Philosophische Fakultät

Christin Schellhardt | Christoph Schroeder (Hrsg.)

MULTILIT

Manual, criteria of transcription and analysis for German, Turkish and English

Universität Potsdam

MULTILIT

Manual, criteria of transcription and analysis for German, Turkish and English

Christin Schellhardt | Christoph Schroeder (Hrsg.)

MULTILIT

Manual, criteria of transcription and analysis for German, Turkish and English

Universität Potsdam

Universität Potsdam 2015
Institut für Germanistik
Am Neuen Palais 10, 14469 Potsdam

Redaktionell betreut durch Christin Schellhardt

Unter Mitarbeit von Mehmet-Ali Akıncı | Meral Dollnick | Ginesa Dux | Esin Işıl Gülbeyaz | Anne Jähnert | Ceren Koç-Gültürk | Patrick Kühmstedt | Florian Kuhn | Verena Mezger | Carol Pfaff | Christin Schellhardt | Christoph Schroeder | Betül Sena Ürkmez

Online veröffentlicht auf dem Publikationsserver der Universität Potsdam
URN urn:nbn:de:kobv:517-opus4- 80390
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4- 80390

Contents

1 MULTILIT - project description and overall research questions

This paper presents an overview of the linguistic analyses developed in the DFG-funded MULTILIT project[1] and the processing of the oral and written texts collected. The project investigates the language abilities of multilingual children and adolescents, in particular, those who have Turkish as a first language. A further aim of the project is to examine from a psycholinguistic and sociolinguistic perspective the extent to which competence in academic registers is achieved on the basis of the languages spoken by the children, including the language(s) spoken at the home, the language of the country of residence and the first foreign language learned at school. To be able to examine these questions using corpus linguistic parameters, we created categories of analysis in MULTILIT.

The data collection comprises texts from bilingual and monolingual children and adolescents in Germany in their first language Turkish, their second language German und their foreign language English. Pupils aged between nine and twenty years produced monologue oral and written texts in the two genres of narrative and discursive. On the basis of these samples, we examine linguistic features such as lexical expression (lexical density, lexical diversity), syntactic complexity (syntactic and discursive packaging), morphological complexity, as well as orthography in the written texts, with the aim of investigating the pupils' growing mastery of these features particularly in academic registers.

To this end the raw data have been transcribed by the use of transcription conventions developed especially for the needs of the MULTILIT data. They are based on the commonly used HIAT and GAT transcription conventions and supplemented with conventions that provide additional information such as features at the graphic level.

The categories of analysis comprise a large number of linguistic categories such as parts of speech, syntax, noun phrase complexity, complex verbal morphology, direct speech and text structures. We also annotate errors and norm deviations at a wide range of levels (orthographic, morphological, lexical, syntactic and textual). In view of the different language systems, these criteria are considered separately for all languages investigated in the project.

For further information, please visit the MULTILIT homepage:
http://www.uni-potsdam.de/daf/projekte/multilit.html

[1] MULTILIT was funded between 2010 and 2013 by the German Research Foundation (DFG) an its French counterpart ANR. Members of the French team participated in the development of this manual. However, we concentrate on the data collected in Germany and the research questions concerning these data.

Research questions:

- In general: language abilities of multilingual children and adolescents with immigrant backgrounds in Germany
- In particular abilities in:
  - German as school language, language of the social environment and majority language
  - Turkish as the language of the home, language of the social environment and (for most) language of a school subject at certain stages of the school career
  - English as a foreign language learned at school, but also present in the media and social environment
- In different constellations:
  - media (spoken, written)
  - genres (narrative, expository)
  - communicative situations (monologue and conversational)

Specific empirical questions to be addressed through observation of the collected data:

- Relation between spoken and written text production in the different languages
  - in the different genres
  - with respect to processes of literacy-oriented complexity[2]
- Development of linguistic competences (pseudo-longitudinal)
  - in the different languages, genres, media
  - in terms of increasing complexity and diversity – structurally and in the lexicon
  - with respect to possible interdependencies between grammatical domains
- Relation between the competences in the different languages
  - with respect to processes of transfer (conceptual, structural) and the dynamics of language contact
  - in terms of competence development (see above)
- Relations between the text productions of Turkish-German multilinguals with the text productions of German and Turkish monolinguals
  - with respect to the relation between spoken and written text production
  - with respect to the development of linguistic competences
  - with respect to the relation between school/majority language (German) and foreign language (English)
- Correlation with biographical, sociolinguistic and sociological data (as collected by means of a questionnaire)

[2] In the sense of *Sprachausbau*, cf. Maas (2010).

2 MULTILIT corpus

2.1 Constitution

The German MULTILIT corpus contains 1,826 oral and written texts produced by 167 pupils from four different grades: pupils from the fifth grade (52 pupils), seventh grade (40 pupils), tenth grade (27 pupils) and twelfth grade (48 pupils) participated in the MULTILIT project. Data collection across four different grades permits a pseudo-longitudinal interpretation of the data.

Data collection took place at schools in various districts of Berlin and was based on the elicitation technique of Berman/Verhoeven (2002): First the pupils saw a nonverbal video film with sequences of daily problems at school produced by Ruth Berman (Tel Aviv). After watching the film the pupils were asked to produce oral and written texts in two genres – narrative and expository. The task for the narrative oral and written texts was presented as follows: “In the film you saw some different problems. Based on your own experience, please recount something similar that you saw or went through.” This task was first to be completed orally and subsequently in writing. The pupils were free to tell different stories in the two modes. For the production of expository texts they were asked to take a stance on their experiences – again firstly in oral and subsequently in written mode. There was a gap of one to two weeks between the data collection in the three project languages. The elicitation started with the language spoken at home (the pupils’ first language Turkish), continued with their second language German and concluded with English, their first foreign language learnt at school. The data were collected at two different primary schools, a grammar school and a comprehensive secondary school. The schools also take a different approach to the multilingualism of their pupils (see also 2.2.3).

A part of the MULTILIT corpus is annotated with all the following levels of annotation:

- syntactic (sentences and clauses)
- noun phrase structure
- parts of speech
- morphology
- textual (openings and closings)
- code mixing/switching
- communicative mode (dialogue vs. monologue)
- direct vs. indirect speech
- ‘norm deviations’ (cross-cutting through the above levels in accordance with the target hypothesis)

This subcorpus contains oral and written texts produced by 28 multilingual pupils in four different grades: 13 pupils from the fifth grade and five pupils each from every other grade – the seventh, tenth and twelfth grades. The full corpus is POS-tagged, thereby enabling us to carry out a large amount of analysis.

2.2 Transcriptions

With regard to the transcriptions, we draw a distinction between the written and oral texts. However, there are also some instructions about how to deal with the transcriptions in general:

- Literary transcription
- Phenomena which are not represented in the transcriptions should be mentioned in the comment tier.

The written transcription conventions are based on the common HIAT conventions (Rehbein et. al. 2004) which are geared to the orthographic system. We additionally note graphically prominent notions. We use these conventions in all languages of the project and apply them to the orthographic systems of the individual languages. In English we use British English.

The oral transcription conventions are also based on the common HIAT conventions, although the punctuation marks are not used here to show the boundary of a sentence but instead refer to the boundary of an utterance. We therefore use a reduced version of the intonational phrase marker described in GAT2 (Selting et. al. 2009).

The transcriptions are compiled with the Partitur Editor of EXMARaLDA, a transcription and analysis tool.[3] With respect to this program and the further work with its analysis tool we also take into account some additional transcription conventions:

- Each word is defined as an event (each word gets a separate event).
- We insert a space after each entry in an event.
- Punctuation constitutes a separate event.

2.2.1 Transcription conventions for written texts

- In general the transcriptions should represent the original texts as closely as possible. In terms of the written texts this implies that the original notation of the pupils will be represented.
- To note unusual notations of the pupils we create separate tiers for transcription (verbal tier) and comments (user-defined tier).

[3] http://www.exmaralda.org/

Convention	**Notation of the pupils**	**Transcription**
Capitalization	e.g. *Das klauen*	e.g. *Das klauen*
Foreign language material	e.g. *im Chat, außenseiter, oldum*	e.g. *im Chat, außenseiter, oldum*
Oral language usage	e.g. *was* (instead of *etwas*), *gidiyom* (instead of *gidiyorum*)	e.g. *was, gidiyom*
Grammatical mistakes	e.g. *in mein Leben, beni vurdu*	e.g. *in mein Leben, beni vurdu*
Cancelled words are marked with double XX_XX	e.g. *~~gelau~~ gegangen*	e.g. *XXgelauXX gegangen*
Cancelled letters within a word are marked with X_X	e.g. *ge~~e~~laufen*	e.g. *geXeXlaufen*
All letters which are not legible are marked with @	e.g. *friendɗ, lauɷghed, arkadaşım©*	e.g. *friend@, lau@ghed arkadaşım@*
Syllabification is preserved	e.g. *laugh-ed, ge-gangen*	e.g. *laugh-ed, ge-gangen*
Punctuation: Every orthographic sign should be in a separate event (.) (,) (?) (!) (…) (“).	e.g. *He thought nobody saw him.*	e.g. *He thought nobody saw him.*
Marking of the end of a row with a slash in a separate event after the last event of the row	e.g. *We went* *inside.*	e.g. *We went / inside.* Note: In combination with hyphenation, e.g. *gelau-/fen.*
Marking of paragraphs with a double slash in a separate event after the last event of the paragraph	e.g. *That's all.* *We fought too.*	e.g. *That's all. // We fought too.*
Inserted words are marked with < in front of the word in the same event as the inserted word	e.g. *Last* ^{year} *I saw a fight.*	e.g. *Last <year I saw a fight.* Note: We note the mode of insertion in the comment tier.
For more than one inserted word, use << in front of each word in the same event as the inserted word	e.g. *Because* $^{of\ the}$ *fight, Ben* ^{de} ^{okula} *gittim*	e.g. *Because <<of <<the fight, Ben <<de <<okula gittim* Note: We note the mode of insertion in the comment tier.
Small or large indent (centred text): § or §§ in an extra event in front of the first indented word	e.g. *I saw a fight last year in the seventh grade.*	e.g. *§§ I saw a fight last year in the seventh grade.* Note: One § marks a slight indent, §§ marks a strong indent.
Large gap between words in continuous text: = in front of the word concerned	e.g. *I saw many fights*	e.g. *I =saw =many fights*
To mark early line break (no new paragraph) use /& in a separate event after the last event of the row	e.g. *I saw a* *fight last year in school.*	e.g. *I saw a /& fight last year in school.*

2.2.2 Transcription conventions for oral texts

- We use the capitalization of each language according to its orthographic rules.
- The end of an utterance is marked with (,) (.) or (?).
- Breaks are noted.
- Numbers are written out in full.

