INSTITUT FÜR SPRACHWISSENSCHAFT
UNIVERSITÄT ZU KÖLN

# WORKING WITH THE *CHILDES* TOOLS:

# TRANSCRIPTION, CODING AND ANALYSIS

Ursula Stephany and Conny Bast*

Table of Contents

Introduction: The *CHILDES* Project ..... 4
PART I: CHAT Transcription ..... 5
1 Theoretical Considerations ..... 5
2 MinCHAT ..... 6
3 Form of Files ..... 7
4 File Headers ..... 8
4.1 Obligatory Headers ..... 8
4.2 Constant Headers ..... 8
4.3 Changeable Headers ..... 9
5 Main Speaker Tier (Main Line) ..... 10
6 CHAT Conventions for the Main Line ..... 11
Actions without Speech ..... 11
Comments ..... 11
Completion ..... 12
Deviant Forms ..... 12
Direct Speech ..... 14
Doubtful Material ..... 14
Explanation ..... 15

* This paper is based on two unpublished manuscripts by the first author: "CHAT Transcription and Coding" (1999) and "Language Analysis Using CHILDES" (1998). We would like to thank the members of the research project "Zweitspracherwerb des Deutschen durch Lernende mit der Muttersprache Russisch" (sponsored by the Max Planck Institute for Psycholinguistics, Nijmegen, NL; Prof. Dr. Wolfgang Klein) Christian Bauer, Sonja Bertram, Sebastian Fränk, and Dr. Maria Voeikova as well as the members of the project "Der Altersfaktor im Erwerb des Deutschen als Zweitsprache" (sponsored by the German Science Foundation (DFG), Bonn, Germany) Dr. Christine Dimroth, Kristian Isringhaus, Ulrike Kösterke, Katrin Lehmann, and Achim Schumacher for helpful suggestions. Prof. Dr. Brian MacWhinney, Carnegie Mellon University, Pittsburgh, NJ, USA, helped us with some tricky CLAN searches. We especially thank Prof. Dr. Steven Gillis, University of Antwerp, Belgium, who, in collaboration with Gert Durieux, provided the original version of a lexicon-based automatic coding system for languages with richer morphologies than English, such as Modern Greek and German. We thank Sarah Downing for correcting our English.

Foreign Words ..... 15
Interjections ..... 16
Interruption ..... 16
Irrelevant Material ..... 17
Metalinguistic Material ..... 17
Omitted Parts of Words ..... 17
Omitted Words ..... 18
Onomatopoeic Forms ..... 18
Overlap ..... 18
Paralinguistic Material ..... 19
Pauses ..... 19
Phonological Fragment ..... 19
Proper Nouns and Titles ..... 20
Punctuation ..... 20
Retracing ..... 21
Scoped Symbols ..... 22
Special Passages ..... 22
Trailing off ..... 23
Unintelligible Speech ..... 23
7 Dependent Tiers ..... 23
Part II: Grammatical Coding ..... 24
1 Lexicon-Based Automatic Morphological Coding of Transcripts ..... 24
1.1 Introduction ..... 24
1.2 How to Create a Language-Specific Lexicon ..... 25
1.3 Generating the %mor Tier ..... 27
1.4 How to Enlarge Your Coded Lexicon and Code Further Files ..... 28
1.5 Creating and Working with Several Lexicons ..... 29
2 Coding Grammatical Errors and Self-Repairs ..... 29
3 Syntactic Coding of Transcripts ..... 30
Part III: CLAN Analysis ..... 32
1 Overview of the Most Widely Used CLAN Programs ..... 32
2 Analyzing Transcripts Using the CLAN Programs ..... 35
2.1 Introduction ..... 35
2.2 COMBO ..... 37

2.3 COOCCUR ..... 38
2.4 DATES ..... 38
2.5 FREQ ..... 39
2.6 GEM ..... 40
2.7 KWAL ..... 40
2.8 LINES ..... 42
2.9 MAXWD ..... 42
2.10 MLT ..... 42
2.11 MLU ..... 43
2.12 MODREP ..... 43
2.13 WDLEN ..... 43
3 Further Practical Hints for Working with the CLAN Programs ..... 44
3.1 How to Create Include Files ..... 44
3.2 Working on Groups of Data ..... 45
Part IV: Examples of Transcribed and Coded Data ..... 45
References ..... 50

## Introduction: The *CHILDES* Project

The present instructions are intended for beginners and are designed to complement the third edition of MacWhinney's reference volumes *The CHILDES Project: Tools for Analyzing Talk* (2000) by offering a pedagogical tool for transcribing, coding, and analyzing learner varieties of German or other languages.

The Child Language Data Exchange System (CHILDES) consists of Codes for the Human Analysis of Transcripts (CHAT),[1] Computerized Language Analysis (CLAN),[2] and a database.[3] There is also an online manual which includes the CHILDES bibliography, the database, and the CHAT conventions as well as the CLAN instructions.

The first three parts of this paper concern the CHAT format of transcription, grammatical coding, and analyzing transcripts by using the CLAN programs. The fourth part shows examples of transcribed and coded data.

Here are some addresses that may come in handy when working with the CHILDES tools:

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 |
| Issues relating to language acquisition: | info-childes@mail.talkbank.org |
| Subscribe (Message: subscribe info-childes): | requests@mail.talkbank.org |
| Technical questions concerning the use of CHAT/CLAN: | info-chibolts@mail.talkbank.org |
| Subscribe (Message: subscribe info-chibolts): | requests@mail.talkbank.org |

Prof. Dr. Steven **Gillis**, Dept. Linguistiek - GER, Universiteit Antwerpen - U.I.A., Universiteitsplein 1, B-2610 Antwerpen-Wilrijk;
E-mail: gillis@uia.ua.ac.be
Server of the Carnegie Mellon University computer for Europe:
http://atila-www.uia.ac.be/childes/index.html

[1] See MacWhinney 2000/I/1:§§ 1 ff. (I/1 refers to vol. I, part 1).
[2] See MacWhinney 2000/I/2:§§ 1 ff. (I/2 refers to vol. I, part 2).
[3] See MacWhinney 2000/II.

Prof. Dr. Ursula **Stephany**, Institut für Sprachwissenschaft, Universität zu Köln, D-50923 Köln;
E-mail: stephany@uni-koeln.de

Conny **Bast**, M.A., Institut für Sprachwissenschaft, Universität zu Köln, D-50923 Köln;
E-mail: Conny.Bast@uni-koeln.de
http://www.uni-koeln.de/phil-fak/ifs/projekte/d_daz-af.htm

# PART I: CHAT Transcription

## 1 Theoretical Considerations

A transcription is a written representation of spoken language. Its ultimate goal is to allow a linguistic analysis of the spoken material transcribed. Therefore, the transcript should represent what the participants actually said as closely as possible. Two problems to be avoided are overregularization and underregularization of a learner's spoken forms. This is not an easy task! The decisions which have to be made when transcribing vocal material "inevitably turn transcription into a theoretical enterprise" (MacWhinney 2000/I/1:11; also see Ochs 1979).

> ***Overnormalization***: "mapping a learner's spoken form onto an adult [or standard - U.St., C.B.] form when, in fact, there was no real correspondence" (MacWhinney 2000/I/1:11).
> ***Undernormalization***: "failing to map a learner's spoken form onto an adult form when, in fact, there is a correspondence" (MacWhinney 2000/I/1:11).

A transcription system "needs to be clear in its markings of categories, while still preserving readability and ease of transcription" (MacWhinney 2000/I/1:14). A computerized transcription must be suitable for dealing not only with a human audience, but also with the digital computer and its programs.

***Clarity***: "Every symbol used in the coding system should have some clear and definable real-world referent. The relation between the referent and the symbol should be consistent and reliable" (MacWhinney 2000/I/1:14).

***Systematicity***: "Codes, words, and symbols must be used in a consistent manner across transcripts" (MacWhinney 2000/I/1:14f.).

The mode of transcription described in the present paper follows the CHAT conventions of MacWhinney (2000/I, Part 1), with certain modifications caused by the needs of the research project "The age factor in the acquisition of German as a second language".

## 2 MinCHAT[4]

1. Write your file in your common text editor or in CED (CHILDES Editor included in CHILDES). Files written in a common text editor must be saved as (unformatted) ASCII files (Text only...) and will take the extension *.TXT or *.DOC. In order to be used with the CLAN programs, files must be saved in CED and should carry the extension *.CHA.
2. In order for the CLAN programs to work correctly, every file must begin with the @Begin line and end with the @End line.
3. Every line must be ended by a carriage return (ENTER).
4. On line 2 the participants must be introduced by @Participants: followed by a tab. This line lists three-letter codes for each participant (e.g. NAS) and the participant's name (e.g. Nastja) and role (e.g. learner).
5. Further header lines may indicate the learner's @Age: (e.g. 8;2.13), @Birth: (e.g. 08-JUN-1989), date of recording @Date: (e.g. 25-MAR-1997), @Filename:, @Situation: etc.
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. *NAS:→).

[4] See also MacWhinney 2000/I/1: chapter 3.

NB. In order to analyze several files by using a common command it may be useful to code all target speakers by using a common code (e.g. *LEA:, *CHI:, *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 part II below.)

As transcription, coding and analysis by the CLAN programs are interdependent, users should learn the basics of CLAN and try out the results of provisional transcription and coding before doing large-scale work.

## 3 Form of Files

The general form of files in CHAT format is the following:

```
@Begin
@Participants
[other headers]
*LEA:      [spoken material]
%mor:      [morphosyntactic coding]
*INT:      [spoken material]
%mor:      [morphosyntactic coding]
@End
```

# 4 File Headers[5]

## 4.1 Obligatory Headers

Obligatory headers are @Begin, @Participants, and @End. @Begin and @End must be placed on the first and last line of the transcript, respectively, and @Participants on the second line.

Headers of the speaker tier (main/speaker line) as well as dependent tiers consist of three letters followed by a colon and a tab:

*NAS: komm her .
%mor: V|kommen:IMP:SG PTL:LOC|her
%com: Nastja addresses dog

The only headers not followed by a colon and a tab are **@Begin** and **@End**.

