Interdisciplinary Studies on Information Structure Vol. 7

UP

HUB

Information Structure in Cross-Linguistic Corpora:

Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure

Information structure

Michael Götze[1], *Thomas Weskott*[1], *Cornelia Endriss*[1], *Ines Fiedler*[2], *Stefan Hinterwimmer*[2], *Svetlana Petrova*[2], *Anne Schwarz*[2], *Stavros Skopeteas*[1], *Ruben Stoel*[1]

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

The guidelines for Information Structure include instructions for the annotation of Information Status (or 'givenness'), Topic, and Focus, building upon a basic syntactic annotation of nominal phrases and sentences. A procedure for the annotation of these features is proposed.

1 Preliminaries

These guidelines are designed for the annotation of information structural features in typologically diverse languages. The main objectives of these guidelines are i) language independence, ii) openness towards different theories, and iii) reliability of annotation.

These objectives resulted in a number of decisions that were implicitly made in the guidelines, the most relevant being the following:

- Annotation instructions rely mainly on functional tests, rather than tests involving linguistic form.
- Possibly different dimensions of information structure are annotated independently from each other, postulating no relation between these different features (as one could do e.g. for topic and focus).
- Most tagsets offer an obligatory tagset (or 'Core Annotation Scheme')

Interdisciplinary Studies on Information Structure 07 (2007): 147–187
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

and a tagset with optional tags (or 'Extended Annotation Scheme'), where the Core Annotation Scheme enables a more reliable and quick annotation and the Extended Annotation Scheme offers more detailed descriptions of the data.

The guidelines are structured as follows: in the next sections, annotation instructions for three different dimensions of information structure, Information Status (Section 2), Topic (Section 3), and Focus (Section 4) are provided. In Section 5, an annotation procedure is proposed and described.

2 Tagset Declaration

2.1 Core Annotation Scheme for Information Structure

Table 1: Tags of the Core Annotation Scheme for Information Structure

Layer	Tags	Short description
Information Status	giv	given
	acc	accessible
	new	new
Topic	ab	aboutness topic
	fs	frame setting topic
Focus	nf	new-information-focus
	cf	contrastive focus

2.2 Extended Annotation Scheme for Information Structure

Table 2: Tags of the Extended Annotation Scheme for Information Structure

Layer	Tags	Short description
Information Status	giv	given (underspecified)

	giv-active	active
	giv-inactive	inactive
	acc	accessible (underspecified)
	acc-sit	situationally accessible
	acc-aggr	aggregation
	acc-inf	inferable
	acc-gen	general
	new	new
Topic	ab	aboutness topic
	fs	frame setting topic
Focus	nf	new-information-focus (underspecified)
	nf-sol	Solicited new-information focus
	nf-unsol	unsolicited new-information focus
	cf	contrastive focus (underspecified)
	cf-repl	replacing
	cf-sel	selection
	cf-part	partiality
	cf-impl	implication
	cf-ver	truth-value (verum)
Note:	…+op	All kinds of foci given above can occur as bound by focus operators like the particles *only, even, also etc.* as well as negation operators. In this case, the tags are supplied with the additional marking *+op*

(cf. 4.5).

3 Layer I: Information Status

3.1 Introduction

The objective of this annotation layer is to annotate discourse referents for their information status in the discourse. "Discourse referents" are meant to comprise entities of many different types, that is individuals, places, times, events and situations, and sometimes even propositions. All these can be picked up by anaphoric expressions.

Their information status[17] reflects their "retrievability", which is meant to be understood as the difficulty of accessing the antecedent referent: a referent mentioned in the last sentence is easily accessible or "given", whereas one that has to be inferred from world knowledge is only "accessible" to the degree that the inference relation is shared between speaker and hearer. A discourse referent which lacks an antecedent in the previous discourse, isn't part of the discourse situation, nor is accessible via some relational reasoning has to be assumed to be "new".

The annotation scheme for information status proposed here consists of 1) a core annotation scheme for the obligatory tags ('giv', 'acc', 'new'), 2) an extended annotation scheme for optional tags ('giv', 'giv-active', 'giv-inactive', 'acc', 'acc-sit', 'acc-aggr', 'acc-inf', 'acc-gen', 'new'), and 3) a recommended annotation procedure.[18]

[17] Related and widely used terms are 'activation', 'retrievability', 'cognitive status', 'givenness', etc.

