University of Essex

# A CHAT-Based Annotation Scheme for Case and Noun-Phrase Inflection in Child Language Data

*Sonja Eisenbeiss*
University of Essex
seisen@essex.ac.uk
*Ingrid Sonnenstuhl*
Düsseldorfer Akademie

Essex Research Reports in Linguistics

Volume 60

Number 3

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. 'A CHAT-Based Annotation Scheme for Case and Noun-Phrase Inflection in Child Language Data', Essex Research Reports in Linguistics, Vol. 60.3. Dept. of Language and Linguistics, University of Essex, Colchester, UK, Jan, 2011.
http://www.essex.ac.uk/linguistics/publications/errl/errl60-3.pdf

# A CHAT-Based Annotation Scheme for Case and Noun-Phrase Inflection in Child Language Data

**Sonja Eisenbeiss[1] and Ingrid Sonnenstuhl[2]**

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

## Abstract

This paper describes a coding scheme and a set of semi-automatic procedures for the annotation of complex noun phrases and their morpho-syntactic properties in child language data. These tools are based on the CHAT 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). The coding scheme presented here focuses on the order and grammatical category of the individual elements in the noun phrase and their gender, number and case marking. It also provides information about the category and lexical identity of the element that assigns case to the respective noun phrase (e.g. the dative preposition *mit* 'with'). The coding scheme was developed for German child language, but it can be adapted to other languages and populations.

## 1. Introduction

This paper describes a coding scheme and a set of semi-automatic procedures for the annotation of complex noun phrases and their morpho-syntactic properties in child language data. All conventions and procedures described in this document were developed on the basis of the CHAT 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) and modifications for German (Stephany and Bast 1999: http://childes.psy.cmu.edu/intro/stephany.pdf, Heike Behrens 2006, pc., http://childes.psy.cmu.edu/manuals/07germanic.doc#_Ref131136188). We modified the CHAT conventions to achieve a targeted coding scheme for the domain of case, noun-phrase structure and nominal inflection; based on coding schemes used for earlier publications

(Clahsen et al. 1994, 1996, Eisenbeiss 1994, 2003, Eisenbeiss et al. 2005/6). Our coding scheme focuses on the order and grammatical category of the individual elements in the noun phrase and their gender, number and case marking. It also provides information about the category and lexical identity of the element that assigns case to the respective noun phrase (e.g. the dative preposition *mit* 'with'). The coding scheme was developed for German child language, but it can be adapted to other languages and populations. The coding scheme can be combined with CHAT-based transcription conventions (see MacWhinney 2000, http://childes.psy.cmu.edu/manuals/chat.pdf, Stephany and Bast 1999, Heike Behrens 2006, pc., http://childes.psy.cmu.edu/manuals/07germanic.doc#_Ref131136188, Eisenbeiss and Sonnenstuhl, this volume).

## 2. Background

To date, there is no fully developed (semi-)automatic annotation tool that creates optimal annotations for a systematic analysis of case-marking, noun-phrase internal agreement and the morpho-syntactic properties of complex noun phrases. Developing such a tool is particularly difficult for a language like German as German has a distinction between two types of cases (Eisenbeiss et al. 2005/6): Structural cases are associated with particular syntactic functions or positions, e.g. the accusative default case for direct objects. Lexical or idiosyncratic case cannot be predicted on the basis of syntactic positions or thematic roles; e.g. the dative assigned by verbs like *passen* 'fit'. Moreover, the same case can be assigned by a verb and a preposition within the same sentence; e.g.

(1) *Ich* *gebe* *dem Mann* *von meiner Schwester*
*I* *give* [*the husband*] $_{DAT}$ *of* [*my sister* ] $_{DAT}$

*den Schluessel* *fuer den Keller*
[*the key*]$_{ACC}$ for the cellar]$_{ACC}$
'I give my sister's husband the key for the cellar'

The availability of lexical case and the potential to have more than one instance of the same case in one utterance make it necessary to encode links between case-assigners and case-marked elements. Another difficulty for coding schemes results from the way in which case is morphologically realised in German: German has both regularly and irregularly inflected nominal forms and a broad range of portmanteau forms that encode a combination of case,

number and gender distinctions. Finally, German has massive syncretism in noun, determiner and adjective paradigms. For instance, the definite article form *der* 'the'could be analysed as a nominative masculine singular form, a dative feminine singular form, a genitive feminine singular form or a genitive plural form. Thus, one cannot simply replace each case form in a transcript with a unique code for a particular case/gender/number inflection. One either has to make a decision for each phrase or one has to use codes with several alternatives. Coding child language poses the additional problem that the word forms the child produces may not be target-like. For such deviations from the target, both the non-target-like child form and the corresponding target form have to be provided.

