Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion (ISSN 1617-1837)
Ausgabe 11 (2010), Seite 154-173 (www.gespraechsforschung-ozs.de)

Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository

Brian MacWhinney / Johannes Wagner

1. Introduction

To get a flavor of CLAN's potential, we advise the reader to test the output of the program by looking at the TalkBank database on the net.[1] We suggest that you

- go to www.talkbank.org,
- open the *Browsable Database* folder,
- proceed to *CABank*, and then
- open the folder *Examples* in your browser.

When you open one of the files in the folder and the transcription has come up on your screen, click on a transcript line and the browser will play sound or video while highlighting the transcript line by line. Note that it is not necessary to install and run CLAN to do play transcriptions and to work with the data.

However, to produce transcriptions, the CLAN program has to be installed. It currently runs under Windows and Macintosh, and files are movable across the platforms. In the remainder of this paper we will describe 7 different issues related to the transcription and analysis process based on the CLAN tools developed by Leonid Spektor at Carnegie Mellon[2] University:

- Recording and Data formats: This process involves the use of camcorders and audio recorders to record interactions. Because the interactional contexts involved vary widely, a variety of different recording methods are required.

---

1 You will need a recent version of a modern browser - i.e. Firefox, Chrome, Opera, or Safari.

2 CLAN, TalkBank and CHILDES Projects have received working support from NIH and NSF.

- Transcription: When transcribing, the researcher listens to the recording and produces a transcript. The preferred practices of transcribing vary in the research community and, consequently, the CLAN editor supports several styles of transcribing. The transcription is directly linked to the data file.
- Analysis: Once a basic transcription is complete, a researcher will work on the data to identify interactional phenomena to be studied in detail. The researcher may then wish to add additional features, codes, and annotations to the transcript.
- Searching: In this operation, a researcher attempts to locate specific interactional phenomena across multiple transcripts. Searching may focus on activities, keywords, phrases, or codes. It is also possible to track the co-occurrences between words and conversational features such as overlap, interruption, loudness or whispering.
- Collections: To further support searching and detailed analysis, researchers build up collections of transcripts or transcript fragments that illustrate particular interactional phenomena.
- Publication: Researchers typically use digital editors like Microsoft Word to compose articles for publication. These articles will often include segments from transcripts.
- Sharing: Once an interaction has been transcribed, the researcher may wish to share both the transcript and the media in a research project, with students and colleagues.

2. CLAN Overview

CLAN has two parts: the programs and the editor. The programs include 24 command-line analysis and search programs, 7 programs for morpho-syntactic analysis, and 35 utility programs for data reformatting (MacWhinney 2000). We will discuss the programs further in the sections 5 and 6 on *Ancillary Analysis* and *Searching*. The file format in CLAN is called CHAT and all files are saved as xxx.cha.

CLAN has existed for more than two decades and has a large user community in several fields. The program is developed according to needs suggested by its users. CLAN uses a sweeping update system. New versions with new, expanded and de-bugged features are continuously uploaded on the CHILDES website. All changes are documented in the CLAN manual which can be downloaded from http://childes.psy.cmu.edu/clan.

For transcription and playback, the relevant part of CLAN is the editor. The editor uses many of the same conventions as Microsoft Word. However, unlike Word, it allows the researcher to link individual segments of the transcript directly to the audio or video media. This form of linkage to the media is crucial in terms of allowing users to playback transcriptions to verify their accuracy. The idea of linking transcripts directly to media was introduced to systems like CLAN, COALA, and SyncWriter in the early 1990s. Since then, new editors such as ELAN, Transcriber, ANVIL, EXMARaLDA, Praat, Phon, and TransAna have all

incorporated this same feature. In this regard, these digital transcription editors differ markedly from Microsoft Word, since Word does not allow the user to link the transcript to the media.

CLAN has four major features that recommend it to research on interaction. These strengths include:

- Flexible playback and transcription: CLAN provides several methods for transcribing data and for linking transcripts to audio and video files, which we will discuss in detail below.
- Full structuring: A major strength of CLAN is that it provides a complete system for converting the Jeffersonian transcription conventions to an unambiguous and semantically accurate format. This format is based on the XML coding scheme that is slowly replacing HTML as the basic method for computerized structuring of documents and data on the web. The important thing for our current purposes is that, using additional facilities, XML can be extremely well structured and computationally explicit. Using these methods, CLAN provides the only method currently available for full digital structuring of CA-style transcriptions.
- Interoperability: CLAN includes utility programs to convert CHAT into the formats required for ELAN, EXMARaLDA, and Praat. Moreover, it can convert from each of these formats back to CHAT and we can show that data will go through this forward and backward conversion without any change or error. Because it is not possible to convert to and from TransAna, we do not recommend use of TransAna.
- Database linkage: CLAN relies on the CHAT format that is used throughout the TalkBank (talkbank.org) and CHILDES databases (childes.psy.cmu.edu) – the two largest databases for spoken language. Because all of these data are in CHAT, users of CLAN have good access to these databases for playback and further analysis.

The major challenges of CLAN for its users derive from its strengths. The XML schema provides an explicit and unambiguous characterization of how talk is composed of words, turns, pauses, overlaps, prosodies, and other features of a detailed transcription. These explicit definitions and formats must all be used accurately and unambiguously. New symbols and characters can be added, but this must be done in a systematic fashion throughout the database and programs. To use the system correctly, users must learn how to format each transcription symbol in precise accordance with the CHAT manual. A second challenge for users of CLAN is that the search programs currently have no user-friendly graphic user interface (GUI). Because of these two weaknesses, initial learning of CLAN can require some days. By comparison, users can learn a relatively unstructured system as TransAna in a day or less. However, the transcripts produced by such unstructured systems are far more limited in terms of the support they provide for analysis, interoperability and data maintenance.

We will now examine in greater detail the use of CLAN for interaction analysis.

3. Digital Recording and Data Formats

Because the methods and instruments for digital recording are continually improving, we have found that it is best to document these processes through electronic documents posted on the web that can be continually updated. The reader can find our survey of the state of the art for digital audio recording on the web at http://talkbank.org/da. Similarly, our recommended methods for digital video recording can be found at http://talkbank.org/dv. Because CLAN relies on QuickTime, digitization for CLAN must use one of the many formats supported by QuickTime. For audio, these formats currently include WAV, AIFF, AIFC, and MP3. For video, they currently include MOV, MPEG1, MPEG2, and MPEG4. Within these video formats, QuickTime supports virtually all compression codecs. However, MPEG video produced before about 2000 can sometimes cause problems. In some cases, the problem involves the way in which audio was combined with the video channel using the process of MUX-ing. Also, QuickTime cannot currently play Windows AVI and the older Flash FLV format.

4. Digital Transcription and Linkage

4.1. Overview

CLAN provides four methods for transcribing audio and video.

