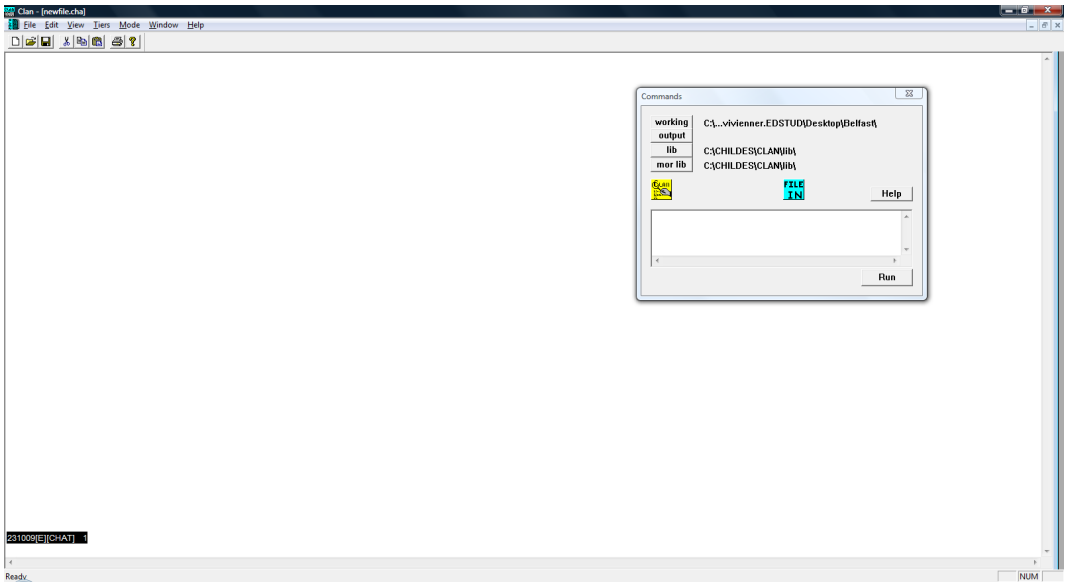
4) Select the database you would like to view. You may need to view a few before you find the one most relevant to your research.
5) For example, click on “Ionin” to view files relating to English – Russian bilingual children. You can click on each file to view.
6) The command line allows you to perform analyses on these files as a group or individually. See later section on performing analyses for more information.

4.2 How to download files to use on your own computer

1) Open the childes website http://childes.psy.cmu.edu
2) Click on “Downloadable database” – on the right hand side of the screen.
3) You can then choose between downloading transcripts or downloading audio and video.
4) Select downloadable transcripts. You will see the following screen



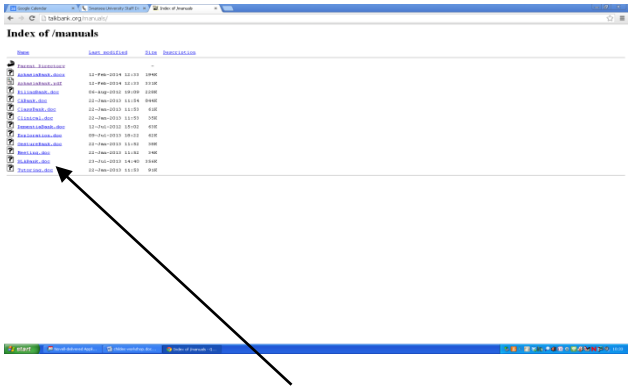
5) Click on bilinguals and download the zipped file containing the database you want. For example, if you want to download the bilingual English-Russian data collected by Tania Ionin, click on "Ionin".
6) Save this file locally on your computer.
7) Remember that you will need to use the program "CLAN" to view these files. You can download CLAN from the CHILDES website or it is already installed on the computers in seminar room J.



