Syntax

Joanna Blaszczak[1], *Stefanie Dipper*[1], *Gisbert Fanselow*[1], *Shinishiro Ishihara*[1], *Svetlana Petrova*[2], *Stavros Skopeteas*[1], *Thomas Weskott*[1], *Malte Zimmermann*[1]

University of Potsdam ([1]) and Humboldt University Berlin ([2])

The guidelines for syntactic annotation contain the layers that are especially relevant for queries related to the interaction of information structure with syntax. The layers of this level are constituent structure, grammatical functions, and semantic roles.

1 Preliminaries

The following guidelines are the original product of the collaboration among different projects within the SFB 632. They are only partially related to other syntactic annotation standards (e.g. Penn Treebank (Santorini 1990), GNOME (Poesio 2000), TIGER corpus (Albert et al. 2003), Verbmobil (Stegmann et al. 2000)). Our main goal is to annotate the most important syntactic information in a theory neutral way. In this sense, the guidelines contain a systematic list of linguistic categories that allow for retrieval of syntactic information from a cross-linguistic corpus. Users who need more fine-grained distinctions may declare further categories as long as this corresponds to the general logic of the guidelines.

Interdisciplinary Studies on Information Structure 07 (2007): 95–133
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

2 Layer declaration

Table 1: Layers

Layer	Name
Constituent structure	CONST
Grammatical functions	FUNCTION
Semantic roles	ROLE

Constituents, their functions, and their roles are annotated within single cells in the hierarchical tiers CS1... CS*n*. They are given in the order: constituent categorial label (e.g. NP) – grammatical function (e.g. SUBJ) – semantic role (e.g. AG).

(1) English

<WORDS>	I	saw	the	boy	who	ate	the	mango	.
<CS1>					NP-SUBJ-AG	V	NP-OBJ-THE		
<CS2>					S-ATTR				
<CS3>	NP-SUBJ-EXP	V	NP-OBJ-TH						
<CS4>	S-MAIN								

3 Layer I: Constituent structure (CS1… CS*n*)

3.1 Introduction

Since labeling of constituent structures always involves embedding, we will use multiple layers on EXMARaLDA for constituent structures. In principle, there is no limit for the number of constituent structure (i.e., one can create as many layers as he/she needs). Each layer will be named ‘CS1’, ‘CS2’, ..., ‘CS*n*’. The ordering of the numbers of annotation layers proceeds from the embedded layer ‘CS*j*’ to the embedding layer ‘CS*j*+1’ (see examples in §3.3). If one needs a

relatively complex and deeply embedded structure, it would be better to annotate the constituent structure using tools like ANNOTATE.

The table form of the constituent structure annotation looks like a reversed syntax tree. Sister constituents are annotated in the same table-line. Daughter constituents are annotated in the higher line.

3.2 Tagset declaration

Our annotation system defines a restricted number of phrasal constituents as declared in Table 2, focusing on the most important syntactic components.

Table 2: Tagset declaration for constituent structure

tag	meaning
AP	adjectival phrase
NP	noun phrase
PP	prepositional phrase
V	verbal head
S	sentence/clause

3.3 Instructions and illustrative examples

There are four obligatory labels for the annotation of constituent structure: NP, PP, V, and S. Verbs and arguments are directly dominated by the S node, i.e. there is no explicit VP node. The difference between internal and external arguments can be retrieved through the layers FUNCTION/ROLE. Only lexical verbs and copular verbs (i.e. units annotated as VLEX or VCOP at the POS layer) are annotated as V; modals and auxiliaries are not marked at the syntactic layer (they can be retrieved through the layer POS; see below). Final punctuation marks are part of the matrix S:

(2) English

<WORDS>	Peter	bought	apples	.
<CS1>	NP	V	NP	
<CS2>	S			

3.3.1 Noun phrase (NP)

An NP consists of a head noun plus any modifying or determining material, i.e. adjectives, relative clauses, determiners, demonstratives, etc.

- NPs typically occur as complements to verbs or prepositions/postpositions:

(3) English

<WORDS>	Peter	followed	the	elephant	.
<CS1>	NP	V	NP		
<CS2>	S				

(4) English

<WORDS>	the	ball	is	on	the	ground	.
<CS1>					NP		
<CS2>	NP		V	PP			
<CS3>	S						

(5) Japanese

<WORDS>	Taro-ga	hana-o	kat-ta	.
<GLOSS>	Taro-NOM	flowers-ACC	buy-PST	
<CS1>	NP	NP	V	
<CS2>	S			

- Substantive pronouns (*he*, *she*, *it, this, that, someone*, *anyone,* etc.) are NPs:

(6) English

<WORDS>	He	knows	that	.
<CS1>	NP	V	NP	
<CS2>	S			

(7) German

<WORDS>	Alles	klar
<GLOSS>	all:NOM.SG.N	clear
<CS1>	NP	
<CS2>	S	

- NPs can be embedded within another NP; note that the part of the NP *das Buch* is not annotated as such:

(8) German

<WORDS>	das	Buch	des	Lehrers
<GLOSS>	DEF: NOM.SG.N	book [NOM.SG.N]	DEF: GEN.SG.M	teacher: GEN.SG.M
<CS1>			NP	
<CS2>	NP			

- The syntactic structure of complex names may be ignored.