[18] Many principles of this annotation scheme are closely related to Nissim et al. 2004. A more detailed discussion of the annotation scheme will follow. The figure below indicates how our annotation scheme relates to notions such as discourse and hearer status.

This section is structured as follows: after the tagset declaration, instructions for annotating information structure are provided. In the last section, a procedure for applying these instructions is recommended.

3.1.1 Tagset Declaration

Table 3: Information status tags

Annotation layer:	Information Status	
Description:	Information status ("activation") of the discourse referents	
Unit:	A constituent which refers to a discourse entity; mostly referential NPs or PPs, or their pronominal counterparts, unless part of an idiom; see Section 2.2.1 Referring expressions.	
Core Annotation Scheme		
Tags:	giv	given
	acc	accessible
	new	new

information status	giv	acc	new
discourse status	discourse-old	discourse-new	discourse-new
hearer status	hearer-old	hearer-old	hearer-new

Extended Annotation Scheme		
Tags:	giv	given (underspecified)
	giv-active	active
	giv-inactive	inactive
	acc	accessible (underspecified)
	acc-sit	situationally accessible
	acc-aggr	aggregation
	acc-inf	inferable
	acc-gen	general
	new	new

3.2 Instructions for Annotating Information Status

In this section, instructions for annotating information status are provided. A procedure for applying these instructions can be found in the next section.

For annotating according to the 'Core Annotation Scheme', the sections 2.2.1 Referring expressions, 2.2.2 Given (giv), 2.2.4 Acc (acc), and 2.2.6 New (new) are relevant. However, the examples in the remaining sections might be helpful as well. For annotating according to the 'Extended Annotation Scheme', all sections have to be considered.

3.2.1 Referring expressions

At this annotation layer, we restrict ourselves to the annotation of discourse referents that are referred to by referential expressions. Among other things, this means that we don't annotate NPs or PPs that don't refer to discourse referents.

Examples for NPs/PPs that don't refer in this sense are

- "There" in sentences such as "**There** is a fly in my soup."
- expletive "it", as in "**It** always rains on Sundays."
- or (parts of) idiomatic phrases such as "on (**the other hand**)", "for (**some**

reason)", "as (**a result**)".

Further examples are given in (1) and (2), which are annotated only for illustrative purposes:

(1)

<WORDS>	Peter	kicked	the	bucket	.
<CS>	NP		NP		
<INFOSTAT>	new		***idiom***		

(2)

<WORDS>	Hans	warf	die	Flinte	ins	Korn	.
<CS>	NP		NP			NP	
<INFOSTAT>	new		***idiom***			***idiom***	
<TRANS>	Hans threw the rifle into the cornfield. (= Hans threw in the towel.)						

3.2.2 Given (*giv*)

The expression has an explicitly mentioned antecedent in the previous discourse: the referent has already been mentioned and is picked up again. In most cases, it is sufficient to check the preceding 5 sentences for an antecedent, but sometimes, anaphoric relations may stretch even across paragraphs.

- *IMPORTANT:* The referent must be referred to *explicitly* in the preceding discourse! That means that there must be expressions that refer to this discourse referent.

Note that referents can be of propositional type as in example (4). There, the first sentence introduces a referent, which the word 'that' in the second sentence refers to - this referent is given in this case.

(3)

<WORDS>	Peter	went	into	the	garden.	He	was	happy	.
<CS>	NP			NP		NP			
<INFOSTAT>	new			new		**giv**			

(4)

<WORDS>	Peter	liked	Tom.	But	this	cat	wouldn't	believe	that	.
<CS>	NP		NP		NP				NP	
<INFOSTAT>	new		new		**giv**				**giv**	

3.2.3 Extended Annotation Scheme: Subcategories of given (*giv*)

We differentiate two subcategories of 'giv', 'giv-active' and 'giv-inactive'.

Note: If you annotate tags at this layer, be as specific as possible. Only if you are not sure about which sub-tag (either 'giv-active' or 'giv-inactive') to choose, choose the less specific tag, i.e. 'giv'.

