

Computerized Language Analysis (CLAN) from The CHILDES Project

Reviewed by FELICITY MEAKINS, *University of Melbourne*

CLAN is an annotation¹ and statistical analysis tool that has a large community of users in the fields of first language acquisition and conversation analysis. It is also a potentially useful tool for the documentation of endangered languages, and offers some functions that are not provided by more popular software, such as ELAN. However, since it was not developed specifically for language documentation, it lacks a number of essential features. Drawing on personal experience, I discuss the pros and cons of this software with respect to its use as a tool for the documentation of endangered languages. It must be noted that any criticisms I have of the software are not intended as commentary on its adequacy as a tool for studies of language acquisition or conversation analysis in well-documented languages.

I began using CLAN while a research assistant with the Aboriginal Child Language project (ACLA²). This project is based at the University of Melbourne, but also has ties with the University of Sydney and the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). The aim of the ACLA project has been to investigate the kind of language input children receive in remote Australian Aboriginal communities. We have been documenting the types of language environments in which children acquire language and the implications for language change. Four Aboriginal communities have been involved in the project, each demonstrating a different degree of language shift and loss. The language environments range from Lajamanu in Central Australia, where Warlpiri people continue to speak their traditional language in conjunction with a mixed language, Light Warlpiri (O'Shannessy 2005), to Yakanarra in the Kimberley region of Western Australia, where Walmajarri people are now predominantly speakers of an English-lexifier creole language, Kriol. My own work has been based in Kalkaringi, a Gurindji community in northern Australia, where a mixed language, Gurindji Kriol, has become the main language of the community, with Gurindji now spoken only by older people (McConvell and Meakins 2005). Although the traditional languages of the communities are relatively well described, little work has been done on the contact languages, and, consequently, the documentation of these languages has formed one arm of the ACLA project. In this respect, our project is situated within both the fields of child language studies and language documentation.

Choosing a language annotation tool for the ACLA project was not a straightforward exercise, partly because our wish list outstripped the functionality of any software available when our project started in 2002. We required a tool that would allow us to transcribe and annotate video recordings, and link the subsequent transcripts with the video files. We also wanted to be able to code our transcripts for various features in order to perform relatively

1 Here I follow Bird and Liberman 2001 and use the term “annotation” to refer to any information associated with a communicative event. It may include the transcription itself, as well as ethnographic information, the physical context, and paralinguistic and extralinguistic cues such as prosody and gesture.

2 <http://www.linguistics.unimelb.edu.au/research/projects/ACLA/index.html>

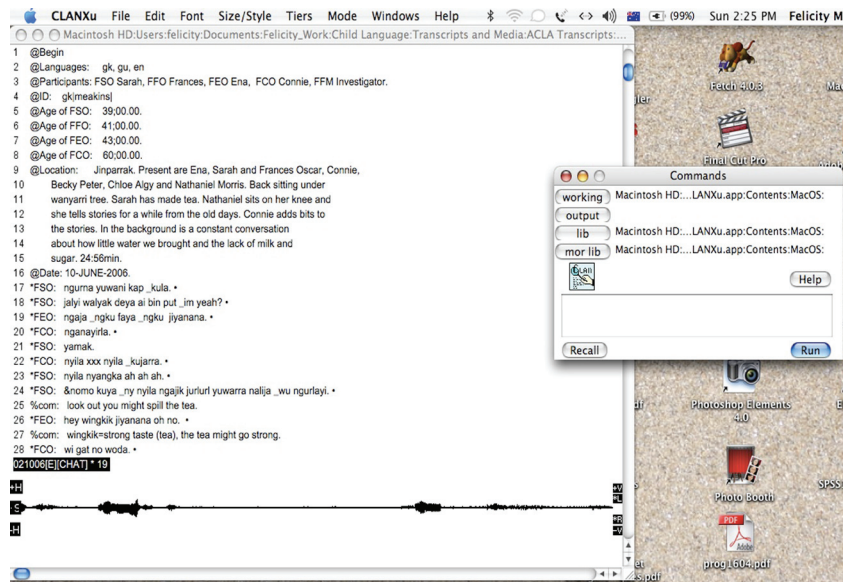
sophisticated combinatory searches and statistical analyses. Finally, it was important that the software we chose would run with few problems on a Macintosh computer and had developers who were committed to the performance of the program on this platform. At that time, Transcriber did not offer an option to transcribe video files, leaving CLAN and ELAN as the main choices. ELAN was an annotation tool being developed by the MPI (Nijmegen) specifically for the documentation of endangered languages. In 2002, most field linguists were working only with audio, using a combination of Transcriber and Shoebox to annotate recordings. Linguists who were using ELAN were reporting the usual sorts of bugs found with new software, many of which have since been remedied. At this stage, CLAN had had a longer history of use and development and was more stable, particularly on Macintoshes. CLAN also provided a powerful set of statistical software, and the MPI (Nijmegen) offered us some excellent technical support in developing the lexicon files that are necessary when using these statistical tools. Thus we opted to use CLAN; however, it must be noted that some of the reasons we chose CLAN in favor of ELAN no longer apply.