- Transcriber Mode: This first method imitates the single key press method used in an older and less sophisticated editor, Transcriber. The user presses the F5 key to begin the process. When a media file has been identified by the program, the user hits the spacebar at the end of any unit and a bullet with hidden time values is inserted in the transcript. This bullet encodes the duration of the utterance as the time between the end of the previous utterance and the time of the pressing of the spacebar. The Transcriber Mode works best with easily segmented data, e.g. interviews and other well-ordered talk with little or no overlaps. When the first run has been completed in which the data are sliced into segments, the transcriber would listen to each bullet and transcribe it, eventually by opening Sonic Mode (see below). In Sonic Mode, inserted bullets can be adjusted at any time. The Transcriber Mode is also useful for linking an existing transcript to media.
- Walker Mode: This method imitates the machines that could rewind audiocassettes and replay through a foot pedal controller. In this new digital version of the old process, rewinding and forward progression can be controlled either from the keyboard or from an attached USB foot controller using a special device driver. The Walker Mode is useful for a ‘quick and dirty’ transcription. A transcriber can write while listening to the continuous play and replay of the same segment and ‘walk’ his/her way through the file. In Walker Mode, however, no bullets can be inserted into the transcription.
- Sonic Mode: This method focuses on use of a waveform drawn at the bottom of the editor screen for precise demarcation of utterances and insertion of be-

gin-end times. Sonic Mode is the prime mode for highly detailed transcriptions.

- Hand editing: Experienced users will find that they can also control linkage and overlap marking by directly editing the time values that are usually hidden in the time bullets. This method is particularly useful for precise annotation of overlaps.

4.2. File headers, speaker lines and dependent tiers

Each transcription files needs to be opened by a number of file header lines, which deliver crucial information to the program. File header lines begin with the @-symbol. The first line of any transcription file would be

@Begin

and, consequently, the last line of the transcript will be

@End.

The lines

2	@Languages:	da
3	@Participants:	AST Asta_Hansen Adult, LIS Lisa_Jensen Adult
4	@Options:	CA
5	@Media:	samfundskrise, video

follow the beginning line.

- The *@Language* line indicates the language used in the transcript (here Danish).
- The *@Participants* line indicates the speaker ID, name, and role for each speaker. In our example we haven chosen *Adult* as a default role. Role indications, language and other metadata in the header line are relevant for being able to search the database. Other metadata are implemented and explained in the CLAN manual. The *participant* line provides as well shortcuts for speaker IDs in the transcription process.
- The *@Options* line defines the strictness of the data format. The code *CA* allows use of the transcription symbols provided by CLAN (see below). These symbols do not conflict with a possible XML export. Another option would be *heritage.* The word ‘heritage’ on the @Options header line marks the fact that this transcript permits a more flexible form of CHAT. Without the *heritage* option, CHAT files must maintain strict XML compatibility. With the heritage option, XML feature checking is turned off for everything inside the utterances themselves. Only the header lines, speaker ID fields, and time markings are checked. Because they are less fully structured, heritage files ("CLAN light") will produce less consistent results during search and analysis.

The example above shows the minimal set of metadata information every file must carry. Other metadata can be indicated and are described in the CLAN manual.

Transcription lines are indicated by a *-symbol before a speaker ID. In addition to the header and speaker lines, there is a third type of line which is marked by an initial percentage symbol, as in %com. These lines, called 'dependent tiers', allow the user to insert comments, translations, codes and other information relevant to the material in the speaker line above them. The user can choose to hide dependent tiers to improve the readability of the transcript and to foreground the actual interaction, rather than commentary on the interaction.

4.3. Linkage to media files

The links to the sound or video media are inscribed in the transcription in the form of bullets. Here are some examples that were extracted directly from a CHAT transcript using the cut-and-paste function:

*Nix: ... please⌈ e, •

*Op: ⌊ Thank ↓yo. •

Notice the small dark bullet at the end of each of these two utterances. When opened, using the escape-A function, the bullet shows the link to the specific milliseconds of a sound (or video file)

*Nix: ... please⌈ e, •%snd:"nh2"_3990_4086•

*Op: ⌊ Thank ↓yo. •%snd:"nh2"_4050_4260

Clicking[3] on the bullet plays the segment referred to in the bullet. In addition to playback of individual segments, CLAN also provides a method for continuous playback of the whole transcript. You start Continuous Playback by placing your cursor at the point where you wish playback to start and then typing escape-8. CLAN then highlights each utterance as it is played and turns the pages of the transcript when necessary. You can halt this playback by a mouse click in the transcript window.

4.4. Transcriptions of audio and video files

We will now explain a bit more about how one transcribes in CLAN. For this example, we will first describe transcription using the Sonic Mode. The transcription can be done directly in the editor window. In Figure 1 (below), the transcriber has opened the sound panel and marked a segment that is played automatically and can be replayed at any time. The marked segment in the sound panel covers lines 13 and 14. Directly above the sound panel, a small window indicates location of the segment in the file and the length of the segment.

[3] CTRL + click in Windows, Apple + click in Apple's OS X system.

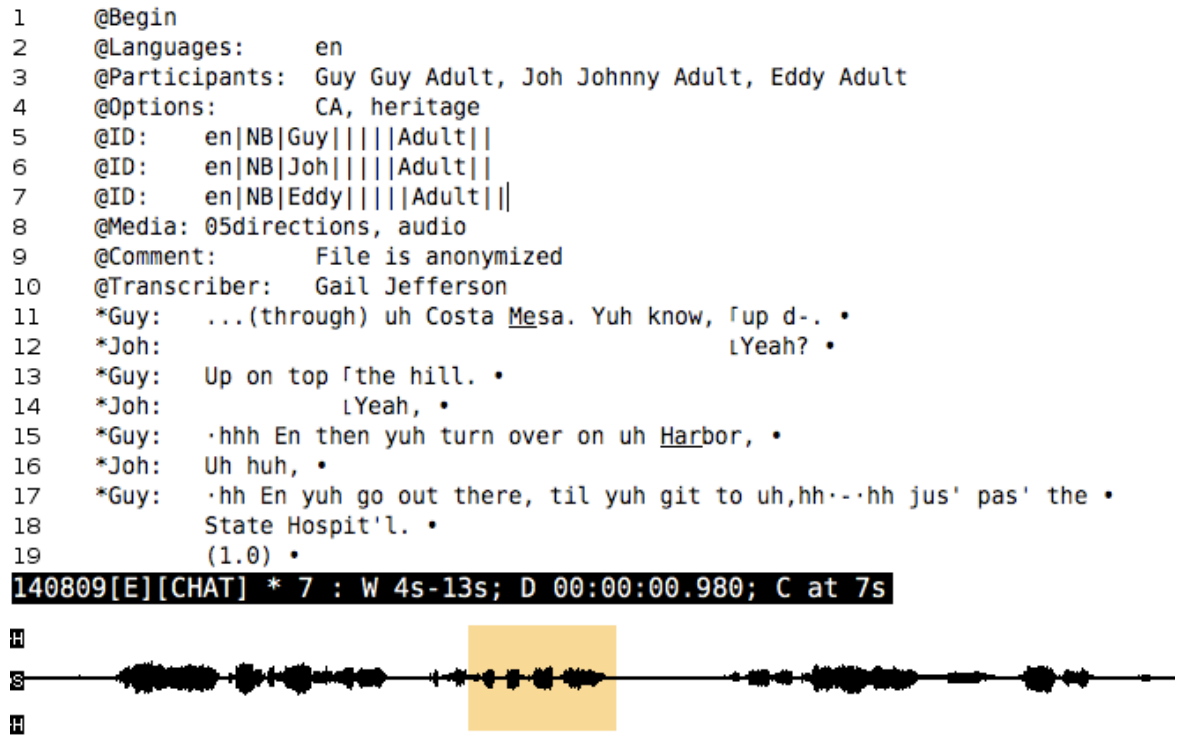

Figure 1: Audio Transcription from a Waveform

Transcription and linkage using Sonic Mode is fairly time-consuming, because this mode emphasizes precision and careful linkage. In order to speed up the process of transcription, many transcribers rely first on the Walker Mode to create an initial rough transcription. Below is a screenshot of a transcript being prepared in Walker Mode. The example shows an audio file but the walker mode works with video files as well.