**Participants** are identified by three-letter codes to be used on the main (speaker) line:

@Participants: DAS Dascha Learner, INT Christian Bauer Interviewer

The IDs for the two target subjects in the research project "The acquisition of German as a second language" are DAS (Dascha) and NAS (Nastja), respectively.

## 4.2 Constant Headers

Headers constant throughout the file are not obligatory, but contain useful information:

@Birth of Learner: 22-NOV-1983
@Age of Learner: 14;2.6
@Time spent in Germany: 0;0.25
@Birth of Interviewer: 19-JAN-1987
@Age of Interviewer: 11;0.10
@Date: 29-JAN-1998

[5] See MacWhinney 2000/I/1:21-30.

@Length of recording: 60 min.
@Tape: Tape 2 recto (A) Sony Digital
@CD: DASCHA nu. 1
@Filename: DAS01-03.CHA
@Language: German as a second language
@Transcriber: Christian Bauer
@Location: Fabienne's room in Cologne
@Comment: Fabienne and Dascha are left by themselves after U. Stephany, Claudia and Mascha (the children's mothers) have left the room. Both girls attend the Apostelgymnasium in Cologne; Fabienne is in the 5th and Dascha in the 8th grade.
@Warning: Transcript has not been double-checked

**Birth** and **Age of interviewer** should only be indicated if s/he is a child; otherwise indicate "adult".

**Filenames** consist of the learner's name (DAS or NAS), indication of the month after her first visit to Germany (date of first visit: 04-JAN-1998) and the running number of the file. The extension of filenames is **.cha**:

DAS08-35.CHA Dascha's 35th file recorded during her 8th month in Germany

### 4.3 Changeable Headers

Changeable headers may occur at the beginning of a file along with the constant headers or in the body of the file:

@Situation: telling the horse story, one of Hickmann's picture stories
@Activities: completing puzzle
@Room Layout: kitchen, window opposite door
@Bg: reading[6]
@Eg: reading

[6] See p.22f. below.

Headers such as @Situation or @Activities are helpful for structuring long transcripts.

## 5 Main Speaker Tier (Main Line)[7]

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

*NAS: ja.

Transcription is generally in orthographic form and lower case Latin letters. Initial capital letters are only used for proper nouns (see p. 20 below), but not for common nouns or sentence-initial letters. Names of scientific disciplines are treated as common nouns. For German, the old spelling conventions are used; however, '**ß**' is replaced by '**ss**' throughout. Numbers are spelled out ('zehn', not '10').

Phonetically deviant forms are normalized on the main tier. Their phonetic form is indicated on a %pho dependent tier in cases of clearly recognizable strong deviance (e.g. use of a glide instead of a fricative):[8]

*NAS: Wilde_Maus.
%pho: Uilde_maus

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 p. 20 below).

For the purposes of the German L2 project, morpheme boundaries are not indicated.

*INT: in Paris warst du in den ferien?
*NAS: ja, ich war dort fünf tage.

[7] See MacWhinney 2000/I/1:§§5-8.
[8] On morphophonologically deviant forms see p.12 below.

*NAS: aber wir waren mit Dascha schon ganz müde.
*NAS: wir hatten hunger, und wir haben nich(t) sehr viel gesehen.

## 6 CHAT Conventions for the Main Line (in alphabetical order)

**Actions without Speech[9] 0**

Action performed by a speaker and unaccompanied by speech is marked by zero '0'.[10] The relevant action may either be marked on the main tier or on a dependent %act tier.

*INT: warst du schon einmal in Sankt_Petersburg?
*LEA: 0 [=! lächelt].

*INT: zeig mir mal den tanz, den ihr heute in der tanzschule gelernt habt.
*DAS: 0.
%act: steht auf und tanzt

**Comments [% text]**

The exact interpretation of the speaker's intention is crucial for a correct linguistic analysis of the data. Therefore, comments should be made whenever necessary and possible. Sometimes non-participants in the interaction, who are however familiar with the extralinguistic context referred to, such as the girls' mother, may be able to help.

Whenever a participant in the interaction is not identifiable, the transcriber should make a best guess and add his/her doubt or an alternative on the %com line (e.g. *PAS: %com: vielleicht nicht Pascal, sondern Nastja). In cases where no best guess is possible, the participant in question is identified as *UNK (unknown). Placing a dollar sign '[$]' at the end of such speaker lines (after the punctuation mark and a space) facilitates their retrieval, when the transcript is to be double-checked later on.

[9] See MacWhinney 2000/I/1:70f.
[10] See also p.19 below on paralinguistic material.

Short comments consisting of a single word may be placed on the main line, for longer comments a **%com** tier should be added (MacWhinney 2000/I/1:73).

*INT: auf dem Eiffelturm warst du ?
*NAS: ja [% bestätigend].

*DAS: und meine [*] beine [*] war und [?] kaputt .
%com: Dascha hatte sich nur ein bein verletzt

**Completion:**[11] **Self-Completion + Other-Completion ++**

Self-completion is marked by '+,' and other-completion by '++' at the beginning of the line on which the completed utterance is transcribed. The utterance which is interrupted (by the speaker him-/herself or by another speaker) is marked by '+...' at the end of the line:[12]

*NAS: und dann sahen wir da einen +...
*INT: ja was?
*NAS: +, elefanten .

*NAS: und dann sahen wir da einen +...
*INT: ++ elefanten .

'+...' is also used when the speaker makes a pause without finishing his/her utterance (see p. 23 on trailing off).

**Deviant Forms**[13] **[*] [: ]**

Deviant (ungrammatical) forms are followed by an asterisk in square brackets. In cases of **(morpho)phonologically deviant forms**, the target form is added in square brackets after a colon and a blank:

*NAS: ich treffte [*] [: traf] ihn.

[11] See MacWhinney 2000/I/1:69.
[12] Also see p.16 below on interruption/self-interruption.
[13] See MacWhinney 2000/I/1:72, 78.

*DAS: ich eh@fp fahre [//] fahrden [*] [: fuhr] rad .
*DAS: und ich fallt [*] [: fiel] .

The same procedure is followed with reading errors, where the target form can be retrieved from the written text.

NB. The CLAN programs accept [*] [: ] and [: ] [*] in either order.

Indication of the target form in square brackets (e.g. [: fuhr]) is important when it comes to automatic coding of the speakers' utterances, since the clan program **MOR** will only code the forms given in brackets and not the deviant forms actually said.

Incomplete, but acceptable **colloquial forms** are not marked by an asterisk, but the (non-reduced) standard form is added in square brackets after a colon and a blank:

*INT: ham [: haben] sich (eine)n eimer geholt .

**Lexical errors** as well as errors concerning the choice of function words or the function of grammatical forms (e.g. wrong case use, tense errors, agreement errors) are marked by an asterisk, but without indication of the target form(s). Errors such as these will be further specified on the grammatical coding tier %mor (see Part II below).

*NAS: das ist eine # bibel [*] .
*INT: ein biber .

*DAS: ja und wann [*] ich war sehr klein ein hund eh@fp # eh@fp +...
*INT: hatte dich da ein hund gebissen?
*DAS: ja, war hier.
*DAS: ich bin [*] ein jahr alt.

Sometimes extralinguistic information is needed in order to decide on the type of error (e.g. a lexially based (gender) error or a grammatical error of agreement). In such cases the decisive information should be provided on the %com line if possible:

*DAS: und meine [*] beine [*] war und [?] kaputt.
%com: Dascha only hurt one of her legs

**Syntactic errors**

Syntactic errors, such as wrong word order and omitted words, are left unmarked.

*INT: wo wart ihr gestern?
*LEA: gestern wir waren im Phantasialand.

*INT: hatte dich da ein hund gebissen?
*DAS: ja, war hier.

**Direct Speech**[14] **+"**

Direct speech is marked by ‘+" ’ at the beginning of the line:

*LEA: da sagte meine schwester zu mir .
*LEA: +" gib mir mal den rotstift .
*LEA: +" ich will etwas anstreichen .

*LEA: +" gib mir mal den rotstift .
*LEA: sagte meine schwester zu mir .

**Doubtful Material**[15] **[?], [=? text]**

If there is uncertainty about what is heard on the tape, the word or group of words serving as a best guess may be marked by ‘[?]’.

*NAS: ich war im zweite etage auf [?] # Eiffelnturm [*] [: Eiffelturm] .
*NAS: ich war im zweite <etage auf> [?] # Eiffelnturm [*] [: Eiffelturm] .

When it is difficult to choose between two possible transcriptions, the alternative transcription may be enclosed in square brackets:

*DAS: wir waren im [=? in] Phantasialand .

[14] See MacWhinney 2000/I/1:68.
[15] See MacWhinney 2000/I/1:38, 73, 74.

NB. In automatic coding, **MOR** does not take [?] or [=? text] into consideration, i.e. doubtful material is transcribed as if it were not doubtful.

**Explanation**[16] **[= text]**

Brief explanations of the situation at hand may be given on the main line, longer ones on the %com tier. Context information must be reliable and may either result from independent knowledge of the situation (e.g. the transcriber has been present during the recording) or from acoustic cues given on the tape.

*DAS: so wie dieser [= sessel] hier.
*DAS: so wie dieser hier.
%com: schaut zu einem sessel im wohnzimmer

Actions accompanying utterances which are important for understanding their meaning should be indicated on the %com tier, whereas the %act tier should be reserved for activites replacing speech (see p. 11 on action without speech).

**Foreign Words** **@e, @r**

Words originating from languages other than German (e.g. English, Russian) and/or pronounced in an English or Russian way are marked by '@e' and '@r', respectively. German translations of Russian words should be provided.[17]

Examples: Saint_Petersburg@e
da@r [:=r ja]

NB. English or Russian loan words (e.g. *computer*, *chatten*, *datscha*) are not marked as foreign words.

[16] See MacWhinney 2000/I/1:71.
[17] See MacWhinney 2000/I/1:72f.

**Interjections**[18] **@i**

Interjections are marked by '@i', which is added to the interjection (e.g. hm@i, oh@i). The precise form of the interjections is not indicated: Varieties such as *ähm, ähem, mmm* are all transcribed as 'hm@i', while interjections expressing surprise, such as *aha, ah, boah, po, ho, och*, are all rendered by 'oh@i'. The form 'hey' is transcribed as 'hey@i' and it is not considered an English word.