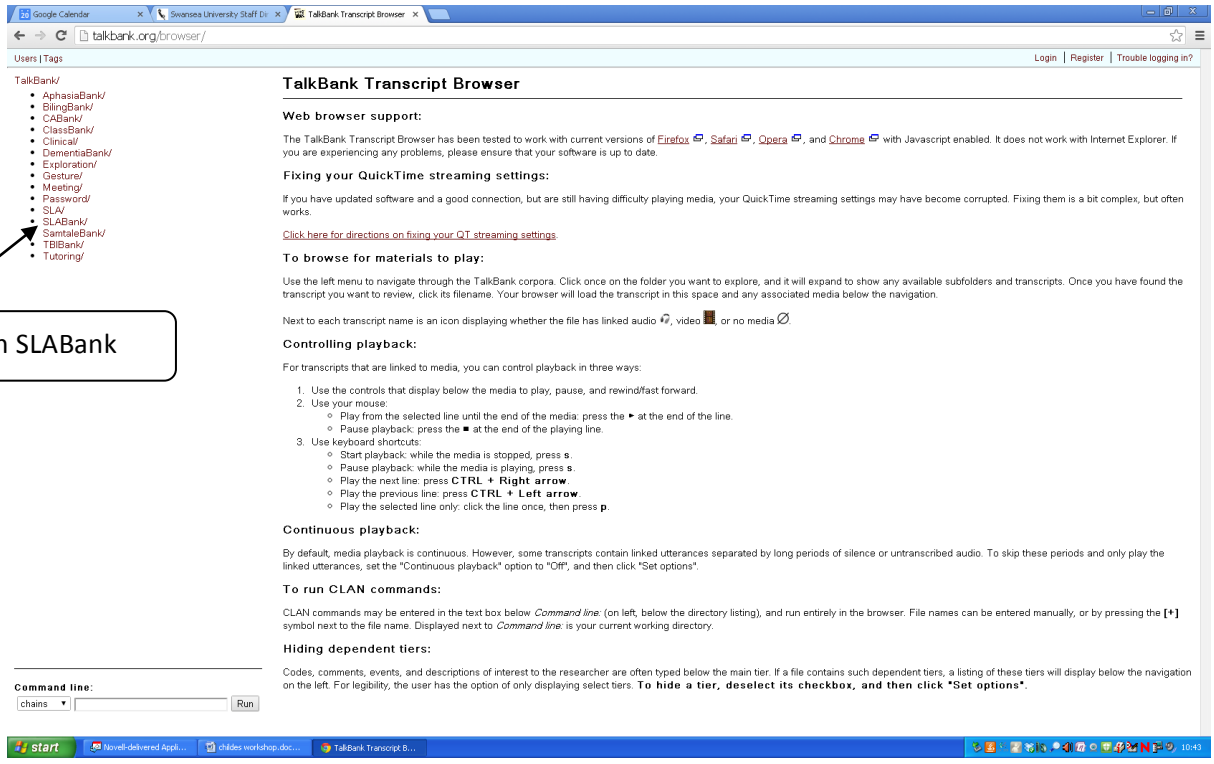
4.3 Accessing the databases for L2 data.

- Using Google Chrome/Firefox/Safari (not IE), go to talkbank.org for adult data.
- The site is divided into 6 sections. You will need to look at the DATA section.
- In the Data section, you will use the BROWSABLE DATABASES.
- But first you need to download the database manuals. (see below).

- Once you have clicked on DATABASE MANUALS, the following screen will appear.



- Click on SLABank.doc.
- Save this file. It will tell you about the content of the different corpora available.
- You may also be interested in BilingBank as it contains some German-English materials. You may wish to download this file for further details as well.
- Next we should start looking at the transcripts. Return to the main talkbank.org page. Click on BROWSABLE TRANSCRIPTS.



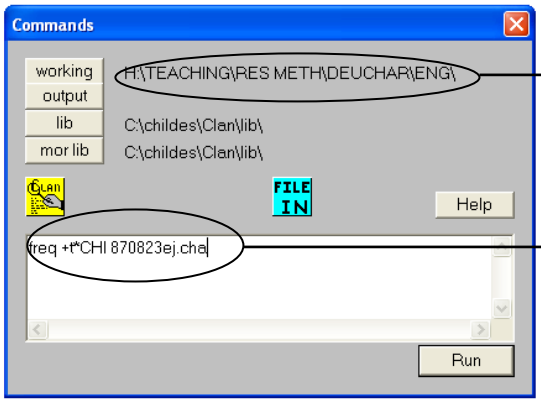
Click on SLABank

- Click on SLABank. (You might also be interested in the BilingBank)
- This brings up a list of all the corpora available. These are coded according to the name of the project, e.g. BELC is the Barcelona English Language Corpus.