Convention	**Transcription**	**Remarks**
Capitalization	The transcription represents the capitalization of the language concerned.	
Foreign language material	Transcribed as it is spelt (no translation).	
Unusual pronunciation regarding the default language	If pronunciation is adapted from another language, enter "Pron. Ger/En/…" (for German and English) in the comment tier	
Pronunciations deviating from the standard are noted in their correct forms in the comment tier	e.g. *kloziz* (=clothes), *happenedh* (= happen-ed)	
Merged forms are noted as they are spelt. Reduced syllables are noted as they are spelt.	e.g. *sone* (= *so eine*), *hab ich ein Tadel bekommen* (= *einen Tadel*), *goin* (= going), *bi tane* (= *bir tane*)	
Contractions such as *don't, I'm, Ayse´ye oder Berlin´e*	To be marked in one event	
Short break of oral fluency	•	Taken from keyboard
Break of up to half a second	• •	
Break of up to three quarters of a second	• • •	
Break of longer than a second	((2.5))	
Breaks within a word are marked with a dash	e.g. *ge-laufen, gel-medim*	The dash gets a separate event.
Hesitation markers get a fixed transliteration	German/Turkish: *Ähm* = em; *Mh* = mh; *E* = ee English: *um/ähm/mmm* = um; *Äh/uh/eh* = uh	Unusually long hesitations are marked with additional vowels e.g. *uuh/uuum* (up to 0.5 s) *uuuh/uuum* (up to 0.75 s) *um (1.5s)* (longer than 1.0s)
Laughing	(laughing)	Written in a separate event
(Audible) thinking	Hm	
Whispering	(whispering)	
Agreement	Mhm	
Negation/rejection	mm, eheh	

Incomprehensible utterances	(incomprehensible)	For indications of the length of the incomprehensible utterances we use one box for each unintelligible word
Self-corrections are noted with a slash	/	
Rising intonation of utterance	?	
Semi-rising or semi-falling intonation of utterance	,	
Falling intonation of utterance	.	

2.2.3 Anonymisation of the data

In order to prevent personal identification of the participants in the MULTILIT project, all the data are anonymised. A unique pseudonym is therefore assigned to each individual pupil. Abbreviations used for data storage are derived from the pseudonyms. The pseudonyms are chosen freely and do not contain any characteristics of the participants.

The participating schools are also renamed. Their renaming is based on the school's language policy. The participating primary schools handle literacy acquisition differently – one uses the concept of bilingual literacy acquisition (multi_bialpha) while the other adopts a monolingual approach to literacy (multi_monoalpha). The grammar and comprehensive schools also act differently. While one supports the acquisition of the pupils' first languages by offering CLIL subject teaching in their first language (multi_bi), the other school provides heritage language instruction (multi_fr).

2.3 Meta data

The data collection contains an extensive questionnaire from each pupil. Because of the various age groups we used different versions of the questionnaire. Pupils from the seventh, tenth and twelfth grades got the full version which investigates five different domains. In addition to general information about the pupils and their parents' education, the pupils were asked about their language skills, language use and literacy-related activities. Pupils from the fifth grade received a contracted version of the questionnaire which contained all the items of the full version but with less detail and was reworded in order to ensure the comprehension of primary school pupils. Some sections of the questionnaire could be used for correlations with the linguistic analysis. The questions from these sections therefore had to be condensed into indices (see Extra/Yağmur 2008 and Fürstenau/Yağmur 2003).

The index for the use of languages with other persons is divided into two parts – one for the first language Turkish and another for the second language German. The pupils are asked about their language use with their parents, older and younger siblings, friends (divided into those of the same and different origin) and grandparents (see table below). For each entry concerning the language in question, points are awarded depending on how many languages are used with the respective person. Two points are given if the respective language is the only language used with this person. One point is given if the respective language is one of a group of languages used with this person. No points are given if the language in question is not used at all with the respective person.

In order to facilitate a better understanding, our approach is illustrated in two examples: For Turkish (TR), bilingual pupil A below gets one point for communication with his mother, older and younger siblings and friends of the same origin because he reports the use of both languages Turkish and German (GR) in these relationships. Consequently, one point is also given for German in each of these relationships. For Turkish, the same pupil gets two points for communication with his father and grandparents because he reports exclusive use of Turkish in communication with these relatives, and he receives no points for communication with friends of different origin because he claims not to use Turkish (but instead German) in his communication with them. For German, he therefore receives zero points for communication with his father and grandparents and two points for communication with friends of different origin. In total, pupil A gets eight points for Turkish and six points for German. Pupil B, on the other hand, does not indicate any communicative behavior with younger and older siblings so that no points are awarded here.

	Mother	**Father**	**Older siblings**	**Younger siblings**	**Friends same origin**	**Friends diff. origin**	**Grandparents**	**Points**	**Turkish/German**	**Group**
Pupil A	TR,GR	TR	TR,GR	TR,GR	TR,GR	GR	TR		1.33	3
Points **TR**	1	2	1	1	1	0	2	**8**		
Points **DE**	1	0	1	1	1	2	0	**6**		
Pupil B	TR,GR	TR,GR	ns	ns	GR,TR	GR	TR		**1**	**2**
Points **TR**	1	1	-	-	1	0	2	**5**		
Points **DE**	1	1	-	-	1	2	0	**5**		

Index for the use of language with other persons (ns = not specified)

To compare the indices of different pupils or the indices of different languages for one pupil it is necessary to relate the values of Turkish and German. The relation between the scores for Turkish and German is divided into three different groups. In the first group German is used more often than Turkish and this group contains all relations with a value of up to 0.7. The second group shows a balanced relationship between the use of German and Turkish (0.8 to 1.3) and in the third group Turkish is the language used more frequently (1.3 and above). The results for pupil A show values of 8 (Turkish) and 6 (German). That means that the first language dominates with a relation of 1.33. The results for pupil B show a balanced relation between Turkish and German – he therefore belongs to group 2.

The index for the use of media involves the use of oral-based media such as radio and television but especially also the use of written-based media such as text messages, e-mails and books. The awarding of points is carried out in a similar manner to the index for the use of languages with other persons. Points are given from zero (no use of the respective media) to two points (use of the respective media exclusively in one language). Again, the relations of the first and second languages are used to derive groups of different language users (see table below).

	TV	**Radio**	**Music**	**E-mails**	**Networks**	**Text messages**	**Books**	**Writing outside school**	**Points**	**Turkish/German**	**Group**
Pupil A	GR,TR	ns	GR,EN TR	GR,TR	GR,TR	GR	GR	ns		0.5	1
Points **TR**	1	-	1	1	1	0	0	-	4		
Points **DE**	1	-	1	1	1	2	2	-	8		
Pupil B	GR	GR	EN,TR	GR	GR,TR	GR,TR	GR	GR		0.25	1
Points **TR**	0	0	1	0	1	1	0	0	3		
Points **DE**	2	2	0	2	1	1	2	2	12		

Index for the use of media (ns = not specified)

The development of these indices makes it possible to compare the different pupils and to group them and their text products according to their language use. The condensed questionnaires are implemented in the corpus data.

3 Formal linguistic criteria

3.1 Approach to the (ideal) selection of formal criteria:

- Must serve to gain an insight into the dynamics of linguistic elaboration, (growing) complexity and diversity (syntactic, morphological, lexical, textual criteria)
- Must serve to gain an insight into pupils' command of text norms (textual criteria)

- Must serve to gain an insight into pupils' analytical capabilities (orthographic criteria, neologisms, coinages)
- Must serve to gain an insight into processes of transfer and the dynamics of language contact ('norm deviations' – but not just these)
- Must permit inter-language comparison
- Must be clear-cut, non-controversial and succinct
- Must be formulated in such a way that they can be processed by the software

3.2 Possible further levels of analysis in relation to the MULTILIT research questions (postponed)

- Syntax: clausal complexity
  - existence/non-existence of non-obligatory (~ adjunct) phrases in the clause
  - type and token of prepositional phrases
- Parts of speech
  - distinction between different types of adverbs
  - distinction between different types of communicative markers
- Morphology
  - complex verbal morphology in German (e.g. subjunctive)
  - predominant tense/mood/aspect (e.g. anchor time)
- For written texts: indicators of orthographic competence
  - German: 'Schärfungsschreibung'
  - Turkish: the "soft g" < ğ >
  - English: to be discussed
- For written texts: indications of text-editing (crossings-out, added words and phrases, orthographical corrections, …)
- More specific textual criteria
  - coherence, cohesion
  - teachers' assessments
  - distinction between implicit and explicit arguments

The annotations were reviewed repeatedly. Problems arising during the process of annotation and ambiguities regarding the analytic categories were discussed in regular project meetings, and the categories were adjusted accordingly. In a correction process, existing annotations were carefully checked by other annotaters which had participated in the discussion. Through this, we hope to have arrived at a high level of agreement between the annotators.

3.3 Technicalities

- Create tiers (for annotation) with title as indicated.
- Comments on particular levels can be inserted in a comment tier for the level in question (e.g. “syn1-com”).
- For further details concerning the use of EXMARaLDA, see http://www.exmaralda.org/.

3.4 Basic rules of annotation

- In general, we only code non-dialogical parts and not the interviewer’s utterances or direct questions of the pupil. However, dialogical parts are annotated with DIAL1 only in the MODE-tier and not coded on other levels (see 3.5.8).
- Even if a construction is grammatically incorrect or divergent, we annotate it on the basis of what the respective pupil has written or uttered. The criteria for norm deviations in the form of incorrect or divergent constructions are provided in section 3.5.5.
- For written texts only: Crossed-out elements (words, phrases, sentences…) are not annotated.