CLAN was developed in 1984 by Brian MacWhinney (Carnegie Mellon University) as a part of the CHILDES (Child Language Data Exchange System) project. The main goal of this project was to build a large database of homogeneously formatted, accessible, and analyzable transcripts from a variety of languages, specifically to address research questions in the area of child language studies. Currently, there are around 130 corpora in the CHILDES database, covering a range of languages including English, German, Afrikaans, Catalan, French, Japanese, and Cantonese. Access to these corpora is via the web and relatively free of restrictions. CLAN itself consists of two main components: the annotation software, CHAT, and the associated statistical package, also called CLAN. CHAT can be used without CLAN; however, CLAN is dependent on well-formatted CHAT transcripts.

CHAT is a relatively easy-to-begin and easy-to-use annotation tool. Information on how to begin using CHAT is given in the summary below. The only information required is a command line and a header that contains metadata about the file (see FIGURE 1). There are a number of ways to begin transcriptions. First, the media file can be played through CHAT and utterance boundaries marked before beginning to transcribe. These tasks can also be performed simultaneously, which is my preference. As I previously noted, one of the advantages of this program is being able to transcribe from video. However, I found that transcribing from video proved to be very time-consuming and difficult. In the end, I resorted to transcribing from audio only. It was easier and quicker, because the sound wave is visible, thereby providing clear cues for where utterances begin and end. I transcribed in CHAT with a separate video window open in Quicktime to provide some context when an utterance was unclear. In fact, video transcription is more time consuming than audio transcription in all annotation programs, including ELAN and Transcriber, so this is not necessarily a disadvantage when using CLAN.

A CHAT window showing the beginning of a CHAT file with header, followed by the transcription is given in Figure 1. The sound links are the bullets next to the lines of transcriptions, which can be expanded to show the actual position in the audio/video file (%snd:"FM045"_17780_20984).

FIGURE 1: A CHAT file including header and transcription



In addition to transcription, many other types of annotation can be linked to the speaker tier. CHAT offers a range of additional annotation tiers for phonetic transcriptions, speech acts, extra-linguistic information, etc. These tiers are dependent on the speaker tier. The most important tier in the CHAT file is the %mor tier. Without this field, many of the more sophisticated statistical analyses offered by CLAN are not possible. The %mor tier allows each morpheme to be coded with information about word class, allomorphy, and meaning. %mor draws this information from a lexicon file, which already exists for many languages, though undoubtedly not for the language you are documenting! (See disadvantages below).

Once a CHAT file has been created, the CLAN statistical tools provide many ways to analyze the data. For example, CLAN allows the user to perform quite complex and specific searches across CHAT files, such as COMBO searches that find combinations of word classes and morpheme types defined by the user. *FREQ* is also a commonly used tool that calculates the frequency of particular morphemes, word classes, and combinations of these. CLAN can also process information from the header, such as calculating ages of speakers from dates. The CLAN manual describes the syntax required for these kinds of searches. These functions are only a sample of the potential CLAN provides for analyzing CHAT transcripts. CLAN and CHAT have a number of features that make the software package preferable to ELAN. First, unlike in ELAN, it is easy in CLAN and CHAT to begin transcribing immediately, without the need to define tiers or fields, and to make conceptual decisions about linking tiers and how the data are to be analyzed. Tiers can be added later, allowing more flexibility for the user overall. Another feature which I see as an advantage over ELAN is the primary direction of transcription, which is horizontal in ELAN, but

vertical in CHAT. I find the horizontal transcription used in ELAN difficult to read and to scroll through, though it must be noted that ELAN now offers a secondary vertical view. CHAT uses only a vertical view, and the transcripts read much like a play. Finally, ELAN does not provide the statistical tools offered by CLAN. If one of the aims of a research project requires complex searches and analyses, then CLAN is the tool of choice.