If you want to download a database of L2 speech, then follow the same instructions as given for childes.psy.cmu.edu in section 4.2.

5 Performing analyses on the files: on your own computer

Once you have selected the files you want to work with, you will want to perform analyses on them. If you are browsing the files on the internet, use the command line in the bottom half of the screen (see point 6). If you are using CLAN on your computer, then you will need to open the file and use the commands window.



The folder where you have saved your files.

The command you are using

Some common analytical tools that can be used include:

a) FREQ

FREQ is a program for frequency analysis. FREQ constructs a frequency word count for user-specified files. A frequency word count is the calculation of the number of times a word, as delimited by a punctuation set, occurs in a file or set of files. FREQ produces a list of all the words used in the file, along with their frequency counts, and calculates a type–token ratio. The type–token ratio is found by calculating the total number of unique words used by a selected speaker (or speakers) and dividing that number by the total number of words used by the same speaker(s). It is generally used as a rough measure of lexical diversity. The command is: [command] [filename].cha

Research question: **What are the lexical range and frequency of the chosen file?**

Structure : [command] [filename].cha

Command: freq 061101ANN.cha

This is a space. Spacing and “” are very important in these commands.

Research question: **What are the lexical range and frequency of the bilingual child in the chosen file?**

Structure : [command] +t*[speaker code] [filename].cha

Command: freq +t*ANN 061101ANN .cha

The +t switch allows you to specify which speaker you want to run the analysis on.

Research question: **What lexical items are in the "wrong" language?**

Structure : [command] +t*[speaker code] +s"*[code used for mixing] [filename].cha

Command: freq +t*ANN +s"*@1" 061101ANN.cha

The transcriber had indicated all lexical items in the « wrong language » with @1.

Research question: **How many instances of the verb "want" have been produced by the child in that transcript?**

Structure: [command] +t*[speaker code] +s"[word]*" [filename].cha

Command: freq +t*CHI +s"want*" 870823ej.cha

The +s switch allows you to specify a word or string. Here we have decided to look for the verb want (and any of its derived forms (e.g. wants, wanted) as specified by the wildcard *).

b) MLU

The MLU[1] program calculates the mean length of utterances in a selected file or files.

Research question: **What is the mean length of utterance of the selected L2 learner of French?**

Structure: [command] +t*[speaker code] [filename].cha

Command: mlu +t*L14 L14MAL13.cha

Research question: **How does the MLU of the selected learner compare with others?**

Structure: [command] +t*[speaker code] [task code]*. cha

Command : mlu +t*L L*.cha

The +t*L switch allows us, here, to search for all learners (whose code starts with *L) in all of the files (which start with L).

c) COMBO

COMBO provides the user with ways of composing search strings to match patterns of letters, words, or groups of words in the data files. This program is particularly important for researchers who are interested in syntactic analysis.

[1] By default, MLU is calculated on the MOR tier (see below), to run MLU on the speaker tier, the speaker code is necessary (+t switch).

Research question: **How is the verb "mirar" (=to look) used in the selected L2 learner of Spanish?**

Structure: [command] +t*[speaker code] +s"[word]*" [filename].cha

Command: combo +t*L50 +s"mira*" L50MJA13.cha

This command locates all transcript lines including any form of the verb "mirar", because of the inclusion of the wildcard symbol *.

The symbol ^ is used to search for two words with no intervening material. The following command searches for any instances where the verb "mirar" is followed by the noun "monstruo" (=monster).

Structure: [command] +t*[speaker code] +s"[word]*^*^[word]" [filename].cha

Command: combo +t*L50 +s"mira*^*^monstruo" L50MJA13.cha

As you can see, this command allows searching for any article used in between the verb and the noun.

d) KWAL

The KWAL program searches for data for user-specified words, and it outputs those keywords in the context. That context (or cluster) provided is a combination if the main tier and any dependent tiers (if any). The –w and +w switches are used to specify the number of transcript lines to be included preceding and following the line containing the target word.