3.5 The MULTILIT criteria

3.5.1 Formal linguistic criteria: clausal syntax (SYN)

Remarks:

- In the annotation, we proceed at multiple levels depending on the degree of embedding:
  - SYN1: specifies the complex sentence with the main clause (MCD, MCI, MCIMP). The sentence might or might not contain further predications. If a main clause contains an embedding, this is marked with (+) following the predication which contains the embedding. The embedding might be subordinate (SC), co-subordinate (MCweil, SCki) or elliptical (SCELL).In any case the embedded element itself is not marked in SYN1, but in SYN2, SYN3 or any other (deeper) level. Also the depth of the embedding is indicated (e.g. “+3”). Disjunct/unembedded subordinations (SCDISJ) and incomplete or discontinuous main clauses (INCOMPL) are also annotated in SYN1.
  - SYN2, SYN3, … specifies the type of the embedded predication. For example, an ellipsis is always indicated in the syntax tier below the construction it belongs to.
  - Again, if an embedded predication contains a further embedding, this is marked with (+) following the predication which contains the embedding.
- Clauses which are direct speech are not marked as complements but as what they are (MCD, …). Direct and indirect speech are not considered syntactic dimensions. They are considered in an separate annotation tier (see 0).
- Sentences comprise fully saturated finite verbs with embedded adjuncts and (further) predications. In spoken texts, pauses and shifts of pitch level might be markers of sentence boundaries, but they can also occur in other contexts. In written texts, full

stops and the like might be markers of sentence boundaries, but in the case of wrong punctuation the sentence can extend beyond the punctuation of the pupil.

- We do not exclude the possibility of sentences (MCD) which are loosely embedded serving as complements of other sentences (e.g. in direct speech). These are marked as MCD on SYN2 (or a further SYN level as applicable).
- Incomplete or discontinuous clauses for which it is clear which type of subordinate/embedded clause they represent (adverbial, relative, …) are coded in a combination of INCOMPL and the type of clause at the applicable level (e.g. INCOMPLSCREL or SCRELINCOMPL).
- (Incomplete) clauses which are repeated later are coded with a W following the annotation of the clause, e.g. INCOMPLW.
- (Incomplete) clauses which are corrected later are coded with an SK following the annotation of the corrected clause, e.g. SCELSK.
- We do not annotate utterances with clausal status (e.g. *Yes*) within the syntactic tier. These are annotated in the tier for parts of speech (POS) as CO.

3.5.1.1 Formal linguistic criteria: clausal syntax – German

Label	**Category**
SYN1	
MCD	Independent declarative clause
MCI	Independent interrogative clause
MCIMP	Independent imperative clause
+X{2, 3, …}	When following any of the above annotations, "+" indicates that the main clause has further predications. We add the number of the syntactic tiers, e.g. MCD+3 for a complex sentence which is annotated up to syntactic tier SYN3
SCDISJ	Disjunct/non-embedded/afterthought (finite) clause with subordinate clause word order
INCOMPL	Incomplete or discontinuous clause (in terms of written standard)
SYN2, SYN3, …	
SCREL	Embedded (finite) subordinate clause – relative clause
SCADV	Embedded (finite) subordinate clause – adverbial clause
SCCOMP	Embedded (finite) subordinate clause – complement clause
SCINF	Embedded infinitive clause
MCWEIL	*weil*-clause with SVO(V) word order
SCELL	Well-formed ellipsis (in terms of written standard)
XXINCOMPL	Incomplete or discontinuous subordinate clause (specify type)
INSERT	Non-clausal insertion
INSERTMCD	Clausal insertion (e.g. *letztes Jahr, es war Mai, hab' ich einen Stein auf den Kopf bekommen*) – add "+" for clausal insertions which are complex and specify type (e.g. *letztes Jahr, ich glaube, dass es Mai war, hab' ich einen Stein auf den Kopf bekommen)*

3.5.1.2 Formal linguistic criteria: clausal syntax – Turkish

Label	**Category**
SYN1	
MCDS	Finite declarative clause with pronominal or lexical subject
MCDZ	Finite declarative clause with zero anaphora subject (“pro drop”)
MCI	Finite interrogative clause
MCIMP	Finite imperative clause
+	When following any of the above annotations, “+” indicates that the main clause has further predications
SCDISJ	Disjunct/non-embedded subordinate clause
INCOMPL	Incomplete clause (in terms of written standard)
SYN2, SYN3, …	
SCREL	(Non-finite) subordinate clause – participle (relative clause)
SCADV	(Non-finite) subordinate clause – converb (adverbial clause, e.g. *-ken*, *-ErEk, -IncE, -Ir, -mEz*; incl. *-dIktan sonra*, *-DIĞI icin, - DIĞI zaman, -mEsInE rağmen…*).
SCCOMP	(Non-finite) subordinate clause – nominalization (complement clause)
SCELL	Finite clause connected through (suffix) ellipsis (e.g. *eve gelir seni beklerim*)
SCsa	Finite clause with subjunctive (*-sE*)
SCdiye	Finite clause with *diye*
SCki	Finite clause formed with the conjunction *ki*
SCip	(Non-finite) clause formed with the converb *-Ip*
XXINCOMPL	Incomplete or discontinuous subordinate clause (specify type)
INSERT	Non-clausal insertion
INSERTMCD	Clausal insertion (e.g. *bi sefer; bilmiyom; sınıftaki çocukla benim kafama taş attılar; o zaman; nasıl deyim; okul değiştirdi*)

3.5.1.3 Formal linguistic criteria: clausal syntax – English

Label	**Category**
SYN1	
MCD	Independent declarative clause
MCI	Independent interrogative clause
MCIMP	Independent imperative clause
+{2, 3, …}	When following any of the above annotations, “+” indicates that the main clause has further predications. We add the number of the syntactic tiers, e.g. MCD+3 for a complex sentence which is annotated up to syntactic tier SYN3
SCDISJ	Disjunct/non-embedded (finite or non-finite) subordinate clause
INCOMPL	Incomplete or discontinuous clause (in terms of written standard)

SYN2, SYN3, …	
SCREL	Embedded (finite) subordinate clause – relative clause
SCRELing	Attributive/relative clause formed without relative pronoun and with *-ing* predicate (e.g. *the boy standing over there*)
SCADV	Embedded (finite) subordinate clause - adverbial clause with *because, if, when*
SCCOMP	Embedded (finite) subordinate clause – complement clause (e.g. wh-clauses), complement clauses without (*that*) complementizer
SCCOMPthat	Embedded (finite) subordinate clauses – complement clause with *that*
SCINFto	Embedded non-finite subordinate (complement) clause: *to*-clause (e.g. *I don't want to do my homework*)
SCINFwh	Embedded non-finite subordinate (complement) clause formed with *to* and wh-complementizer (e.g. *I don't know what to do*)
SCINFing	Embedded non-finite subordinate (complement) clause: V-*ing*-clause (e.g. *I don't like doing homework. That's why they don't know how to act in difficult situations without being aggressive.*)
SCINFpart	Embedded non-finite participle clause (e.g. *So I arrived banned for a lesson of the teacher*).
SCELL	Well-formed ellipsis (in terms of written standard)
XXINCOMPL	Incomplete or discontinuous subordinate clause (specify type)
INSERT	Non-clausal insertion
INSERTMCD	Clausal insertion (e.g. *letztes Jahr, es war Mai, hab' ich einen Stein auf den Kopf bekommen*) – add "+" for clausal insertions which are complex and specify type (e.g. *letztes Jahr, ich glaube, dass es Mai war, hab' ich einen Stein auf den Kopf bekommen*)[4]

3.5.2 Formal linguistic criteria: noun phrase complexity (NP)

Remarks:

- At this point, the criteria of noun phrase complexity serve as the only criteria concerning clause-internal complexity.
- Nouns which cannot be expanded into a phrase are not annotated (e.g. incorporated nouns, nouns in light verb constructions, also spatial nouns in Turkish, etc.).
- Proper names are not annotated as NPs, neither are pronouns (e.g. *ich*, *man*, *jemand*, *einer*, …).
- We distinguish between different levels of NPs. They can co-occur at one level or be part of a superior NP. NPs with more than one different type of extension on one level are coded as NPMULT (as described below). Where NPs occur at different levels, only the highest level is coded as an NP while parts of NPs which are also NPs (e.g. NPs in relative clauses) are not coded as such. However, all clauses which are embedded into NPs are coded in the SYN-tier.

[4] We have not (yet) found an example for INSERTMCD or INSERTMCD+ in the English data.

- German: Amalgamated forms (preposition plus article) are coded as parts of NPs. However, prepositions (if occurring in isolation) are NOT coded as parts of NPs.
- We do not include subordinate clausal structures which are nominalized in the sense that they occupy noun phrase positions in the sentence – these are regarded as complement clauses and coded under the criteria of clausal syntax.
  - For Turkish, this means that we only code NPs with heads that are noun-based or derived nouns (e.g. with - *Iş*) but do not count syntactic nominalizations because these are regarded as clauses/subordinations.
  - However, for English, nominalizations with *-ing* are annotated as NPs.
- Extraposed parts of NPs (e.g. extraposed relative clauses) are not marked in the NP tier although the head NP is coded with the respective tag (e.g. NPREL) and has an added "+": NPREL+.
- NPs which are incomplete are coded as INCOMPL.
- If an NP has more than one expansion of a different type at the same level (except for NPC and NPCO), it is tagged as NPMULT. This also applies to coordinated parts where the coordinated parts altogether have more than one expansion.
- If an NP has coordination at any level, we add "CO" to the annotation (note that for the coordination of heads, not all NPs are represented in the annotation). Thus, for instance, *frischer Tee und frischer Kaffee* is annotated as NPMULTCO, while *frischer Tee und Kaffee* is NPADJCO.
- If an NP has more than one expansion of the same type, we add the number following the label of NP (e.g. NPADJ2).