You might have noticed that the transcript in figure 1 uses the standard Jeffersonian transcription symbols while the one in Figure 2 uses another set of symbols. To comply with the rigidity prescribed by XML, we have replaced a number of the Jeffersonian symbols with more iconic symbols, which only cover one single function each. An overview is found at http://talkbank.org/CABank/codes.html. As already mentioned, using the heritage option switches off strict XML checking and permits the use of all kinds of symbols. However, if this option is used, then the various search programs may yield inaccurate or incomplete results. Using the CA option complies with the new transcription symbols, which in CLAN are available through keyboard shortcuts and are implemented in the fixed-width CAfont (c.f. chapter 8 as well). A notable feature of the new transcription set is that CLAN uses four different characters to mark overlaps. These four symbols mark the beginning and end of the first speaker turn and the beginning and end of the second or other speaker turn. Using these marks and the CAFont (see below) CLAN can automatically insert the correct number of spaces for proper overlap alignment.

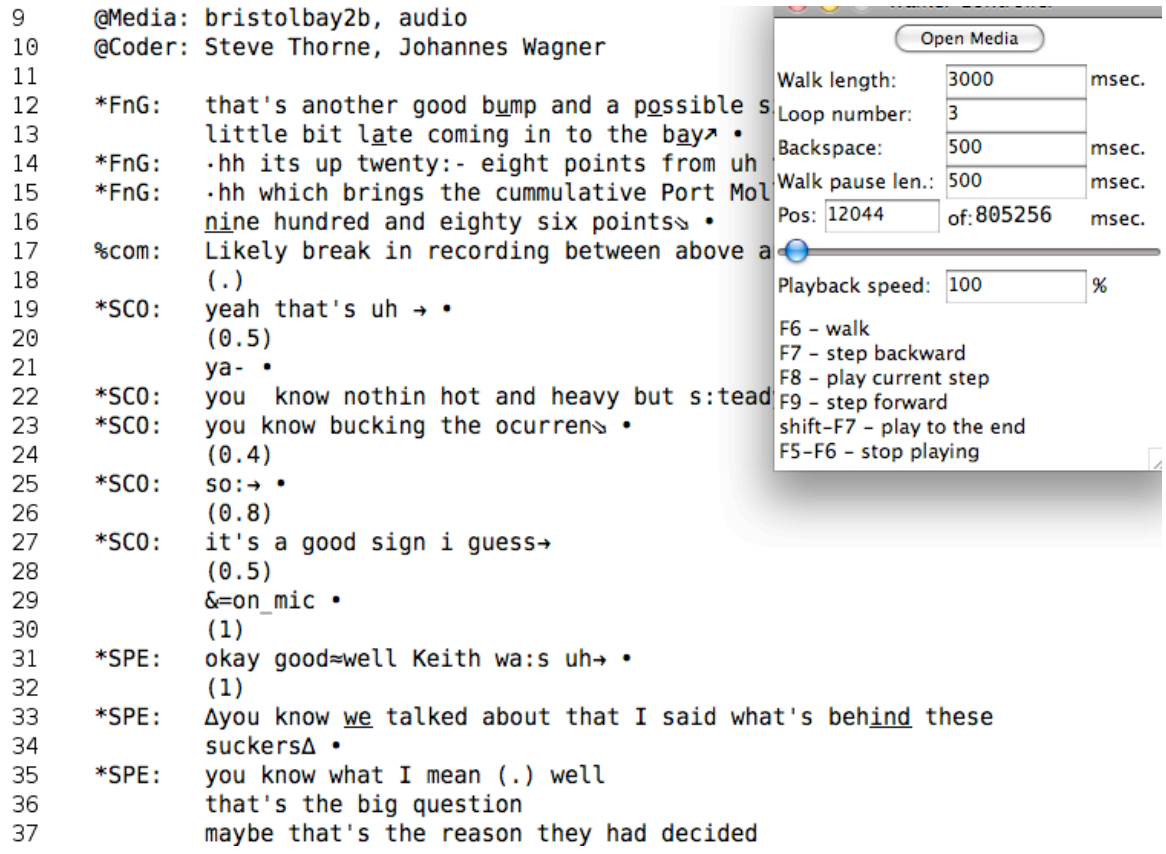

Figure 2: Audio Transcription in Walker Mode

Video transcription (MacWhinney, 2007) follows the same basic principles as audio transcription. Figure 3 illustrates how video files are transcribed in CLAN. The editor opens a QuickTime video player. On the bottom of the player window, segment size, replay and linking can be controlled. As in audio transcriptions, CLAN inserts bullets with hidden time values to encode the start and stop times for utterances. For further details of transcribing, the CLAN manual can be downloaded from http://childes.psy.cmu.edu/clan.

The different transcription modes can easily be combined. When Sonic Mode is activated in the video transcription, CLAN generates a sound file out of the video file and presents the waveform at the bottom of the screen. The video player can now be controlled through the Sonic panel.