(9) English

<WORDS>	Noam	Chomsky
<CS1>	NP	

- In the case of discontinuous constituents, such as Split NPs or extraposed relative clauses, label both parts of the NP with an index number :

(10) German

<WORDS>	Autos	kaufte	Hans	blaue	.
<GLOSS>	car: N.ACC.PL	buy: PRT.3.SG	Hans: M.NOM.SG	blue: N.ACC.PL	
<CS1>	NP_1	V	NP	_1	
<CS2>	S				

(11) German

<WORDS>	A	boy	came	yesterday	who	ate	the	mango	.
<POS>			VLEX			VLEX			
<CS1>					NP	V	NP		
<CS2>					S				
<CS3>	NP_1		V		_1				
<CS4>	S								

- Expletive subjects are annotated as NPs:

(12) English

<WORDS>	It	is	raining	.
<POS>	PRONEXPL	VAUX	VLEX	
<CS1>	NP		V	
<CS2>	S			

- In languages like German and Dutch, expletives can occupy the first position in the sentence (so called pre-field) without being the subject. These expletive elements are not annotated at the syntactic layer.

(13) German

<WORDS>	Es	hat	ein	Mann	angerufen	.
<GLOSS>	3.SG.N	have: 3.SG	INDEF: M.NOM.SG	man: M.NOM.SG	call: PRF.PTCP	
<POS>	PRONEXPL	VAUX	DET	NCOM	VLEX	
<CS1>			NP		V	
<CS2>	S					

IMPORTANT NOTE: All kinds of pronouns are annotated as NP! Exceptions: expletives in the pre-field in German; non-substantive pronouns, e.g., possessive pronouns (which do not substitute for a complete NP but for a determiner only).

3.3.2 Prepositional phrase (PP)

A PP consists of a prepositional/postpositional head and its NP-complement (see (3) above), plus optional modifiers.

(14) German

<WORDS>	exactly	in	the	middle
<CS1>			NP	
<CS2>	PP			

- In the case of Preposition stranding, label both parts of the PP (preposed NP and the head P) with the same index number (just like the Split NP case).

(15) English

<WORDS>	Who	did	you	give	the	book	to	?
<CS1>	NP							
<CS2>	PP_1		NP	V	NP		_1	
<CS3>	S							

- Pronominal adverbs are also PP constituents.

(16) German

<WORDS>	Ich	warte	darauf	.
<GLOSS>	1.SG:NOM	wait:1.SG	there.on	
<POS>	P	VLEX	ADV	
<CS1>	NP	V	PP	
<CS2>	S			

- Constituents containing a subordinating conjunction such as “as” governing an NP are annotated as PPs.

(17) German

<WORDS>	wie	Hans
<GLOSS>	as	Hans.M[SG.NOM]
<POS>	SUB	NPRP
<CS1>		NP
<CS2>	PP	

(18) German

<WORDS>	als	erstes
<GLOSS>	as	first.N[SG.NOM]
<POS>	SUB	NPRP
<CS1>		NP
<CS2>	PP	

- PPs may be embedded within higher PPs.

(19) German

<WORDS>	bis	zum	Auto
<MORPH>	bis	zu-m	Auto
<GLOSS>	1.SG:NOM	to-DEF:[N]SG.DAT	car.N[SG.DAT]
<POS>	P	P-DET	NCOM
<CS1>			NP
<CS2>		PP	
<CS3>	PP		

- In the following cases, the annotation just marks a flat PP.

(20) German

<WORDS>	von	hinten
<GLOSS>	from	back
<POS>	P	ADV
<CS1>	PP	

(21) German

<WORDS>	zum	Pferd	zurück
<MORPH>	zu-m	Pferd	zurück
<GLOSS>	to-[N]SG.DAT	horse:[N]SG.DAT	back
<POS>	P-DET	NCOM	ADV
<CS1>		NP	
<CS2>	PP		

3.3.3 Verb (V)

A V at the syntax layer is either a lexical (VLEX) or a copula verb (VCOP) at POS layer. Modal verbs and auxiliaries are not annotated in the constituent structure. The verb and its arguments are placed at the same CS*n* layer.

(22) English

<WORDS>	Peter	enthusiastically	sang	a	song	to	Mary	.
<POS>			VLEX					
<CS1>							NP	
<CS2>	NP		V	NP		PP		
<CS3>	S							

(23) English

<WORDS>	I	saw	the	boy	who	ate	the	mango	.
<POS>		VLEX				VLEX			
<CS1>					NP	V	NP		
<CS2>					S				
<CS3>	NP	V	NP						
<CS4>	S								

(24) English

<WORDS>	There	is	a	man	in	the	garden	.
<POS>		VCOP						
<CS1>						NP		
<CS2>		V	NP		PP			
<CS3>	S							

In examples with modal verbs and auxiliaries, the V node is only assigned to the lexical verb.

(25) English

<WORDS>	He	must	go	to	Paris	.
<POS>		VMOD	VLEX			
<CS1>					NP	
<CS2>	NP		V	PP		
<CS3>	S					

Raising and control verbs are treated like ordinary verbs. They subcategorize for a sentential complement as shown in (26) and

(27) below. Compare the annotation of the control verb *intend* and the raising verb *seem* below with the annotation of the modal verb *must* in (25): Note that at the layer CS3 of (26) and (27), infinitival phrases are annotated as S.

