Transcription Conventions for the Eisenbeiss German Child Language Corpora

Sonja Eisenbeiss
University of Essex
seisen@essex.ac.uk

Ingrid Sonnenstuhl
Düsseldorfer Akademie

Essex Research Reports in Linguistics

Volume 60

Number 2

17 Jan, 2011

Dept. of Language and Linguistics,
University of Essex,
Wivenhoe Park,
Colchester, Essex, UK,
CO4 3SQ

http://www.essex.ac.uk/linguistics/publications/errl/

Essex Research Reports in Linguistics present ongoing research activities of the members of the Department of Language and Linguistics.

The main purpose of these reports is to provide a quick publication outlet. They have 'pre-publication status', and most will subsequently appear in revised form as research articles in professional journals or in edited books.

If you have technical problems downloading a paper, or for further information about these reports, please contact the editor:

Doug Arnold: doug@essex.ac.uk.

Citation Information:

Sonja Eisenbeiss and Ingrid Sonnenstuhl. 'Transcription Conventions for the Eisenbeiss German Child Language Corpora', Essex Research Reports in Linguistics, Vol. 60.2. Dept. of Language and Linguistics, University of Essex, Colchester, UK, Jan, 2011.
http://www.essex.ac.uk/linguistics/publications/errl/errl60-2.pdf

Transcription Conventions for the Eisenbeiss German Child Language Corpora

Sonja Eisenbeiss[1] and Ingrid Sonnenstuhl[2]

[1]University of Essex, [2]Düsseldorfer Akademie

Abstract

In this document, we describe the transcription conventions for the Eisenbeiss German child language corpora.[1] These conventions are based on the so-called CHAT transcription conventions of the **Ch**ild **L**anguage **D**ata **E**xchange **S**ystem (MacWhinney 2000; CHILDES: http://childes.psy.cmu.edu/; CHAT: http://childes.psy.cmu.edu/manuals/chat.pdf). We have incorporated modifications and additions of CHAT for German that were suggested by Stephany and Bast (1999; http://childes.psy.cmu.edu/intro/stephany.pdf) and by Heike Behrens (2006; pc.; http://childes.psy.cmu.edu/manuals/07germanic.doc#_Ref131136188).

1. The Eisenbeiss Corpus

This corpus, metadata for individual recordings and further information about the participants are available via the online archive of the Max Planck Institute for Psycholinguistics, Nijmegen: http://corpus1.mpi.nl/ds/imdi_browser/ MPI>Acquisition>L1 >Eisenbeiss. The corpus consists of two sub-corpora.

The first sub-corpus is the Eisenbeiss Elicitation Corpus, which contains several hundred recordings of three to six year old German and Dutch children The data collection involved semi-structured elicitation games that encouraged children to produce simple noun phrases, noun phrases with one or more adjectives and noun phrases with possessor phrases in a broad range of syntactic contexts (subject, direct and indirect object, adverbial phrases, etc.). See Eisenbeiss 2010, this volume, Eisenbeiss et al. 2009, for some of the elicitation tasks and materials used in the collection of the Eisenbeiss elicitation corpus.

The second subcorpus, the so-called L-Family corpus, involves more than 1000 recordings from a two-year observation of a monolingual German family with four children and two adults – the mother of the children and the father of the two younger children. All participants speak dialect-free standard High German; and during the recording period, the two

[1] The collection of both corpora was funded by the Max Planck Society and took place within the Acquisition Group of the Max Planck Institute for Psycholinguistics, headed by Prof. Wolfgang Klein. The development of the transcription conventions and some initial transcriptions of these corpora were funded by the Research Promotion Fund and the Research Endowment Fund of the University of Essex.

adults were involved in higher education and working part-time. An overview of the children is given in Tab.1:

Tab.1. Children involved in the L-Family Corpus

Child	Gender	Age	Year of birth	Day-care	School
Lenny (C1)	Male	5;2-7;8	1993	1997-2000	from 2000
Leon (C2)	Male	2;0-4;6	1996	from 1997	-
Liam (C3)	Male	0-2;5	1999	from 2000	-
Luna (C4)	female	0-0;4	2001	-	-

Two types of data have been obtained: (1) spontaneous speech of children, parents, and guests collected during meals and free play, and (2) semi-spontaneous speech from elicitation games targeted at various types of noun-phrases. The elicitation tasks include the tasks used in the Eisenbeiss Case Elicitation corpus (see Eisenbeiss, this volume). See Slobin et al. (in press) and Eisenbeiss et al (2009) for initial publications based on this corpus.