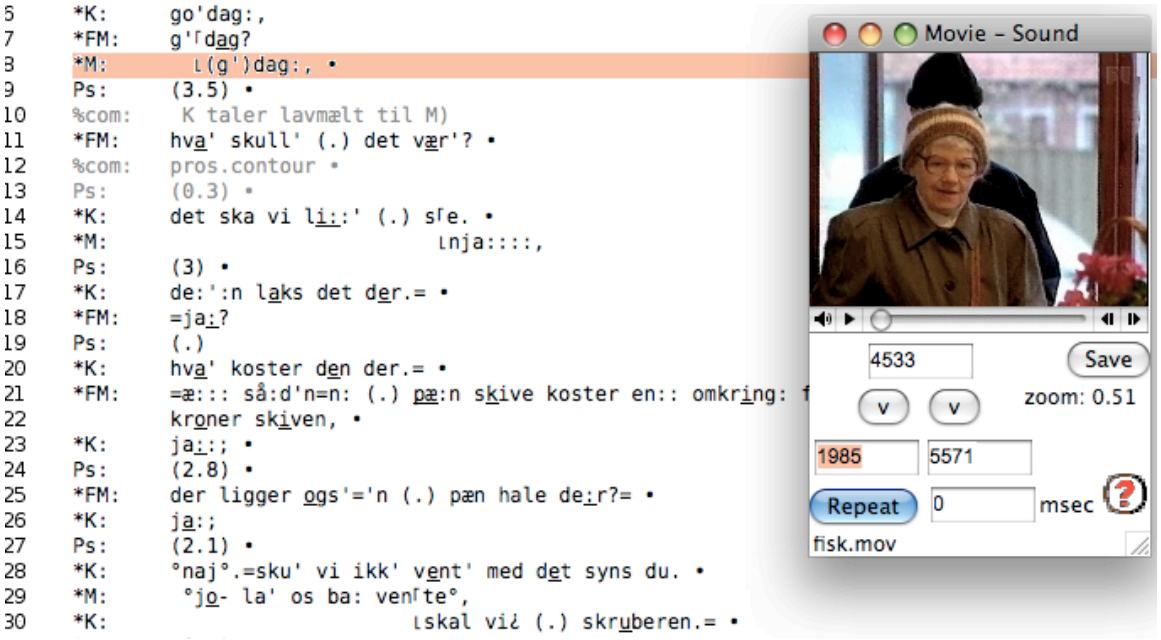

Figure 3: Video transcription

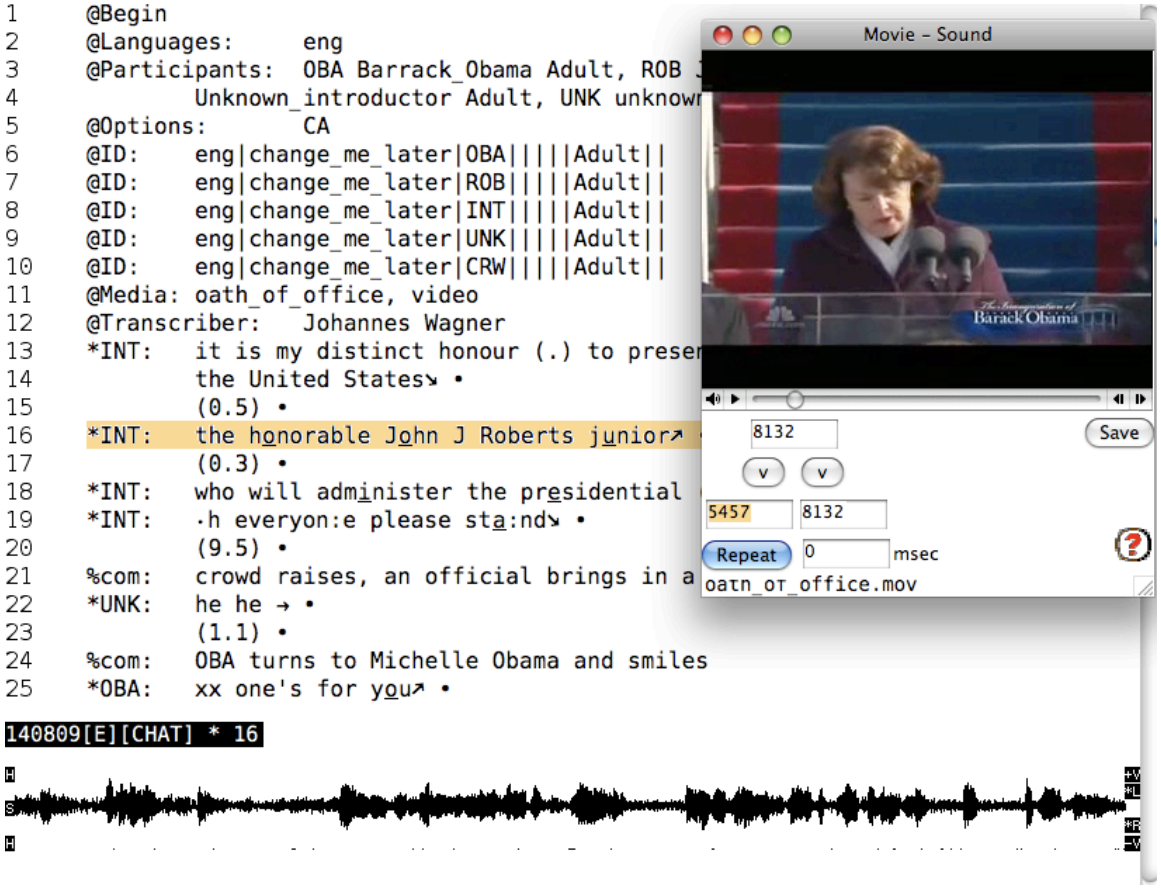

Figure 4: Video transcription with an open Sound Panel

When the transcription has been finished, CLAN offers several ways to check the technical status of the transcription file. A specific command (ESC+L) allows to debug technical errors in the file. Other programs, e.g. the INDET program which aligns overlapping segments in the transcription, indicates technical errors to be sorted out before the program can run. Finally, the CHAT2XML translator points out all features, which block the export to XML.

5. Ancillary Digital Analysis in CLAN

CLAN supports three methods for ancillary digital analysis. These include:

1. Analysis with Praat.
2. Nested files.
3. Export to Partitur editors.

- Analysis with Praat: Any bullet that refers to a sound file can be directly exported to Praat, where segmental and prosodic features of the speech can be analyzed and displayed. The Praat analyses can then be written to picture files, which can be attached to the transcript. The latter is illustrated in Figure 5. In that figure you see four open windows: The large window is the transcription proper. To the right, the ‘Movie-Sound window’ displays the video with its QuickTime controls. At the bottom of the transcription window you see the waveform for the sound that is integrated into the transcription window. To the left below, a picture window displays the prosodic contour in Praat. To produce this window, the user must open Praat concurrently with CLAN. The user then highlights a time bullet and selects the option to "Send to Sound Analyzer". Within Praat, the user selects the spectrogram option and then Praat options to highlight the prosodic contour. CLAN can send sound clips to various programs. However, most users currently rely on Praat for sound analysis.
- Nested Files: The window in the lower right of Figure 5 is a nested picture file linked to a bullet in the transcript. The user inserts bullets for nested files by selecting the "Insert Bullet" function and then navigating to locate the relevant picture or text file. Bullets for nested files can be inserted at any relevant position in the transcript. In Figure 5, the picture window is used to summarize a Praat picture. However, the method of inserting bullets for nested files can be extended for a variety of analytic purposes. For example in studies of classroom interactions, a nested graphics file can be used to display materials being written on the blackboard or forms being displayed to students on a computer screen. To maximize compatibility, the nested file should be in JPEG format. Nested text files can also support the detailed annotation of gestures. Because gestures involve so much positional, configurational, and temporal detail, transcripts become unreadable if this information is coded directly into the transcription lines or dependent tiers. To solve this problem, we create secondary files for gestural breakdown linked to the lines in the main transcript where the gesture sequences occur. Figure 6 illustrates how this works for coding of a gestural sequence in a Danish video sample. This example is

based on a detailed analysis framework constructed by Lone Laursen. In this sample, the speaker on the right, who is leaning forward, is describing an amusing situation regarding a pain therapy experiment to a group of Danish medical workers in the lunchroom. On line 24, she begins a 4-part gestural sequence that continues in lines 28 and then 31. The initial sketch of a gestural coding of this 4-part sequence is given in the left-hand window in Figure 6. This nested window opens when you click on the bullet in line 24. Within this nested window, there are further links from each of the four segments of the gesture sequence for playback of the corresponding video segments. In addition, the text window provides room for listing of the body parts, positions, directions, and classes of the individual components of the longer sequence. Moreover, it is possible to link this file to additional files and pictures.