(26) English

<WORDS>	He	intended	to	go	to	Paris	.
<POS>		VLEX		VLEX			
<CS1>						NP	
<CS2>				V	PP		
<CS3>	NP	V	S				
<CS4>	S						

(27) English

<WORDS>	He	seems	to	go	to	Paris	.
<POS>		VLEX		VLEX			
<CS1>						NP	
<CS2>				V	PP		
<CS3>	NP	V	S				
<CS4>	S						

3.3.4 Adjectival Phrase (AP)

In general, adjectives are not annotated at the syntactic layer. However, there are two exceptions: adjectives (or APs) that function as nominal predicates are annotated with AP. The head of the AP is not labeled; this information can be retrieved from the POS layer.

(28) English

<WORDS>	He	seems	to	be	thick	.
<POS>		VLEX		VLEX	ADJ	
<CS1>				V	AP	
<CS2>	NP	V	S			
<CS3>	S					

Similarly, APs that have arguments are also annotated.

(29) English

<WORDS>	Der	auf	Maria	stolze	Mann	lacht	.
<GLOSS>	DEF: M.NOM.SG	on	Maria: F.SG	proud:M.NOM.SG	man:M.NOM.SG	laugh:3.SG	
<CS1>			NP				
<CS2>		PP					
<CS3>		AP					
<CS4>	NP					V	
<CS5>	S						

3.3.5 Clause (S)

'S' stands for clauses. It marks both main clauses and subordinate clauses. The root S symbol also covers the final punctuation mark.

Here are some examples:

- Relative clause; note that the part without the relative clause (*the boy*) is not annotated as NP.

(30) English

<WORDS>	the	boy	who	ate	the	mango	...
<CS1>			NP	V	NP		
<CS2>			S				
<CS3>	NP						

- Clausal complement (embedded clause)

(31) English

<WORDS>	I	thought	that	John	loves	Mary	.
<CS1>				NP	V	NP	
<CS2>	NP	V	S				
<CS3>	S						

Dependent verb forms (infinitives, gerunds, participles, etc.) are labeled as S:

- Infinitival complements of lexical verbs are annotated as S:

(32) English

<WORDS>	I	intend	to	go	to	Paris	.
<CS1>						NP	
<CS2>				V	PP		
<CS3>	NP	V	S				
<CS4>	S						

- The same holds for nominalized verb forms:

(33) English

<WORDS>	He	likes	eating	apples	.
<CS1>			V	NP	
<CS2>	NP	V	S		
<CS3>	S				

- Verbs used in attributive constructions are annotated as S if they contain arguments or PP adjuncts (compare the examples below):

(34) German

<WORDS>	Der	lachende	Mann	schläft	.
<CS1>	NP			V	
<CS2>	S				

(35) German

<WORDS>	Der	auf	Maria	wartende	Mann	lacht	.
<GLOSS>	DEF: M.NOM.SG	on	Maria: F.SG	wait: PTC:M.NOM.SG	man: M.NOM.SG	laugh: 3.SG	
<CS1>			NP				
<CS2>		PP					
<CS3>		S					
<CS4>	NP					V	
<CS5>	S						

- Like attributive verbs, adverbial forms are annotated as S if they contain arguments or PP adjuncts:

(36) English

<WORDS>	Er	fliegt	auf	einem	Teppich	sitzend	.
<GLOSS>	3.SG: M.NOM	fly: 3.SG	on	INDEF: M.DAT.SG	carpet: M.SG	sit:PTC	
<CS1>				NP			
<CS2>			PP				
<CS3>	NP	V	S				
<CS4>	S						

- Examples for correlative expressions:

(37) German

<WORDS>	Ich	warte	darauf	,	dass	er	kommt	.
<GLOSS>	1.SG:NOM	wait:1.SG	there.on		that	3.SG: NOM.M	come: 3.SG	
<POS>	P	VLEX	ADV		SUBCOM	P	VLEX	
<CS1>						NP	V	
<CS2>					S			
<CS3>	NP	V	PP					
<CS4>	S							

(38) German

<WORDS>	Er	weiß	es	,	dass	sie	kommt	.
<GLOSS>	3.SG: NOM.M	know: 3.SG	3.SG: ACC.N		that	3.SG: NOM.F	come: 3.SG	
<POS>	P	VLEX	P		SUBCOM	P	VLEX	
<CS1>						NP	V	
<CS2>					S			
<CS3>	NP	V	NP					
<CS4>	S							

3.3.6 Coordination

Coordinated constituents are annotated as S if they contain a verb.

(39) English

<WORDS>	John	eats	apples	and	Mary	eats	oranges	.
<CS1>	NP	V	NP		NP	V	NP	
<CS2>	S				S			
<CS3>	S							

(40) English

<WORDS>	John	eats	apples	and	drinks	water	.
<CS1>	NP	V	NP		V	NP	
<CS2>	S				S		
<CS3>	S						

(41) English

<WORDS>	John	eats	apples	and	water	.
<CS1>	NP	V	NP		NP	
<CS2>	S					

If the coordinated constituents belong to different categories, their union is annotated as S.

(42) English

<WORDS>	wo	und	wer	bist	Du	?
<GLOSS>	where	and	who:M.NOM.SG	be:2.SG	2.SG.NOM	?
<CS1>	PP		NP	VCOP	NP	
<CS2>	S		S			
<CS3>	S					

3.3.7 Punctuation marks

In general, punctuation marks are not included in the constituent structure. The only exception exception is the sentence final punctuation (‘.’ or ‘?’, etc.) which

is dominated by root S. This allows for easy retrieval of sentence type (declarative, interrogative, imperative).