2. Tools Used for Transcription and Annotation

Transcriptions can be created and time-linked to the video/audio-stream using the tools provided by CHILDES or the multi-media-annotator ELAN (Wittenburg et al. 2006; http://www.lat-mpi.eu/tools/elan/). We used the latter as it was well supported at the Max Planck Institute, is XML-based, easy to use, and offers a lot of flexibility and excellent functionality for dealing with multi-speaker and multi-media corpora, for instance automatic recognition of silent periods in the recording. Import and export options are available for a range of text formats and software packages. In particular, ELAN allows for import from and export to the CHAT-format of CHILDES. Thus, one can use the so-called CLAN-tools of CHILDES for complex searches and frequency analyses or for semi-automatic morpho-syntactic annotation (MacWhinney 2000; http://childes.psy.cmu.edu/manuals/clan.pdf).

2. Transcription

We do not transcribe phonetically, but orthographically. As explained in more detail below, some non-target-like forms are standardised to facilitate computerised searches, but any deviations from standard German that are related to grammatical marking are transcribed as they are. In the following, examples are presented in German as this guide is targeted at users who want to transcribe or use German data.

2.1. Structure and Naming of ELAN (eaf-)files

For each speaker the ELAN file contains

- one line for the transcription (independent, parent of comment lines, default)
- one line for comments (referring to transcription tier)

Transcription tiers have the type TRANSCRIPTION (independent, default language English);and utterances are time-linked to the audio/video-stream.

Comment tiers have the type COMMENT (referring, symbolic association, default language English) and refer to the respective transcription tier for the speaker. Transcription/annotation files involving these tiers are created and the selection of tiers and their properties are saved as template files; see Tab.2-5 for some examples.

Tab.2: L-Family Template (L_Family_Core.etf)

	Tier Name		
Participant	Transcription	Comment	Coding
first child, Lenny	CH1	CH1_Com	CH1_Cas
second child, Leon	CH2	CH2_Com	CH2_Cas
third child, Liam	CH3	CH3_Com	CH3_Cas
fourth child, Luna	CH4	CH4_Com	CH4_Cas
mother, Natalie	MOT	MOT_Com	MOT_Cas
father, Ole	FAT	FAT_Com	FAT_Cas
researcher, Sonja	SON	SON_Com	SON_Cas
greatgrandmother, (Ma)tilde	TIL	TIL_Com	TIL_Cas

Tab.3: Elicitation Corpus Puzzle Task template (Elicitation_Puzzle.etf)

	Tier Name		
Participant	Transcription	Comment	Coding
Child	CH1	CH1_Com	CH1_Cas
researcher	RES	RES_Com	RES_Cas

Tab.4: Elicitation Corpus Picture-Pairing Template

(Elicitation_picture_pairing_two_children.etf)

	Tier Name		
Participant	Transcription	Comment	Coding
first child	CH1	CH1_Com	CH1_Cas
second child	CH2	CH2_Com	CH2_Cas
Researcher	RES	RES_Com	RES_Cas

Tab.5: Elicitation Corpus Picture-Pairing Template for Recordings with a Researcher (Elicitation_picture_pairing_child_researcher.etf)

	Tier Name		
Participant	Transcription	Comment	Coding
child	CH1	CH1_Com	CH1_Cas
researcher	RES	RES_Com	RES_Cas