In cases where it is easy to decide whether the interjection 'hm@i' has an interrogative, affirmative or negative function (question, agreement, refusal) it may be further specified as

| | |
|---|---|
| interrogative interjection | hm@ii ? |
| affirmative interjection ('yes') | hm@ia . |
| negative interjection ('no') | hm@in . |

**Interruption**[19] **+/.** **Self-Interruption** **+//.**

Uninvited interruptions by another speaker are marked by '**+/.**' at the end of the line. Invited interruptions for completion (prompting) should be marked by '++' at the beginning of the line (see p. 12 on completion).

*INT: wie lange +/ ?
*NAS: hm@i ich bin hier drei monate .

*NAS: ich bin heute +//.
*NAS: nein, ich muss dir etwas anderes erzählen.

[18] See MacWhinney 2000/I/1:32f.
[19] See MacWhinney 2000/I/1:67f.

**Irrelevant Material**[20] **www.**

Stretches of speech by native interlocutors which are unrelated to the dialogue need not be transcribed. This may be the case if at some point during the session some people enter the room and speak to the interviewer:

*LEA: ich glaube da kommt meine mutter .
*INT: ja vielleicht .
*MOT: www.
%add: addressing interviewer

When the participants look at journals and utter *boah* etc. for several minutes, it is sufficient to transcribe this once, then add a speaker tier with 'www.' and a %com line indicating what is going on in the non-transcribed section.

However, untranscribed material should, on the whole, be rather limited.

**Metalinguistic Material**[21] **["]**

Metalinguistic material is followed by '["]':

*INT: weisst du, was suchen ["] ist ?
*INT: <eine tote hose sein > ["] bedeutet <eine niete sein> ["] .

**Omitted Parts of Words**[22] **()**

Omitted parts of words are added in parentheses when the target form is clear.

*DAS: mein(e) [*] kusin(e) [*] Sonja .

[20] See MacWhinney 2000/I/1:36f.
[21] See MacWhinney 2000/I/1:73.
[22] See MacWhinney 2000/I/1:38, 53f.

**Omitted Words**[23]

As adding omitted words to a learner's speech is often not possible unambiguously or would require considerable analytical effort, such omissions are left unmarked, in order not to interfere with the authenticity of the learner's speech and for speed of transcription.

**Onomatopoeic Forms**[24] **@o**

Onomatopoeic forms are marked by '@o':

*INT: das ist eine flasche mit warmem wasser .
*INT: du siehst ja den dampf pff@o pff@o pff@o zum wärmen .
*NAS: das ist # eine # eh@fp +...
*INT: pieppieppiep@o pieppieppieppiep@o macht die .
*INT: pieppieppieppieppiep@o .

**Overlap**[25] **[>], [<], [<>]**

Utterances made by two speakers may overlap in time. Depending on whether the overlap follows, precedes, or both follows and precedes, '[>]', '[<]', '[<>]' are used at the end of those lines (in front of the punctuation mark) in which overlapping stretches of speech occur:

*NAS: wir fahren zum Phantasialand [>] .
*INT: kuck mal dahinten [<] !
*INT: wohin fahrt ihr [>] ?
*NAS: zum Phantasialand [<>] .
*INT: was [<] !

If the interlocutor utters an interjection while the other person continues with his/her utterance, this interjection is best placed after the completed utterance.[26]

[23] For further specification of omissions see MacWhinney 2000/I/1:38ff.
[24] See MacWhinney 2000/I/1:32f.
[25] See MacWhinney 2000/I/1:74f.

**Paralinguistic Material**[27] **[=! text]**

Paralinguistic material, such as laughing or smiling, may be marked by '=!', a space, and then a short text. Longer texts should be put on the %com tier.

*INT: weisst du was <tote hose> ["] bedeutet ?
*DAS: ja [=! lächelt] .

**Pauses**[28] **#, eh@fp**

Utterance-internal **unfilled pauses** are marked by '#' if they last for at least 2 seconds (check the duration of pauses in your sound program!); if they last for at least 8 seconds mark this by '##'.

Conversational pauses are to be distinguished from mere breath-taking. When the interviewer asks a question and the learner takes at least 5 seconds before answering, the pause is marked by '#' in front of the learner's answer.

When a speaker makes a pause of at least 2 seconds while performing some action, such as leafing through a book, this does not count as a self-interruption. It is marked by '#' and the action is indicated on the %com line:

*NAS: ich hole das mal # raus .
%com: blättert in einem buch

**Filled Pauses** (fp) are transcribed by 'eh@fp'. Two filled pauses in a sequence are marked by 'eh@fp eh@fp':

*NAS: das ist # eine # eh@fp +...

**Phonological Fragment**[29] **&**

The ampersand symbol '&' is used at the beginning of false starts:

---

[26] On retrieving overlaps see KWAL (p.42).
[27] See MacWhinney 2000/I/1:70f.
[28] See MacWhinney 2000/I/1:65f.
[29] See MacWhinney 2000/I/1:37f.

*INT: da sind wir nach &mai Nijmegen gefahren.

Strings beginning with '&' are not treated as words by the CLAN programs unless they are specifically searched for.[30]

Repairs are only marked by '[//]' if the word fragment is part of a phrase corrected as a whole; if the phrase is just repeated, this is marked by '[/]':

*NAS: und <am &f> [//] in ferien ich war <in &pa> [/] in Paris .

In cases where false starts are not corrected, the intended form is transcribed and the form actually pronounced is indicated on the %pho tier:

*DAS: nächstes mal möchte ich noch vorschlagen .
%pho: vror # schlagen

**Proper Nouns and Titles[31]**

Proper nouns of people and places as well as titles are transcribed with an initial capital letter. Multiple-word nouns should be joined by '_' in order to output them as single words in the analysis:

| | |
|---|---|
| Sankt_Petersburg | Nsync |
| Russland | Kaspisches_Meer |
| Titanic film | McDonalds |

All words should be written out: 'Sankt' (not 'St.'), 'Doktor' (not 'Dr.').

**Punctuation[32]**

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

, . ; ? ! [ ] < >

[30] See KWAL p.40.
[31] See MacWhinney 2000/I/1:41f.
[32] See MacWhinney 2000/I/1:32, 57ff., 76f. and MacWhinney 2000/I/2:135f.

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 should thus not be placed at the end of the speaker line.

NB. The CLAN programs **MLU**, **MLT**, and **COMBO** count lines terminated by a comma as separate utterances, while counting lines containing more than one clause separated by a comma as single utterances. The following text is counted by MLU and MLT as containing three utterances:

*NAS: wenn ich war in Musee_d'Orsay, wir waren mit Dascha schon ganz müde .
*NAS: und wir hatten hunger .
*NAS: und wir haben nich(t) viel gesehen .

**Retracing**[33] **[/], [//]**

Repetition without correction is marked by '[/]', repetition with correction is marked by '[//]'. If several words are repeated, they are placed in angle brackets. Several repetition marks may be used in one and the same utterance as seen in the following examples:

*LEA: ich fahre [/] fahre rad .
*LEA: <ich fahre> [/] ich fahre rad .
*LEA: <ich fahren> [//] ich eh@fp fahre [//] fahrden [*] [: fuhr] rad .

Filled pauses occurring directly in front of the repeated word(s) are placed after the retracing symbol:

*LEA: der [/] eh@fp der brief .

Non-standard forms occurring within retraced sequences are replaced by standard forms only if they are repeated.

*LEA: <ham ihre sämtlichen> [//] ham [: haben] im supermarkt zahnpasta gekauft .

[33] See MacWhinney 2000/I/1:76ff.

**Scoped Symbols**[34] **< > [ ]**

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

*LEA: <ich fahre> [/] ich fahre rad.

**Special Passages**[35] **@Bg:, @Eg:**

Special passages, such as reading a text or telling a picture story, may be marked by ‘@Bg:’ and ‘@Eg:’. Text reading is introduced by ‘@Bg: reading’ and ended by ‘@Eg: reading’, while telling a picture story is introduced by ‘@Bg: picture story’ and ended by ‘@Eg: picture story’:

*INT: lies mal vor !
@Bg: reading
*LEA: als er ins Musee_d'Orsay kam, war er schon ganz müde .
*LEA: und er hatte hunger .
*LEA: deshalb hat er nicht viel gesehen .
@Eg: reading
*INT: schön .
@Bg: picture story
*LEA: es war einmal ein junge .
*LEA: der hatte einen hund und einen frosch .
[...]
*LEA: da nahmen sie einen kleinen frosch und gingen fröhlich nach hause .
@Eg: picture story

In cases where reading or story-telling passages are interrupted by ordinary speech, the interruptions each have to be excluded from the reading and narrative gems by ‘@Bg’ and ‘@Eg’.

[34] See MacWhinney 2000/I/1:§8.
[35] See MacWhinney 2000/I/1:27f.

**Trailing off**[36] **+...**

Trailing off is marked by '+...':

*LEA: ich gehe gerne zur +...

**Unintelligible Speech**[37] **xxx**

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

## 7 Dependent Tiers[38]

With the exception of the %mor tier, 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:**

A phonetic-phonemic rendering of material deviating from standard pronunciation may be indicated on this tier, using the UNIBET conventions based on SAMPA given in MacWhinney (2000/I:133f., Part 1):

*DAS: die vögel singen.
%pho: siNg@n

[36] See MacWhinney 2000/I/1:66.
[37] See MacWhinney 2000/I/1:36.
[38] See MacWhinney 2000/I/1:§9.

This line may also be used for rendering phonic material without any obvious morphemic analysis (MacWhinney 2000/I/1:36, 86):

*DAS: yyy.
%pho: memis

NB. A %pho tier is only indicated in cases of considerable phonetic deviation from the standard pronunciation.

Although the dependent tier %pho may be included in the main line (e.g. ich traf [%pho: treffte] ihn), this is to be avoided since the CLAN programs will only output such forms together with the information [%pho: ]:

Command: freq +s"[*pho:*]"
Output: [%pho: treffte]

**%syn:**

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 part II. Coding).