(43) English

<WORDS>	He	met	Peter	,	who	read	a	book	.
<CS1>					NP	V	NP		
<CS2>			NP		S				
<CS3>	NP	V	NP						
<CS4>	S								

3.3.8 Ellipsis, traces, etc.

The current guidelines only support annotation for overt information. Elided elements are not annotated as such.

(44) English

<WORDS>	Peter	bought	apples	and	Mary	oranges	.
<CS1>	NP	V	NP		NP	NP	
<CS2>	S				S		
<CS3>	S						

(45) German

<WORDS>	Peter	möchte	ein	rotes	.
<GLOSS>	Peter:M.NOM.SG	want:3.SG	INDEF:N.ACC.SG	red:N.ACC.SG	
<POS>	N	VLEX	DET	ADJ	
<CS1>	NP	V	NP		
<CS2>	S				

(46) German

<WORDS>	Zu	der	Post	?
<GLOSS>	to	DEF:F.DAT.SG	post.office:F.SG[DAT]	
<POS>	P	DET	N	
<CS1>		NP		
<CS2>	PP			
<CS3>	S			

4 Layer II: Grammatical functions (FUNCTION)

4.1 Introduction

This layer encodes the syntactic relations that various syntactic constituents in a clause (NP, PP, AP, S) entertain with respect to the main verb of that clause. Relevant information at this layer relates to the questions of (i) whether a syntactic constituent is an obligatory addition to the verb (*argument*), or whether it is an optional addition that could be easily left out (*adjunct*), (ii) whether the relative status of the different arguments differs and – if so – which of the arguments of a verb (if any) has a prominent status with respect to grammatical processes such as agreement, binding, focus marking etc.

Note that only constituents that are annotated at the CS layers may be labeled for grammatical function.

4.2 Tagset declaration

Table 3: Core annotation scheme

tag	meaning
ADJ	adjunct
ADV	adverbial subordinate clause
ARG	argument

ATTR	relative clause
MAIN	main clause
PRDNOM	predicate nominal

Table 4: Extended annotation scheme

tag	meaning
SUBJ	Subject
OBJ	unspecified object
DO	direct object
IO	indirect object

4.3 Instructions

4.3.1 General

The tags for grammatical functions are given within the layers of constituent structure after the constituent labels.

(47) English

<WORDS>	He	met	Peter	,	who	read	a	book	.
<CS1>					NP-SUBJ	V	NP-DO		
<CS2>	NP-SUBJ	V	NP-DO		S-ATTR				
<CS3>	S-MAIN								

4.3.2 Core vs. Extended Annotation scheme

Not all projects will need equally fine-grained distinctions between the various grammatical functions. For instance, while for some projects it may be sufficient to mark the difference between a syntactic *argument* (ARG) and a syntactic *adjunct* (ADJ), others may want to mark differences between different kinds of arguments, say subject (SUBJ) vs (direct) object (DO). In the absence of further

guidelines from the individual project, the annotators are recommended to restrict their annotation to the core scheme.

4.3.3 Core scheme

ARG.

The category ARG is assigned to those syntactic constituents that appear as obligatory complements to the main verb. This means that they CANNOT be left out without a change in grammaticality or a significant change in meaning. (Notice that the obligatory appearance of an element in some syntactic position, does not necessarily mean, that this element is an argument. It can be that such elements appear in specific syntactic position due to some special syntactic requirements of a given language, e.g., in V2 sentences in German, some element (sometimes an expletive one) must obligatorily appear in the first position, e.g. *danach kommt ein Einhörnchen, ein Einhörnchen kommt danach, *kommt ein Einhörnchen danach, es kommt ein Einhörnchen danach.*)

ARGs mostly (but not always!) refer to (groups) of individuals and are assigned structural case (NOM, ACC, PAR) in case-assigning languages.

As the classical terminology suggests, intransitive verbs such as *sleep* in *John sleeps* only take one argument, namely the NP *John* (note the ungrammaticality of **sleeps*). Transitive verbs such as *criticizes* in *John criticizes a book* take two arguments, namely *John* and *a book* (note that omission of *a book* induces an unspecific generic meaning, along the lines of 'John generally criticizes something or other'). A ditransitive verb such as *give* in *John gave Mary a book* takes three arguments, namely *John*, *Mary*, and *a book*. Again, omission of one or more of the arguments either leads to ungrammaticality (**gave Mary*) or to a change in (verb) meaning (*John gave a book* = 'John donated a book').

(48) English

<WORDS>	John	gave	Mary	a	book	.
<CS1>	NP-ARG	V	NP-ARG	NP-ARG		
<CS2>	S-MAIN					

ADJ.

The category ADJ (=adjunct) is assigned to those constituents that appear as optional additions, be it to the main verb or to a given noun.

This means that they CAN be left out freely without a change in grammaticality or a significant change in meaning. In *John called Mary (from school) (with his cell phone)* the optional additions *from school* and *with his cell phone* are such optional additions that can be left out freely.

Adjuncts are generally used to convey additional information about the time, place, manner, or cause of the event or situation described by the clause (see below). That is, they restrict the class of events/situations described by the clause to a subset. If required the category ADJ can be split up into semantic sub-categories, that are annotated in layer semantic roles (time, location, etc.).