The templates can be used to create new eaf-files, reducing the work load and ensuring consistency. Additional tiers for other participants (for instance occasional visitors to the family) can be added and tiers that are not used can be deleted. To create an ELAN file, transcribers need to start ELAN, and click on FILE > NEW and SELECT MEDIA. Then they must select the video-file (.mpg) and the corresponding sound file (.wav). The SELECT TEMPLATE option allows the transcriber to add the required template. If all files required are in the same folder, this should work easily and the resulting eaf. file can be saved immediately. In order to avoid data loss, files are saved frequently and under different names. All eaf-files created with ELAN for the Eisenbeiss L-Family and the Elicitation Corpus have the following structure:

<name of the wav/video-file>_<initials of the transcriber>_<two-digit running number of saved copy>.<eaf>

For instance, the first saved copy of a transcription file created by Sonja Eisenbeiss on the basis of the video- and audiofiles **ased2001Jul19b.mpg** and **ased2001Jul19b.wav**, would be called **ased2001Jul19b_SE_01.eaf**. The 17th copy would be called **ased2001Jul19b_SE_17.eaf**. When a

transcript has been checked by a second transcriber, the initials of this transcriber will be added, e.g. **ased2001Jul19b_SE_17_IS.eaf**.:

2.2. Special Characters and Capitals

Only ASCII-characters are used for transcriptions. Any special characters of German are transcribed using ASCII-characters: ä = ae; ö = oe; ü = ue; ß = ss.

Capital letters are only used for proper names, i.e. for names of people, animals, locations, etc. (e.g. *Lenny*). The beginning of sentences as well as nouns are not marked by capitals (e.g. *ich habe Lenny in Koeln gesehen.*).

2.3 The Spelling of Stems and Bases

As we are not specifically interested in the development of phonology and articulatory abilities, stems are slightly standardised to aid later computer searches. In particular, if one or more sounds are altered without any effects on grammatical marking, the standard forms are used in the transcription. For instance, both *dunnel* and *tonnel* are transcribed as *tunnel* if the intended word stem can be identified from the context.

(1)

Production	**Transcription**
guggema, guggemal	gucke mal
soen	schoen
tumachen	zumachen
topf	kopf (if the child refers to a head, not a pot)
wer, mer	wir (if the intended meaning is "we")

NOTE: changes of the stem vowel can function as grammatical markers (e.g. umlaut and ablaut). If this is the case, non-target forms of the stem vowel are transcribed as they are produced. The deviations from the target are indicated by an asterisk in square brackets and the target forms are provided in square brackets, after a colon and a blank; see e.g.:

(2)

Production	**Target**	**Transcription**
manner	maenner	manner [*] [: maenner]
er lauft	laeuft	er lauft [*] [: laeuft]
sie helft	hilft	sie helft [*] [: hilft]
huende	hunde	huende [*] [: hunde]

Missing syllables or individual sounds are added in round brackets – as long as grammatical marking is not affected:

(3)	**Production**	**Transcription**
	n	(eine)n
	nane	(ba)nane
	ma ma	ma(ch) ma(l)
	no ma	no(ch) ma(l)

NOTE: if the omission of syllables or individual sounds affects grammatical marking (e.g. *ein* instead of *eine*) and leads to non-target-like forms, no round brackets are used. Rather, the deviation from the target is indicated by an asterix in square brackets; and the target form is provided in square brackets, introduced by a colon; see below.

2.4 The Spelling of Grammatical Markers

Grammatical markers, i.e. inflectional morphemes and derivational morphemes, are never standardised.

Deviations from the target that are not acceptable in colloquial speech, are indicated by an asterisk in square brackets and the target forms are provided in square brackets, introduced by a colon (see Stephany and Bast 1999):

(4)	**Production**	**Transcription**
	ich treffte ihn.	ich treffte [*] [: traf] ihn.
	ich fahre [//] fahrden rad .	ich fahre [//] fahrden [*] [: fuhr] rad .
	ich fallt.	ich fallt [*] [: fiel] .
	da ist eine gelbe haus	da ist eine [*] [: ein] gelbe [*] [: gelbes] haus

Incomplete, but acceptable colloquial forms are not marked by an asterisk. If it is possible to add the missing material, this is done in round brackets. If not, the (nonreduced) standard form is added in square brackets after a colon and a blank (see Stephany and Bast 1999):

(5)	**Production**	**Transcription**
	ich geh	ich geh(e)
	die habn sich was geholt.	die hab(e)n sich was geholt.
	ham sich was geholt .	ham [: haben] sich was geholt .