The CHILDES database offers a preliminary version of a semi-automatic morpho-syntactic tagger for German child language – the so-called German MOR-Grammar (Brian MacWhinney and Heike Behrens, p.c.; Stephany and Bast 1999; see http://childes.psy.cmu.edu/morgrams/). However, this tool is still under development; and there are quite a few remaining problems with the segmentation of inflected words that would need to be addressed, particularly in the annotation of inflected forms. For instance, the high-frequency dative plural form *Kind-er-n* 'children' is not parsed or recognised. Moreover, adjectives are not coded for case, number, and gender inflection; and in our pilot, the application of the MOR-grammar resulted in three alternative analyses for the uninflected adjective form *schwer* 'heavy/difficult': the correct analysis (adj|schwer), an analysis where *schw* is incorrectly treated as the stem and *–er* as a comparative ending (adj:CP|schw-CP) and an analysis, where *schw* is incorrectly analysed as a stem and *–er* as a case/gender/number inflection (adj|schw-er).

Thus, the use of the German MOR-grammar would require a considerable investment of time and money to adapt it and apply it. A lot of this time would be spent on coding and checking codes for aspects of the corpora that are irrelevant for projects on noun-phrase structure and inflection (e.g. codes for adverbs, verbal inflection, etc.). Hence, we developed and piloted an annotation scheme that focuses on noun-phrases and case-assignment and provides additional information for this purpose. This scheme involves codes for article and argument omissions, links between predicates and arguments, and the provision of case, gender, and number information for each argument. Hence, it provides efficient coding for a broad range of analyses in the domain of noun-phrase-internal agreement, argument realisation, determiner realisation, pronoun development, etc. Moreover, this coding scheme employs the standard annotations used in standard MOR-grammars and the CHAT-format of

CHILDES. This allows us to use any coded data sets to contribute to the further development and evaluation of the German MOR-grammar.

## 3. Information Encoded

The coding scheme described below provides information about (i) the individual case assigning lexeme, the class of case-assigner (e.g. dative preposition), (iii) the case-marked phrase, and (iv) individual case-marked word forms within this phrase. We do not attempt a full coding of all elements in a sentence (e.g. adverbs, conjunctions, etc.). However, the coding could easily be augmented by additional (CHAT) codes for elements that are not included in the coding scheme presented here. The individual codes for categories of case-marked and case-assigning elements have been adopted from the CHILDES codes for English (http://childes.psy.cmu.edu/manuals/chat.pdf, p.95), the German MOR-grammar provided on the CHILDES Webpage, and the advice for German offered by Stephany and Bast (1999).

However, we have used the individual codes in a slightly modified way. In the publicly available German MOR grammar, adjectives, nouns, determiners and pronouns are not annotated uniformly. In particular, adjectives, which are highly syncretic in German, are not coded for case, gender, and number. Instead, the affix (e.g. *–en*, *-es*, or *-er*) is simply added, using a hyphen. The same was done for most determiners. For instance, the indefinite article form *einer* 'a' and the possessive pronoun form *meine* 'my' are coded as "det|ein-er" and "det|mein-e", respectively. In contrast, pronouns and definite articles have codes for case, gender and number in the MOR grammar. For instance the definite article form *der* can be coded as "art|der&F" if it is interpreted as a feminine singular form in the nominative or the accusative, or as "art|der&M" if it is analysed as a masculine singular form in the nominative. This treatment of adjectives and determiners is inconsistent. More importantly, coders either have to accept a list of several alternatives for syncretic forms like the definite article form *der* or disambiguate non-target-like forms while they code. This means, for instance, that one has to decide whether a child's substitution of the article form *der* for the article form *dem* produced a case error (nominative instead of dative masculine singular), a gender error (feminine instead of masculine dative singular) or a combined case and number error (genitive plural instead of dative singular). Making such decisions without detailed knowledge of the child's current grammatical system is not only time-consuming, but potentially misleading. Hence, we have generalized the coding method that the MOR-grammar employs for adjectives and many determiners to ALL nominal elements. The only

grammatical features that are encoded for the word forms themselves are (i) features realised by derivational affixes (e.g. comparative affixes on adjectives) and (ii) inherent properties of the lexical elements: their respective grammatical category (part-of-speech; e.g. adjective, determiner or noun) and the gender of nouns (e.g. “mann&M”).