3.5.2.1 Formal linguistic criteria: noun phrase complexity – German

Label	Category
NPO	Non-expanded NP (non-compound N with or without an article, demonstrative or possessive determiner or quantifier)
NPC	Head of NP is compound N+N. NPC is not considered in the context of NPMULT and NPCO
NPCO	NP is coordinated and the coordinated parts do not have any expansion, (e.g. *Tee und Kaffee*)
NPADJ	NP has a prenominal adjectival modifier, including ordinal numbers
NPPRT	NP has a participial modifier
NPEXPRT	NP has an expanded participial modifier
NPGEN	NP has a prenominal genitive
NPPADV	NP has a postnominal adverbial modifier
NPPREP	NP has a postnominal prepositional phrase
NPPGEN	NP has a postnominal genitive
NPPEX	NP has other expanded postnominal modifiers (*wie* phrase, *als* phrase)

NPAPP	NP has postnominal apposition (e.g. *meine Freundin Denise*; *Peter, unser Nachbar; das Unterrichtsfach Physik*)
NPREL	NP has a relative clause
NPINF	NP has an infinitive clause
NPCOMP	NP has a postmodifying complement clause
NPMULT	NP has more than one expansion of different types at one level (except for NPC and NPCO)

3.5.2.2 Formal linguistic criteria: noun phrase complexity – Turkish

Label	**Category**
NPO	NP is a non-expanded NP (non-compound N possible with article, count noun or (other) determiner (e.g. *bir oyuncak; bu oyuncak*; *beş oyuncak*, …)); NOT nominal in constructions with light verbs (e.g. *etmek, yapmak, kılmak*) and verb-incorporated nominals (e.g. *çorap aldım*); spatial nouns are only annotated when they form compounds with lexical nouns (e.g. *evin arkası*)
NPC	Head of NP is compound N+N (e.g. *okul çantası*; *öğrenme imkanı; kitap siparışı*; …) including complex compounds (e.g *gelen var mı sorusu*). NPC is not considered in the context of NPMULT and NPCO.
NPCO	NP is coordinated and the coordinated parts do not have any expansion (e.g. *çay ve kahve*)
NPADJ	NP has a prenominal adjectival modifier (incl. participles with *-mIş*, *-Ir*, *-dIk* [without POSS], *-EcEk* [without POSS]); also incl. ablative attributes (e.g. *demirden bir kapı* and ordinal numbers)
NPADJEXP	NP has an expanded prenominal adjectival modifier (e.g. *çok güzel bir kitap*)
NPGEN	NP has a (prenominal) genitive modifier (e.g. *öğrencinin çantası*), including an agentive genitive in verb-derived nouns (e.g. *öğrencinin sorusu*; *ucağın inişi*) and ablative-marked attributes in partitive constructions (e.g. *pastadan bir dilim; arkadaşlardan biri*) where this may vary with the genitive.
NPLOC	NP has a modifier phrase formed with *-ki* or *-dEki* (e.g. *geçen haftaki ders; masadaki kitap*; …)
NPPRTAN	NP has a non-expanded participle modifier with *-En* (e.g. *gelen adam; kopya çeken öğrenci*)
NPPRTANEXP	NP has an expanded participle modifier with *-En* (e.g. *saat beşte bana gelen adam; yakalanmadan kopya çeken öğrenci*)
NPPRTDIK	NP has a participle modifier with *-dIK* or *-EcEK* (e.g. *dans ettiğimiz salon*; *gideceğimiz yer*)
NPPRTDIKEXP	NP has an expanded participle modifier with *-dIK* [+POSS] or *-EcEK* [+POSS]
NPPOST	NP has a modifier in the form of a postpositional phrase (e.g. *spor sahası gibi bir yer*)
NPMULT	NP has more than one expansion of different types at one level (except for NPC and NPCO) (e.g. *kopya çeken akıllı öğrenci*)

3.5.2.3 Formal linguistic criteria: noun phrase complexity – English

Label	Category
NPO	Non-expanded NP (non-compound N with or without article, demonstrative or possessive determiner or quantifier)
NPC	Head of NP is compound N+N. NPC is not considered in the context of NPMULT and NPCO.
NPCO	NP is coordinated and the coordinated parts do not have any expansion (e.g. *tea and coffee*)
NPADJ	NP has a prenominal adjectival modifier
NPPRT	NP has a participial modifier
NPEXPRT	NP has an expanded participial modifier
NPGEN	NP has a prenominal genitive
NPPADV	NP has a postnominal adverbial modifier
NPPREP	NP has a postnominal prepositional phrase
NPPGEN	NP has a postnominal genitive (*of* genitive)
NPPEX	NP has other expanded postnominal modifiers (*like* phrase, *as* phrase)
NPAPP	NP has postnominal apposition (e.g. *Peter, my friend*)
NPREL	NP has a relative clause
NPINF	NP has an infinitive clause
NPCOMP	NP has a postmodifying complement clause
NPMULT	NP has more than one expansion of different types at one level (except for NPC and NPCO)

3.5.3 Formal linguistic criteria: parts of speech (POS)

Remarks:

- We base this on the STTS-tag table[5] which was originally developed for the annotation of German texts, adopt it for our purposes with a couple of changes, and developed tags from this for English and Turkish. Thus, while for German we basically follow the STTS-approach, for the other languages we only use the labels but always acknowledge the different typology. Both STTS as well as our adaption of it are function-based and part-of-speech-based and this double nature accounts for the high degree of differentiation.
- Standard abbreviations such as *usw., z.B.* and *ok* are annoted in POS as what they would be if spelled out.
- Words which are repeated later are coded with a W following the POS annotation of the word, e.g. ADJAW.

[5] see http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

- Words which are corrected later are coded with an SK following the POS annotation of the word, e.g. PRELSSK.
- Words which are interrupted are coded as INCOMPL.
- If there is more than one alternative with regard to POS in the particular context, we code the alternatives and add a question mark to each, e.g. *wo* KOUS?PRELW?.
- If words are written together, the POS categories of the two words are also written together in the respective event.
- References to be used in case of doubt:
  German: Duden-Grammatik 2006
  English: Longman Grammar 2012
  Turkish: Göksel/Kerslake 2005

3.5.3.1 Formal linguistic criteria: parts of speech – German

Part of speech	Label	Description	Example (German)

Adjective	ADJA	Attributive adjective (not verb-based), including quantifying adjectives (ordinals, *other*, etc.) when in adjective position	*das große Haus;*
			die anderen Kinder;
			die beiden Brüder;
			das ganze Problem;
			die vielen Korrekturen

	ADJD	Adverbial or predicative adjective, not participle	*er fährt schnell;* *er ist schnell*

	ADJV	Attributive or adverbial present or past participle ('Partizip Perfekt')	*das gesunkene Schiff;*
			die weinende Katze;
			lachend sagte er…

Adverb	ADV	Adverb (non-pronominal, non-deictic), including modal and grading adverbs and modal particles	*schon; bald; doch; ganz schön; sehr müde*
	ADVP	Pronominal or deictic adverb	*dafür; dabei; dorthin; deswegen; trotzdem; herunter; selber* Note: *selber* not reflexive intensifying, otherwise ADV.
Adposition	APPR	Preposition in prepositional phrase; left circumposition	*in der Stadt; ohne mich*
	APPRART	Preposition including article	*im Haus; zur Sache*
	APPO	Postposition	*ihm zufolge; der Sache wegen*
	APZR	Right circumposition	*von jetzt an*
Article	ART	Definite or indefinite article	*der; die; das; ein; eine*
Cardinal numeral	CARD	Cardinal numeral (ordinals are tagged as ADJA)	*zwei Männer; im Jahre 1994*

Conjunction	KOUI	Subordinating conjunction requiring a ‘*zu*’+infinitive construction	*um zu leben; anstatt zu fragen*
	KOUS	Subordinating conjunction including complementizer (but not relative pronoun)	*weil; dass; damit; wenn; ob*
	KONS	Non-subordinating, clause-combining conjunction	*Peter kam und/ denn/ oder/ aber Hans kam*
	KONP	Phrase-combining conjunction	*Peter und/ oder Hans gingen/ geht weg*
	KOKOM	Comparative conjunction	*als; wie*
Noun	NN	Ordinary noun	*Tisch; Herr; Achtung*
	NE	Proper name	*Hans; Hamburg; HSV*
	NOM	De-verbal and de-adjectival nominalisation, including de-adjectival nominalisation derived from quantifying adjectives such as *beide* or *andere*	*das Laufen; die Bläue; das Wichtigste; die beiden; das Ganze; der andere*
	NKO	Nominal compound	*Schulbrot; Hausaufgabe*
Pronoun	PDS	Substituting demonstrative pronoun (also textual reference)	*dieser; jener; so (et)was*
	PDSPER	Definite article as demonstrative pronoun referring to persons	*der; die; das;* *die spielen da*
	PDAT	Attributive demonstrative pronoun	*jener Mensch*
	PIS	Substituting indefinite pronoun	*keiner; viele; man; niemand; ein bisschen*
	PIAT	Attributive indefinite pronoun (quantifier) when in determiner position	*kein Mensch; irgendein Glas; wenig Wasser; ein wenig Wasser; ein bisschen Angst*
	PPERX	Personal pronoun Distinguish person (X: first, second, third)	*ich; er; ihm; mich; dir*
	PPOSSX	Substituting possessive pronoun Distinguish person (X: first, second, third)	*meins; deiner*
	PPOSATX	Attributive possessive pronoun Distinguish person (X: first, second, third)	*mein Buch; deine Mutter*

	PRELS	Substituting relative pronoun	*der Hund, der*
	PRELAT	Attributive relative pronoun	*der Mann, dessen Hund*
	PRELW	Interrogative relative pronoun	*das Thema, worüber wir*
	PRFX	Reflexive personal pronoun Distinguish person (X: first, second, third)	*sich; einander; dich; mir*
	PWS	Substituting interrogative pronoun	*wer; was; wann; worüber; wobei*
	PWAT	Attributive interrogative pronoun	*welche Farbe; wessen Hut*
	PWSKOMP	Interrogative pronoun in the function of complementizer	*ich fragte ihn, wann er kommt*
Particle	PTKZU	*zu* in front of infinitive	*zu gehen*
	PTKNEG	Negation particle	*nicht*
	PTKVZ	Separated verb particle	*er kommt an; er fährt rad*
	PTKA	Grading particle with adjective or adverb	*am schönsten; zu schnell; allzu schön*
Discourse marker	CO	Interjection, particle with communicative function, answering particle, hesitation particle	*mhm; ach; tja; ja; nein; danke; undso(weiter); usw.*
Verb	VVFIN	Finite verb with lexical meaning	*du gehst; wir kommen an*
	VVINF	Infinitive form of verb with lexical meaning	*gehen; ankommen*
	VVIZU	Infinitive of lexical verb with integrated *zu*	*anzukommen; loszulassen*
	VVPP	Past participle ('Partizip Perfekt'), verb with lexical meaning, predicative	*gegangen; angekommen*
	VAFIN	Finite auxiliary verb (also copular)	*du bist gegangen; wir werden gehen; er ist wunderbar*
	VAINF	Non-finite auxiliary	*sein; haben*
	VAPP	Past participle auxiliary ('Partizip Perfekt'), predicative	*gewesen*
	VMFIN	Finite modal verb (expecting another full lexical infinitive verb)	*er darf hereinkommen*
	VMINF	Non-finite modal verb	*er hat gehen wollen*
	VMPP	Predicative past participle ('Partizip Perfekt'), modal	*sie hat das sofort gewollt*