# Part II: Grammatical Coding

## 1 Lexicon-Based Automatic Morphological Coding of Transcripts

### 1.1 Introduction

The automatic coding system to be presented is based on the CLAN program MOR (MacWhinney 2000/I/1:104ff.) Steven Gillis has extended this to apply to languages with richer morphologies (see **MinMOR** in CHILDES). While automatic coding systems devised for English by MacWhinney and for Dutch by Gillis use rules to derive morphologically complex forms from their bases, the system created for languages with richer morphologies, such as Greek and German, only has a rudimentary rule component and mainly relies on a lexicon in which both inflectional forms and derivations are listed.

MinMOR contains 3 basic files with the extension .cut (...ar, ...cr, ...lex) which can be used for coding data from any language. MOR requires these files in order to function correctly. The 3 files to be used with German are **gerar.cut**, **gercr.cut**, and **gerlex.cut**. They may be renamed (without changing their content) for use with any language. The lexicon file **gerlex.cut** must be enlarged according to the data you wish to analyze.
The 3 files must be placed in the CHILDES\CLAN\LIB subdirectory. When using the CLAN programs make sure that the lib directory occurs in the Commands window (C:\childes\clan\lib).

Since the CLAN programs, including MOR, only run on ASCII files, be sure to use unformatted versions of your transcripts and your lexicon for automatic coding and analysis.

### 1.2 How to Create a Language-Specific Lexicon

In order to run MOR on your first transcript, you need a rudimentary lexicon, such as gerlex.cut, which contains at least one coded entrance. For German, this entry could read as follows:

```
hund    {[scat N]}      "hund:MASC:NOM/ACC:SG"
```

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:

```
mor +xl +gger +lgerlex.cut +k @
```

+k This option allows MOR to find all proper nouns (transcribed with an initial capital letter). Without this option, MOR will code such words automatically as n:prop precluding further grammatical specification.

NB. You can also set a default grammar for the MOR files. This will be done in the CLAN Editor, using *EDIT:Set default MOR files*. There is then no need to indicate which grammar or lexicon MOR has to use and the command creating a lexicon of uncoded word forms will look as follows:

```
mor +xl +k @
```

Either of the above commands results in a file with the extension 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.

bellt {[scat ?]}
der {[scat ?]}
das {[scat ?]}
hund {[scat ?]}
pferdchen {[scat ?]}

All entries found in this file have to be coded by hand using a text editor. Although users are free to code lexical entries according to the needs of their respective research questions, the following directions are to be observed:

1. Add the appropriate **s**[yntactic]**cat**[egory] replacing the question mark with the major part of speech of the grammatical word form written in the left-hand column (e.g. {[scat N]}). Adhere to the grammatical codes specified at http://childes.psy.cmu.edu/html/morcats.doc and based on MacWhinney (1995, 2000/I) as far as possible. You may add subclasses to the major parts of speech, separating them by a colon (e.g. {[scat N:PROP]}). Parts of speech which cannot be categorized may be coded as "unknown": {[scat unknown]}.
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 (use "straight quotes"). If a word form is grammatically ambiguous, you may add a new line for each grammatical interpretation of the form or use obliques (e.g. NOM/ACC).

After coding has been finished, your file will contain three columns:

sieht {[scat V]} "sehen:PRES:3S"
der {[scat ART:DEF]} "der:MASC:NOM:SG"
der {[scat PRO:PRS]} "der:MASC:NOM:SG"
das {[scat ART:DEF]} "das:NEUT:NOM/ACC:SG"
das {[scat PRO:PRS]} "das:NEUT:NOM/ACC:SG"
hund {[scat N]} "hund:MASC:NOM/ACC:SG"
pferdchen {[scat N]} "pferd:DIM:NEUT:NOM/ACC:SG"

This coded file is the beginning of the coded lexicon for the language you are studying. Integrate it into the appropriate *lex.cut file.

### 1.3 Generating the %mor Tier

After all entries contained in your lexicon have been coded, you are ready to code your transcript morphologically. The following command will add a %mor: tier to each main speaker tier in your transcript:

mor +gger +lgerlex @ (without default setting)
mor @ (with default setting, see NB. above p. 25)

If you want to code only the learner's utterances (e.g. Nastja), use the following command:

mor +gger +lgerlex +t*NAS –t*INT @ (without default setting)
mor +t*NAS -t*INT @ (with default setting)

For the sentence "der hund sieht das pferdchen" the %mor tier will look like this:

*NAS: der hund sieht das pferdchen .
%mor: ART:DEF|der:MASC:NOM:SG^PRO:PRS|der:MASC:NOM:SG
N|hund:MASC:NOM/ACC:SG V|sehen:PRES:3S
ART:DEF|das:NEUT:NOM/ACC:SG^PRO:PRS|das:NEUT:NOM/ACC:SG
N|pferd:DIM:NEUT:NOM/ACC:SG

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 in a specific context, is placed after the vertical bar.

If a word form is associated with two or more codings, all alternatives will be provided, separated by '^'. Files have to be disambiguated by hand. There is a convenient way to do this by using the mode 'disambiguate tier' in CED, the CHILDES editor.[39] The different codings appear at the bottom of the screen. Mark the correct alternative with the mouse (or the arrow keys).

[39] For automatically disambiguating English or French coded transcripts, see MacWhinney (2000/I:121ff., Part 2) on POST.

Clicking twice (or pressing ENTER) will insert the correct option into your transcript. If no option seems to be correct, use the UND (undecided) tag.

After disambiguation, transcripts are ready for morphosyntactic analysis with the help of the CLAN programs.

NB. MOR will treat metalinguistic material marked by '["]' as quotes and code them automatically as n:quote. This may lead to a different number of units on the %mor tier as compared to the speaker tier, so that MODREP[40] will not work.

### 1.4 How to Enlarge Your Coded Lexicon and Code Further Files

When working with a new file you will first want to retrieve all new words not yet included in your lexicon. Use the same command as above, but be sure to work with the enlarged lexicon containing the codings of the first file(s) you worked with:

mor +xl +gger +lgerlex @ (without default setting)
mor +xl @ (with default setting)

This command will create a file containing only word forms not occurring in the first coded file and therefore not yet contained in your coded lexicon. You do not need to proceed file by file, but can also run the above command on several files simultaneously (by placing all of them in the FILE IN window; see part III below).

Unite the output file of the above command with your lexicon, order all entries alphabetically, and code the as yet uncoded new entries.

Use your enlarged lexicon to generate a %mor: tier in the new file(s) by the above command repeated here for convenience:

mor +gger +lgerlex @ (without default setting)
mor @ (with default setting)

[40] See p.43 below.

### 1.5 Creating and Working with Several Lexicons

It may be advisable to create seperate lexicons for different kinds of studies (e.g. analyses of narratives vs. dialogues) or for different parts of the alphabet.

Be sure to indicate all lexicons to be considered in your MOR command, so that all their entries will be simultaneously entered on the %mor tier. To code a file using the two lexicons gerak-lex.cut and gerlz-lex.cut the MOR command could be:

mor +gger +lgerak-lex +lgerlz-lex @ (without default setting)
mor +lgerak-lex +lgerlz-lex @ (with default setting)

## 2 Coding Grammatical Errors and Self-Repairs[41]

In order to do a detailed analysis of errors occurring in learner languages, it is useful to distinguish different types of errors on the %mor line. Here are some suggestions for manually coding errors and self-repairs:

**Morphophonemic errors (#)**

*NAS: er habte [*] [: hatte] einen hund.
%mor: ... V|haben:PAST#=hatte:3S ...

**Wrong use of grammatical categories (*)**

*LEA: dann ist der junge auf einen baum geklettert und guckte in ein grosses loch.
%mor: ... V|gucken:PAST*=PERF:3S ...

**Successful self-repair ($)**

*NAS: er habte [*] [//] hatte einen hund.
%mor: ... V|haben:PAST$:3S ...

or

%mor: ... V|haben:PAST#$:3S ...

[41] For further suggestions see MacWhinney 2000/I/1:§14 and §16.8.

**Unsuccessful self-repair (%)**

*NAS: er gang [*] [//] gehte [*] [: ging] zur schule.
%mor: ... V|gehen:PAST%:3S ...

or

%mor: ... V|gehen:PAST#%:3S ...

Error coding should on the one hand classify the actual forms used by the learner according to the target language and on the other mark the category which would have been appropriate in a given context (e.g. PRO:POSS|mein:FEM:NOM/ACC*=DAT:SG 'ich habe meine(r) mutter das bild gegeben'). This form of error coding enables one to establish the grammatical system of the learner language at a certain stage of development, instead of exclusively relating the learner's speech to the target language.

Because of inflectional homonymy, coding may sometimes cause difficulties. Thus, the adjective 'alte' in 'die alte dinge für den krieg' may either be classified as a strong NOM/ACC plural or a weak NOM/ACC singular. Only the analysis of all relevant examples will show if the learner distinguishes between weak and strong or between singular and plural forms of adjectives. It may well be that at a certain stage of development the learner only possesses a general purpose adjectival form ending in *-e* which is sometimes, but not always, used in a target-like way.

Here is a suggestion for coding such **unclear examples** (§):

*NAS: die alte dinge für den krieg.
%mor: ADJ|alt:NOM/ACC:WEAK/STRONG§:SG/PL§

## 3 Syntactic Coding of Transcripts

For certain types of syntactic analysis a **%syn** dependent tier has to be added to the transcript below the %mor line. This must be done manually.

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

ADJ adjective
ADV adverbial

| | |
|---|---|
| ADV:LOC | locative adverbial |
| 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 |
| TEMP | temporal adverbial |
| V2 | verb second rule observed in main clause |
| V2* | verb second rule not observed |
| V0* | verb wrongly omitted in main clause |
| V2:AUX | auxiliary correctly placed in second position |
| V:PP | past participle of verb |
| VF | verb final position observed in subordinate clause |
| VF* | verb final position not observed |
| VF*:AUX | auxiliary not in final position in subordinate clause |
| VF*:INF | infinitive not in final position |
| VF0* | verb omitted in final position |
| MC | main clause |

# Part III: CLAN Analysis

## 1 Overview of the Most Widely Used CLAN Programs

**CHAINS** Discourse and text analysis;[42] syntactic analysis.

**CHECK** Verifies if transcripts correspond to CHAT conventions.[43]

**CHIP** Analyzes specified pairs of utterances.[44]
Examples of application: spoken language, aphasia, language acquisition, language impairment (input, relationship between input and imitation, individual differences in imitativeness in both normal or language-impaired children).