Hyphens are used to indicate separable affixes, while “&” is employed to indicate inherent features (like the gender of nouns) and morphemes that are not separable. Compare e.g. the gender coding for the noun Mann “mann&M”, the code for the regular comparative from *kleiner* ‘smaller’ “adj:CP|klein-CP” and the code for the irregular comparative form *besser* ‘better’ “adj:CP|gut&CP”. The respective codes are presented in the following sections.

In addition, we code which case, gender and number features the respective noun phrase should exhibit and which element assigns the respective case. For instance, if a child says **Ich gebe der Pferd einen Apfel* ‘I give *the horse an apple’ instead of *Ich gebe dem Pferd einen Apfel*, the codes show that the child should have produced a dative masculine singular form (*dem*) for an indirect object of the ditransitive verb *geben* ‘give’ – but used *der* instead. The respective codes and their combinations are discussed below.

This combination of a code for the case/gender/number-context and a code for the affix that the respective speaker produced, has proved to be highly effective in earlier projects (e.g. Clahsen et al. 1994, 1996, Eisenbeiss 1994, 2003, Eisenbeiss et al. 2005/6); and it does not introduce any biases into the coding process. In earlier work, we have also successfully coded verbal inflection with a distinction between (i) context-codes for the features tense, person, and number, and (ii) codes for the verb form itself, which only indicate its morphological structure, category and lexical properties (e.g. v|les-en or v:aux|hab-t).

In addition to information about case-marked forms and grammatical features of their contexts, our coding scheme also contains information about links between each case-form and the respective case-assigner, which the standard MOR-grammar does not provide. For instance, when we want to investigate dative case assignment by the verb *helfen* ‘help’, standard MOR-annotations would only allow us to search for all utterances with *helfen* and a particular case-marked form, e.g. the definite dative feminine singular article *der*. The output of such a search could contain relevant utterances; i.e. utterances where the article is part of the dative argument; e.g. *Die Frau hilft **der** Tochter bei den Hausaufgaben* ‘The woman is helping **the$_{\mathbf{DAT}}$** daughter with the homework.’. However, such a search would also find the article form *der* when it is part of a nominative subject or a prepositional phrase in a sentence with *helfen* (e.g. ***Der** Mann hilft dem Kind bei den Hausaufgaben.* ‘**The$_{\mathbf{nom}}$** man is helping the daughter with the homework.’ or *Die Frau hilft dem Kind bei **der** Rechenarbeit.* ‘The woman

is helping the child with **the$_{DAT}$** maths work.'). One could of course eliminate all irrelevant hits of these searches by hand. However, this is a time-consuming task and it does not enhance the original annotation of the transcripts for further analyses. Our coding scheme includes information about the case-assigner in the coding for each case-marked form.

## 4. Part-of-Speech Codes

The following codes for grammatical categories were taken from the CHAT manual, but we have added a few more subcategories, highlighted in bold. In particular, we have added codes for definite and indefinite articles, making use of codes from the MOR-grammar and related codes from the CHAT manual. For instance, we use the "indefinite" and "demonstrative" codes not only for pronouns – as in the German MOR grammar, but also for determiners. This allows us to distinguish between definite articles (*der*), indefinite articles (*ein*-) and demonstratives (*dies*- 'this'). The codes for types of determiners were taken from the list of determiners in the German MOR-grammar.

In our coding, we do not distinguish between pronominal forms of determiners and non-pronominal forms of determiners (*Wo ist **der** (Mann)?* 'Who is **the** (man)?'). Pronominal and non-pronominal forms of determiners in German only differ morphologically for determiners that end in *–ein* (e.g. indefinite articles and possessive pronouns), and only in nominative masculine singular and nominative/accusative neuter contexts. In these contexts, the pronominal forms have the so-called "strong" endings *-er* for nominative masculine singular and *–es* for nominative/accusative neuter singular, whereas the non-pronominal forms remain uninflected (e.g. *Da ist **einer***. 'There is one.' vs. *Da ist **ein** Mann* 'There is a/one man.'). Our codes for case-assigners group elements occurring in the same noun phrase together. Thus, one can see whether a determiner is part of a phrase with other elements or occurs as a pronoun. Moreover, our transcripts lines for each speaker contain the error codes provided by CHAT ([*]; see Eisenbeiss and Sonnenstuhl, this volume). Hence, one can also use this code to determine whether an observed uninflected form like *ein* was target-like or not. However, we will use codes to distinguish between articles and relative pronouns, even if they have the same form (***die**$_{article}$ frau, **die**$_{relative\text{-}pronoun}$ ich gesehen hab* 'the woman **that** I have seen'). Personal pronouns that are used as reflexives (*Ich wasche **mich***. 'I wash myself') will be coded as personal pronouns. Only pronoun forms that are unique reflexive forms (*sich* 'himself/herself/itself') will be coded as reflexives. Tab.1 provides part-of-speech codes. Note that ordinal numbers (e.g. *zweite* 'second') are coded as adjectives, in line with CHAT-conventions. Below, we will show how different types of indefinite determiners,

demonstratives, etc. are distinguished by lexical information (see in particular Tab.4). Tab.2 and Tab.3 show the codes for other grammatical features that are used to code nouns or contexts for pronominal or full noun phrases.