Active (*giv-active*)

The referent was referred to *within the last or in the current sentence.*

(5)

<WORDS>	Peter	went	into	the	garden	.	It	was	blooming	.
<CS1>	NP			NP			NP			
<CS2>	S						S			
<INFOSTAT>	new			new			giv-active			

(6)

<WORDS>	Peter	liked	Tom	.	But	Maria	wouldn't
<CS>	NP		NP			NP	
<INFOSTAT>	new		new			new	

<WORDS>	believe	that	.
<CS>		NP	
<INFOSTAT>		giv-active	

(7)

<WORDS>	...	They	laughed	.
<CS>	...	NP		
<INFOSTAT>	...	giv-active		

<WORDS>	And	then	they	fought	each	other	again	.
<CS>			NP		NP			
<INFOSTAT>			giv-active		giv-active			

Inactive (*giv-inactive*)

The referent was referred to *before the last sentence.*

(8)

<WORDS>	Peter	went	into	the	garden	.
<CS>	NP			NP		
<INFOSTAT>	new			new		

<WORDS>	It	was	blooming.	Peter	was	happy.
<CS>	NP					
<INFOSTAT>	giv-active			giv-inactive		

(9)

<WORDS>	Peter	went	into	the	garden	.
<CS>	NP			NP		
<INFOSTAT>	new			new		

<WORDS>	It	was	blooming	.	He	was	happy	.
<CS>	NP							
<INFOSTAT>	giv-active				giv-inactive			

3.2.4 Accessible (*acc*)

The referent of the expression has not been mentioned, but is accessible via some kind of relation to a referent in the previous discourse, in the situative context, or the assumed world knowledge of the hearer, or a combination thereof. In particular, the referent should fulfil one of the criteria in the next section (Section 2.2.5).

E.g. in the example below, the NP "the flowers" refers to a part of the previously introduced discourse referent "the garden".

(10)

<WORDS>	Peter	went	into	the garden	.	The flowers	blossomed	.
<CS>	NP			NP		NP		
<INFOSTAT>	new			new		**acc**		

(11)

<WORDS>	Could	you	pass	the sugar	,	please	?
<CS>		NP		NP			
<INFOSTAT>		**acc**		**acc**			
	(situative context)						

In example (11), both the adressed person and the sugar are part of the situative context of the communication.

(12)

<WORDS>	Peter	loves	violets	,	above	all	.
<CS>			NP				
<INFOSTAT>	giv		**acc**				
	(world knowledge)						

3.2.5 Extended Annotation Scheme: Subcategories of Accessible (*acc*)

The referent of the expression has not been mentioned, but is accessible via some kind of relational information, the situative context, or the assumed world knowledge of the hearer.

- *Note:* If you annotate tags at this layer, be as specific as possible. Only if you are not sure about which sub-tag (either or 'acc-sit', 'acc-aggr', 'acc-inf' or 'acc-gen') to choose, choose the less specific tag, i.e. 'acc'.

Situative (*acc-sit*)

The referent is part of the discourse situation.

(13)

<WORDS>	Could	you	pass	the sugar	,	please	?
<CS>		NP		NP			
<INFOSTAT>		acc-sit		acc-sit			
	(in dialogue during breakfast)						

(14)

<WORDS>	The	kid	hits	the	cow	.
<CS>	NP			NP		
<INFOSTAT>	acc-sit			acc-sit		
	(pointing with the finger at the figures in the book)					

Aggregation (*acc-aggr*)

The referring expression denotes a group consisting of accessible or given discourse referents.

(15)

<WORDS>	Peter	went	shopping	with	Maria	.	They	bought	many	flowers	.
<CS>	NP				NP		NP		NP		
<GIVEN>	new				new		acc-aggr		new		

(16)

<WORDS>	Peter	went	shopping	with	Maria	.	They	bought	many	flowers	.
<CS>	NP			NP			NP		NP		
<GIVEN>	new			new			acc-aggr		new		

Inferable (*acc-inf*)

Since reliably distinguishing various types of inferables[19] appears to be difficult (cf. Nissim et al. 2004), we restrict ourselves to identifying inferables as such and don't annotate their subtypes. However, we provide some types here as a help for recognizing various instances of inferables.

Assign 'acc-inf', if the referent is part of one of the following bridging relations:

- part-whole: The referent is in a part-whole relation to a referent in the preceding discourse.