Despite these advantages, particularly with the analysis tool, the CHAT file itself allows little complexity in relating different types of annotation. Unlike ELAN, CLAN tiers have a direct one-to-one correspondence with the speaker tier, which allows less flexibility in relating the tiers to the media file. ELAN relates tiers directly to the media files, which allows, for example, two gesture events to relate to one utterance. This type of linking is not possible with CHAT. However, perhaps the biggest problem with CLAN is that automated glossing is not possible. It is also not possible in ELAN and Transcriber; however, it is relatively unproblematic to move between Toolbox and these programs. This is not possible with CLAN. Most users of CLAN work with languages that are already well documented, such as English, German, and Japanese, where glossing is not a major concern. A glossing tier is available, but it is not automated, and this task is hugely time consuming as a result. To compound glossing problems, the tiers are not aligned by using tabs, but spaces, which means that it is difficult for the reader to align morphemes visually. The closest CLAN comes to an automated coding line is the %mor tier; however, the morpheme glosses are not particularly transparent and are also only aligned using spaces rather than tabs. Population of the %mor tier is automated; however, it requires writing both a lexicon file for the language and an associative file, and while the lexicon file is not difficult to set up, writing the associative file requires some technical expertise. An example of a line of transcript with an accompanying %mor tier is given below. The result is fairly nightmarish for the linguist who is used to Shoebox glossing. Also, it takes some time to become accustomed to separating morphemes by using a space and an underscore:

```
*FCE: nyawa _ma tu karu baisikul _jawung .
%mor: deml=this@5:nyawa suf:topl=topic@5_ma adjl=two@32:tu
      n:humanl=child@5:karu n:inanimatel=bicycle@32:baijinggul
      der:havingl=having@5:_jawung .
```

As with all annotation software, support is at hand. The online CLAN community is large. A CLAN list exists,³ and people are usually happy to help with technical problems. In particular, the main developer, Brian MacWhinney, responds quickly to queries and incorporates most feedback into ever new versions of CLAN. There is also a manual that is relatively clearly written, but it lacks page numbers, a contents page, and an index. It was intended to be used as an electronic file that can be searched using the “find” function; however, the hard-copy form is frustrating because of this lack of indexing. Despite the wealth of online support, it is advisable to seek on-the-ground support. Having someone to explain the conceptual set-up of the program and to help with trouble-shooting is hugely advantageous.

3 Subscribe to info-childes:requests@mail.talkbank.org, message: [subscribe info-childes](mailto:subscribe%20info-childes)

In general, I would not strongly recommend CLAN as a tool for documenting endangered languages. Because the program was developed to be used for child language studies and conversation, many of the functions that are essential for language documentation, such as automated glossing, are not available. Its best features are the analysis tools, and with a good macro, ELAN or Shoebox files should be transportable into CLAN to take advantage of these tools.

- Pros:** Easy to set up and use; links media and transcriptions; allows other annotation; vertical transcription; contains powerful statistical tools.
- Cons:** Tiers are linked to the speaker tier, not the media file; automatic glossing is not possible; lexicon files are difficult to write; it is not possible to move between CLAN and other software, such as ELAN and Shoebox.
- Primary function:** Annotation and linking of media file, and statistical analysis of annotations, particularly in the area of child language and conversation analysis.
- Platforms:** **PC:** CLANWin is for Windows XP/2000/NT. Windows 95, 98, or ME are no longer supported.
Mac: Users of OS10.4 above should use CLANXu. Users of Macs with older operating systems can use CLANX. OS9 no longer supported.
Unix: For Unix users, the source code for CLAN is available. UnixCLAN only provides the analysis commands of CLAN in the Unix environment. A Unix version of the CLAN editor has not yet been developed.
- Open Source:** Free and downloadable from <http://childes.psy.cmu.edu/clan/>
- Reviewed version:** CLANXu.1 (run on Macintosh OS10.4 above)
- Application size:** 10.4 MB
- Documentation:** MACWHINNEY, BRIAN. 1991. *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

SOKOLOV, JEFFREY L., and CATHERINE E. SNOW. 1994. *Handbook of research in language development using CHILDES*. Hillsdale, NJ: Lawrence Erlbaum Associates.

A good web introduction can be found at:

<http://www.let.uu.nl/~Jacqueline.vanKampen/personal/CHILDES-English/clan-programs.htm>

REFERENCES

- BIRD, STEVEN, and MARK LIBERMAN. 2001. A formal framework for linguistic annotation. *Speech Communication* 33:23–60.
- MC CONVELL, PATRICK, and FELICITY MEAKINS. 2005. Gurindji Kriol: A mixed language emerges from code-switching. *Australian Journal of Linguistics* 25(1):9–30.
- O'SHANNESY, CARMEL. 2005. Light Warlpiri: A new language. *Australian Journal of Linguistics* 25(1):31–57.

Felicity Meakins
fhm@unimelb.edu.au