*Tab.1: Part-of-Speech Codes for Nominal Elements*

| Category | Code | Sub-categories | Code | Example |
|---|---|---|---|---|
| adjective | adj | | | *rot* |
| adjective | adj | participle used as adjective | a:part | *angemalt* |
| determiner | det | possessive pronoun | det:poss | *mein* |
| determiner | det | definite article | det:def | *der* |
| determiner | det | indefinite determiner | det:indef | *eine, kein, irgendwelche,* |
| determiner | det | demonstrative pronoun | det:dem | *diese, jene* |
| determiner | det | interrogative | det:int | *welcher* |
| noun | n | noun | n | *house* |
| noun | n | proper noun | n:prop | *Lenny* |
| number | num | number word | | *zwei* |
| preposition | prep | dative-assigning | p:dat | *mit, bei ,...* |
| preposition | prep | accusative-assigning | p:acc | *fuer, ohne, ...* |
| preposition | prep | genitive-assigning | p:gen | *wegen, angesichts...* |
| preposition | prep | accusative/dative-alternating | p:accdat | *in, auf,...* |
| pronoun | pro | personal pronoun | pro:pers | *ich, wir, sie,...* |
| pronoun | pro | reflexive | pro:refl | *sich* |
| pronoun | pro | relative | pro:rel | *dessen, deren, den,…* |
| pronoun | pro | indefinite pronoun | pro:indef | *irgend(et)was/jemand/wann/wer/wie/wo/wohin, jedermann, jemand, man, niemand,...* |
| pronoun | pro | interrogative pronoun | pro:int | *wer, wen, wem, was, wieviel, wievielte ...* |
| pronoun | pro | negative pronoun | pro:neg | *nichts* |
| quantifier | qn | | | *alle, jeder,...* |

*Tab.2: Codes for Syntactic Contexts*

| Case | Context | Description | Example |
|---|---|---|---|
| NOM | SUB | nominative subject | ***Der Mann*** *gibt dem Bären den Honig(topf).*<br>'**The man** is giving the bear the honey(pot).' |
| | PRED | predicative nominative noun phrase | *Das ist* ***ein/der Mann****.*<br>'That is **a/the man**.' |
| ACC | DO | direct accusative object | *Der Mann gibt dem Bären* ***den Honig****.*<br>'The man is giving the bear **the honey**.' |
| | PRED | predicative accusative noun phrase | Er nannte ihn **einen Idioten**<br>'He called him **an idiot**' |
| | PP | accusative complement of a preposition | … ***auf den Rücken****.*<br>'…**on the back'** |
| | ADV | adverbial accusative | *Er sang* ***den ganzen Abend***<br>'He sang **all evening'** |
| DAT | SA | single dative argument of a one-place predicate | ***Mir*** *ist schlecht*<br>**I$_{DAT}$** is sick<br>'I feel sick' |
| | DO | dative object of a two-place verb | *Der Honig schmeckt* ***dem Bären****.*<br>'The honey tastes good **to the bear**.' |
| | IO | indirect dative object of a three-place verb | *Der Mann gibt* ***dem Bären*** *den Honig.*<br>'The man is giving **the bear** the honey.' |
| | PP | dative complement of a preposition | *...****mit dem Helm****.*<br>'…**with the helmet**' |
| | EXT | "extra" dative argument | *Wasch* ***dir*** *die Hände!*<br>Wasch **you$_{DAT}$**the hands!<br>'Wash your hands!' |
| | ETH | "ethical" dative | *Renn* ***mir*** *nicht so schnell!*<br>Run **me$_{DAT}$** not so fast!<br>'Don't run so fast (I don't approve of fast running)!' |
| GEN | DO | genitive object of two-place verb | Er gedenkt **seiner Mutter**<br>'He is commemorating **his mother**' |
| | IO | genitive indirect object of a three-place verb | Er bezichtigt ihn **des Mordes**.<br>He accuses him **the$_{GEN}$ murder$_{GEN}$**<br>'He accuses him of murder' |
| | PP | genitive complement of a preposition | … *wegen* ***des Regens***<br>… because **the$_{GEN}$ rain$_{GEN}$**<br>'... because of **the rain**' |
| POSS | POSS | -s possessor in an adnominal possessive construction | ***Lennys*** *Haus*<br>'**Lenny's** house' |
| CAS | CON | unclear case context | *xxx* ***das Haus****.*<br>'xxx **the house**' |