(49) English

<WORDS>	Today	John	came	to	school	.
<CS1>					NP-ARG	
<CS2>		NP-ARG	V	PP-ARG		
<CS3>	S-MAIN					

PPs may be either arguments or adjuncts. They are annotated as arguments when they are governed by the verb. Some identifying properties of arguments PPs are that (a) the semantics is not compositional and (b) the choice of preposition depends on the verb totally. The prototypical category are verbs that govern certain prepositions:

(50) English

<WORDS>	I	am	waiting	for	Mary	.
<CS1>					NP-ARG	
<CS2>	NP-SUBJ		V	PP-ARG		
<CS3>	S-MAIN					

PPs that are not governed by the verb are adjuncts.

(51) English

<WORDS>	I	am	sleeping	in	bed	.
<CS1>					NP-ARG	
<CS2>	NP-SUBJ		V	PP-ADJ		
<CS3>	S-MAIN					

Note that NPs may be adjuncts too:

(52) English

<WORDS>	The	other	day	John	came	to	school	.
<CS1>							NP-ARG	
<CS2>	NP-ADJ			NP-ARG	V	PP-ARG		
<CS3>	S-MAIN							

4.3.4 Extended scheme

SUBJ.

The category SUBJ is assigned to a designated argument that is prominent with respect to a number of grammatical relations such (i.) as constituency with the verb, (ii.) agreement, (iii.) and binding, etc. This prominence is often taken to correspond to a prominent position in the syntactic structure of the clause.

(i) Unlike direct objects, subjects do not seem to form a constituent with the verb as shown by the fact that the two cannot be topicalised together in **[**Johann** gesehen] hat den Mann nicht* vs *[Den Mann gesehen] hat Johann nicht*.

(ii) In agreement languages, the subject is that argument that the verb always agrees with (in some languages the verb additionally agrees with the object as well): ***Johann*** *(sg.) sleeps (sg.)* vs. **The boys (pl.) sleeps (sg.)*

(iii) subjects can bind reflexive pronomina: ***Peter*** *blamed himself* vs. **Heself blamed Peter*.

Other properties that may help to identify SUBJs:

- In NOM-ACC-languages, the default case of nominal subjects is the Nominative: ***Der Mann*** *ist gekommen* vs **Den Mann ist gekommen*. This can be formalized in a rule:

  If there is only one nominative constituent in a clause, mark this constituent as SUBJ.

- Subjects are most often expressed by nominal constituents (NPs), but sentential subjects as in ***[That Peter won the race]*** *surprised me* are also possible with certain verbs.

- In languages that do not mark case morphologically, the subject status is coded by word order, i.e. the subject usually occupies a designated (linear) position relative to verb and direct object (if present). E.g., in English or French, the subject precedes the verb (and the direct object) in the default case: ***Peter*** *saw her* vs. **Her saw Peter*.

  Note that this test has to be applied with care. It seems to work fine with transitive SVO-structures, but in intransitive or existential structures the subject may also follow the object: e.g. *There came Peter out of the hall*.

- In languages that mark NOM only sporadically (e.g. on pronouns and full NPs, but not on CPs (German), or only on pronouns, but not on full NPs and CPs), a substitution test combined with considerations of linear order may help in some cases:

If a constituent α is not morphologically marked for case, but if α is in the default position for subjects (this must be independently established on the base of reference grammars) and if α can be replaced with a NOM-marked constituent β, mark α with SUBJ:

a. [$_{\alpha}$ Peter] saw Mary → He$_{nom}$ saw Mary.
   substitution possible → mark α = Peter with SUBJ

b. [$_{\alpha}$ That Peter came] surpised us. → He$_{nom}$ surprised us.
   substitution possible → mark α = [that Peter came] with SUBJ

Warning: It does not follow automatically from the impossibility of substitution that α is NOT a subject. In English, case-marked pronouns cannot be substituted for subject NPs in existential sentences because of their definiteness: *There came **Peter** down the Hill //→ *There came **he** down the hill.*

- Often the subject has the semantic role of AGENT (see 5 below), but this is not a 1:1-correspondence. E.g., in passive structures, non-agentive constituents function as subjects syntactically: ***He** was beaten*. Likewise, the subject of transitive psych-verbs such as *to like* in ***He** likes dogs* does not refer to the agent of a particular event, but rather to the experiencer (EXP) of a particular psychological disposition.

DO.

The category DO is assigned to the second argument of a transitive verb, which is not designated in the sense that it is less prominent than the subject. This rule-of-thumb makes the NP *Bill* in *The boys like **Bill*** the DO, since it does not agree with the main verb in number.

Like subjects, DOs are assigned structural case (ACC/PAR or ABS) in case-assigning languages. Like subjects, DOs have a default base position

relative to verb and subject in languages that do not assign case: In English and French, the DO follows the main verb (and the subject). DOs are generally taken to stand in close syntactic relation with the main verb, which is reflected by the fact that they can be displaced together: *[**Den Mann** gerufen] haben wir*.

Apart from this, DOs are often only identifiable based of the absence of properties typical for subjects. E.g. a DO cannot bind a reflexive in a subject position (see above), and it cannot agree with the verb in the absence of subject-verb-agreement. Other Properties that may help to identify DOs:

- There is a tendency for DOs to express the semantic role of PATIENT/THEME. However, even if all PATIENT/THEME -expressions are DOs the reverse does not hold completely. Consider e.g. *The news surpised John*, where the DO *John* expresses the semantic role of experiencer.
- As with subjects, DOs are most often expressed by nominal constituents (NPs), but sentential DOs are also possible, especially with attitude verbs (*to think*, *to belief*) or verbs of saying: *John said [that Maria had come late again]*.

IO.