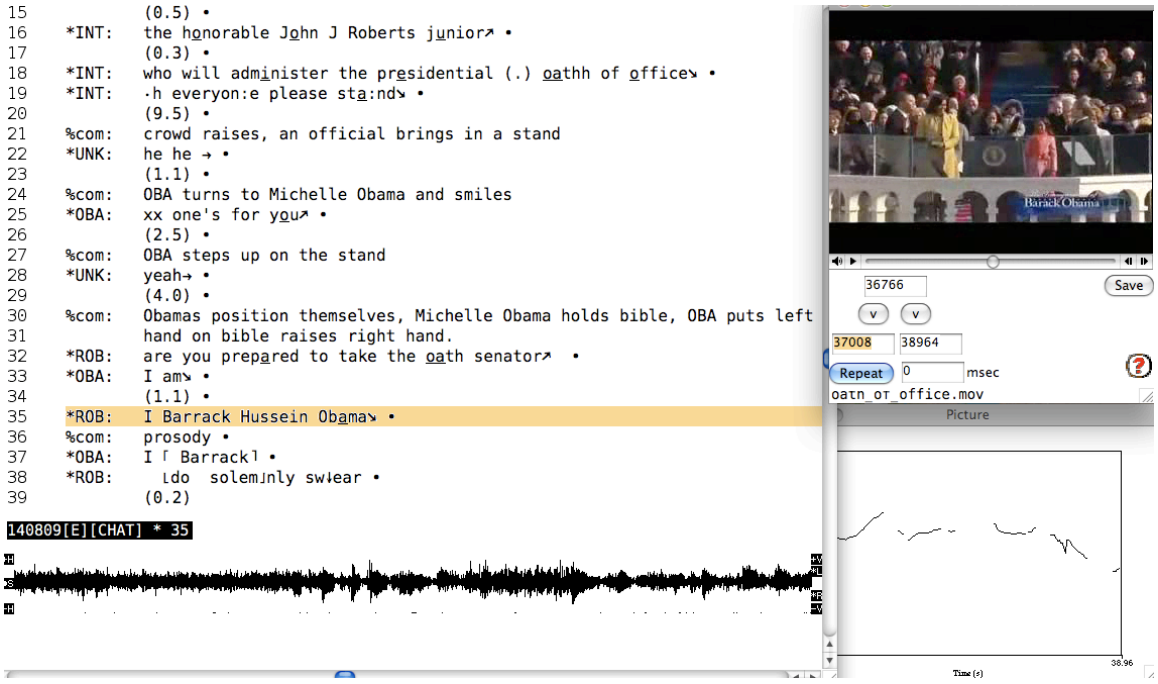

Figure 5: Pitch contour in a segment exported to Praat

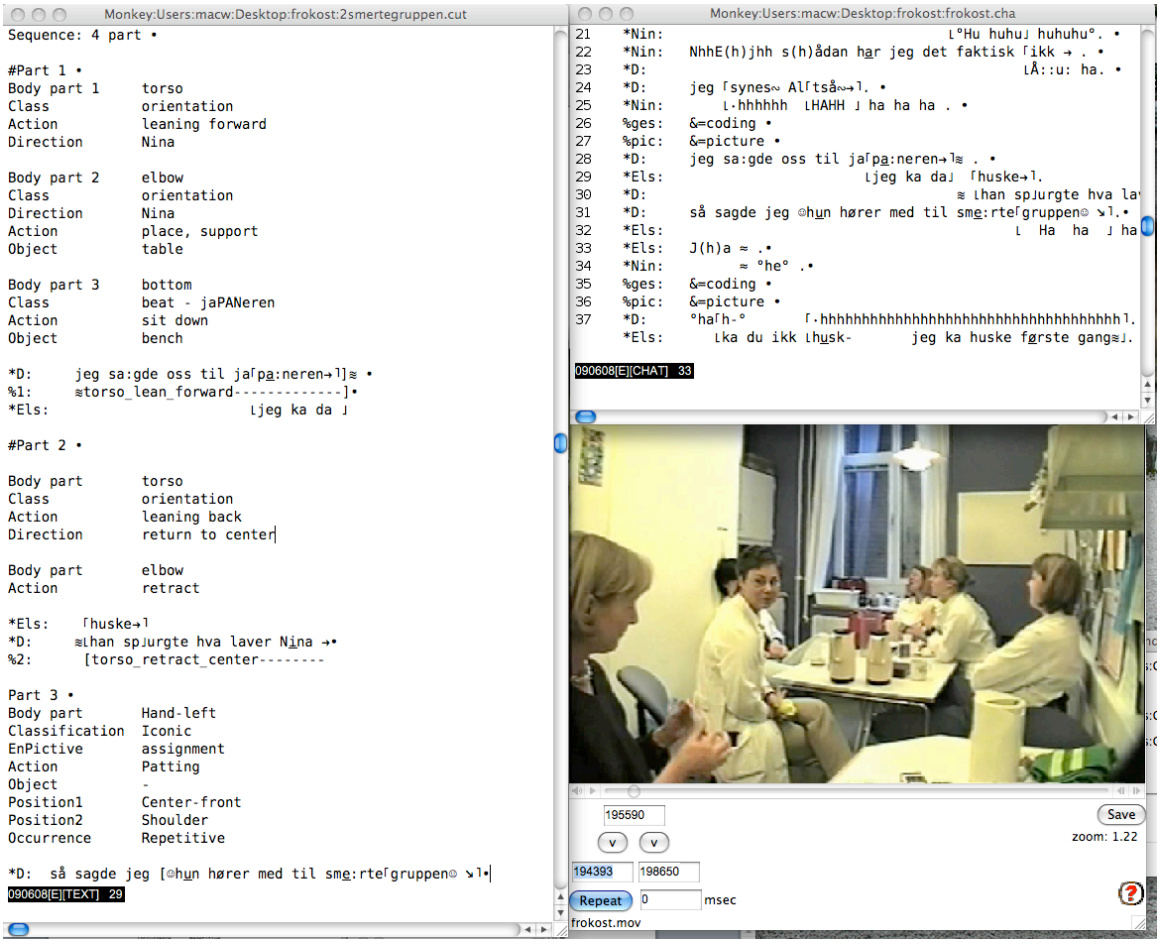

Figure 6: A transcript with video and a nested text file

- Export to Partitur Editors: During the process of analysis, researchers may find that the standard transcript format provided by CLAN fails to properly display crucial overlap features. This is because CHAT files make no explicit display of a timeline. Instead, time values are encoded in bullets and hidden from the researcher. In the Partitur or musical notation format, on the other hand, the left-to-right time scale is fundamental. In order to understand the tradeoff between these two display formats, consider how the same interaction can be displayed in these two contrasting formats. Figure 7 presents a transcript with associated video for the "mytheory" transcript contributed by Tim Koschmann to TalkBank. This is a video of a Problem-based Learning (PBL) analysis of a case of a patient with amnesia and anomia (Koschmann 1999; Koschmann/MacWhinney 2001).