*Tab.3: Codes for Gender and Number Features*

| Category | Feature | Code |
|---|---|---|
| Gender | Masculine | M |
| | Femininine | F |
| | Neuter | N |
| | Unclear | G |
| Number | Plural | Pl |
| | Singular | SG |
| | Unclear | NU |
| Person: Singular (for personal pronouns only) | 1st person | 1S |
| | 2nd person | 2S |
| | 3rd person | 3S |
| Person: Singular (for personal pronouns only) | 1st person | 1P |
| | 2nd person | 2P |
| | 3rd person | 3P |

Note omissions of parts of speech (e.g. article omissions) are encoded directly in the transcription, making use of the respective CHAT-transcription conventions (see MacWhinney 2000 and Eisenbeiss and Sonnenstuhl, this volume, for details). They will be copied into the coding tier.

## 5. Morphological Structure

We use the following conventions from the CHAT-manual to encode the morphological structure of word forms.

(2) prefix#
part-of-speech|
stem
&fusionalsuffix
-suffix

For instance, the form *ungluecklichen* 'unhappy'would be coded as "a|un#glueck-lich-es". As mentioned above, we do not use codes for inflectional morphemes, but the actual form. For fusional inflectional morphology, we will use the code "&umlaut". For instance, the noun

plural *Kuehe*[1] 'cow' would be coded as "n|kuh&umlaut-e". Irregular forms of definite articles can be presented directly after the "|" symbol as the lexeme can be identified using the subcategory code, e.g. "det:def|das" and "det:def|die".

German has many amalgams of prepositions and articles (e.g. *im=in dem* 'in the', *ins=in das* 'into the' or *aufs=auf-das* 'onto the'). We will use the conventions of the German MOR-grammar and link the codes for the preposition and the codes for the article, e.g. *im* prep:accdat|in~det:def|dem 'in the'. Note that the code for the case refers to the case assignment properties of the preposition, not to the actual form of the amalgam. I.e., if the child incorrectly combined the dative article form with the accusative-assigning preposition *fuer* 'for' (**fuerm*), we would code this as prep:acc|fuer~det:def|dem. Recall that the non-target-like form would be marked by "[*]" on the transcription tier and the target form would be provided in square brackets.

In addition to codes used in the CHAT-manual and MOR-grammars, we introduced markers for nouns that require morphological case-markers, for instance so-called weak masculine nouns like *Junge* 'boy', which take an *–en* marker in non-nominative and plural contexts (see below).

## 6. Annotations for Pronouns, Determiners, Quantifiers and Numerals

As explained above, all pronouns, determiners, quantifiers and numerals are only coded for their categorical features and their morphological structure. Case and number contexts and syntactic role are encoded with a separate set of codes. These codes are added to the codes for the individual word forms, as explained below. Otherwise, we mostly follow the conventions described in the CHAT manual. The distinction between different types of determiners and quantifiers that is shown in Tab.4 was adopted from the German MOR-grammar.

In line with the CHAT manual, but in contrast to the German MOR-grammar, we did not distinguish between determiners and articles. We also did not distinguish between pronominal and non-pronominal forms of determiners. This makes any automatic coding easier. Moreover, it is in line with the German MOR-Grammar and reflects the observation that pronominal and non-pronominal forms only differ for indefinite articles, possessive pronouns, in nominative masculine singular and nominative/accusative. In these contexts, the pronominal forms have the "strong" endings *-er* for nominative masculine singular and *–es* for nominative/accusative neuter singular, whereas the non-pronominal forms remain uninflected. Our codes for case-assigners group elements occurring in the same noun phrase

[1] We do not use umlaut symbols in our transcriptions; see Eisenbeiss and Sonnenstuhl, this volume.

together as they have the same case-assigner. Thus, one can see whether a determiner is part of a phrase with other elements or occurs as a pronoun. Moreover, our transcripts contain CHAT-format error codes ([*]). Hence, one can also use this code to determine whether an inflected or uninflected form was selected appropriately. Note that suppletive forms like definite articles are coded using the actual morphological form, while affixed forms are decomposed into stems and affixes.