**COLUMNS** Displays dialogues in two (or three) columns (left column: child/learner; right column: adult interlocutor) separating the contributions of the child/learner from those of the other interlocutors.[45]

**COMBO** Finds combinations of keywords.[46]
Examples of application: inflectional and derivational morphology, syntactic analysis (e.g. word order), discontinuous morphemes, questions, negations; picture stories, experimental data.

**COOCCUR** Syntactic analysis.[47]
Examples of application: collocations.

**DATES** Takes two time values and computes the third (e.g. the age of child/learner is computed on the basis of date of birth and current date).[48]

[42] See MacWhinney 2000/I/2:37ff; Sokolov/Snow (eds.) 1994.
[43] See MacWhinney 2000/I/2:41ff.
[44] See MacWhinney 2000/I/2:45ff; Sokolov & MacWhinney 1990, Sokolov & Moreton 1994:174-209.
[45] See MacWhinney 2000/I/2:53ff.
[46] See MacWhinney 2000/I/2:56ff.
[47] See MacWhinney 2000/I/2:63f.
[48] See MacWhinney 2000/I/2:64.

**DIST** This program computes the number of utterances existing between occurrences of a specified key word or code.[49]
Examples of application: relationship between input and learner's speech.

**FLO** Creates a more legible version of the speaker tier ignoring all special CHAT coding information.[50]

**FREQ** Frequency analysis and type/token ratio.[51]
Examples of application: alphabetical list of words (or morphemes) indicating frequency of each word form (morpheme), frequency of grammatical categories, occurrence of suffixes (on the basis of a reverse concordance).

**GEM** Finds comparable parts of transcripts in the same or different files.[52]
Examples of application: reading passages or story telling as opposed to dialogue passages.

**KWAL** Finds words (grammatical forms, lexemes, grammatical categories) and lists all examples with a given keyword.[53]
Examples of application: morphological, lexical and syntactic analysis.

**LINES** Adds line numbers to transcripts.[54]

**MAXWD** Finds and measures the longest word or longest utterance.[55]
Examples of application: assessment in language acquisition (piping with MLU: e.g. mean length of the five longest utterances).[56]

[49] See MacWhinney 2000/I/2:65.
[50] See MacWhinney 2000/I/2:72.
[51] See MacWhinney 2000/I/2:72ff.
[52] See MacWhinney 2000/I/2:81ff.
[53] See MacWhinney 2000/I/2:87ff.
[54] See MacWhinney 2000/I/2:89.
[55] See MacWhinney 2000/I/2:91f.
[56] See Pan 1994:39.

**MLT** Computes mean length of speaker turns.[57]
Examples of application: language acquisition, discourse analysis, gender-specific language, aphasia.

**MLU** Computes mean length of utterance and number of utterances.[58]
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.[59]
Examples of application: phonetic (or graphic) variation in language acquisition, language impairment, aphasia, unimpaired speech.

**MOR** Provides automatic morphological coding, generating a %mor: tier for all (or selected) speaker tiers.[60]

**PHONFREQ** Produces an inventory of initial, medial, and final phonological elements on the %pho line, indicating frequencies.[61]
Examples of application: acquisition of phonology.

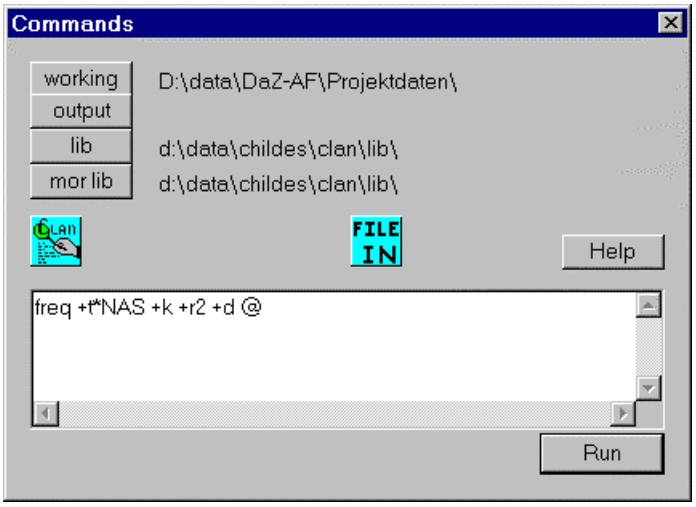
**TEXTIN** Converts transcripts in paragraph form to CHAT files.[62]

**WDLEN** Measures word length (in letters) and utterance length (in words) and produces histograms of their frequencies.[63]
Examples of application: speakers' profiles.

[57] See MacWhinney 2000/I/2:92ff.
[58] See MacWhinney 2000/I/2:95ff.
[59] See MacWhinney 2000/I/2:101ff.
[60] See MacWhinney 2000/I/2:104ff.
[61] See MacWhinney 2000/I/2:120f.
[62] See MacWhinney 2000/I/2:127.
[63] See MacWhinney 2000/I/2:132f.

# 2 Analyzing Transcripts Using the CLAN Programs

## 2.1 Introduction




Before starting an 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 ASCII files for automatic morphological coding are located (e.g. 'childes\clan\lib'). In order to set the directories, press the button 'working' (or 'lib'), locate the desired directory and press the 'Select directory' button.

Then proceed as follows:

Press the button 'CLAN' and select the appropriate program.

Type in the desired options.

Select the file (or files) to be analyzed by pressing the button 'File in'. Locate the file, mark and double-click it so that it appears in the right-hand window.

Press 'Done' and then 'Run'.

| | | |
|---|---|---|
| Some options: | +t*NAS | includes speaker tier NAS |
| | +k | makes the analysis case sensitive |
| | +r2 | marks parentheses as found in transcript |
| | +f | sends output to a file |

To see a list of the options for each CLAN program, just type the name of the program or select it (e.g. FREQ, KWAL, etc.).

Sample file NAS04-13.CHA coded in CHAT format:

```
@Begin
@Participants:          NAS Nastja target-child, INT Christian Bauer
@Birth of Learner:      8-JUN-1989
@Age of Learner:        8;10.9
@Time spent in Germany:     0;3.13
@Date:          17-APR-1998
@Tape:          8A (8 recto)
@Filename:      NAS04-13.TXT
@Transcriber: Christian Bauer
@Situation:     NAS and INT in NAS' home in Cologne. INT introducing session
@Comment:       Only the learner´s data have been coded.
*INT:           so, Nastja, kannst du dich noch an mich erinnern ?
*NAS:           eh@fp +...
*INT:           wann haben wir uns schon mal gesehen ?
*NAS:           eh@fp # eh@fp # im Max_Planck_Institut .
%mor:           PREP:LOC|in:DEF:NEUT:DAT:SG N:PROP|Max_Planck_Institut .
*INT:           genau, im Max_Planck_Institut .
*INT:           da sind wir zusammen mit frau # Stephany nach &mij [//] Nijmegen
                gefahren .
*NAS:           ja .
%mor:           PTL|ja .
*INT:           und # damals warst du gerade neu in Deutschland angekommen .
*INT:           wie lange +...
*NAS:           eh@fp # ich bin hier drei monate .
%mor:           PRO:PRS|ich:1S:NOM V:COP|sein:PRES:1S ADV|hier NUM|drei
                N|monat:MASC:NOM/ACC:PL .
*INT:           hm@i .
*NAS:           das ist viernten [*] [: vierter] monat [/] monat .
%mor:           PRO:DEM|das:NEUT:NOM:SG V:COP|sein:PRES:3S
                NUM|viert:MASC:ACC*=NOM:SG N|monat:MASC:NOM:SG .
*NAS:           und <am &f> [//] in ferien ich war <in &pa> [/] in Paris.
%mor:           CONJ:COO|und PREP:LOC|in N|ferien:ACC:PL PRO:PRS|ich:1S:NOM
                V:COP|sein:PAST:1S PREP:LOC|in N:PROP|Paris .
@End
```

Here are some useful notes on analyses concerning the CLAN programs in general:

Material included in brackets containing a colon '[: text]' will automatically replace the immediately preceding word or sequence of words throughout (including **MOR**). Using the switch **+r5** makes such material invisible to the CLAN programs. However, this is an invalid option for **MOR**. The error marker [*] is invisible to the CLAN programs by default and will therefore not disturb the process of replacement.

## 2.2 COMBO

combo +t%mor +s"V:COP*^ADJ*" +k +r2 @

Lists all utterances containing a copula immediately followed by an adjective.

combo +t%mor +s"V:COP*^ADJ*" +x +k +r2 @

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

combo +t%mor +s"V:COP*^*^ADJ*" +k +r2 @

Lists all utterances containing a copula eventually followed by an adjective.

combo +t%mor +s"ADJ*^N|*" +k +r2 @

Searches adjectives immediately followed by a noun.

combo +t%mor +s"ADJ*^*^N|*" +k +r2 @

Searches adjectives immediately or eventually followed by a noun.

combo +t%mor +s"!V*" +k +r2 @

Lists all verbless utterances.

combo +t%mor +s"V*^*^(PRO*+N*)" +k +r2 +x @

Lists all utterances comprising a verb and pronoun or noun, in any order.
NB. Parentheses must not be immediately preceded by "*".

combo +t%mor +s"V*^(PRO*NOM*^*N|*ACC*)" +x +k +r2 @

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

combo +t%mor +s"PRO*^(V*^*^*N|*)" +k +r2 @

Lists all utterances containing a pronoun directly followed by a verb, and a noun eventually following the verb.
NB. The +x option is invalid with commands comprising the string "*^*^*".

combo +t%mor +s"V*3*^*N|*NOM*" +x +k +r2 @

Lists all utterances containing a third person verb form and a noun in the nominative case, in any order.

combo +t%mor +t*NAS +s"eins^*^%mor:^*^PRO*" +k +r2 @

Lists all utterances in which *eins* is used as a pronoun and not as the homophonous numeral.

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

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

combo +t*MOT +s"klein*^@dim.mot @

Lists all utterances in mother's speech in which a form of the adjective *klein* 'small' is immediately followed by a diminutive included in the include file dim.mot containing *chen and *lein.

## 2.3 COOCCUR

cooccur + t*NAS +n5 @

Lists cooccurrences of words in the learner's utterances.

+n5 Combines 5 words in a cluster (max. 20 words).