The category IO is assigned to that argument of a (ditransitive) verb that is not assigned the status of SUBJ nor DO. In case-languages, IOs are often assigned the Dative. Semantically, the IO is often used to express the receiver or beneficient/maleficient of an event, such as the NP *John* in *Mary gave **John** a book/ kiss*.

Unlike SUBJs and DOs, IOs seem to always refer to individuals and must be expressed by a nominal constituent.

(53) English

<WORDS>	John	gave	Mary	a	book	.
<CS1>	NP-SUBJ	V	NP-ARG	NP-ARG		
<CS2>	S-MAIN					

Prepositional objects are annotated with the generic label OBJ:

(54) German

<WORDS>	Ich	warte	auf	Hans	.
<GLOSS>	1.SG:NOM	wait:1.SG	on	Hans:ACC.SG.M	
<CS1>				NP-ARG	
<CS2>	NP-SUBJ	V	PP-OBJ		
<CS3>	S				

4.3.5 Nominal Predicates

PRDNOM: a nominal predicate (noun or adjective), either with or without copula.

(55) English

<WORDS>	He	is	thick	.
<CS1>	NP-SUBJ	V	AP-PRDNOM	
<CS2>	S-MAIN			

(56) English

<WORDS>	He	is	the boss	.
<CS1>	NP-SUBJ	V	NP-PRDNOM	
<CS2>	S-MAIN			

(57) Russian

<WORDS>	Ona	studentka	.
<CS1>	NP-SUBJ	NP-PRDNOM	
<CS2>	S-MAIN		

The term nominal predicate may be used for the complements of further copulative verbs (cf. small clauses), e.g. *consider, call,* etc.

(58) English

<WORDS>	He	considers	him	a	thief	.
<CS1>	NP-SUBJ	V	NP-OBJ	NP-PRDNOM		
<CS2>	S-MAIN					

4.3.6 Sentences and clauses

Sentences and clauses are annotated in four categories:

- The tag MAIN is used for main clauses.

(59) English

<WORDS>	John	sleeps	.
<CS1>	NP-SUBJ	V	
<CS2>	S-MAIN		

- Relative clauses are annotated as ATTR.

(60) English

<WORDS>	I	saw	the	boy	who	ate	the	mango	.
<POS>		VLEX				VLEX			
<CS1>					NP-SUBJ	V	NP-OBJ		
<CS2>					S-ATTR				
<CS3>	NP-ARG	V	NP-OBJ						
<CS4>	S-MAIN								

- Subordinate clauses with the function of an argument (subject or object) are annotated as ARG.

(61) English

<WORDS>	Mary	thinks	that	he	came	.
<CS1>				NP-SUBJ	V	
<CS2>	NP-SUBJ	V	S-ARG			
<CS3>	S-MAIN					

- Subordinate clauses with adverbial function are annotated as ADV.

(62) English

<WORDS>	Tom	sleeps	when	the	sun	rises	.
<CS1>				NP-SUBJ		V	
<CS2>	NP-SUBJ	V	S-ADV				
<CS3>	S-MAIN						

4.3.7 Non-annotated syntactic functions

The following elements are not annotated for grammatical function:

- particles, e.g., German *ja*, *jawohl*, *doch*, etc.: these elements do not have a grammatical function, but rather they express speaker's attitudes towards the proposition.
- conjunctions, e.g., *and, but, because,* etc.
- adjectives in attributive use, e.g. *a nice boy*: the attributive function may be inferred by the fact that the adjective is part of the entire NP.

5 Layer III: Semantic roles (ROLE)

5.1 Introduction

Lexical heads not only require a certain number of arguments but also determine the semantic properties of these arguments depending on how these are involved

in the state of affairs described by the lexical head. This means that the syntactic arguments enter certain *semantic* (also called *thematic* or *theta*-) roles, which are pre-established by the selecting properties of the lexical head. The relationship between a lexical head and its arguments can be explained by the use of a small finite set of universally applicable notions which indicate whether a certain argument is the performer of an action, just undergoes an action etc. Note that only constituents that are annotated at the CS and FUNCTION layers may be labeled for semantic role.

5.2 Tagset declaration

The tags of semantic roles are not given in separate layers. They are inserted in the layers of constituent structure after the labels of grammatical functions.

Table 5: Core annotation scheme

tag	meaning
AG	Agent
CAUSE	Cause
COM	Comitative
EXPER	Experiencer
GOAL	Goal
INSTR	Instrument
LOC	Location
MAN	Manner
POSS	Possessor
THEME	Theme
TIME	Time

5.3 Instructions

5.3.1 General

The tags for semantic roles are given within the layers of constituent structure after the grammatical functions.

(63) English

<WORDS>	He	met	Peter	,	who	read	a	book	.
<CS1>					NP-SUBJ-AG	V	NP-DO-THEME		
<CS2>					S-ATTR				
<CS3>	NP-SUBJ-AG	V	NP-DO-THEME						
<CS4>	S-MAIN								

The tags for semantic roles are used with NPs, PPs, or S-ARGs that function either as arguments of verbs (*John sleeps*), or as adjuncts (*in Athens…*), or as dependents of NPs (*the house on the hill*). Not all constituents are annotated for semantic role, e.g. NP arguments of prepositions, relative clauses, etc. are not labeled for this layer.

5.3.2 Agent

NPs that refer to the entities that cause actions, either animates or inanimates, are annotated as agents.