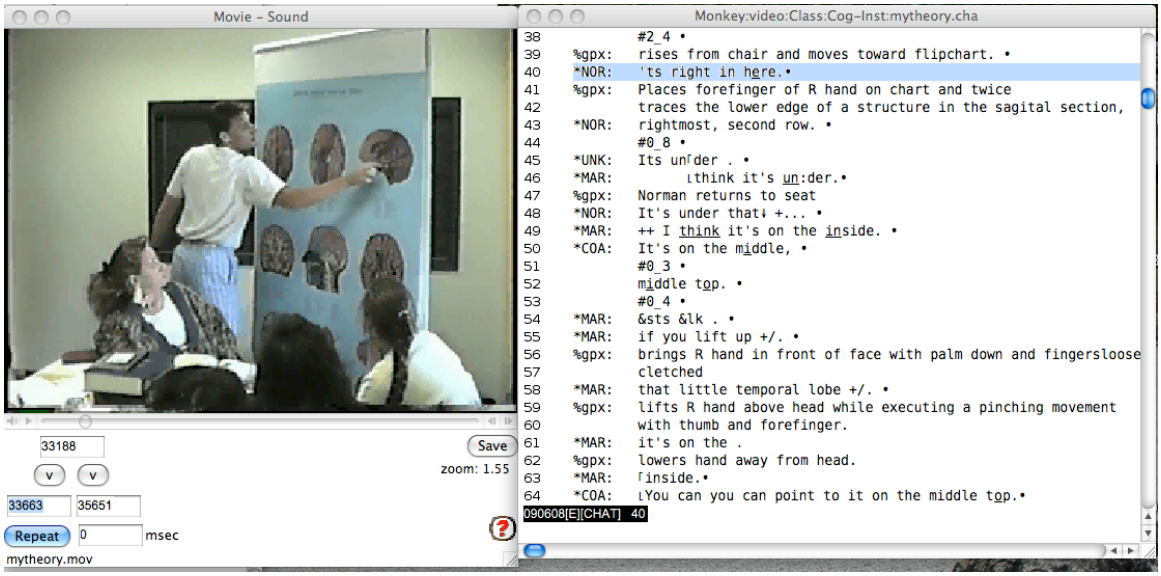

Figure 7: The MyTheory interaction in the CHAT format

The CHAT file displays the basic verbal interaction quite clearly, but it fails to encode the precise synchrony between Norman's finger pointing and his utterance in line 40.

For a clearer visual alignment of the gesture and speech, we can export from CLAN to any of the four major Partitur editors: ELAN, EXMARaLDA, Praat, and Anvil. Taking ELAN as an example, the export command is "chat2elan +e.mov mytheory.cha. The output file called mytheory.eaf can then be opened by ELAN. The resultant display of mytheory.eaf inside ELAN is given in Figure 8. In this Partitur display, we can see that Norman's action of pointing at the sagittal brain section in the upper right of the chart (transcribed in %gpx@NOR) is largely synchronous with the utterance 'ts right in here' (transcribed in *NOR). If we then wish to align the pieces of this utterance more accurately inside ELAN, we can do this by dividing the utterance and the gesture into smaller segments that we can then align. We can use QuickTime to back and forth through the video frame by frame to achieve the closest possible alignment. When we are finished with this additional analytic work, we can then use ELAN2CHAT to reformat the data back into CHAT.

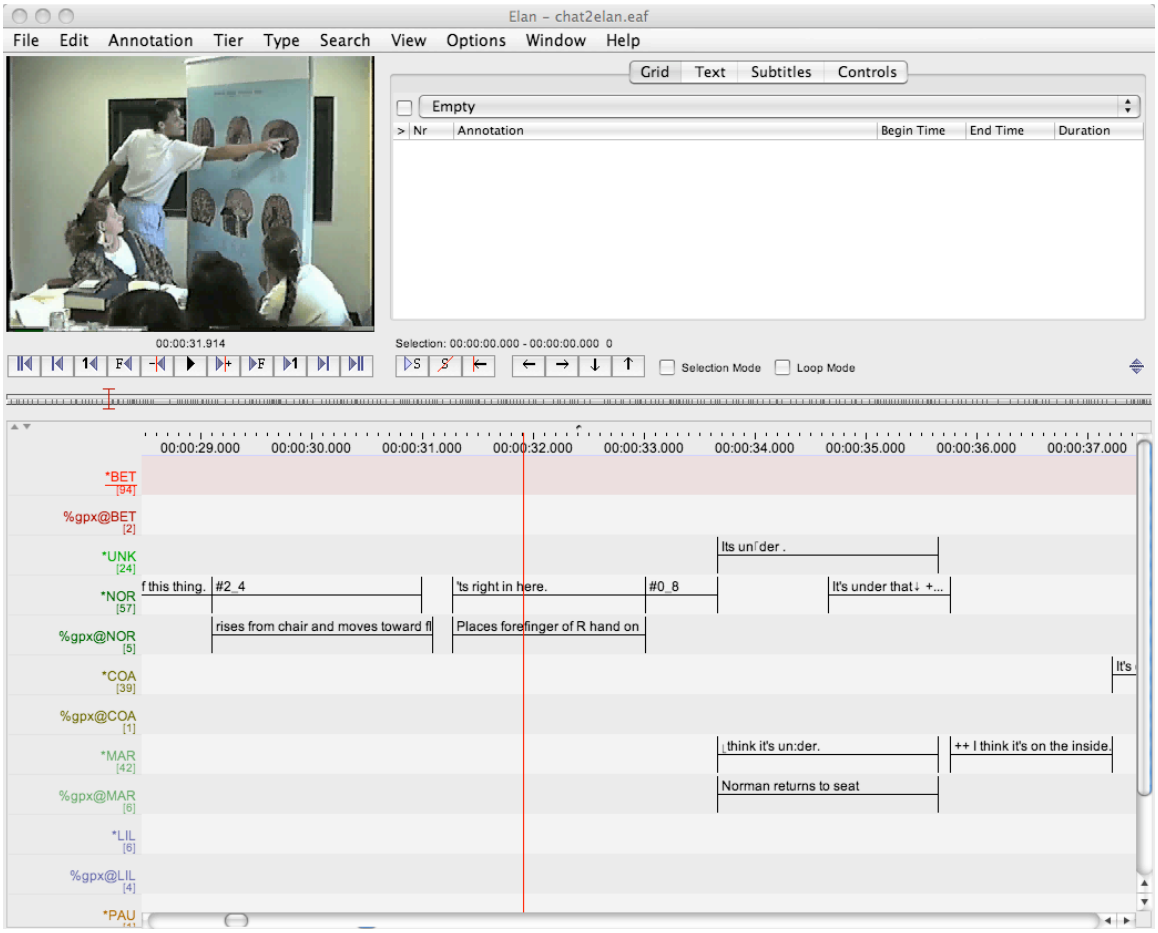

Figure 8: The MyTheory interaction in the Elan format

6. Searching in CLAN

CLAN provides 24 programs for searching and analysis. However, some of these programs, such as those that focus on morphosyntactic analysis are largely irrelevant for interaction analysis. The CLAN programs of most interest to interaction analysis include CHIP, COMBO, GEM, KWAL, and TIMEDUR. COMBO and KWAL allow users to search for all types of word and symbol combinations. The output is sent to a summary file with each match on its own line. The user can then triple-click the match and directly open up the original file and play back the interaction from the file. We are currently engaged in a process of extending and refocusing the CLAN search facilities in programs such as KWAL and COMBO to provide searches that are more fully tailored to the interests of interaction research. This extended search facility will focus on features such as turn length, pause length, overlap duration, and tone movements. Using regular expression (RegEx) pattern matching, it will be possible to match these features with each other and with particular lexical items. Here is a list of the possible search types we intend to support:

- Conversational markers: Words like *well*, *still*, or *yet* and sequences like *yes but* play important roles in marking alignment, sequencing, and topic shifting. FREQ, KWAL, and COMBO make it easy to search for these forms. To the

extent that these words are transcribed in eye-dialect, it is important to include each possible variant. To facilitate this, CLAN automatically filters out words from marks such as colons for lengthening or up arrows for pitch jump. For example, CLAN recognizes the form °↑ye:ah° (with marks for lengthening, high pitch, and soft voice) as just *yeah.* CLAN can also be set to ignore capitalization, so that words are located with and without emphasis.

- Interjections, fillers, and other words: Searches for fillers like *um* or interjections like *WOO* or *uh-hmm* are fundamentally no different from searches for other word forms. In general, CLAN can be tuned to locate all consistent patterns, as long as the spelling is within a specified set.
- Transcription symbols: All new transcription symbols (c.f. chapter 4, above) can be located directly.
- Feature combinations: CLAN can locate sequences of symbols, such as smile voice co-occurring with high pitch or speed-up.
- Turn position: Each of these symbols can be located in terms of their position within the turn. The relevant positions include: turn begin, TCU begin, turn end, TCU end.
- Overlap position: Words and features can be located in respect to overlap features. For example, one can check to see whether certain fillers are located close to overlaps.
- Terminations: The transcription conventions supported by CLAN specify six different utterance terminators, five of them being based on prosody. In addition, dashes can be used to mark word break-off. These various markers can be located individually or in the context of other features and overlaps.
- Pause duration: CLAN can track and profile the frequency and durations of pauses within and between utterances and turns.
- Gesture profiling: The system for gesture analysis discussed earlier is in its very initial stages of development. However, we envision methods of searching the nested text files that code gestural features using the descriptors in those files for body parts and gesture type. This type of analysis will also support study of gesture-speech synchrony.
- Speaker selections: It is also easy to combine any of the search types mentioned above with selections based on particular speakers, or – through the use of metadata in the ID lines (above) – group of speakers with defined roles. In this way, CLAN can support additional sociolinguistic analyses that can supplement interaction analyses.

It is currently possible to run all CLAN analyses over the web, using CLAN programs running on the TalkBank server. However, this facility requires users to already know the basic syntax and functionality of CLAN commands. We are also developing methods that will allow local work groups to develop their own custom search routines using XQuery searches of the TalkBank XML database. They will also be able to design their own web interfaces to provide additional support for new users. Along these lines, the ICOR group at the University of Lyon 2 has succeeded in adapting their existing CLAPI search machine (http://clapi.univ-

lyon2.fr/) to perform searches on XML compatible CLAN corpora. This search engine runs directly from web pages, using a very accessible interface. In this sense, CLAPI points out the direction for the creation of new search facilities for interactional CHAT data.

7. Collections in CLAN

Studies in interaction are often based on collections of instantiations of interactional phenomena. Typically, these instances are collected in a special folder. When the folder holds a sizable collection, the researcher would run single case analysis on all examples and identify the core examples in the collection as well as the deviant cases – which are the most interesting ones for the argument.

Cutting segments out of the master data and putting them into a separate folder sometimes runs into two practical problems. The first has to do with any changes in the master file that needs updates in those segments that have been sorted in different collections. The second problem arises if the researcher – while working on a segment in a collection – needs access to more transcription lines than are displayed in the segment, i.e. needs direct access to the master data. CLAN solves these problems by keeping the collection inside the master data and allowing search commands to instantly establish the complete collection in a separate file. For example, a student of interactions with second language learners may be interested in studying repairs, corrections, and misunderstandings. CLAN provides a simple method for building such collections called GEM-marking. To mark a specific area of a transcript as relevant for a collection of misunderstandings, the researcher would insert two header lines into the transcript, as in this example:

@Bg: misunderstanding

*Jen: the actual material of the interaction would go here.

@Eg: misunderstanding

In this example, @Bg marks the beginning of the gem and @Eg marks the end of the gem. CLAN programs like GEMLIST, GEMFREQ, and GEM can then be used to collect together all of these examples into a file. Like all of the CLAN programs, GEM can be run on hundreds or even thousands of files at once, with the output all going to a single unified new file. An important feature of such GEM output files, as well as the KWAL and COMBO output files discussed in the previous section is that they include special lines listing the original file and line number. It is then possible to triple-click on the relevant example in the output file and CLAN will then reopen the original file. At this point, if the file is linked to audio or video, the researcher can directly replay the original interaction for further fine-grained analysis of the phenomenon. Repeated use of this method allows the researcher full access to the raw data underlying each collection.

The researchers can chose as many categories for the GEMs as necessary. The GEMLIST command will at any time search the data and list all GEMs used to avoid inconsistencies in categorization.

8. Publication

When submitting papers for publication, researchers need to make sure that the publisher is able to accurately replicate the special characters, underlining, and spacing in the included conversational segments. To format these features, CLAN uses a specially designed Unicode font called CAfont. Unlike most Unicode fonts, CAFont uses fixed-width spacing, so that each character takes up the same space. This makes it easier to align overlap marks in CHAT files. CAFont font works equally well for both Windows and Macintosh and can be downloaded for free from the CHILDES website. This font can also be used in Word and other text editors. However, publishers are unlikely to be aware of the unique nature of CA-Font. Rather than placing your examples into the proper format, they may place them into a Unicode font with proportional spacing such as Arial Unicode. This will lead to a misalignment of overlaps and other features. To solve this problem, you may wish to ship the publisher the CAFont when you submit your paper. Alternatively, you can produce a screen shot of the section in the transcript and integrate it into the Word file. We have used this technique in Figures 1-8, above.

9. Digital Sharing through CLAN

One of the most exciting opportunities opened up by the digital revolution relates to the process of data sharing. In the past, data sharing involved photocopying of typewritten transcripts and copying of audiotapes and videotapes. This type of data sharing has been fundamental to interaction analysis ever since Gail Jefferson transcribed the GTS and Newport Beach data and shared her transcriptions with her colleagues. However, in the non-digital world, the copying of tapes was problematic, since it led to marked degradation in quality. With digital transcripts and media, copying and distribution is extremely easy and there is no data loss.

The fundamental idea underlying the TalkBank approach is that data should be as openly accessible as possible. To support this, the various TalkBank websites and programs allow researchers and students to download complete corpora with both transcripts and media. Because CLAN supports interoperability, these data can be viewed in a wide variety of editors and formats and can be subjected to the ancillary analyses discussed above. And, since CLAN commands can be run from web pages, the data do not need to be downloaded for inspection and searches, but can be played across the net.