2.5 Compound nouns and other Complex Nouns

All components of morphological compounds are separated by '+':

(6) bahn+hof
butter+brot
mecker+tante

An underscore can be used to ensure that several nominal elements are analysed as one unit by search-programs, even though these elements form a phrasal combination, not a compound, e.g.:

(7) Neuss_Norf
Mickey_Mouse
Doctor_Mueller
Raupe_Nimmersatt
null_komma_nix
rucki_zucki

NOTE: All words are written out: ‘Sankt’ (not ‘St.’), ‘Doktor’ (not ‘Dr.’) (see Stephany and Bast 1999).

2.6. Punctuation and Apostrophies

The following punctuation characters can be used: **, . ; ? !** The end of each utterance has to be marked by a full stop, a question mark or an exclamation mark (see Stephany and Bast 1999). Commas and semicolons can be used within an utterance. Clitics are separated from preceeding words by a blank and an apostrophy:

(8) die schaffen 's schon

2.7. Doubtful Material and Unintelligible Speech

If it is not entirely clear what the speaker actually said, the word or group of words serving as a best guess may be marked by '[?]' (s. Stephany and Bast 1999):

(9) die kommt da rauf [?]

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

(10) die kommt da rauf [=? raus]

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

2.8. Scoped Symbols

Symbols placed in square brackets ('[ ]', e.g. [*] or [?]) can refer to single words or to more material. Then, the entire sequence must be surrounded by pointed brackets ('< >'; s. Stephany and Bast 1999):

(11) <ich fahre> [/] ich fahre rad.
ich muss <da rauf> [?]
ich muss <da rauf> [=? das rauf]

2.9. Onomatopoeic Forms, Variants of Colloquial Forms, Interjections, etc.

Onomatopoeic forms are marked by '@o' (s. Stephany and Bast 1999):

(22) kikeriki@o macht der.

In written German, many colloquial forms, interjections, and swear words can be spelt in different ways. For the sake of consistency, we are using only one spelling version. Interjections such as *auweia* are spelt as one word as they will not be analysed further.

(13)

Variants	**Transcription**
Kuck mal, guck mal, gucke mal	guck(e) mal
Auweia, au weia, auweija	auweia
oweh, o weh	oweh
o gott, oh gott, ogott,	ogott
hurra, hurrah,	hurra

ba, bah,	ba
baeh, baechs	baeh
hoppla, hoppela, hoppala	hoppla
tschuess, tschuessi, tschuehuess	tschuess
bums, bumms,	bums
prost, proscht, prosit	prost
o la la, oh lah lah, olala	olala
aha, ahah	aha
ach	ach
och	och
pst, pscht	pst
boing	boing
okay, ok	ok
jo, ja,	ja
pfui	pfui
nee, noe,	nee
scheiss, scheisse	scheiss(e)
herrgottnochmal, herrgottnochemal	
herr gott noch einmal	herrgottnoch(ein)mal

For standardisation purposes, each transcriber records the current list of all variants and their transcriptions in the document variants_<year in 4 digits>_<month in two digits>_<day in two digits>_ <transcriber initials>.doc (e.g. *variants_2007_10_03_SE.doc*).