(64) English

<WORDS>	The	boy	opens	the	window	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

(65) English

<WORDS>	The	wind	opens	the	window	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

5.3.3 Theme

Theme is a general term covering the notions of:

- *Patient*: an entity affected by the action

(66) English

<WORDS>	The	girl	paints	the	fence	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

- *Result*: an entity effected by the action, i.e. which emerges out of the action:

(67) English

<WORDS>	The	woman	built	a	house	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

- *Theme*: an entity effected by the action, i.e. which emerges out of the action:

(68) English

<WORDS>	Akropolis	is	in	Athens	.
<CS1>				NP-ARG	
<CS2>	NP-SUBJ-THEME	V	PP-ARG-LOC		
<CS3>	S-MAIN				

5.3.4 Experiencer

Experiencer is the sentient being that participates in a state/event of emotion (*love, hate,* etc.), volition (*wish, want, etc.*), cognition (*think, remember*, etc.), perception (*see*, *hear*, etc.) or bodily sensation (*feel cold, feel hungry*, etc.).

(69) English

<WORDS>	Mary	enjoys	algebra	.
<CS1>	NP-SUBJ-EXPER	V	NP-OBJ-THEME	
<CS2>	S-MAIN			

(70) English

<WORDS>	Algebra	interests	John	.
<CS1>	NP-SUBJ-THEME	V	NP-OBJ-EXPER	
<CS2>	S-MAIN			

5.3.5 Goal

Goal is a general term covering the notions of:

- *Recipient*: an entity which receives something:

(71) English

<WORDS>	John	gave	Mary	a	book	.
<CS1>	NP-SUBJ-AG	V	NP-IO-GOAL	NP-ARG-THEME		
<CS2>	S-MAIN					

- *Benefactive*: an entity to whose advantage an action is performed (or *malefactive*: an entity to whose disadvantage an action is performed):

(72) English

<WORDS>	John	bought	flowers	for	Mary	.
<CS1>					NP-ARG	
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME	PP-ADJ-GOAL		
<CS3>	S-MAIN					

- *Purpose*: the intension for which an action is performed:

(73) English

<WORDS>	John	said	it	for	more	clarity	.
<CS1>					NP-ARG		
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME	PP-ADJ-GOAL			
<CS3>	S-MAIN						

5.3.6 Instrument

Instruments are means with the help of which the action is carried out.

(74) English

<WORDS>	John	opened	the	door	with	the	keys	.
<CS1>						NP-ARG		
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME		PP-ADJ-INSTR			
<CS3>	S-MAIN							

5.3.7 Possessor

Possessor is the entity who owns something.

(75) English

<WORDS>	Bill	has	a	new	car	.
<CS1>	NP-SUBJ-POSS	V	NP-OBJ-THEME			
<CS2>	S-MAIN					

(76) English

<WORDS>	Bill's	car
<CS1>	NP-POSS	NP
<CS2>	NP	

5.3.8 Location

Location covers the spatial relations of:

- static spatial location:

(77) English

<WORDS>	Mary	is	in	New	York	.
<CS1>				NP-ARG		
<CS2>	NP-SUBJ-THEME	V	PP-ARG-LOC			
<CS3>	S-MAIN					

- direction of movement (do not mistake *direction* with *goal*, the latter being preserved for the intended target of an action not necessarily connected with spatial movement, see 5.3.5):

(78) English

<WORDS>	He	rushed	to	the	street	.
<CS1>				NP-ARG		
<CS2>	NP-SUBJ-AG	V	PP-ARG-LOC			
<CS3>	S-MAIN					

(79) English

<WORDS>	He	put	the	money	into	his	pocket	.
<CS1>						NP-ARG		
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME		PP-ARG-LOC			
<CS3>	S-MAIN							

- source: indicating the origin of movement

(80) English

<WORDS>	The	gold	falls	from	the	sky	.
<CS1>					NP-ARG		
<CS2>	NP-SUBJ-THEME		V	PP-ARG-LOC			
<CS3>	S-MAIN						

- path: indicating a place through which the movement takes place.

(81) English

<WORDS>	He	ran	through	the	door	.
<CS1>				NP-ARG		
<CS2>	NP-SUBJ-THEME	V	PP-ARG-LOC			
<CS3>	S-MAIN					

5.3.9 Time

Time covers a point or an interval of time at which the action takes place.

(82) English

<WORDS>	He	came	at	noon	.
<CS1>				NP-ARG	
<CS2>	NP-SUBJ-AG	V	PP-ADJ-TIME		
<CS3>	S-MAIN				

(83) English

<WORDS>	He	worked	all	night	long	.
<CS1>	NP-SUBJ-AG	V	NP-ADJ-TIME			
<CS2>	S-MAIN					

5.3.10 Cause

Cause indicates the reason why something happens and is often expressed by a PP (*because of, with, through* etc.). Sometimes this role is close to the role of Instrument. The criterion for the choice of tag CAUSE is if the expression can be paraphrased through a clausal subordinate clause:

(84) He convinced me <u>with his honesty.</u> ↔ He convinced me because he was honest.

(85) He climbed <u>with a hammer.</u> ≠ He climbed because he had a hammer.

(86) English

<WORDS>	He	stroke	me	with	his	originality	.
<CS1>					NP-ARG		
<CS2>	NP-SUBJ-THEME	V	NP-OBJ-EXPER	PP-ADJ-CAUSE			
<CS3>	S-MAIN						

(87) English

<WORDS>	I	worked	because	he	liked	it	.
<CS1>				NP-SUBJ-EXP	V	NP-OBJ-THEME	
<CS2>	NP-SUBJ-AG	V	S-ADJ-CAUSE				
<CS3>	S-MAIN						

(88) English

<WORDS>	Why	did	it	happen	?
<CS1>	ADJ-CAUSE		NP-SBJ-THEME	V	
<CS2>	S-MAIN				

5.3.11 Manner

Manner applies to constituents that denote how something is carried out.