Additional tags			
Linguistic material from other languages	XXAMDT	Turkish in German text XX: define POS	
	XXAMDE	English in German text XX: define POS	
First constituent part of a truncated compound coordination	TRUNC	Words ending with a hyphen where the hyphen substitutes the right constituent part (head) of a compound	*An- und Abreise*
Non-word with special characters	XY	To be annotated as one part of speech in one event (even if transcribed in more than one event)	*3:7; H2O; D2XW3*

3.5.3.2 Formal linguistic criteria: parts of speech – Turkish

Part of speech	**Label**	**Description**	**Example (Turkish)**
Adjective	ADJA	Attributive adjective (not verb-based), including quantifying adjectives (ordinals, *other*, etc.) when in adjective position	*büyük ev; ikinci öğrenci*
	ADJD	Adverbial adjective (predicative adjectives fall under ADJDVCOP) (see VCOP)	*hızlı gidiyor*
	ADJV	Attributive participle (non-clausal, only subject participle, not -*An*)	*batmış gemi; içilir su; beklenmedik olay; ekilecek tarla*
Adverb	ADV	Adverb (non-pronominal, non-deictic), including modal adverbs (most of these are nominals but we consider the function here)	*yakında; her gün; apar topar; yine de; belki; mutlaka; bir kere*
	ADVP	Pronominal or deictic adverb	*oraya gitti; ileri; geri; asağıya; içine; o kadar; o zaman*
	ADVV	Verb-based adverb (converb), non-clausal/reduced clause	*güle-oynaya çıktık; gülerek yaklaştı*
Postposition	APPO	Postposition	*senin için; eve kadar; ayı gibi; arkadaşın olarak; vidyo sayesinde*
Article	ART	Indefinite article	*bir*
Cardinal numeral	CARD	Cardinal numeral (ordinals are tagged as ADJA)	*iki adam* *1994 senesinde*

Conjunction	KON	Conjunction (with following or preceding finite sentence)	*ki; ve; çünkü; halbuki; ama; diye; hem güler hem ağlar*
	KONP	Phrase-combining conjunction	*Mehmet ve/ ya da/ ile Ayşe geldi; hem sen hem ben*
Noun	NN	Ordinary noun, including derived noun and locational or temporal noun (if in nominal position)	*masa; sevgi; bayan; kitaplık; önü; arka*
	NE	Proper name	*Mehmet; Ayşe,*
	NATTR	Attributive noun (with ablative case [*-dAn*] or with *-dA(ki)*)	*köşedeki ev; demirden bir kapı*
	NA	Marked noun of Arabic or Persian origin	*misbah; zamir*
	NOM	De-verbal and de-adjectival nominalization, non-clausal	*gidiş kolay oldu*
	NKO	Nominal compound, including compound formed with locational or temporal noun	*ev ödevi; eşek şakası; sınav sırasında*
	Nth	Passe-partout word	*şey*
Pronoun	PDS	Substituting demonstrative pronouns *bu* and *şu* (also textual reference) (*o* is PPER)	*bu olmadı; şu olmadı*
	PDSPER	Substituting demonstrative pronouns *bu* and *şu*, used to refer to person (also textual)	*bu gitti; şunu gördün mü?*
	PDAT	Attributive demonstrative pronoun	*bu adam*
	PIS	Substituting indefinite pronoun	*kimse; biri(si); bazı(ları) gitti; hiçbiri*
	PIAT	Attributive indefinite pronoun (~ indefinite quantifier)	*bazı arkadaşlar; herhangi bir adam; bir takım arkadaşlar; çok kavga olur*
	PPERX	Personal pronoun Distinguish person (X: first, second, third)	*ben; bana; sen; seni; bizi; onu; onları*
	PPOSSX	Substituting possessive pronoun Distinguish person (X: first, second, third)	*bizimki; benimki*
	PPOSATX	Attributive possessive pronoun Distinguish person (X: first, second, third)	*benim kitabım; onun çocuğu*

	PREC	Substituting reciprocal pronoun	*birbirine*
	PRFS	Substituting reflexive pronoun	*kendisi gitti; kendisine zarar verir*
	PRFAT	Attributive reflexive pronoun	*kendi çocuğu*
	PWS	Substituting interrogative pronoun	*kim; ne; hangisi; ne zaman; neden; nerede; neler; kime; neyi*
	PWAT	Attributive interrogative pronoun	*hangi renk; kaç çocuk*
	PWSKOMP	Interrogative pronoun in the function of complementizer	*ona kim olduğunu sordum; nereye gittiği merak ettim*
	PCLS	Substituting numeral classifier	*bir tanesini gördüm*
	PCLAT	Attributive numeral classifier	*bir tane adam gördüm*
Particle	PTKNEG	Negation particle	*hiç*
	PTKA	Grading particle and particle with adjective, cardinal number, adverb	*daha güzel; en güzel; tek bir kez yaşadım*
	PTKFOC	Focus particle	*ben de geldim; ben ise geldim; ben bile geldim*
	PTKI	Interrogative particle (only if not with verb)	*kalem mi*
Discourse marker	CO	Interjection, particle, particle with communicative function, answering particle, hesitation particle	*ha; mm; eyvah; yani; evet; hayır; bence; sanırım; bana kalırsa*
Verb	VVFIN	Finite verb with lexical meaning	*sen gittin; biz düşündük, gerekiyor*
	VVINF	Non-finite verb with lexical meaning in subordinate non-finite clause (but distinguish from ADJV, ADVV and NOM where the construction is non-clausal)	*gitmeyi; gidince; gitmek; gittiğini; giden; gidip; gittikten sonra*
	VCOFIN	Finite construction of N+V where V is a light verb (e.g. *yap-, et-*) or the construction is a frozen phrase such as in *zarar ver-, çak ver-* or *kopya çek-*. VCOFIN+ has to be written under the predicate (e.g. *gelmedi*) if the VCOFIN is separated by other elements such as *hiç* or *de*, e.g. *azar da işittik*.	
	VCOINF	Non-finite construction of N+V	
	VCOMFIN	Finite complex verb consisting of a lexical verb plus auxiliary *ol-*	*gitmiş olacak; dönüyor olduk*
	VCOMINF	Non-finite complex verb consisting of a lexical verb plus auxiliary *ol-*	*gitmiş olacakken; dönüyor olup*

	VCOP	Copular (suffix) in nominal clause (coded together with predicative nominal, adverb or adjective) – annotated also in the case of zero.	*Emreydi* (NEVCOP); *kırmızıydı* (ADJDVCOP); *evdeymişler* (NNCCOP); *öğretmenimdiniz* (NNVCOP); *buradaydim* (ADVPVCOP); *güzeldir* (ADJDVCOP); *hevesliysek* (ADJDVCOP); *şoförse* (NNVCOP)
	VAFIN	Finite auxiliary verb *ol-* if not in complex construction	*neler oluyordu; bugün işte ocağım/olmalıyım/ olabilirdim; farkında oldular*
	VAINF	Non-finite auxiliary verb *ol-* if not in complex construction	*evde olduğunu; öğretmen olmak; gerekli olduğunda; yorgun/haklı olduğunu*
Non-verbal predicates	PEXIST	Finite existential verbs (*var*, *yok*), any finite form of *var* and *yok* with different TAM markers	*elma var/varmış/vardır; elma yok/yokmuş/yoktur*
	PNEG	Negation predicate (not coded with VCOP)	*yorgun değil/değildi; değilmiş*
Additional tags			
Linguistic material from other languages	XXAMTD	German in Turkish text XX: define POS	*klären yapılmış*
	XXAMTE	English in Turkish text XX: define POS	*Team work yaptık*
Non-word with special characters	XY	To be annotated as one part of speech in one event (even if transcribed in more than one event)	*3:7; H2O; D2XW3*

3.5.3.3 Formal linguistic criteria: parts of speech – English

Part of speech	Label	Description	Example (English)
Adjective	ADJA	Attributive adjective (not verb-based), including quantifying adjectives (ordinals, *other*, etc.) when in adjective position	*the big house; the other girl; the second call*
	ADJD	Predicative adjective	*he is fast; it is nice*
	ADJV	Attributive or adverbial participle or *-ing* form	*the sinking ship; a broken bottle; goes home laughing*

Adverb	ADV	Adverb (non-pronominal, non-deictic), including modal and grading adverbs and modal particles	*soon; still tired*
	ADVD	Adjective-based adverb	*he drives quickly; fast*
	ADVP	Pronominal or deictic adverb	*there; here*
Adposition	APPR	Preposition in prepositional phrase	*in the town; without me*
	APPRC	Complex preposition	*because of; due to*
	APPRV	Preposition as part of verb	*walks on; step in; jump down*
Article	ART	Definite or indefinite article	*the; a*
Cardinal numeral	CARD	Cardinal numeral (ordinals are tagged as ADJA)	*six men; in the year 1994*
Conjunction	KOUI	Subordinating conjunction requiring a '*to*' + infinitive construction	*in order to leave the room*
	KOUS	Subordinating conjunction including complementizers (but not relative pronoun)	*because; while; if*
	KONS	Non-subordinating, clause-combining conjunction	*Peter left and/ but/ or Mary left*
	KONP	Phrase-combining conjunction	*Peter and/ or Mary left*
	KOKOM	Comparative conjunction	*as*
Noun	NN	Ordinary noun	*table; mister; attention*
	NE	Proper name	*Hans; Berlin; England*
	NOM	De-verbal nominalization (with *-ing* or *to*), also de-adjectival nominalization[6]	*hiking is a nice hobby; to err is human*
	NKO	Nominal compound	*homework; school bag*
Pronoun	PDS	Substituting demonstrative pronoun (also textual reference)	*that was silly*
	PDAT	Attributive demonstrative pronoun	*this guy; that thing*
	PIS	Substituting indefinite pronoun	*nobody; anybody; many came; none*
	PIAT	Attributive indefinite pronoun (~ indefinite quantifier) Note: *another* is ARTPIAT	*no man; every child; some books; some of the books*
	PPERX	Non-reflexive personal pronoun Distinguish person (X: first, second, third)	*I; you; he; she; it; we; you; they*

[6] Note that the difference between *-ing* (NOM) and *-ing* (VVING) is one of constructional complexity: *-ing* forms which are annotated as VVING are heads of clausal constructions while *-ing* forms which are NOM are non-clausal.