2.10. Interjections and Direct Speech

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 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 (s. Stephany and Bast 1999):

(12)	**Variants**	**Transcription**
	interrogative interjection	hm@ii
	affirmative interjection (‘yes’)	hm@ia
	negative interjection (‘no’)	hm@in

Direct speech is marked by a plus and double quotes at the beginning and double quotes at the end:

(15) und dann hat Ben gesagt +" du bist doof" und ist rausgerannt

2.11. Omitted Words and Constituents

Though the coding of word and constituent omissions is an extremely difficult and unreliable process, it is very helpful for analyses of the feedback that children receive for non-target-like utterances. Hence, we code omissions of words by placing a combination of the zero symbol with the missing word on the transcription tier for the respective speaker. If what is important is not the actual word omitted, but its grammatical category (part-of-speech), or if it is not possible to determine the missing word then a code for the category can follow the zero:

- 0art: article omission

  This code is only used if the lack of an article leads to an unacceptable noun phrase (e.g. *da ist haus*). Note that a noun without an article is acceptable if there is another determiner (*ich will dieses haus*), a possessive pronoun (*das ist mein haus*) or possessive noun (*das ist Susis haus*), a plural context (*da sind huehner*) or when the noun is a mass noun (*ich will wasser*) or a name (*da ist Susi*). Thus, NO article omission is coded in these circumstances, even when an article would be acceptable in the respective context (e.g. *da ist die Susi*) .
- 0aux: auxiliary omission

  This code is used if the sentence contains a past participle without an auxiliary and it is not clear which auxiliary is missing (e.g. Max *ist/hat geschwommen*). If the auxiliary form can be identified, the zero is combined with the respective missing auxiliary form.
- 0subj: subject omission

  This code is used if the subject is missing in a sentence that would require a subject in the target language. I.e., this code is not used in imperatives, even though they lack overt

subjects. Sometimes it cannot be determined where the missing subject would occur. This is not important, as long as 0subj is positioned somewhere in the sentence.

- 0do: omission of direct object

  This code is only used if the sentence is clearly unacceptable without the direct object (e.g. *ich bewache jetzt aber* or *ich hab angemalt*). It is not used if the sentence is acceptable without a direct object (e.g. *ich esse gerade*).

- 0io: omission of indirect object

  This code is only used if the sentence is ungrammatical without the indirect object. It is not used if the omission would be acceptable (e.g. *ich geb jetzt mal* during a card game).

- 0prep: omission of preposition

  This code for a missing preposion is only used if it is clear that a preposition is missing, but it is not possible to determine which specific preposition is missing. When the preposition can be determined, zero is combined with this preposition.

(14)

Production	**Transcription**
da ist haus	da ist 0art haus
ich weggegangen	ich 0bin weggegangen
Max geschwommen	Max 0aux geschwommen
da singt jetzt	da singt 0subj jetzt
ich hab angemalt	ich hab 0do angemalt
ich geb die blume	ich geb 0io die blume
ich will mama	ich will 0zur mama
ich gehe schule	ich gehe 0p 0art schule (NOTE: while one can assume that a preposition and an article are missing, one cannot determine the exact preposition and article as the target utterance might be *in die* or *zur schule*)

2.12. Pauses, Retracing, Interruptions and Completions

Utterance-internal **unfilled pauses** are marked by '#' if they last for at least 2 seconds; if they last for at least 8 seconds, this is indicated by '##'. **Filled Pauses** (fp) are transcribed by 'eh@fp'. Two filled pauses in a sequence are marked by 'eh@fp eh@fp' (s. Stephany and Bast 1999):

(16) das ist ## ein # eh@fp huhn

Repetition without correction is marked by '[/]'; repetition with correction is indicated by '[//]'. If several words are repeated, they are placed in pointed brackets ('< >'). Several repetition marks may be used in one and the same utterance:

(17) ich fahre [/] fahre rad .
<ich fahre> [/] ich fahre rad .
<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 (s. Stephany and Bast 1999):

(18) der [/] eh@fp der brief.

Uninvited interruptions by other speakers are marked by '**+/**' at the end of the annotation, followed by clause end punctuation. Self-interruptions are indicated by '**+//**' (s. Stephany and Bast 1999):

(19) MOT wie oft muss ich dir +/ ?
CH1 doch.

(21) MOT und jetzt nimmst du +//.
MOT nee, so nicht!

Indicating completions in the transcript is crucial for studies of interactions between the child and other speakers. Hence, self-completion is marked by '+,' and completion by others is indicated by '++' at the beginning of the annotation that contains the completion. The utterance that is interrupted (by the speakers themselves or by other speakers) is marked by '+...' at the end of the annotation.