(17)

<WORDS>	The	garden	beautiful	.	Its	entrance	is	just	across	this	river	.
<CS>	NP				NP					NP		
<GIVEN>	giv-act				acc-inf					acc-sit		

- set-rel: The referent is part of a set relation (i.e. subset, superset, member-of-the-same-set) to a referent in the preceding discourse.

(18)

<WORDS>	The	flowers	in	the	garden	blossom	.
<CS>	NP						
<GIVEN>	giv-inactive						

[19] or Bridging expressions.

<WORDS>	The	flowers	near	the	gate	blossom	violet	.
<CS>	NP							
<GIVEN>	acc-inf							

(19)

<WORDS>	The	children	swam	in	the	lake	.
<CS>	NP						
<GIVEN>	giv-inactive						

<WORDS>	The	famliy	experienced	a	beautiful	day	.
<CS>	NP			NP			
<GIVEN>	acc-inf			acc-gen			

- entity-attribute: The referent is constitutes an attribute of a referent in the preceding discourse.

(20)

<WORDS>	The	flowers	enchanted	Peter	.	Their	scent	was	wonderful	.
<CS>	NP					NP				
<GIVEN>	acc-new			giv-inactive		acc-inf				

General (*acc-gen*)

The speaker can assume that the hearer knows the referent from his or her world knowledge. Note that the expression can take on different forms (i.e. indefinite, definite, or bare NP).

- Type: The referent of the expression is a set or kind of objects.

(21)

<WORDS>	The	lion	is	dangerous	,	when	she	has	children	.
<CS>	NP						NP		NP	
<GIVEN>	acc-gen						giv-active		acc-gen	

- Token: The referent of the expression is a unique object which is assumed to be part of world knowledge.

(22)

<WORDS>	The	sun	set	.	Pele	scored	his	second	goal	.
<CS>	NP				NP		NP			
<GIVEN>	acc-gen				acc-gen		new			

3.2.6 New (*new*)

The referent is new to the hearer and to the discourse.

(23)

<WORDS>	Peter	went	into	the	garden.	Another	man	appeared.
<CS>	NP			NP		NP		
<INFOSTAT>	new			new		new		

3.3 Annotation Procedure

Please follow the following steps for every referring NP or PP in the discourse:

Q1: Has the referent been mentioned in the previous discourse?

- yes: label expression as **giv**!

  If you annotate with the Extended Annotation Scheme:

  Q1.1: Was the referent referred to within the last sentence?

  yes: label expression as **giv-active**

  no: label expression as **giv-inactive**

- no: go to *Q2*!

Q2: Is the referent a physical part of the utterance situation?

- yes: label expression as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-sit**!

- no: go to *Q3*!

Q3: Is the referent accessible (1) via some kind of relation to other referents in the previous discourse, (2) from assumed world knowledge, or (3) by denoting a group consisting of accessible or given discourse referents?

- yes: go to *Q4*!
- no: label expression as **new**!

Q4: Does the referring expression denote a group consisting of accessible or given discourse referents?

- yes: label element as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-aggr**!
- no: go to *Q5*!

Q5: Is the referent inferable from a referent in the previous discourse by some relation as specified in section 2.2.5 under 'Inferable (acc-inf)'?

- yes: label element as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-inf**!
- no: go to *Q6*!

Q6: Is the referent assumed to be inferable from assumed world knowledge?

- yes: label element as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-gen**!
- no: go back to *Q1* and start all over again! You must have missed something.

4 Layer II: Topic

4.1 Introduction

In its current version, the annotation scheme for Topic consists solely of the Core Annotation Scheme.

4.1.1 Tagset Declaration

Table 3: Topic tags

Annotation Layer:	Topic	
Description:	Sentence or Clause topics	
Unit:	XP	
Core Annotation Scheme		
Tags:	**ab**	aboutness topic: > what the sentence is about
	fs	frame-setting topic > frame within which the main predication holds
Note:		Topics may be nested within a focus.

4.2 Core Annotation Scheme for Topic

Topics come in two varieties: aboutness topics and frame setting topics. The two categories are not exclusive, i. e. a sentence can have an aboutness topic as well as one or several frame setting topics.

Note that not all sentences have topics (see 4.2.1 below). In some languages topics are marked overtly (either by a morphological marker or by a designated position in the syntax), while in others, topics can be identified only indirectly, i. e. via clause-internal or contextual information.