Research question: **In what contexts is the word "there" used by this two-year old bilingual?**

Structure: [command] +t*[speaker code] +s"[word] –w[lines above] +w[lines below] [filename].cha

Structure: kwal +t*CHI +s"there" -w2 +w2 870823ej.cha

e) Others

There are a large number of other analyses tools available within the CLAN programs. On the next page, we have reproduced the list of some of the main commands from the CLAN program manual with the corresponding page numbers.

f) Options/switches

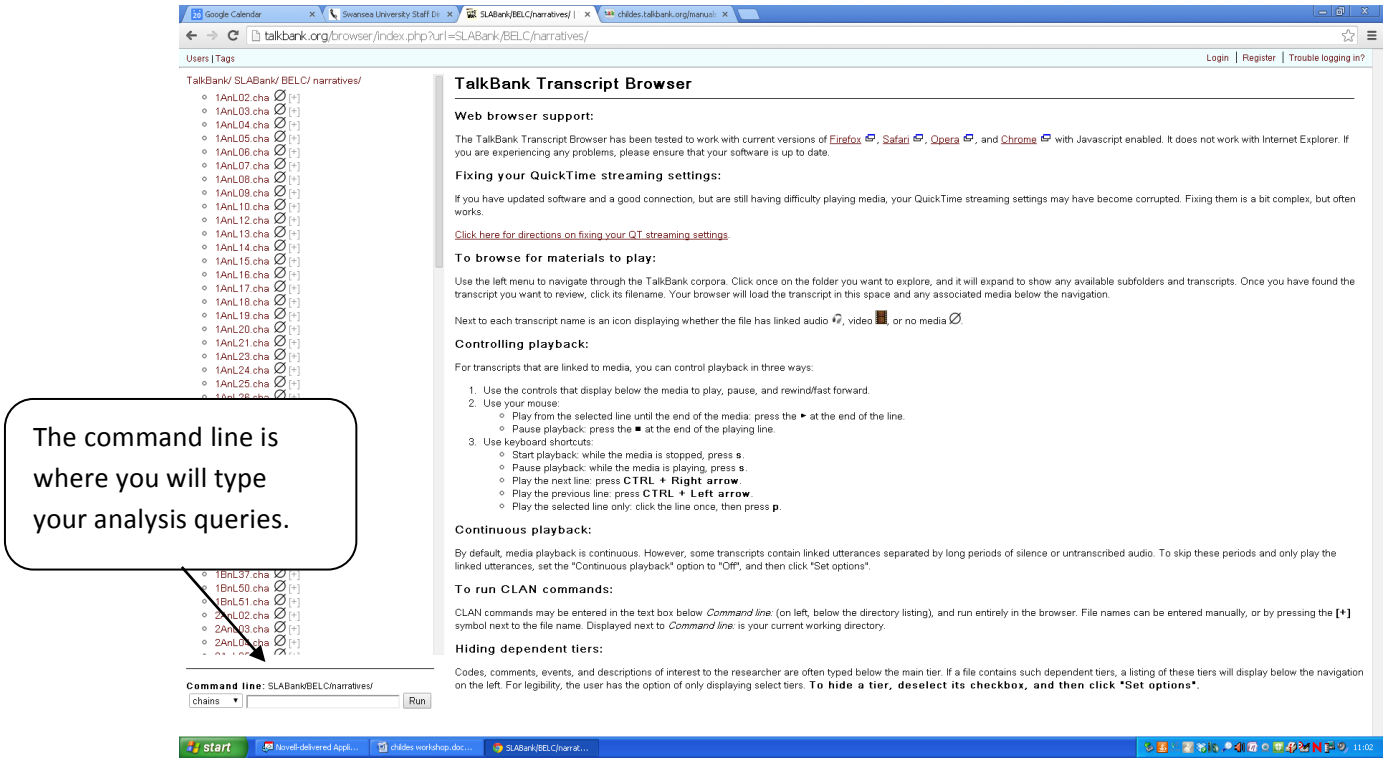
We have mentioned above, a number of switches (e.g. +f, +t, etc) which are used to limit the analyses/searches. There are more switches available. These are listed from page 119 of the CLAN programs manual.