	PPOSSX	Substituting possessive pronoun Distinguish person (X: first, second, third)	*mine is here; yours is there*
	PPOSATX	Attributive possessive pronoun Distinguish person (X: first, second, third)	*my book; your mother*
	PRELS	Substituting relative pronoun	*the guy whom I met*
	PRELAT	Attributive relative pronoun	*the guy whose dog*
	PRFX	Reflexive personal pronoun Distinguish person (X: first, second, third)	*each other; himself; myself*
	PWS	Substituting interrogative pronoun	*who; what; when; about what*
	PWAT	Attributive interrogative pronoun	*which colour; whose hat*
	PWSKOMP	*wh* pronoun in the function of complementizer	*I asked him what he thought*
Particle	PTKto	*to* in front of infinitive	*to go*
	PTKNEG	Negation particle (not when amalgamated with copular or modal verb, see below)	*not*
	PTKA	Grading particle with adjective or adverb	*most pleasant; too good*
Discourse marker	CO	Interjection, particle with communicative function, answering particle, hesitation particle	*well; okay; yes; no; hm*
Verb	VVFIN	Finite verb with lexical meaning Note: a verb which lacks third person singular *-s* is still coded as VVFIN on the POS tier	*you went; we arrived*
	VVINF	Infinitive form of verb with lexical meaning	*he wanted to go*
	VVING	Continuous (*-ing)* form of verb with lexical meaning (in predicative function, with aux) but also including *going to* future and heads of *-ing* complements and relative clauses (see footnote 5)	*he was going; is walking; the boy standing over there; he likes playing in the garden*
	VVPP	Past participle, verb with lexical meaning, predicative	*he has seen; he has fought*
	VAFIN	Finite verb, aux or copular Add NEG for amalgamated negative (VAFINNEG)	*you have; you are; we will; you haven't*

	VAINF	Non-finite auxiliary	*have*
	VMFIN	Finite verb, modal (expecting another lexical verb) Add NEG for amalgamated negative (VAFINNEG)	*may; can't*
	VMINF	Non-finite verb, modal (expecting another lexical verb)	*wanting to go;* *to want to go*
Additional tags			
Linguistic material from other languages	XXAMED	German in English text XX: define POS	
	XXAMET	Turkish in English text XX: define POS	
Non-word with special characters	XY	To be annotated as one part of speech in one event (even if transcribed in more than one event)	*3:7; H2O; D2XW3*

3.5.3.4 Automatic POS tagging

To handle the enormous amount of data in the MULTILIT corpus, we used the reliable automatic part-of-speech tagging software developed by Schmid (1994) for all three languages in the project – English, German and Turkish.

First of all the verbal tier of written and spoken interview data of the pupils is extracted from the transcriptions. Following this, each entry of the verbal tier is allocated a part-of-speech tag. A lexicon for every POS tag based on a well-checked data set is therefore needed. The tagged verbal tier has to be reconverted to the .exb data format of the EXMARaLDA transcription. The annotated data can be used with the software tools of EXMARaLDA in the same way as all other data. The six-step procedure is detailed below:

- Corpus normalization
  - As pointed out above, the entire corpus was prepared using the EXMARaLDA format, with tier elements of speaker productions containing event elements.
  - Unlike Westpfahl & Schmidt 2013, we did not normalize the orthographic form of the tokens since the HIAT annotation scheme used already takes into account synalepha and similar phenomena.
  - In addition, we decided only to use the transcription tier tokens of the productions (written and oral) as training data for the TreeTagger parameter file despite the fact that manual annotators suggested a target hypothesis in a separate tier for each verbal production.

- Used tagger:
  - We chose the TreeTagger for our purposes as it has already been used both for languages of a flectional nature such as English and German and those with agglutinating characteristics such as Turkish, Swahili and Finnish.
  - Another feature is that TreeTagger performs well on small training sets.
- Extraction of training-relevant data:
  - To extract verbal and POS tiers from the EXMARaLDA data of the pupil productions that were needed for training the tagger, we used the lxml python bindings to libxml2, a robust, fast processing and commonly used C library for XML parsing.
  - Each token of the verbal tiers was first extracted and then aligned according to its timestamp attribute to produce a (token, POS tag) tuple that was subsequently used to feed the TreeTagger for training purposes.
- Tag set normalization:
  - The tag set used for German is based on the STTS tag set created by Schiller 1999 but modified to meet the needs of part-of-speech annotation for the MULTILIT Corpus.
  - The tag sets for English and Turkish were developed by the MULTILIT team. We used the STTS tag set as a starting point in order to have parallels as far as the typological similarities of the three languages go, but nevertheless had to create numerous different and/or additional tags, in particular for Turkish, where typological differences were obvious.
  - Furthermore, we used a compositional feature for the tag set to cover phenomena of everyday spontaneous language where two tokens are merged. While this is an approach that suitably maps spontaneous language to a fixed tag set, there is of course a risk of error in the case of automatic prediction due to the divergence of POS tag combinations that are not explicitly covered in the core tag set.
  - The consequence is a need for more training data in order to achieve reliable results.
- Lexicon:
  - While there might be a risk of agglutinating language corpora generating an oversized lexicon, this was not the case with our study and its Turkish lexicon comprising written and spoken productions amounting to 2,152 tokens compared to 2,109 German tokens and just 1,012 English tokens.
  - Note that we simply consider the lexicon as the set of tokens, that is, the collection of unique tokens in the corpus.
- Training of the tagger:

  Training was carried out separately for oral and written data, with samples picked randomly for evaluation at a ratio of 0:3 (training data vs. evaluation data).

- Because of the small subset, the written training data amounted to just 52 documents, while the size of the evaluation set was 13 documents.
- In order to achieve a more reliable result, we also trained oral and written data in a joint set with 162 training files and 70 evaluation files (0:3)

A list of the officially available parameter files can be found at http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

3.5.4 Formal linguistic criteria: complex verbal morphology (MORPH)

Remarks:

- Only for Turkish and English
- Concentration on complex TAM markers (and a few others – passive, causative, reflexive, reciprocal)
- English: present continuous (*-ing*) is coded in the POS tier (POS: VVING)

3.5.4.1 Formal linguistic criteria: complex verbal morphology – Turkish

Label	**Marker**	**Name**	**Examples**
Simple TAM markers			
COND	*-sA*	Conditional	*yaparsa*
GM	*-Dir*	Generalizing/factitive modality	*güzeldir; davranışlardır*
PSB	*-(y)Abil, -AmA*	(Im)possibility	*gelebiliriz*; *gelemeyiz*
OBLG	*-malI*	Obligative	*göstermeli; beklenmeli*
IMPF2	*-mAktA*	Imperfective 2	*aramaktayız*
Complex TAM markers			
IMPFPRET	*-(I)yordu*	Past imperfective	*çekiyorlardı*
IMPFEVID	*-(I)yormuş*	Imperfective evidential	*biliyormuş*
PFPRET	*-mIştI*	Past perfect	*kaybetmişti*
PRETPRET	*-dIydI*	Double preterite (colloquial)	*gittiydi*
PFGM	*-mIştIr*	Perfective factitive	*yaşamıştır*
IMPF2PRET	*-mAktAydI*	Past imperfective	*gösterilmekteydi*
IMPF2GM	*-mAktAdIr*	Imperfective factitive	*gitmektedir*
Others			
PASS	-Il	Passive	*anlatıldı; gösterilmekteydi*
CAUS	-tIr	Causative	*yaptırdı*

REFL	-n	Reflexive	*yıkandı*
RECI	-Is	Reciprocal	*gülüştüler*

3.5.4.2 Formal linguistic criteria: complex verbal morphology – English

Label	**Name**	**Examples**
PFPRET	Past perfect	*had gone; had said*
SUBJ	Subjunctive	*if he went; if he were tired*
PASS	Passive	*was driven home; had been driven home*

3.5.5 Formal linguistic criteria: norm deviations (ERR)

Remarks:

We need to draw a preliminary distinction between error and deviation:

- Def. "error": an inappropriate construction in any communicative context
- Def. "deviation": an inappropriate construction in the given communicative context.

In principle, we say that whatever we annotate on the ERR level has a correspondence (what we think is the correct version) in a target hypothesis (ZH – "Zielhypothese"). The target hypotheses serve the purpose to make our hypotheses about norms and deviations visible.

We have one central target hypothesis, ZH1, which is the minimum target hypothesis, giving only the grammatically correct (not necessarily also the contextually correct) version of the entire text. Furthermore, the target hypothesis ZH1 is also used for constructions which are acceptable in spoken language (e.g. contractions and elisions) but need to be 'corrected' in the ZH1 in order to make the construction in question accessible to automatic parsing[7]. In this case, there is no correspondence to a deviation marked in an ERR tier. A second target hypothesis, ZH2, is developed at places where we annotate deviations. Section 3 ("Zielhypothesen") of the FALKO manual[8] serves as a guide for distinguishing between ZH1 and ZH2.[9]

The ZHs do not include the dialogue parts of a text (DIAL1, see 3.5.8), repetitions, the wrong parts of self-corrections and incomplete utterances at any level where it is not clear what they

[7] By training the parser for automatic parsing with the original verbal data we got more than 90% correct annotations. Therefore we did not need the ZH1 to make the applicable text accessible to automatic parsing.

[8] See Reznicek, Lüdeling, Krummes & Schwantuschke (2012).

[9] For a discussion see Lüdeling (2008) and Reznicek, Lüdeling & Hirschmann, (2014).

are. If a deviation only corresponds to one target hypothesis (ZH1 or ZH2), we annotate this by adding "ZH1" or "ZH2" as applicable to the annotation in the ERR tier (e.g. WODEVZH2). We do not add "ZH1"/"ZH2" if the deviation corresponds to both target hypotheses.