Concerning complex sentences, choose the following strategy: check whether the whole sentence has an aboutness and/or a frame setting topic. Then check for each single finite clause contained within the complex sentence – with

the exception of restrictive relative clauses – whether it has an aboutness or a frame setting topic.

4.2.1 Topicless sentences

All-new or event sentences do not have a topic. *(The informant is shown a picture of a burning house, and is asked: What happens?)*

(24)

<WORDS>	A	house	is	on	fire	.
<TOPIC>						

4.2.2 Aboutness Topic (*ab*)

The aboutness topic is the entity about which the sentence under discussion makes a predication. In general, aboutness topics tend to be fronted crosslinguistically.

The only expressions that can denote aboutness topics are:

(i.) referential NPs (i. e. definite descriptions and proper names),

(ii.) indefinite NPs with specific and generic interpretations, and indefinites in adverbially quantified sentences that show Quantificational Variability Effects,

(iii.) bare plurals with generic interpretations, and bare plurals in adverbially quantified sentences that show Quantificational Variability Effects, and

(iv.) finite clauses denoting concrete facts about which the subsequent clause predicates (see below).

Note 1 (Specificity)

- Specificity can be tested as follows: If the respective indefinite can be preceded by "a certain …" without forcing a different interpretation, it gets interpreted as a specific indefinite.

Note 2 (Genericity)

- Genericity can be tested as follows: If a sentence containing an indefinite or a bare plural is roughly equivalent to a universal quantification over the set of individuals that satisfy the respective NP-predicate, it is a generic sentence. Examples: (25a) below is roughly equivalent to (25b) and (26a) is roughly equivalent to (26b).

(25) a. *A dog* is smart.
b. *All dogs* are smart.

(26) a. *Cats* are snooty.
b. *All cats* are snooty.

Note 3 (Quantificational Variability Effects)

- Quantificational Variability Effects can be defined as follows: An adverbially quantified sentence that contains an indefinite NP or a bare plural is roughly equivalent to a sentence where the combination Q-adverb + indefinite NP/bare plural has been replaced by a quantificational NP with corresponding quantificational force. Examples: (27a) is roughly equivalent to (27b), and (28a) is roughly equivalent to (28b).

(27) a. *A dog* is *often* smart.
b. *Many dogs* are smart.

(28) a. *Cats* are *usually* snooty.
b. *Most cats* are snooty.

Quantificational NPs other than indefinites and other kinds of XPs can *never* be aboutness topics. In general, NPs marked as given or accessible on the information status layer are often aboutness topics.

Whether an NP (with the exception of specifically interpreted indefinites) should be marked as the aboutness topic of a sentence can be tested in the following way:

Test for Aboutness Topics

An NP *X* is the aboutness topic of a sentence *S* containing *X* if

- *S* would be a natural continuation to the announcement *Let me tell you something about X*
- *S* would be a good answer to the question
  What about X?
- *S* could be naturally transformed into the sentence
  Concerning X, S′
  or into the sentence
  Concerning X, S′,
  where *S′* differs from *S* only insofar as *X* has been replaced by a suitable pronoun.

Note that in the case of generic sentences and adverbially quantified sentences that contain singular indefinites, the first occurrence of *X* in the tests above must be replaced by a corresponding bare plural.
(See the examples below.)

Whether a specific indefinite should be marked as the aboutness topic of a sentence can be tested in the following way:

Test for Aboutness Topics for Specific Indefinites

A specific indefinite *X* is the aboutness topic of a sentence *S* containing *X* if the following transformation of *S* sounds natural:

- Within *S*, replace the indefinite article in *X* by *this* or *that*
- Transform the resulting sentence *S′* into *Concerning X, S′*.

(See example 33 below.)

(29) {The informant is shown a picture of a burning house, and is asked: What about the house?}

<WORDS>	The	house	is	on	fire	.
<TOPIC>	ab					

(30) {Yesterday I met Peter and Anne in London.}

<WORDS>	Peter	was	wearing	red	socks	.
<TOPIC>	ab					

Transforming S into “Concerning Peter, he is wearing red socks” or testing the sentence in the context “Let me tell you something about Peter” sounds natural.

(31) {A dog is often smart.}

<WORDS>	A	dog	is	often	smart	.
<TOPIC>	**ab**					

Transforming S into “Concerning dogs, a dog is often smart” or preposing “Let me tell you something about dogs” sounds natural.