Although TalkBank tries to maximize data access and interoperability, some corpora need to be password protected due to the private nature of the talk and other restrictions. For example, doctor-patient communications cannot be opened up to public viewing because of serious privacy issues. In other data sets, the discussion of personal topics may make full data-sharing inappropriate. However, even for these protected corpora, some access to the conversations is possible if data owners agree to run searches for other users on the protected data. This procedure can provide useful information about the frequency of certain phenomena without giving external users access to the data themselves.

To promote deepening of the empirical basis especially of conversation-analytic research, a publicly accessible CABank has been established as part of TalkBank. This corpus is being continually expanded. Currently, it includes these

corpora: LDC's CallFriend, CMU student conversations, Gulf War radio talkshow discussions, Gail Jefferson's Watergate and Newport Beach transcriptions, Curtis LeBaron's *Journal of Communication* samples, the MOVIN database, the Sakura collection of videotaped student discussions from Susanne Miyata, the Danish SamtaleBank corpus, the Santa Barbara Corpus of Spoken American English, the Stuttgart Corpus of Spoken English, and some samples of narratives in Yiddish.

Detailed transcriptions are extremely labor intensive. When completed and made available, however, they can be extremely useful for a broad research community. Publicly available data will eventually lead to an overall improvement in research methodology across fields. For example, language teaching, be it in the first or in second language, is extremely text-based in its tradition, goals, training methods, and assessment procedures (Firth/Wagner 1997). This skewed emphasis is caused, at least in part, by the fact that access to authentic spoken language is scarce and difficult. Easy access to multimedia recordings of authentic target language interactions linked to carefully produced transcripts could radically improve teacher education, teaching materials, and student projects. Because it can produce such resources, interaction analysis has a great deal to offer to allied fields in many areas of the social sciences. However, to achieve this potential, researchers must join together in the creation of large databases that support full open access. We therefore invite readers to discuss possibilities for further data sharing with the authors of this paper.

10. Priorities for Future Development

The top priority for future development is the extension of the publicly available database with detailed transcriptions of data in as many languages as possible. However, this is a development over which we have only partial control. We have much greater control over the shaping of the digital infrastructure for CLAN. In this regard, there are nine high-level priorities:

- We need to build a user-friendly interface to the CLAN programs. This interface should have the same shape both locally and across the web. It should be designed in a way that allows programmers at local facilities to build custom search interfaces for local research groups.
- We need further to develop search possibilities either inside or outside CLAN that are tailored better to the needs of the research community than existing search engines.
- We need to explore methods for speeding up the initial transcription process using the computer to "presegment" the transcript by detecting pauses and perhaps overlaps in the waveform.
- Web-based versions of CLAN need to support methods for "collaborative commentary" on transcripts (MacWhinney et al. 2004).
- CLAN needs to have a separate graphic utility for allowing users to enter participant names, roles, and ages without directly typing into CHAT files.
- The CLAN menus should be made open to internationalization, so that users can see the commands in their native languages.

`

11. Summary

CLAN is a transcription editor with a large array of functions. CLAN supports several way of transcribing audio and video data, allows searching, exporting and importing data and, as we have shown in the introduction, can be used over the web. CLAN can link several kinds of documents to a transcription: data segments, pictures, texts and other transcriptions. The CLAN manual documents the variety of function in CLAN, several of which we not even have shown here (morpho-syntactic analysis, coding). CLAN is constantly developed further and new features are documented in the manual.

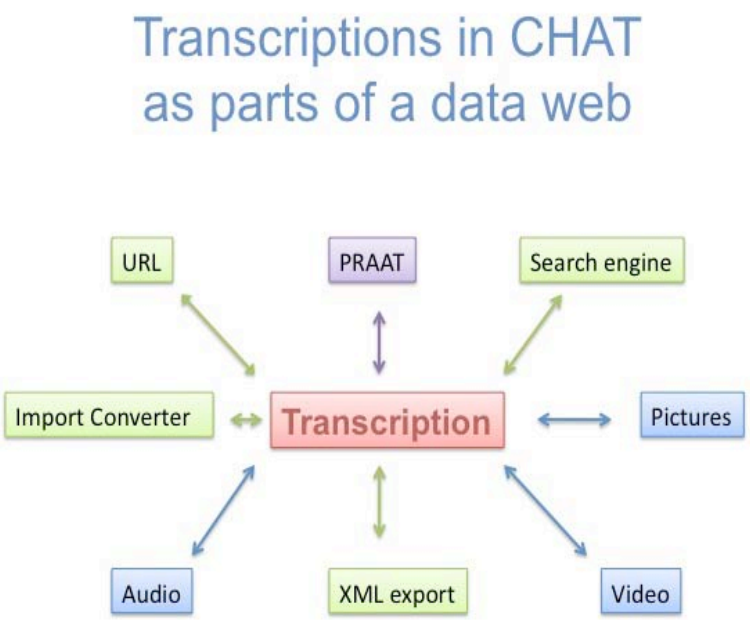

Figure 9: An overview of CLAN functions

12. Useful links

http://childes.psy.cmu.edu is the homepage of the CHILDES project within CLAN has been developed and maintained. Apart from the program itself, other software resources are available here, as well as the manuals for the program.

http://www.talkbank.org gives access to a large number of resources (editors, manuals...). TalkBank offers a very large number of digital corpora, among them detailed transcriptions in different languages. Talkbank hosts transcriptions done by Gail Jefferson.

http://www.conversation-analysis.net is the webpage of the Danish MOVIN network. It links directly to the Danish database on interaction data, http://samtalebanken.hum.sdu.dk/.

http://clapi.univ-lyon2.fr hosts sociolinguistic and interactional data with a GUI search machine.

13. References

Firth, Alan / Wagner, Johannes (1997): On discourse, communication, and (some) fundamental concepts in Second Language Acquisition research. In: Modern Language Journal 81, 285-300.

Koschmann, Timothy (1999): Special Issue: Meaning making. In: Discourse Processes 27(2), 98-167.

Koschmann, Timothy / MacWhinney, Brian (2001): Opening up the black box: Why we need a PBL TalkBank. In: Teaching and Learning in Medicine 13, 145-147.

MacWhinney, Brian (2000): The CHILDES Project: Tools for Analyzing Talk. Mahwah: Lawrence Erlbaum Associates.

MacWhinney, Brian (2007): Opening up video databases to collaborative commentary. In: Goldman, Ricki / Pea, Roy / Barron, Brigid / Derry, Sharon (eds.), Video research in the learning sciences. Mahwah: Lawrence Erlbaum Associates, 537-546.

MacWhinney, Brian / Martell, Craig / Schmidt, Thomas / Wagner, Johannes / Wittenburg, Peter / Brugman, Hennie et al. (2004): Collaborative commentary: Opening up spoken language databases. In: LREC, 11-15.

Schegloff, Emanuel A. (1987): Analyzing single episodes of interaction: An exercise in Conversation Analysis. In: Social Psychology Quarterly 50, 101-114.

Prof. Brian MacWhinney
Carnegie Mellon University &
University of Southern Denmark
macw@cmu.edu

Prof. Johannes Wagner
University of Southern Denmark
jwa@sitkom.sdu.dk

Veröffentlicht am 6.10.2010