- It is necessary to establish a tier for each dimension (ERRTXT, ERRSYN, ERRMORPH, ERRLEX, ERRORTHO) since categories may overlap (e.g. morphological mistake in word order deviation).
- Self-corrected deviations/errors are not annotated under ERR but marked with SK in the POS tier.
- In ERRLEX, the POS is annotated to the applicable error. We code this as what it should be (our interpretation) rather than what it is.
- In order to code missing elements (see ERRSYN), split the event immediately before or after the place where the respective element should be placed and code it in the new event thereby created. Note that the event has to be split in all tiers.
- A strong accent (German accent in a Turkish or English text, Turkish accent in a German text) should be noted in the comment tier.
- Orthography:
  - In the case of multiple orthographical errors, e.g. PHON and WORD, we note both, e.g. PHONWORD. If there is more than one orthographical error of the same category within one word, this is not additionally annotated (not: PHONPHON).
  - In principle, norm deviations (in particular in Turkish and English) are to be understood as creative and analytical approaches towards bridging gaps in standard knowledge. The analysis therefore tries to understand the (orthographical) interim system ('interlanguage') developed in the written texts.
  - Turkish: Constant non-deployment of Turkish diacritics and use of substitute variants without diacritics (e.g. <g> for <ğ>, <i> for <ı>, <c> for <ç>, <s> for <ş>) is to be marked in the form of a general comment.
  - We use REG (register deviation) for written and spoken texts with respect to the situation of text production and the register which was expected (for example REG deviations in German are some amalgamated forms (preposition plus article, e.g. *ins Kino*) also found in written texts, dialectic forms and very informal vocabulary in spoken texts (e.g. *ich bin nach hause* <u>*jejangen*</u>; *das* <u>*Kloppen*</u> *ist eigentlich was sehr Schlimmes*)).

3.5.5.1 Formal linguistic criteria: norm deviations – German

Label	Description	Notes	Examples
Text (ERRTXT)			
TXT	Problem of reference tracking	For oral and written texts	

Syntax (ERRSYN)			
WO	Word order, subtypes:		
WOERR	Word order error	Merge all events which are affected by the wrong word order (and its correction in ZH1)	*Oder ich finde auch nicht Streiten gut*
WODEV	Word order deviation	Might be acceptable in other contexts	*weil* clauses with SVO in written texts
Z	Missing element, subtypes:	Does not count for interrupted sentences	
ZS	Missing subject		*Erst fängt so mit Schimpfen an.*
ZO	Missing object		*Ein Mädchen fand das Geld sah auch das von der Tasche der Frau fiel und gab trotzdem nicht zurück.*
ZXX	Missing other element	XX: define according to part of speech	*da hatte jemand aus meiner Klasse auf em n Kopf Stein geworfen.* (ZART)
CONGR	S-V congruence	Problems with congruence between subject and verb	*Prügeln ist auch nicht gut, weil man den anderen wehtust und auch sich*
Morphology (ERRMORPH)			
N	Wrong nominal inflection (case, number, gender) concerning elements of NP (adjective, determiner, noun) and pronouns, also missing compound *-s*	No detailed distinction (case, number, gender, which part of NP affected, …)	*So nahm meine Fußballkarriere sein Ende;* *Ich denke darüber das es nicht gut ist ein Mittelfinger zeigt;* *Dann habe ich ihr gefragt;* *Umgebungkontrolle*
V	Wrong verbal inflection: temporal and other		*... wie sie sich gegenseitig hälften*
Vgibs	The form *gibs* with redundant *-s*		*...aber es gibs ja noch mal die technologische Reihe*
VDEV	Stylistically inappropriate tempus/modus/aspect		*Wenn ich diskriminiert würde, hätte ich wahrscheinlich Tag und Nacht nur geheult.*
Lexicon (ERRLEX)			
ERRXX	Inappropriate, redundant or wrong word (incl. collocations, idioms, frozen phrases) (not informal register and not case, number or redundant pronoun)	XX: define according to part of speech as well as COLL for collocations and PHRASE for idioms and frozen phrases	*etwas persönlich finden;* *Mit dem Geld zu klauen habe ich gesehen;* *ich habe gewartet bis sie fertig war und am richtigen Moment wo meine Lehrerin nicht geguckt hat habe ich abgeschrieben;*

NEOLXX	Neologisms (formally wrong but showing linguistic creativity)	XX: define according to part of speech	*Akzeptierung; Außereinanderhaltungen*
REG	Stylistic or register deviation	Any kind of stylistic or register deviation, also in spoken texts	*ick*
Orthography (ERRORTHO) (for written texts only)			
GRAPH	Graphic level	Any mistake pertaining to the use of a grapheme from another language, e.g. deployment of Turkish diacritics	
PHON	Phonographic level	Any (other) mistake (supposedly) pertaining to wrong, missing or additional graphemes e.g. *falln gelasen; dan; Mittschüler; Viedo; als der Lehrer das bemärkte /*	
SYLL	Syllable level	Wrong syllabification	
WORD	Word level	Wrongly separated or merged word forms, wrong or missing in-sentence capitalization e.g. *daraufgetreten; ein mal; darauf hin; zu stande kommen; vorallem*	
SEN	Sentence level	Wrong sentential or clausal punctuation or missing capital at beginning of sentence	

3.5.5.2 Formal linguistic criteria: norm deviations – Turkish

Label	Description	Notes	Example
Text (ERRTXT)			
TXT	Problem of reference tracking	For oral and written texts	
Syntax (ERRSYN)			
WO	Word order, subtypes:		
WOERR	Word order error	Merge all events which are affected by the wrong word order (and its correction in ZH1)	*Yani kavgam hiç olmadı*
WODEV	Word order deviation	Might be acceptable in other contexts	*kötü bir olay benim açıdan bir oğrencimizi dışlamak*
Z	Missing element, subtypes:	Does not count for interrupted sentences	
ZS	Missing subject	(as error – this does not pertain to pro-drop)	*Klassenfotoda da komik, ama iyi değil*
ZO	Missing object		*yani çalıştım ama unuttum*
ZXX	Missing other element	XX: define according to part of speech	
CONGR	S-V congruence	Problems with congruence between subject and verb	*Biz kämpfen konuşdum.*

ERRXX	Redundant pronoun in argument position e.g. subject pronoun, reflexive pronoun	XX: define element according to part of speech, add S if in subject position, OBJ for object position, O for other	*iki kişi birbirlerine bakmalarına onu yaşadım*
ERRXX	Inappropriate or wrong type of subordination or linking	XX: define construction as in SYN2	*çünkü daha başarılıydı diye*
Morphology (ERRMORPH)			
N	Wrong nominal inflection (case, number, possessive)	No detailed distinction (case, number, …)	*baskalari kopya çekmelerini gördüm;* *Ben ve arkadaşım kursumuz giderken … ;* *Arkadaşlara böyle şeyler yaparsanız hiç bir arkadaşlarınız olmaz*
V	Wrong verbal inflection: TMA and voice		*para düşmeklen ve aufheben yapmakla*
DER	Derivational morphology		*ve şu kavga hiç gerek deyildi.*
Lexicon (ERRLEX)			
ERRXX	Inappropriate, redundant or wrong word (incl. collocations, idioms, frozen phrases) (not informal register and not case, number or redundant pronoun)	XX: define the element (→ classification of parts of speech as well as COLL for collocations and PHRASE for idioms and frozen phrases	*o çocukların sırasında olmak istemedim;* *böyle bi olaylara gerek yok* (ERRART)
NEOLXX	Neologisms (formally wrong but showing linguistic creativity)	XX: define according to part of speech	postposition: *ilen* verb: *kopyalama*
REG	Stylistic or register deviation	Any kind of stylistic or register deviation, also in spoken texts	*bi tane küçük kağıda yazıp gizlicene öğretmenin arkasından bakmak*
Orthography (ERRORTHO) (for written texts only)			
GRAPH	Graphic level	Deployment of specific German graphemes/grapheme combinations (<sch>, <ä>, <eu>, <tsch>, …) instead of Turkish ones (<ş>, <e>, <oy>, <ç>) in Turkish words. Note: constant non-deployment of Turkish diacritics (e.g. <g> for <ğ>, <i> for <ı>, <c> for <ç>, <s> for <ş>) is only to be marked in the form of a general comment.	
PHON	Phonological level	Any (other) mistake (supposedly) pertaining to wrong, missing or additional graphemes	
SYLL	Syllable level	Wrong syllabification	

WORD	Word level	Wrongly separated or merged word forms, wrong or missing in-sentence capitalization
SEN	Sentence level	Wrong sentential or clausal punctuation or missing capital at beginning of sentence

3.5.5.3 Formal linguistic criteria: norm deviations – English

Label	Description	Notes	Example
Text (ERRTXT)			
TXT	Problem of reference tracking	For oral and written texts	
Syntax (ERRSYN)			
WO	Word order, subtypes:		
WOSV	Inversion of subject & verb		*In the class are the childrens*
WONEG	Wrong placement of negator		*I not have this*
WOERR	Any other wrong word order apart from the above		*Listen in the film it is problem many.*
WODEV	Word order deviation	Might be acceptable in other contexts	*I don't like the write from a children.*
Z	Missing element, subtypes:	Does not count for interrupted sentences	
ZCOP	Missing copular		*I from Berlin.*
ZART	Missing article		*We are big family.*
ZA	Missing auxiliary, subtypes:		
ZAdo	No *do*-support in Q or NEG		*I'm not speak English.*
ZAbe	No passive aux *be*		*people shouldn't mobbed*
ZS	Missing subject		*I say is bad*
ZO	Missing object		*Some one children not showed his friends*
ZXX	Missing other element	XX: define according to part of speech	*The girls laughing the boys.* (ZAPPR)
Morphology (ERRMORPH)			
N	Wrong nominal inflection	No detailed distinction (case, number, which part of NP affected, …)	*feets*
V	Wrong verbal inflection, also missing third person -*s*,		*gaved* (instead of *gave*); *mean* (instead of *means*)

Lexicon (ERRLEX)			
ERRXX	Inappropriate, redundant or wrong word (incl. collocations, idioms, frozen phrases) (not informal register and not case, number or redundant pronoun)	XX: define the element (→ classification of parts of speech as well as COLL for collocations and PHRASE for idioms and frozen phrases	*The fight was solve from a neutral person.* (ERRAPPR)
NEOLXX	Neologisms (formally wrong but showing linguistic creativity), not neologisms formed with material from different languages (see language mixing 3.5.6.3)	XX: define according to part of speech	*give the money to a "help-organisation"* (NEOLNN)
FFXX	False friend	XX: define according to part of speech	*Mobbing*
REG	Stylistic or register deviation	Any kind of stylistic or register deviation, also in spoken texts	*ok; Bey! End; I have 3 Brothers and 1 sister;*
Orthography (ERRORTHO) (for written texts only)			
GRAPH	Graphic level	Deployment of specific German graphemes/grapheme combinations (<sch>, <ä>, <eu>, <tsch>, …) instead of English ones	
PHON	Phonological level	Any (other) mistake (supposedly) pertaining to wrong, missing or additional graphemes	
SYLL	Syllable level	wrong syllabification	
WORD	Word level	Wrongly separated or merged word forms, wrong or missing in-sentence capital	
SEN	Sentence level	Wrong sentential or clausal punctuation or missing capital at beginning of sentence	

3.5.6 Formal linguistic criteria: Language mixing (MIX)

Remarks:

- Only clear examples of ‘other’ language use are regarded as language mixing.
- Do not create a ‘MIX’ tier if there is no mixing in the text.
- We do not annotate the part of speech of language mixing here; this is done in the part of speech criteria.
- False friends (e.g. *mobbing* in English texts) are not annotated here and neither are innovations which might be based on transfer of the derivational procedure (see Norm deviations *3.5.5.3*).