(32) {Cats are snooty.}

<WORDS>	Cats	are	snooty	.
<TOPIC>	**ab**			

Transforming S into “Concerning cats, cats are snooty” or preposing “Let me tell you something about cats” sounds natural.

(33) German

<WORDS>	Einen	Hund	mag	Peter	wirklich	.
<GLOSS>	A/One-ACC	dog	likes	Peter	really	
<TOPIC>	**ab**					
<TRANS>	Peter really likes one/a certain dog.					

Specificity: “A dog” can be replaced by “A certain dog”. (Aboutness-) Topicality: S can be transformed into “Concerning a certain dog, Peter really likes that dog”.

(34)

<WORDS>	That	Maria	is	still	alive	is	pleasing	.
<TOPIC>		ab						
<TOPIC>	ab							

Transforming the matrix sentence S into “Concerning the fact that Maria is still alive, S” is possible. Concerning the subordinate clause S´, the proper name “Maria” is the aboutness topic of this clause, as this clause can be transformed into the sentence “Concerning Maria, she is still alive”.

4.2.3 Frame Setting (*fs*)

Frame setting topics constitute the frame within which the main predication of the respective sentence has to be interpreted. They often specify the time or the location at which the event/state denoted by the rest of the clause takes place/holds. Temporal or locative PPs, adverbial phrases and subordinate

clauses denoting (sets of) spatial or temporal locations are therefore typical frame setting topics crosslinguistically.

Note, however, that not every such phrase is a frame setting topic: Frame-setting topics are typically fronted, and the spatial or temporal locations denoted by them are often already part of the shared background of the discourse participants, or can at least be inferred easily.

Furthermore, fronted adverbials denoting domains against which the subsequently reported fact is to be evaluated can be frame setting topics, too (Typical examples are adverbs like *physically*, *mentally* etc. in sentences like *Physically, Peter is doing fine*).

In some languages (e.g. Chinese, Vietnamese) the choice is even wider: There, for example, constituents denoting supersets of the entities of which something is predicated in the subsequent clause can also be frame setting topics (see the Chinese example below). In languages like German and English, on the other hand, the same meaning can only be expressed by employing special constructions like *Concerning X*, *S*, or *As for X*, *S* (where X is the frame setting topic, and S the subsequent clause).

Note: In contrast to aboutness topics, with frame setting topics there is never a direct predication relation between the frame setting topic and the subsequent clause.

(35) Vietnamese

<WORDS>	Đi	chợ	Mỗi	Tuần	Tôi	đi	ba	lần	.
<GLOSS>	Go	market	Every	week	1.SG	go	three	time	
<TOPIC>	fs		fs		ab				
<TRANS>	As for going to the market, every week I go three times.								

(36) Manado Malay: {They told me she was waiting for me at my home.}

<WORDS>	Kita	pe	pulang	dia	so	Pigi	.
<GLOSS>	1.SG	POSS	come home	she			
<TOPIC>	fs			ab			
<TRANS>	When I came home, she had already left. (My coming home ...)						

(37) German

<WORDS>	Gestern	abend	haben	wir	Skat	gespielt	.
<GLOSS>	Yesterday	evening	have	we	Skat	played	
<TOPIC>	fs			ab			
<TRANS>	Yesterday evening, we played Skat.						

(38) German

<WORDS>	Körperlich	geht	es	Peter	sehr	gut	.
<GLOSS>	Physically	goes	it	Peter	very	well	
<TOPIC>	fs			ab			
<TRANS>	Physically, Peter is doing very well.						

(39) Chinese

<WORDS>	Yie.sheng	Dong.wu	Wo	zui	xi.huan	Shi	zi	.
<GLOSS>	Wild	animal	I	very	like	lion	Suffix	
<TOPIC>	fs		ab					
<TRANS>	Concerning wild animals, I really like lions.							

(40)

<WORDS>	In	Berlin	haben	die	Verhandlungspartner	…
<GLOSS>	In	Berlin	have	the	negotiating partners	
<TOPIC>	fs			ab		
<TRANS>	In Berlin, the negotiating partners did not pay attention to one rule.					

<WORDS>	…	eine	Regel	nicht	beachtet	.
<GLOSS>	…	one	rule	not	paid-attention-to	
<TOPIC>	…					
<TRANS>	In Berlin, the negotiating partners did not pay attention to one rule.					