(23) CH1: und dann war da so ein +...
MOT: ein was?
CH1: + haus.
MOT : und dann war da so ein +...
CH1: ++ haus.

2.13. Explanations and Comments

Brief explanations can be given on the transcription tier for the respective speaker, indicated by [= text]. The comment tiers can be used to provide information that is necessary for the interpretation of the utterances, whether this information can be retrieved from the video itself or not. For instance, references to movies, fictional characters, etc. are explained whenever possible (asking the participants in the recordings might be required for this). Context information must be reliable, whether it is based on information that was provided by the video or by consultations with participants in the recording.

(24) CH1 ich will den gelben [=ball] hier.
CH1_Com He is pointing to the red ball. There is no yellow ball in the room.

Acknowledgements

We would like to thank the University of Essex Research Promotion Fund and Research Endowment Fund for supporting our work on child language corpora, which was the basis for this paper. We would also like to thank the Max Planck Society, Wolfgang Klein, and the Technical Group at the Max Planck Institute for Psycholinguistics in Nijmegen for their support in the development of the Eisenbeiss corpora.

References

Behrens, H. (2006). The input-output relationship in first language acquisition. *Language and Cognitive Processes*, 21, 2-24.

Clahsen, H., Eisenbeiss, S., and Penke, M. 1996. Lexical Learning in early syntactic development. In: Harald Clahsen (ed.), *Generative perspectives on language acquisition*. *empirical findings, theoretical considerations and crosslinguistic comparisons.* Amsterdam: John Benjamins, 129-159.

Clahsen, H., Eisenbeiss, S., and Vainikka, A. 1994. The seeds of structure. A syntactic analysis of the acquisition of case marking. In: Hoekstra, T. and Schwartz, B.D. (eds.). *Language acquisition studies in generative grammar*. Amsterdam: John Benjamins, 85-118.

Eisenbeiss, S. 1994. Kasus und Wortstellungsvariation im deutschen Mittelfeld. Theoretische Überlegungen und Untersuchungen zum Erstspracherwerb. In: Haftka, B. (ed.), *Was*

determiniert Wortstellungsvariation? *Studien zu einem Interaktionsfeld von Grammatik, Pragmatik und Sprachtypologie.* Opladen: Westdeutscher Verlag, 277-298.

Eisenbeiss, S. 2003. *Merkmalsgesteuerter Grammatikerwerb.* Dissertation University of Düsseldorf. (downloadable from http://deposit.ddb.de/cgi-bin/dokserv?idn=97646330x&dok_var=d1&dok_ext=pdf&filename=97646330x.pdf)

Eisenbeiss, S. 2010. Production methods in language acquisition research. In: Blom, E. and Unsworth, S. (eds.) *Experimental methods*. Amsterdam: John Benjamins, 11-34.

Eisenbeiss, S., Bartke, S., and Clahsen, H. 2005/2006. Structural and lexical case in child German: evidence from language-impaired and typically-developing children. *Language Acquisition* 13:3-32.

Eisenbeiss, S., Matsuo, A., Sonnenstuhl, I. 2009. Learning to encode possession. In: McGregor, W.B. (ed.): *The expression of possession*. Berlin: deGruyter, 143-211.

MacWhinney, B. 2000) *The CHILDES project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Slobin, D.I., Bowerman, M., Brown, P., Eisenbeiss, S., Narasimhan, P. (in press). Putting things in places: developmental consequences of linguistic typology. To appear in: Bohnemeyer, J. and E. Pederson, *Event representation in language: Encoding events at the language-cognition interface*. Cambridge: Cambridge University Press.

Stephany, U. and Bast, C. 1999. *Working With The Childes Tools: Transcription, Coding And Analysis.* (downloadable from http://childes.psy.cmu.edu/intro/stephany.pdf)

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H. 2006. ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation. http://www.lat-mpi.eu/papers/papers-2006/elan-paper-final.pdf