*Tab.4: Annotations for Determiners and Quantifiers*

| Category | Example | Code Template | Example code |
|---|---|---|---|
| definite article (also if used as a pronoun) | *der* | det:def\|FORM | det:def\|der |
| definite article (also if used as a pronoun) | *das* | det:def\|FORM | det:def\|das |
| demonstrative pronoun, inflected | *jenes* | det:dem\|LEXEME-AFFIX | det:dem\|jen-es |
| demonstrative pronoun, uninflected | *dies* | det:dem\|LEXEME | det:dem\|dies |
| indefinite article, inflected | *einer* | det:indef\|LEXEME-AFFIX | det:indef\|ein-er |
| indefinite article, uninflected | *ein* | det:indef\|LEXEME | det:indef\|ein |
| indefinite determiner, inflected | *keinen* | det\|LEXEME-AFFIX | det:indef\|kein-en |
| indefinite determiner, uninflected | *kein* | det\|LEXEME | det:indef\|kein |
| indefinite pronoun | *man* | pro:indef\|FORM | pro:indef\|man |
| interrogative determiner, inflected | *welches* | det:int\|LEXEME-AFFIX | det:int \|welch-es |
| interrogative determiner, uninflected | *welch* | det:int\|LEXEME | det:int\|welch |
| interrogative pronoun, irregularly inflected | *wer* | pro:int\|FORM | pro:int\|wer |
| interrogative pronoun, regularly inflected | *wievielte* | pro:int\|LEXEME-AFFIX | pro:int\|wievielt-e |
| irgend+wh-pronoun, inflected | *irgendwelche* | det:indef\|LEXEME-AFFIX | det:indef\|irgendwelch-e |
| negative pronoun | *nichts* | pro:neg\|FORM | pro:neg\|nichts |
| number, cardinal | *zwei* | num\|LEXEME | num\|zwei |
| number, ordinal | *zweite* | adj\|LEXEME-AFFIX | adj\|zweit-e |
| personal pronoun | *ich* | pro:per\|FORM | pro:per\|ich |
| possessive pronoun, inflected | *seiner* | det:poss\|LEXEME-AFFIX | det:poss\|sein-er |
| possessive pronoun, uninflected | *mein* | det:poss\|LEXEME | det:poss\|mein |
| quantifier | *alle* | qn\|LEXEME-AFFIX | qn\|all-e |
| *selb-*, inflected | *selber* | det\|LEXEME-AFFIX | det\|selb-e |

## 7. Annotations for Adjectives

Like determiners, pronouns, quantifiers and numerals, adjectives are only coded for categorical features and their morphological structure. Case and number contexts and syntactic role are encoded with a separate set of codes. These codes are added to the codes for the individual word forms, as discussed below. For adjectives, we are using the conventions of the CHAT-manual and the German MOR-grammar. For participles that are used as adjectives, the additional subcategory code “part” is used.

*Tab.5: Annotations for Adjectives*

| Category | Example | Code Template | Example code |
|---|---|---|---|
| adjective | *gut* | adj\|LEXEME | adj\|gut |
| adjective, inflected | *kleines* | adj\|LEXEME-AFFIX | adj\|klein-er |
| adjective, irregular comparative | *besser* | adj \|LEXEME&CP | adj \|gut&CP |
| adjective, regular comparative | *kleiner* | adj\|LEXEME-CP | adj \|klein-CP |
| adjective, irregular superlative | *besten* | adj \|LEXEME&SP | adj \|gut&SP |
| adjective, regular superlative, | *kleinsten* | adj \|LEXEME-SP | adj \|klein-SP |
| adjective, irregular comparative, inflected | *besseres* | adj \|LEXEME&CP-AFFIX | adj \|gut&CP-es |
| adjective, regular comparative, inflected | *kleineres* | adj \|LEXEME-CP AFFIX | adj \|klein-CP-es |
| adjective, irregular superlative, inflected | *bestes* | adj \|LEXEME&SP-AFFIX | adj \|gut&SP-es |
| adjective, regular superlative, inflected | *kleinstes* | adj \|LEXEME-SP-AFFIX | adj \|klein-SP-es |
| adjective with derivational affixes | *unglueckl ich* | adj\|PREFIX#LEXEME-AFFIX | adj\|un#glueck-lich |
| participle used as adjective, with prefix and inflectional affix | *angemalte* | adj:part\|PREFIX#LEXEME-AFFIX | adj\|an#ge#malt-e |