(89) English

<WORDS>	Handle	with	care	!
<CS1>			NP-ARG	
<CS2>	V	PP-ADJ-MAN		
<CS3>	S-MAIN			

Adverbs may also denote manner, however, they are not annotated at any of the syntactic layers.

(90) English

<WORDS>	Ann	drove	quickly	.
<CS1>	NP-SUBJ-AG	V		
<CS2>	S-MAIN			

5.3.12 Comitative

Comitative applies to an animate entity that accompanies a participant of the action.

(91) English

<WORDS>	Peter	walked	with	Bill	.
<CS1>				NP	
<CS2>	NP-SUBJ-AG	V	PP-ADJ-COM		
<CS3>	S-MAIN				

6 Problematic cases

6.1 Sentence fragments

As a rule of the thumb: Provide the maximum information for what you see. For instance, in case of fragmentary answers to yes/no questions (compare (92)), annotate *yes* or *no* as S. In case of fragmentary answers to constituent questions (compare (93)), annotate the fragment according to its syntactic category and function; note that the fragment is also annotated as S.

(92) English

<WORDS1>	Are	you	hungry	?		
<WORDS2>					yes	.
<CS1>	V	NP-SUBJ-THEME	PRDNOM		S-MAIN	
<CS2>	S-MAIN					

(93) English

<WORDS1>	Who	ate	beef	?		
<WORDS2>					John	.
<CS1>	NP-SUBJ-AG	V	NP-OBJ-THEME		NP-SUBJ	
<CS2>	S-MAIN				S-MAIN	

6.2 Correction and breaks

Corrections by the speaker, i.e., words or sequences of words that serve to correct erroneous utterances, are marked with the symbol "!". This indicates that a constituent which has already been introduced is updated/corrected. Breaks and break fillers are not annotated in the constituent structure.

(94) English

<WORDS>	John	...	eh	...	no	Peter	laughed	.
<CS1>	NP-SUBJ-AG					!NP-SUBJ-AG	V	
<CS2>	S-MAIN							

(95) English

<WORDS>	Mary	saw	John	...	no	Peter	.
<CS1>	NP-SUBJ-EXP	V	NP-OBJ-THEME			!NP-OBJ-THEME	
<CS2>	S-MAIN						

In case only parts of constituents are corrected, only the corrected version (the complete constituent) is annotated in the constituent structure layer.

(96) English

<WORDS>	A	...	eh	...	the	woman	comes	.
<CS1>					NP-SUBJ-AG		V	
<CS2>	S-MAIN							

6.3 Non-grammatical sequences

In the case of ungrammatical information, if it is obvious to the annotator what the speaker actually wanted to say, the ungrammatical feature is marked with the symbol "#". The symbol is annotated at the layer at which the error arises, e.g. with incorrect case, at the morphological transcription, as in (75), or with incorrect word order, at the constituent structure, as in (76), (77). In case of errors in the constituent structure, the error should be marked as locally as possible, i.e., at the smallest erroneous constituent, compare (76) and (77).

(75) German

<WORDS>	Hans	sah	mir	.
<GLOSS>	Hans:NOM.SG.M	see:PAST.3.SG	1.SG:#DAT	
<CS1>	NP-SUBJ-EXP	V	NP-OBJ-THEME	
<CS2>	S-MAIN			

(76) German

<WORDS>	Ich	will	essen	Nudeln	.
<GLOSS>	1.SG.NOM	want:3.SG	eat:INF	spaghetti-ACC.PL	
<CS1>	NP-SUBJ-EXP		V	NP-OBJ-THEME	
<CS2>	#S-MAIN				

(77) German

<WORDS>	Er	trinkt	Bier	ein	.
<GLOSS>	3.SG.NOM	drink:3.SG	beer[ACC.SG.N]	DEF:ACC.SG.N	
<CS1>	NP-SUBJ-EXP	V	#NP-OBJ-THEME		
<CS2>	S-MAIN				

Often it might be difficult to know for sure what the intended utterance would have been. If it is not obvious to the annotator how to reconstruct the grammatical, intended utterance, only grammatical fragments of the sentence are annotated as usually, whereas questionable fragments are marked by "#" , to

mark their ungrammaticality. Note that no constituents dominating such questionable fragments are annotated, i.e., there is no "S" annotation in (78).

(78) German

<WORDS>	Er	denkt	Bier	ein	.
<GLOSS>	3.SG.NOM	thinks:3.SG	beer[ACC.SG.N]	DEF:ACC.SG.N	
<CS1>	NP-SUBJ-EXP	V	#	#	

7 References

Albert, Stefanie et al. 2003. *TIGER Annotationsschema*. Draft. Univerisities of Saarbrücken, Stuttgart, and Potsdam.

Poesio, Massimo. 2000. T*he GNOME Annotation Scheme Manual*. http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm.

Santorini, Beatrice 1990. *Annotation Manual for the Penn Treebank Project*. Technical Report, University of Pennsylvania.

Stegmann, R., H.Telljohann, and E. W. Hinrichs. 2000. *Stylebook for the German Treebank in VERBMOBIL*. Technical Report 239. Verbmobil.