3.5.6.1 Formal linguistic criteria: language mixing in German texts

Label	Category
ALT	Alternation (phrase or clause)
LEX	Lexical insertion, morphologically integrated
OK	Add ‘OK’ to the annotation (e.g. ALTOK; LEXOK) if the insertion is indicated as belonging to the ‘other’ language (e.g. by way of quotation marks, comments or other means).

3.5.6.2 Formal linguistic criteria: language mixing in Turkish texts

Label	Category
ALT	Alternation (phrase or clause)
LEXINT	Lexical insertion, morphologically integrated (i.e. with Turkish morphology)
LEXNINT	Lexical insertion, not morphologically integrated (i.e. missing Turkish morphology which should be there)
OK	Add ‘OK’ to the annotation (e.g. ALTOK; LEXOK) if the insertion is indicated as belonging to the ‘other’ language (e.g. by way of quotation marks, comments or other means).

3.5.6.3 Formal linguistic criteria: language mixing in English texts

Label	Category
ALT	Alternation (phrase or clause)
LEX	Lexical insertion
NEOL	Neologism, word formed with one part EN and one part DE or TR
OK	Add ‘OK’ to the annotation (e.g. ALTOK; LEXOK) if the insertion is indicated as belonging to the ‘other’ language (e.g. by way of quotation marks, comments or other means).

3.5.7 Formal linguistic criteria: textual

Remarks:

In accordance with Tolchinsky et al. (2002) we concentrate on the openings and closings in the monologue parts of the oral texts or the primarily monologue-based written texts.

- Focus with regard to the formation of criteria
  - The aspect of stance (“Haltung”/positioning in Tolchinsky et al. 2002), i.e. the frame of reference used by the speaker or writer to open and close his or her topic.
  - The aspect of function, i.e. the role the opening or closing plays in the text or the function within the narrative or expository text.
- Research aims
  - To find similarities and differences in modality and genre
  - To observe the degree of linguistic development and correspondences between orality and literacy in different age groups and languages

- Procedure
  - A distinction is made between opening utterances (“o”) and closing utterances (“c”) (see below for exceptions).
  - All texts including extremely short ones are annotated according to the textual criteria. In texts consisting of only one utterance, the distinction between opening and closing is omitted. In these texts, FUNC may also be omitted.
  - Clarification: the structural boundary of an opening/closing is the first and last main clause or complex sentence from a linguistic point of view (MCD or MCD+x).
  - It is possible for an opening or closing to be coded with more than one criterion if necessary (e.g. in FUNC: argument (ARG) and evaluation (EVAL) for one opening or closing). In this case, both codings are combined, e.g. ARGEVAL.
  - Ambiguous types are coded with a slash between the two annotations.
  - Headings are annotated in the MODE tier (see 3.5.8).
  - We do not distinguish implicit and explicit arguments.
  - Particularly oral texts but sometimes also written texts are further split according to the interviewer’s questions and comments so that new narratives appear. In this case, we would have to annotate several events that would be marked with a number behind the coding (e.g. MORo1, MORo2 etc.). At this stage, however, we annotate every text as one narrative or discussion as applicable

Label	**Term**	**Explanation**
Stance (STANCE) / Positioning		
MORo MORc	Moral	Prescriptive, evaluative, desiderative
DISCo DISCc	Discursive	Explicit reference to own act of speaking or writing
EPISo EPISc	Episodic	Explicit reference to a concrete episode, regardless of whether or not this comes from the video
GENo GENc	Generalization (synoptic)	General statement (non-moral)
Othero Otherc	Other	None of the above categories is applicable
Nono Nonc	No stance (no position)	Opening or closing without positioning/stance
Function (FUNC)		
ORIo ORIc	Orientation	Explicit spatial or temporal setting or reference to event, relation to speaker or writer’s experience (or what they are talking/writing about)

INTROo	Introduction	General statement
EPISo EISc	Episode	Could be at any point: an episode itself without introduction, setting or other context
ARGo ARGc	Argument	Implicit or explicit argument
EVALo EVALc	Evaluation	Explicit opinion, point of view or judgement about an event or character resolution – pertaining to action/conflict in the narrative
CONC	Conclusion	Summary and condensation of the material presented in the text. Pertains to closing
CODA	Coda	Formulaic or non-formulaic relation of events to the state of affairs at the time of narration Pertains to closing
Otherc Othero	Other	Not clearly any of the above
Nono Nonc	No function	Opening or closing without function

3.5.8 Formal linguistic criteria: communicative mode (MODE)

Remarks:

- We distinguish between oral texts, where we have DIAL1 and DIAL2, and written texts, where headings are annotated. However, if we have DIAL1 or DIAL2 in a written text, or (something like) a heading in a spoken text, we annotate this and add an exclamation mark (!).
- We distinguish dialogue utterances from the rest which we expect to be of a monologue nature.

	Label	**Category**
Oral texts	DIAL1	Direct addressing of interviewer (answer to a question, question or other)
	DIAL2	Repetition or paraphrasing of interviewer’s question or statement
Written texts	HEAD	Heading (annotate for POS and NP but not for SYN)
Written texts	FORM	Formulaic text conclusion

3.5.9 Formal linguistic criteria: direct and indirect speech (QT)

(No remarks)

Label	Category	Notes	Examples
DIR	Direct speech		*He said go away! ;He said “ok”*
IND	Indirect speech		*he said that I should go away*
IP	(explicit) speech-introducing phrase	No distinction between introduction of direct vs. indirect speech	*He said go away!;* *He said “ok”;* *he said that I should go away*

4 References

Berman, R.; Verhoeven, L. (2002). Developing text-production abilities across languages, genre and modality. In: Written Languages and Literacy, 5, 1.

Dudenredaktion (eds.) (2006). Duden – die Grammatik: unentbehrlich für richtiges Deutsch, Mannheim: Dudenverlag.

Göksel, A.; Kerslake, C. (2005). Turkish. A comprehensive grammar. (Comprehensive Grammars) London: Routledge.

Biber, D. (2012). Longman Grammar of spoken and written English. Harlow: Longman.

Extra, G. & Yağmur, K. (2008). Immigrant minority languages in Europe: cross-national and cross-linguistic perspectives. In: Extra & Gorter (eds.): Multilingual Europe: Facts and Policies. Berlin, New York: De Gruyter. 315-336.

Fürstenau, S.; Yağmur, K. (2003): Sprachen übergreifender Vergleich. In: Fürstenau, S.; Gogolin, I.; Yağmur, K. (eds.): Mehrsprachigkeit in Hamburg. Ergebnisse einer Sprachenerhebung an den Grundschulen in Hamburg. Münster: Waxmann. 123-137.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: Walter, M.; Grommes, P. (eds.): *Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung.* (Linguistische Arbeiten, 520) Tübingen: Niemeyer, 119-140.

Maas, U. (2010). Literat und orat. Grundbegriffe der Analyse geschriebener und gesprochener Sprache. *Grazer Linguistische Studien. 73*, 21-150.

Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F.; Herkenrath, A. (2004). Handbuch für das computergestützte Transkribieren nach HIAT. *Arbeiten zur Mehrsprachigkeit* Folge B (Nr. 56). Universität Hamburg: Sonderforschungsbereich Mehrsprachigkeit. Available online at: http://www1.uni-hamburg.de/exmaralda/Daten/6D-HIAT/AzM_56-extern-audios.pdf (last access: 2014-11-04).

Reznicek, M.; Lüdeling, A.; Krummes, C.; Schwantuschke, F. (2012). Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Available online at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko (last access: 2014-11-04).

Reznicek, M.; Lüdeling, A.; Hirschmann, H. (2014). Competing Target Hypotheses in the Falko Corpus. A Flexible Multi-layer Corpus Architecture. In: Díaz-Negrillo, A. (ed.):

Automatic treatment and analysis of Learner Corpus Data. Amsterdam: Benjamins. Available online at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen-en/marc/ReznicekEA_toAppear.pdf (last access: 2014-11-04)

Schiller, A., Teufel, S., Stöckert, C. und Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht, Universitäten Stuttgart und Tübingen. Available online at http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf. (last access 2015-01-19).

Schmid, H. (1994): Probabilistic part-of-speech tagging using decision trees. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer. Available online at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.2255 (last access: 2015-01-19).

Selting, M.; Auer, P.; Couper-Kuhlen, E.. et al. (2009). Gesprächsanalytisches Transkriptionssystem (GAT 2). In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10. Available online at: http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf (last access: 2014-11-04).

Tolchinsky, L.; Johansson, V.; Zamora, A. (2002). Text openings and closings in writing and speech: Autonomy and differentiation. In: *Special issue of written language and literacy*, Volumes 5 (1) & 5 (2). Amsterdam: Benjamins, 219-253.

Westpfahl, S., & Schmidt, T. (2013). POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch 1. *JLCL*, *28*(1), 139–153.