## 8. Annotations for Nouns

As explained above, nouns are only coded for their inherent categorical and gender features, their morphological structure and any special requirements of their declension class for overt case-marker. Case and number contexts and syntactic role are encoded with a separate set of codes. These codes are added to the codes for the individual word forms, as described in

sections 3.2 and 3.8. Weak masculine nouns like *Junge* 'boy', which require an *–(e)n* in all contexts except nominative singular, are marked as they are the nouns that take nominative and accusative markers in the singular. In addition, we have introduced a special code for nouns that exhibit a dative-plural form ending in *–(e)n* that is distinct from the plural form for other contexts (e,g. *Kind-er-n* 'children'). As these codes refer to inherent properties of the respective nouns, they are combined with the gender information for this noun. The codes for the inherent properties precede any affixes.

*Tab.6: Annotations for Nouns*

| Category | Example | Code Template | Example code |
|---|---|---|---|
| noun, feminine | *Blume* | n\|LEXEME&GENDER | n\|blume&F |
| noun, weak masculine | *Junge* | n:wk\|LEXEME&GENDER | n:wk\|junge&M |
| noun, requiring distinct dative plural marker | *Kind* | n:dp\|LEXEME&GENDER | n:dp\|kind&N |
| noun, proper | *Lenny* | n:prop\|LEXEME&GENDER | n:prop\|Lenny&N |
| noun, one affix | *Frauen* | n\|LEXEME&GENDER -AFFIX | n\|frau&F-en |
| noun, weak masculine, one affix | *Jungen* | n:wk\|LEXEME&GENDER -AFFIX | n:wk\|frau&F-en |
| noun, two affixes, involving distinct dative plural marker | *Kindern* | n:dp\|LEXEME&GENDER -AFFIX-AFFIX | n:dp\|kind&N-er-n |
| noun, umlaut, requiring distinct dative plural marker | *Muetter* | n:dp\|LEXEME&GENDER &UMLAUT | n:dp\|mutter&F&umlaut |
| noun, requiring distinct dative plural marker, with umlaut and one number affix | *Kuehe* | n:dp\|LEXEME&GENDER &UMLAUT -AFFIX | n:dp\|kuh&F&UMLAUT-e |
| noun, requiring distinct dative plural marker, with umlaut and number and case affix | *Kuehen* | n\|LEXEME&GENDER &UMLAUT –AFFIX-AFFIX | n:dp\|kuh&F&UMLAUT-e-en |
| noun, possessive | *Lennys* | n\|lexeme&GENDER-s | n:prop\|Lenny&M-s |

## 9. Combinations of Codes

Codes for the inherent grammatical features of the individual case-marked element and codes for its context features and the respective case-assigning element are separated by colons, just like codes for other grammatical features in CHAT-format. Note that gender context features are only specified for elements that show anaphoric gender agreement (personal pronouns or determiners used as pronouns) or gender concord (in particular determiners, quantifiers, possessive pronouns and adjectives that agree with nouns) . Gender features for nouns are inherent and are hence attached to the noun itself, using the "&" symbol. The template for combining codes is:

(3)
<GRAMMATICAL CATEGORY OF CASE-MARKED ELEMENT> |
<LEXICAL/MORPHOLOGICAL INFORMATION FOR CASE-MARKED ELEMENT> :
<CONTEXT INFORMATION: CASE> :
<CONTEXT INFORMATION: SYNTACTIC FUNCTION> :
<LEXICAL IDENTITY OF CASE-ASSIGNING ELEMENT> :
<CONTEXT INFORMATION: GENDER> :
<CONTEXT INFORMATION: NUMBER>

Take for instance the coding for the nominal elements in the following sentence:

(4) *Der Junge* *legt* *dem Pferd* *den Sattel* *auf den Ruecken*
[The boy]$_{NOM}$ puts [the horse]$_{DAT}$ [the saddle]$_{ACC}$ on [the back]$_{ACC}$
'The boy puts the saddle on the horse's back'

**Tab.7: Codes for Nominal Elements in Example (4)**

| | |
|---|---|
| det:def\|der:NOM:SUB:legen:M:SG | *der* |
| n:wk\|junge&M:NOM:SUB:legen:SG | *Junge* |
| det:def\|dem:DAT:EXT:N:SG | *dem* |
| n:dp\|pferd&N:DAT:EXT:SG | *Pferd* |
| det:def\|den:ACC:DO:legen.M:SG | *den* |
| n:dp\|sattel&M:ACC:DO:legen:SG | *Sattel* |
| det:def\|den:ACC:PP:auf:M:SG | *den* |
| n\|ruecken&M:ACC:PP:auf:SG | *Ruecken* |