Command	Page	Function
CHAINS	49	Tracks sequences of interactional codes across speakers.
CHECK	53	Verifies the correct use of CHAT format.
CHIP	56	Examines parent-child repetition and expansion.
COMBO	62	Searches for complex string patterns.
COOCUR	70	Examines patterns of co-occurence between words.
DIST	71	Examines patterns of separation between speech act codes.
DSS	72	Computes the Developmental Sentence Score.
FREQ	78	Computes the frequencies of the words in a file or files.
FREQMERG	87	Combines the outputs of various runs of FREQ.
FREQPOS	88	Tracks the frequencies in various utterance positions.
GEM	89	Finds areas of text that were marked with GEM markers.
GEMFREQ	91	Computes frequencies for words inside GEM markers.
GEMLIST	92	Lists the pattern of GEM markers in a file or files.
KEYMAP	92	Lists the frequencies of codes that follow a target code.
KWAL	93	Searches for word patterns and prints the line.
MAXWD	96	Finds the longest words in a file.
MLT	98	Computes the mean length of turn.
MLU	100	Computes the mean length of utterance.
MODREP	105	Matches the child's phonology to the parental model.
PHONFREQ	108	Computes the frequency of phonemes in various positions.
RELY	110	Measures reliability across two transcriptions.
STATFREQ	111	Formats the output of FREQ for statistical analysis.
TIMEDUR	112	Uses the numbers in sonic bullets to compute overlaps.
VOCD	112	Computes the VOCD lexical diversity measure.
WDLEN	118	Computes the length of utterances in words.

Full details in the CLAN manual (warning it's 215 pages long)

6 Performing analyses on the files: on the browsable transcripts

- In the BROWSABLE TRANSCRIPTS view, there is a box at the bottom of the page, which allows you to perform many different types of analysis. It is called the COMMAND LINE.
- Open the corpus folder you want to work on so you have a list of all the files in the box at the left of the screen. In this shot I have selected BELC and then narratives.
- Once you have decided which analysis you want to run, then you need to tell the command line what to do. From the drop-down menu, select your analysis.
- In the window, then either type in the name of the file you want to run the analysis on. Alternatively, you can ask it to run on all the files by typing *.cha
    - * is called the wildcard and you can use it to replace parts of the filename.
    - For example, if you wanted all the participants who are in the beginner group (who have 1 at the beginning of the filename) you could type 1*.cha.
    - Another example, this corpus has coded the age that the learner started learning English by using a letter (you can find full details about this in the SLABank.doc manual). So if you wanted to track progress over time, you might want to compare the A learners at the different times (time 1, time 2 etc). So you could type *A*.cha