## 2.4 DATES

dates +b 05-JUN-1999 +d 19-APR-2001

+b followed by a blank indicates the birth date in day-month-year format.
+d followed by a blank indicates the current date in day-month-year format.
+a followed by a blank indicates the age in CHAT format.

The output of this command is as follows:
@Birth of CHILD: 5-JUN-1999
@Date: 19-APR-2001
@Age of CHILD: 1;10.14

dates +a 1;10.14 +d 19-APR-2001

@Age of CHILD: 1;10.14
@Date: 19-APR-2001
@Birth of CHILD: 5-JUN-1999

## 2.5 FREQ

freq +t*NAS +k +r2 @

Lists all of Nastja's word forms in alphabetical order, indicating their frequencies.

freq +t*NAS +k +r2 +o +d0 @

| | |
|---|---|
| +o | Sorts output by descending frequency. |
| +d0 | Indicates the numbers of the lines where each word form is located. |

freq +t*NAS +r2 +k +d0 +o1 @

| | |
|---|---|
| +o1 | Creates a reverse concordance (i.e. sorts output alphabetically according to word endings). |

freq +t*NAS +r2 +k +d0 +s"[*]" @

| | |
|---|---|
| +s"..." | Searches particular strings within word boundaries. |
| +s"[*]" | Finds all forms containing an error marked by [*]. |

freq +t*NAS +r2 +k +d0 +s"*lein" @

| | |
|---|---|
| +s"*lein" | Finds all diminutives ending in *-lein*. |

freq +t%mor -t* +k +d0 @

Produces an alphabetical list of lexemes according to their parts of speech and grammatical coding indicating frequencies.

freq +t%mor -t* +s"*GEN*" +k +d0 @

Finds all forms in the genitive singular or plural.

freq +t%mor +s"*GEN:PL" +k +d0 @

Finds all forms in the genitive plural.

freq +t%mor +o +k +d0 +s"ART:DEF*" @

Lists all forms of the definite article by descending frequency.

freq +t%mor +k +d0 +s"ART:DEF%" @

Indicates frequency of definite articles ignoring grammatical subdivisions.

freq +t*NAS +k +d0 +s"*0*" @

Lists omitted words coded by *0 (e.g. das ist *0ein hund.).

freq +t*NAS +s"*(*)" +k +d0 +r2 @

Lists word forms with missing endings.

freq +t*NAS +s"*r" +k +d0 +r2 @

Finds all of Nastja's words ending in /-r/ (such as *der, dieser, einer*).

### 2.6 GEM

gem +sreading +d @

+s Retrieves all reading passages.
+d Outputs the file in CHAT format for further analysis.

NB. Since the switch '-s' cannot be used with GEM, one possibility is to entirely divide the transcripts into gem sections so that the analysis of the reading passages can be compared with that of free conversation. Another possibility is to use a text editor to manually eliminate all non-gem passages.

### 2.7 KWAL

kwal +t*NAS +s"*?" +k +r2 @

Lists all interrogative clauses.

kwal +t%mor +s"*NEG*" +k +r2 @

Lists all utterances containing a negative particle.

kwal +t*NAS +s"[*]" +k +r2 @

Lists all utterances containing incorrect forms marked by '*' on the main tier.

kwal +t%mor +s"***" +k +r2 @

Lists all utterances containing incorrect forms marked by '*' on the %mor tier.

kwal +t*NAS +s"*(*)*" +r2 +k @

Lists all utterances containing forms with omitted parts.

kwal +t*NAS +s"&*" @

Lists all utterances containing false starts.

kwal +t*NAS +s"\+…" @

Finds all trailing-off passages.

kwal +t*NAS +s"er" +k +r2 -w2 +w2 @

Lists all utterances containing the pronoun *er*, including the two preceding utterances (-w2) and the two succeeding ones (+w2).

kwal +t*NAS +s"%chen" +s"%lein" +k +r2 @

Lists all examples containing diminutives ending in *-chen* or *-lein*. In contrast to the option +s"*chen", keywords only consist of the suffix and not the individual lexemes used.

kwal +t*NAS +s@diminut.cha +k +r2 @

Lists all utterances containing a diminutive suffix contained in the include file diminut.cha.

kwal +t*NAS +s@locprep.cha +k +r2 @

Lists all utterances containing a locative preposition contained in the include file locprep.cha.

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

Lists all of Nastja's examples containing a personal pronoun.
NB. By adding the option +t*NAS, you avoid finding examples of the other speakers once the %mor tier has been provided for all speakers.

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.

This command is useful for printing transcripts without dependent tiers.
NB. Feeding the output file into LINES does not produce the line numbers referred to by the CLAN programs analyzing the original file.

kwal +t*NAS +s@negation.cha +k +d @ | mlu

Calculates the MLU of Nastja's utterances containing a negative particle included in the file negation.cha ("piping" of KWAL and MLU).

kwal +t*NAS -s@negation.cha +k +d @ | mlu

Calculates the MLU of Nastja's utterances not containing a negative particle included in the file negation.cha.

kwal +pnewpunct.cut +s"[>]" @

Finds all utterances containing overlaps.
NB. The include file newpunct.cut has to be defined by omitting '<' and '>'.[64]

kwal +s@quotes.cut @

Finds all utterances contained in the include file quotes.cut (see p. 45).

### 2.8 LINES

lines @

Inserts line numbers referred to by the CLAN programs.

lines +t* -t% @

Counts only speaker tiers (main tiers).

### 2.9 MAXWD

maxwd +t*NAS +c7 +r3 @

+c7 Finds and measures Nastja's 7 longest words in letters.
+r3 Material included in parentheses is removed.

maxwd +t*NAS +c7 +r3 +g2 @

+g2 Finds the 7 longest utterances, calculated in words.

maxwd +t*NAS +c7 +r3 +g2 +d1 @ | mlu

+d1 Outputs the transcript in legal CHAT format (to be used for piping).
+c7 Calculates MLU of 7 longest utterances.

### 2.10 MLT

mlt @

Calculates ratios of words to turns, utterances to turns, and words to utterances.

[64] See MacWhinney 2000/I/2:135f. and p.44 below.

## 2.11 MLU

mlu @

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

mlu -s"*@i*" @

Outputs a word/utterance ratio, disregarding interjections.

mlu -s"*@i" -b- @

Outputs a word/utterance ratio if morphemes have been hyphenated, disregarding interjections.

## 2.12 MODREP

modrep +b*NAS +c%pho +k @

+b Sets the model tier name.
+c Sets the replica tier name.

Compares a learner's variable renderings on the %pho tier to the forms indicated on the speaker tier *NAS.

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

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

modrep + b%mor +c*DAS +o*pl @

Finds all forms on the main line which match forms on the %mor line carrying a plural marker and indicates their frequency.

+o Limits the output to a particular string.

NB. Modrep only works if the model tier and the replica tier contain exactly the same number of items.

## 2.13 WDLEN

wdlen +t*NAS @

Produces histograms of the frequencies of word lengths (in characters) and of utterance lengths (in words).

# 3 Further Practical Hints for Working with the CLAN Programs

## 3.1 How to Create Include Files

Only use one item per line. End each line (including the last line) with a carriage return (ENTER). Save the include file in ASCII (unformatted) (text only).

Include file DIMINUTIVE.CUT for German:

*lein*
*chen*

This type of include files is useful for detecting lexemes with certain types of affixes in uncoded files. Include files may also be established for certain semantic fields, such as motion verbs.

Include files may either contain the metacharacters '*' or '%' with the following difference:

DIMINUTIVE1.CUT *lein*
*chen*

freq +t*NAS +s@diminutive1.cut @

+s@ The search will be effected by the include file indicated.

FREQ will count *tischlein, tischleins, pferdchen, pferdchens* as 4 different types of lexemes or grammatical forms of words.

DIMINUTIVE2.CUT *lein%
*chen%

freq +t*NAS +s@diminutive2.cut @

FREQ will count *tischlein* and *tischleins* as different tokens of one and the same type.

An include file *quotes.cut* must be created in order to retrieve occurrences of direct speech. This file looks as follows:

```
+"
+"/
+/.
```

Use the option +s@quotes.cut with KWAL.

### 3.2 Working on Groups of Data

If you want to work on groups of files, you must put them all into the 'File in' window.

freq +t*NAS +u @

+u Unites the output of all chosen files (e.g. NAS02-04.cha, NAS02-05.cha, NAS02-06.cha).

NB. If the option +u is not chosen, the CLAN programs will analyze each file separately.

## Part IV: Examples of Transcribed and Coded Data

These examples contain the specific characteristics of the transcripts of the German L2 Project.

```
@Begin
@Participants:    NAS Nastja target-child, INT Christian Bauer
@Birth of Learner:          8-JUN-1989
@Age of Learner:          8;10.9
@Time spent in Germany:          0;3.13
@Date:  17-APR-1998
@Tape: 8A (8 recto)
@Filename:       NAS04-13.TXT
@Transcriber: Christian Bauer
@Situation:       NAS and INT in NAS' home in Cologne. INT introducing session
@Comment:       Only the learner´s data have been coded.
*INT:    so, Nastja, kannst du dich noch an mich erinnern ?
*NAS:   eh@fp +...
*INT:    wann haben wir uns schon mal gesehen ?
*NAS:   eh@fp # eh@fp # im Max_Planck_Institut .
%mor:   PREP:LOC|in:DEF:NEUT:DAT:SG N:PROP|Max_Planck_Institut .
*INT:    genau, im Max_Planck_Institut .
*INT:    da sind wir zusammen mit frau # Stephany nach &mij [//] Nijmegen gefahren .
*NAS:   ja .
%mor:   PTL|ja .
*INT:    und # damals warst du gerade neu in Deutschland angekommen .
*INT:    wie lange +...
*NAS:   eh@fp # ich bin hier drei monate .
%mor:   PRO:PRS|ich:1S:NOM V:COP|sein:PRES:1S ADV|hier NUM|drei N|monat:MASC:NOM/ACC:PL.
```