Note that in some utterances, one can determine the case context, but not identify a particular case designer: the case-marked phrase might be an extra argument (as in example (4)) or a non-prepositional adverbial like *einen ganzen Monat* '(for) an entire month'. In these situations, the case context that is specified (e.g. "dat:ext" or "acc:adv") will make it clear that no lexical case-assigner could be identified. For other utterances, the category of the case assigning lexeme might be identifiable, but parts of the case-assigning element might not be intelligible – as in *Der Junge hat das Pferd geXXX.* 'The boy has XXXed the horse.', where it is clear that XXX stands for a verb, but it is not clear which verb the child used. In these contexts, we will use the placeholder "lexeme" instead of the case-assigning verb.

## 10. Coding and Checking Procedures

The basis for the annotations described here are transcripts that follow CHAT conventions (MacWhinney 2000; http://childes.psy.cmu.edu/manuals/chat.pdf; see Eisenbeiss and Sonnestuhl, this volume, for an adaptation to German). This can be achieved using the CHILDES database tools or any other tool that has a CHAT export function; e.g. the multi-media annotator ELAN and its CHAT export function (Wittenburg et al. 2006). We copy the transcription tier content to a new tier, labelled "%cas", where we will replace the relevant word forms with their codes.

In order to do this, we will first need a list of all relevant word forms, we make use of tools provided by CHILDES. Using the CHAT-format allows us to use the FREQ-command of the CLAN tool (http://childes.psy.cmu.edu/manuals/clan.pdf) to create lists of all word forms and their frequencies – either for the entire corpus or for individual recordings, either for all speakers or for individual speakers. We can then semi-automatically replace each nominal inflected word from the word list with its code. This can be done using PERL scripts, CLAN tools, or by searching the transcript files for each case-marked word form in the list and pasting the code for the respective word-form onto the coding tier, in the appropriate linear position for this word.

We then search for codes for case-marked elements (determiners, nouns, etc.) and add the appropriate context-features by hand: case, syntactic function, gender, number, case assigner. For complex noun phrases, where several elements appear in the same case context this can be speeded up by using copy and paste functions. For instance for an accusative direct object noun phrase that contains four case-marked word forms - an article, two adjectives and a noun - one code for case assigner, case, gender, and number context can be used for all four

case-marked word forms. The only modification is that the noun won't have a gender context feature as gender is coded as an inherent feature for nouns.

During this step, we check the transcripts for general correctness and ensure that the transcript contains appropriate codes ("[*]") for non-target-like forms. Once all case-marked forms have been coded, we search the transcripts for all forms of case-assigners on the word lists and check that all forms that are assigned case by this case-assigner are coded correctly.

Thus, we semi-automatically code inherent properties and the morphological structure of case-marked forms; and we add context codes by hand, checking earlier steps as we go. When all codes have been entered, a frequency list for all codes is created. This list can then be used for further checks. In particular, we can check for errors in coding format and potential mismatches between context features and codes for case-assigners. For instance, if a case-marked form has the dative verb *helfen* 'help' as a case-assigner, it can either be in a direct dative object context or in a nominative subject context, but not in a direct accusative object context. After all annotations and checks have been made, potential conflicts are resolved, if necessary involving a third person. Then, an independent annotator checks 10 percent of all transcriptions and annotations and reliability checks are carried out.

Counts and utterance lists for particular (classes of) case-marked forms and case-assigners and their combinations can be created on the basis of coded CHAT-files, using the commands FREQ, COMBO and KWAL. These files can also be imported to ELAN, databases or statistical software for further analysis.

## 11. Summary

Our transcriptions conventions follow the CHAT-conventions of the CHILDES database and earlier work for German in this format. However, our annotation scheme differs from the standard MOR-annotations in CHILDES by focusing on case-marking and adding more information that is relevant for analyses of noun phrase structure and inflection:

- It only contains codes for the relevant properties of case-assigners and case-marked forms, not for other elements and their properties (e.g. uninflected quantifiers, adverbs, etc.).
- It contains information about links between each case-form and the respective case-assigner.
- Unlike the standard MOR-grammar for German, our annotation scheme does not require users to attribute grammatical features to non-target-like forms during the

coding stage. Instead we encode the features required by the context and the surface form provided by the speaker.

The semi-automatic coding procedures speed up the coding process; and the coding process provides several checks of transcriptions and codes, which enhances accuracy and reliability.

## Acknowledgements

We would like to thank the University of Essex Research Promotion/Endowment Fund for supporting our work on child language corpora. 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 supporting 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., 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.

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

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