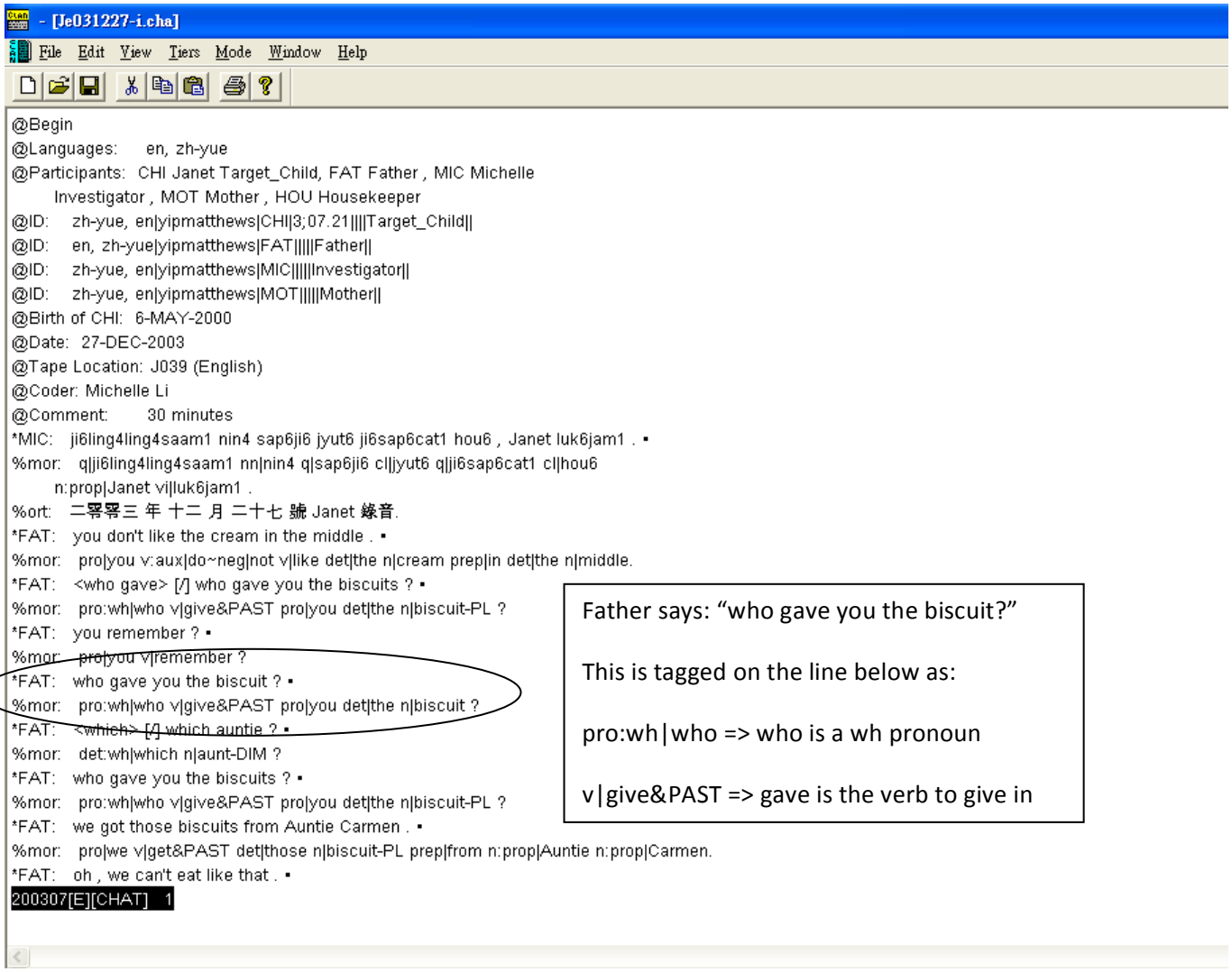
- Many of the files have been coded morpho-syntactically (i.e. they have tagged the parts of speech). These are on a dependent %mor tier. For some analysis you don't want it to calculate the results on the %mor tier but on the main tier (starts with the participant code, e.g. *PAR). Therefore, you need to tell the command line not to look at that tier. You do this by using a +/- then t for tier and then the name of the tier, e.g. %mor.
    - This gives you–t%mor.
- The command line will also automatically perform the analysis on all the speakers in the file(s). This can lead to really long output files with extra information about the investigator that you might not be interested in. You can tell the command line to only look at particular people. You do this using +t. So if you only want to look at the participant you would use:
    - +t*PAR **It is important that you use the right code.**

7 Morphosyntactic (parts-of-speech) tagging

This is a complex subject which can only be explained in outline here. Full details of the process are explained in the CLAN programs manual.

Within CLAN, the MOR program can analyse a transcript and provide a list of the morphological properties of each word class it identifies. For example, it can tag words as verbs, and provide additional information on tense, person and number. For nouns, it provides information of gender, number, etc.

The program does this by adding a dependent tier (so-called MOR tier), headed %mor:. For example:



The program has two main components: a parser (which works similarly for all languages) and a language-specific lexicon. To this day, lexicons are available in the following languages: Cantonese, Chinese, Dutch, English, French, German, Hebrew, Japanese, Italian and Spanish. Others are currently being developed.