*INT: hm@i .
*NAS: das ist viernten [*] [: vierter] monat [/] monat .
%mor: PRO:DEM|das:NEUT:NOM:SG V:COP|sein:PRES:3S NUM|viert:MASC:ACC*=NOM:SG N|monat:MASC:NOM:SG .
*NAS: und <am &f> [//] in ferien ich war <in &pa> [/] in Paris.
%mor: CONJ:COO|und PREP:LOC|in N|ferien:ACC:PL PRO:PRS|ich:1S:NOM V:COP|sein:PAST:1S PREP:LOC|in N:PROP|Paris .
*INT: hm@i, in Paris warst du in den ferien ?
*NAS: ja, ich war dort # eh@fp fünf tage .
%mor: PTL|ja PRO:PRS|ich:1S:NOM V:COP|sein:PAST:1S ADV|dort NUM|fünf N|tag:MASC:ACC:PL .
*INT: hm@i .
*NAS: und eh@fp # diese stadt ist ganz [*] schön .
%mor: CONJ:COO|und PRO:DEM|dies:FEM:NOM:SG N|stadt:FEM:CAS:SG V:COP|sein:PRES:3S ADV|*ganz=sehr ADJ|schön .
*INT: ja, hat es dir gefallen ?
*NAS: ja .
%mor: PTL|ja .
*INT: hm@i .
*NAS: ich war im [*] zweite(n) [*] etage auf [?] # Eiffelnturm [*] [: Eiffelturm] .
%mor: PRO:PRS|ich:1S:NOM V:COP|sein:PAST:1S PREP:LOC|*in=auf:DEF:MASC/NEUT:DAT:SG NUM|zweit:FEM:NOM/ACC*=DAT:SG N|etage:FEM:DAT:SG PREP:LOC|auf N:PROP|Eiffelturm .
*INT: hm@i .
*NAS: ja .
%mor: PTL|ja .
*INT: auf dem Eiffelturm warst du ?
*NAS: ja .
%mor: PTL|ja .
*INT: hm@i .
*NAS: und # ich war noch im [/] # eh@fp # im Musee_d'orsay .
%mor: CONJ:COO|und PRO:PRS|ich:1S:NOM V:COP|sein:PAST:1S ADV|noch PREP:LOC|in:DEF:NEUT:DAT:SG N:PROP|Musee_d'orsay .
*INT: hm@i .
*NAS: aber # wenn [*] ich war in Musee_d'orsay wir waren mit [*] Dascha schon ganz müde.
%mor: CONJ|aber CONJ:SUBOR|*wenn=als PRO:PRS|ich:1S:NOM V:COP|sein:PAST:1S PREP:LOC|in N:PROP|Musee_d'orsay PRO:PRS|wir:1P:NOM V:COP|sein:PAST:1P PREP|*mit N:PROP|Dascha:FEM:SG ADV|schon ADJ|ganz ADJ|müde
%err: wir waren mit Dascha = waren Dascha und ich
*NAS: wir hatten hunger und wir haben nich(t) sehr viel gesehen.
%mor: PRO:PRS|wir:1P:NOM V|haben:PAST:3P N|hunger:MASC:ACC:SG CONJ:COO|und PRO:PRS|wir:1P:NOM V:AUX|haben:PRES:1P PTL:NEG|nicht ADV|sehr QUANT|viel V|sehen:PP .
*INT: hm@i .
*NAS: und # &w ich war im Louvre .
%mor: CONJ:COO|und PRO:PRS|ich:1S:NOM V:COP|sein:PAST:1S PREP:LOC|in:DEF:MASC:DAT:SG N:PROP|Louvre .
*INT: hm@i .
*NAS: da ist ganz [*] schön .
%mor: ADV|da V:COP|sein:PRES:3S ADV|*ganz=sehr ADJ|schön .
*INT: ja, was hat dir am besten gefallen im Louvre ?
*NAS: eh@fp # eh@fp # da ist ganz [*] schöne bilder .
%mor: ADV|da V:COP|sein:PRES:3S*=3P ADV|*ganz=sehr ADJ|schön:NOM:PL N|bild:NEUT:NOM:PL.
*INT: hm@i .
*NAS: und # das ist +...
%mor: CONJ:COO|und PRO:DEM|*das=da:NEUT:NOM:SG V:COP|sein:PRES:3S +...
*NAS: und da ist ein zimmer .
%mor: CONJ:COO|und ADV|da$ V:COP|sein:PRES:3S ART:INDEF|ein:NEUT:NOM:SG N|zimmer:NEUT:NOM:SG .
*NAS: dort ist ganz dunkel .
%mor: ADV|dort V:COP|sein:PRES:3S ADV|ganz ADJ|dunkel .
*INT: hm@i .

*NAS: und eh@fp # wir haben dort fotografier(e)n [*].
%mor: CONJ:COO|und PRO:PRS|wir:1P:NOM V|haben:PRES:1P ADV|dort V|fotografieren:INF*=PP
*NAS: aber das darf man nicht .
%mor: CONJ|aber PRO:DEM|das:NEUT:ACC:SG V:MDL|dürfen:PRES:3S PRO:INDEF|man PTL:NEG|nicht .
*INT: oh@i, da habt ihr heimlich fotografiert ?
*NAS: ja .
%mor: PTL|ja .
*NAS: aber # wir wissen [*] es nicht .
%mor: CONJ|aber PRO:PRS|wir:1P:NOM V|wissen:PRES*=PAST:1P PRO:PRS|es:3S:NEUT:ACC PTL:NEG|nicht.
*NAS: und wenn wir haben schon fotografiert ,
%mor: CONJ:COO|und CONJ:SUBOR|*wenn=als PRO:PRS|wir:1P:NOM V:AUX|haben:PRES:1P ADV|schon V|fotografieren:PP
*NAS: eh@fp # kommt [*] eine frau und sie hatte [*] gesagt <dass # man &ni> [//] das darf man nicht .
%mor: V|kommen:PRES*=PAST:3S ART:INDEF|ein:FEM:NOM:SG N|frau:FEM:NOM:SG CONJ:COO|und PRO:PRS|sie:3S:FEM:NOM V:AUX|haben:PAST*=PRES:3S V|sagen:PP PRO:DEM|das:NEUT:ACC:SG V:MDL|dürfen:PRES:3S PRO:INDEF|man PTL:NEG|nicht .
*NAS: da ist ganz [*] schön.
%mor: ADV|da V:COP|sein:PRES:3S ADV|*ganz=sehr ADJ|schön
*NAS: da # da ist &ge [/] gerp # heisst das, gerp .
%mor: ADV|da ADV|da V:COP|sein:PRES:3S U|gerp V|heissen:PRES:3S ART:DEF|das:NEUT:NOM:SG U|gerp .
*INT: hm@i .
*NAS: und da [/] <da ist viel> [/] eh@fp da ist viel interessantes.
%mor: CONJ:COO|und ADV|da V:COP|sein:PRES:3S QUANT|viel ADJ|interessant:NEUT:NOM:SG
*NAS: eh@fp und [/] # und dort war die [//] auf zweite(n) etage oder dritte(n) eh@fp die statue,
%mor: CONJ:COO|und ADV|dort V:COP|sein:PAST:3S PREP:LOC|auf NUM|zweit:FEM:NOM/ACC*=DAT:SG N|etage:FEM:CAS:SG CONJ:COO|oder NUM|dritt:FEM:NOM/ACC*=DAT:SG ART:DEF|die:FEM:NOM:SG N|statue:FEM:NOM:SG
*NAS: und es war eh@fp # uniforma@r [: uniform] .
%mor: CONJ:COO|und PRO:PRS|es:3S:NEUT:NOM V:COP|sein:PAST:3S N|*uniforma@r=uniform:FEM:NOM:SG .
*INT: hm@i .
*NAS: eine uniforma [*] [: uniform] .
%mor: ART:INDEF|$0ein:FEM:NOM:SG N|*uniforma@r=uniform:FEM:NOM:SG .
*NAS: eh@fp wir war(e)n ganz lange zeit zurück .
%mor: PRO:PRS|wir:1P:NOM V:COP|sein:PAST:1P ADV|ganz ADJ|lang:FEM:ACC:SG N|zeit:FEM:ACC:SG ADV|zurück .
*NAS: eh@fp +...
*INT: ganz lange zeit zurück .
*NAS: ja .
%mor: PTL|ja .
*INT: +, das habe ich nicht verstanden .
*INT: was <meinst du> [>] ?
*NAS: eh@fp [<] .
*NAS: ich eh@fp meine das [/] <das ist> [/] # eh@fp # eh@fp # das ist eh@fp eine [/] eine # rote hüte [*]+...
%mor: PRO:PRS|ich:1S:NOM V|meinen:PRES:1S PRO:DEM|das:NEUT:NOM:SG V:COP|sein:PRES:3S ART:INDEF|ein:FEM:NOM:SG ADJ|rot:FEM:NOM:SG N|hut:MASC:NOM/ACC:PL*=SG +...
*INT: hm@i .
*NAS: +, mit +...
%mor: + PREP|mit +...
*NAS: eine schöne hüte [*], so, so .
%mor: ART:INDEF|ein:FEM:NOM:SG ADJ|schön:FEM:NOM:SG N|hut:MASC:NOM/ACC:PL*=SG ADV|so ADV|so .
%com: zeigt mit den Händen einen Dreispitz
*INT: ja .
*NAS: und +...
%mor: CONJ:COO|und +...
*INT: hm@i .

```
*NAS:   +, es ist # eh@fp ganz [*] schön und die kleide(r) auch ganz [*] schön .
%mor:   + PRO:PRS|es:3S:NEUT:NOM V:COP|sein:PRES:3S ADV|*ganz=sehr ADJ|schön
        CONJ:COO|und ART:DEF|die:NOM:PL N|kleid:NEUT:NOM:PL# ADV|auch ADV|*ganz=sehr
        ADJ|schön .
*INT:   hm@i .
*NAS:   eh@fp # dort war(en) [?] offizieren .
%mor:   ADV|dort V:COP|sein:PAST:§3P N|offizier:DAT*=NOM:PL .
*INT:   hm@i .
*NAS:   eine [*] statuen  von offizieren .
%mor:   ART:INDEF|ein:FEM:NOM:SG*=PL N|statue:FEM:NOM:PL PREP|von N|offizier:DAT:PL .
*INT:   hm@i .
*NAS:   und # eh@fp auf diese(n) statuen waren dieses [/] diese kleider und diese hüte .
%mor:   CONJ:COO|und PREP:LOC|auf PRO:DEM|dies:NOM/ACC*=DAT:PL N|statue:FEM:DAT:PL
        V:COP|sein:PAST:3P PRO:DEM|dies:NOM:PL$ N|kleid:NEUT:NOM:PL CONJ:COO|und
        PRO:DEM|dies:NOM:PL N|hut:MASC:NOM:PL .
*INT:   hm@i .
*NAS:   und dort waren noch # eh@fp # eh@fp eine alte dinge .
%mor:   CONJ:COO|und ADV|dort V:COP|sein:PAST:3P ADV|noch ART:INDEF|ein:FEM:NOM:SG*=PL
        ADJ|alt:NOM:PL N|ding:NEUT:NOM:PL .
*INT:   hm@i .
*NAS:   ganz alte .
%mor:   ADV|ganz ADJ|alt:NOM:PL .
*NAS:   eh@fp # und # in den [*] hof +/.
%mor:   CONJ:COO|und PREP:LOC|in ART:DEF|der:MASC:ACC*=DAT:SG N|hof:MASC:CAS:SG +/.
*INT:   was waren da für ganz alte dinge ?
*NAS:   eh@fp .
*INT:   was waren das für dinge ?
@End
```

The following **syntactic coding** is useful for studying word order with respect to verb forms in main and subordinate clauses:

```
[from NAS04-13.cha]
*NAS:   eh@fp # ich bin hier drei monate .
%syn:   S V2 ADV:LOC ADV:TEMP [MC]
*NAS:   das ist viernten [*] [: vierter] monat [/] monat .
%syn:   S V2 PRED:N [MC]
*NAS:   und <am &f> [//] in ferien ich war <in &pa> [/] in Paris .
%syn:   ADV S V2* ADV [MC]
*NAS:   ja, ich war dort # eh@fp fünf tage .
%syn:   S V2 ADV ADV [MC]
*NAS:   und eh@fp # diese stadt ist ganz [*] schön .
%syn:   S V2 PRED:ADJ [MC]
*NAS:   ja .
*NAS:   ich war im [*] zweite(n) [*] etage auf [?] # Eiffelnturm [*] [: Eiffelturm] .
%syn:   S V2 ADV ADV [MC]
*NAS:   und # ich war noch im [/] # eh@fp # im Musee_d'orsay .
%syn:   S V2 ADV ADV [MC]
*NAS:   aber # wenn [*] ich war in Musee_d'orsay wir waren mit [*] Dascha schon ganz müde.
%syn:   CONJ:TEMP* S VF* ADV [C:TEMP] S V2* PP* ADV PRED:ADJ [MC]
*NAS:   wir hatten hunger und wir haben nich(t) sehr viel gesehen .
%syn:   S V2 DO CONJ:COO S V:AUX2 NEG DO V:PP [MC]
*NAS:   und # &w ich war im Louvre .
%syn:   S V2 LOC [MC]
*NAS:   da ist ganz [*] schön .
%syn:   LOC V2 S0* ADV PRED:ADJ [MC]
*NAS:   eh@fp # eh@fp # da ist [*] ganz [*] schöne bilder .
%syn:   LOC V2 ADV ADJ PRED:N [MC]
```

*NAS: und # das ist +...
%syn: S V2 +... [MC]
*NAS: und da ist ein zimmer .
%syn: LOC V2 PRED:N [MC]
*NAS: dort ist ganz dunkel .
%syn: LOC V2 S0* ADV PRED:ADJ [MC]
*NAS: und eh@fp # wir haben dort fotografier(e)n [*].
%syn: S V2:AUX LOC V:PP [MC]
*NAS: aber das darf man nicht .
%syn: DO V2 S NEG [MC]
*NAS: ja .
*NAS: aber # wir wissen [*] es nicht .
%syn: S V2 DO NEG [MC]
*NAS: und wenn wir haben schon fotografiert .
%syn: CONJ:TEMP* S VF*:AUX ADV V:PP [C:TEMP]
*NAS: eh@fp # kommt [*] eine frau und sie hatte [*] gesagt <dass # man &ni> [//] das darf man nicht .
%syn: V2 S CONJ:COO S V2:AUX V:PP [MC] DO V2 S NEG [MC]
*NAS: da ist ganz [*] schön,
%syn: LOC V2 S0* ADV PRED:ADJ [MC]
*NAS: da # da ist &ge [/] gerp # heisst das, gerp .
%syn: LOC V2 S [MC]
*NAS: und da [/] <da ist viel> [/] eh@fp da ist viel interessantes .
%syn: LOC V2 S [MC]
*NAS: eh@fp und [/] # und dort war die [//] auf zweite(n) etage oder dritte(n) eh@fp die statue.
%syn: LOC V2 LOC S [MC]
*NAS: und es war eh@fp # uniforma@r [: uniform] .
%syn: S V2 PRED:N [MC]
*NAS: eh@fp wir war(e)n ganz lange zeit zurück .
%syn: S V2 ADV [MC]
*NAS: ich eh@fp meine das [/] <das ist> [/] # eh@fp # eh@fp # das ist eh@fp eine [/] eine # rote hüte [*] +...
%syn: S V2 [MC] S V2 PRED:N [MC]
*NAS: +, es ist # eh@fp ganz [*] schön und die kleide(r) auch ganz [*] schön .
%syn: S V2 ADV PRED:ADJ CONJ:COO S V20* ADV ADV PRED:ADJ [MC]
*NAS: eh@fp # dort war(en) [?] offizieren .
%syn: LOC V2 S [MC]
*NAS: und # eh@fp auf diese(n) statuen waren dieses [/] diese kleider und diese hüte .
%syn: LOC V2 S CONJ:COO S [MC]
*NAS: und dort waren noch # eh@fp # eh@fp eine [*] alte dinge .
%syn: LOC V2 ADV S [MC]
*NAS: das ist +...
%syn: S V2 +... [MC]
*NAS: wie heisst es <wenn # die> [//] wenn # ist eh@fp ein # kreit [*] [: krieg] ja, oder kreig [: krieg] .
%syn: Q V2 S [MC] C:TEMP VF* S [C:TEMP]
*NAS: ah ich weiss nicht .
%syn: S V2 NEG [MC]
*NAS: wenn [/] wenn die menschen sind böse .
%syn: C:TEMP S VF* PRED:ADJ [C:TEMP]
*NAS: und da ist [*] soldaten .
%syn: LOC V2 PRED:N [MC]
*NAS: eh@fp # eh@fp aber das ist (ei)n altes [*] krieg .
%syn: S V2 PRED:N [MC]
*NAS: und eh@fp auf dem hof waren die # eh@fp # schon nicht für ganz alter [*] krieg .
%syn: LOC V2 S PP [MC]
*NAS: in den [*] hof war [*] ganz viel [*] diese rohr [*] .
%syn: LOC V2 S [MC]
*NAS: und # wir haben eine [*] fotos [*].
%syn: S V2 D [MC]
*NAS: +, wenn [*] ich sitze auf eine [*] rohe [*] [: rohr] .
%syn: CONJ:TEMP S VF* LOC [C:TEMP]
*NAS: und [/] und ich mache so +...
%syn: S V2 ADV [MC]

*NAS: +, und Dascha hat mi(r) [?] fotografiert .
%syn: S V2:AUX D V:PP [MC]
*NAS: eh@fp # es macht # eh@fp wenn ein [*] kugel dort +/.
%syn: S V2 [MC] CONJ:COND S LOC VF0* [C:COND]
*NAS: +, ist, dann [/] eh@fp &na eh@fp dann es macht so und <es ist # ganz> [//] da <ist &gan> [/] ist ganz &de die [//] # an die häuser waren kaputt .
%syn: TEMP S V2* ADV [MC] CONJ:COO LOC V2 PRED:ADJ [MC]
*NAS: das kann alles <machen kaputt> [>] .
%syn: S V2:AUX D VF*:INF ADJ [MC]
*NAS: und # in Musee_d'orsay war # eine Gioconda # heisst das .
%syn: LOC V2 PRED:N [MC] V2 S [MC]
*NAS: und da ist eine frau .
%syn: LOC V2 S [MC]
*NAS: und eh@fp # alle wissen nicht frau das oder mann.
%syn: S V2 NEG D [MC]
*NAS: und eh@fp # es hat so ein &ge gemält [*] [: gemalt], dass # ein augen [*] ist <auf &li> [/] auf link [*] [: links].
%syn: S V2 ADV V:PP [MC] CONJ:SUBOR S VF* LOC [C:COMP]

## References


MacWhinney, Brian & Snow, Catherine E. (1985). The child language data exchange system. *Journal of Child Language* 12: 271-295.

MacWhinney, Brian & Snow, Catherine E. (1990). The child language data exchange system: An update. *Journal of Child Language* 17: 457-472.

MacWhinney, Brian (1994). New horizons for CHILDES research. In J.L. Sokolov & C.E. Snow (eds.) 1994: 408-452.

MacWhinney, Brian (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum. 2 vols.

Ochs, Elinor (1979). Transcription as Theory. In E. Ochs & B. Schieffelin (eds.), *Developmental pragmatics*. New York: Academic Press.

Pan, Barbara A. (1994). Basic measures of child language. In J.L. Sokolov & C.E. Snow (eds.) 1994: 26-49.

Sánchez-Martínez, Juan Carlos (1994). Untersuchungen zu Tempus und Aspekt bei spanisch-deutsch bilingualen Kindern. Romanisches Seminar, Universität zu Köln. Ms.

Sokolov, Jeffrey L. & MacWhinney, Brian (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. & Moreton, Joy (1994). Individual differences in linguistic imitativeness. In J.L. Sokolov & C.E. Snow (eds.) 1994: 174-209.

Sokolov, Jeffrey L. & Snow, Catherine E. (eds.) (1994). *Handbook of Research in Language Development Using CHILDES*. Hillsdale, NJ: Lawrence Erlbaum.

Sokolov, Jeffrey L. & Snow, Catherine E. (1994). Transcript analysis using the Child Language Data Exchange System. In J.L. Sokolov & C.E. Snow (eds.) 1994: 1-25.

Thomas, Margaret (1994). Young children's hypotheses about English reflexives. In J.L. Sokolov & C.E. Snow (eds.) 1994: 254-285.