5 Layer III: Focus

5.1 Introduction

The annotation guidelines for Focus consist of a *Core Annotation Scheme* and an *Extended Annotation Scheme* which differ with respect to size and detailedness.

5.1.1 Tagset Declaration

Table 4: Focus tags

Annotation Layer:	Focus
Definition:	That part of an expression which provides the most relevant information in a particular context as opposed to the (not so relevant) rest of information making up the *background* of the utterance. Typically, focus on a subexpression indicates that it is selected from possible alternatives that are either implicit or given explicitly, whereas the background can be derived from the context of the utterance.
Unit:	Focus can extend over different domains in the utterance (like affixes, words, clause constituents, whole clause) and can be discontinuous as well. One expression can contain more than one focus.
Core Annotation Scheme	

Tags:	nf	new-information focus
	cf	contrastive focus
Extended Annotation Scheme		
Tags:	nf	new-information focus
	nf-sol	solicited new-information focus
	nf-unsol	unsolicited new-information focus
	cf	contrastive focus
	cf-repl	replacement
	cf-sel	selection
	cf-part	partiality
	cf-impl	implication
	cf-ver	truth value (verum)
Note:	…+op	All kinds of foci given above can occur as bound by focus operators like the particles *only, even, also etc.* as well as negation operators. In this case, the tags are supplied with the additional marking +*op* (cf. 4.5).

5.1.2 Some preliminaries

The Core Annotation Scheme is designed for basic annotation of focus phenomena in large amounts of language data. It aims at high inter-annotator agreement.

There are at least two ways for a part of an utterance to gain information structural relevance over the rest of the sentence:

(a) it provides new information and/or information which carries the discourse forward.

(b) it is contrasted with a semantically and/or syntactically parallel constituent in the particular discourse.

Based on this, we distinguish between the following general types of focus: new-information focus (*nf*) and contrastive focus (*cf*).

- We assume that *nf* and *cf* are not mutually exclusive but may apply within one and the same domain. For this purpose, two separate tiers for focus annotation are provided.
- Information structure plays a role not only in declaratives as answers to wh-questions but in interrogatives and imperatives as well, so that focus is also annotated there. If there is no special context indicated for a wh-question, it can be assumed that *nf* is made up by the interrogative element (cf. ex. 41 versus ex. 68).

On the basis of the Core Annotation Scheme, further sub-types of focus can be distinguished as shown in the Extended Annotation Scheme.

5.2 New-information focus (*nf*)

5.2.1 Core Annotation Scheme

New-information focus (*nf*) is that part of the utterance providing the new and missing information which serves to develop the discourse.

(41)

<WORDS>	Who	is	reading	a	book	?
<NFocus>	nf					
<CFocus>						

<WORDS>	Mary	is	reading	a	book	.
<NFocus>	nf					
<CFocus>						

5.2.2 Extended Annotation Scheme: Subcategories of new-information focus (nf)

In defining the new-information focus domain of a sentence, we propose two strategies according to the major distinction between question-answer sequences and running texts. For these two cases, we use *nf-sol* and *nf-unsol* in the Extended Annotation Scheme, respectively.

Note: If you annotate tags at this layer, be as specific as possible. Only if you are not sure about which sub-tag (either *nf-sol* or *nf-unsol*) to choose, choose the less specific tag, i.e. *nf*.

Solicited new-information focus (*nf-sol*)

The *solicited new-information focus* is that part of a sentence that carries information explicitly requested by another discourse participant.

Comment: Note that the focus domain in the answer differs according to the information already presupposed by the question. The following examples illustrate this test for various focus domains.

- all-focus sentences: answers to questions like *"What's new?"*, *"What's going on?"*

(42)

<WORDS>	What	's	that	smell	?
<NFocus>	nf				
<CFocus>					

<WORDS>	The	kitchen	is	burning	.
<NFocus>	nf-sol				
<CFocus>					

Non-biased yes-no questions (also known as polar questions) and their answers are also cases of all-focus sentences since they are expressed to identify the truth-value of the entire proposition.

(43)

<WORDS>	Is	this	book	in	German	?
<NFocus>	nf					
<CFocus>						

<WORDS>	Yes	,	it	is	.
<NFocus>	nf-sol				
<CFocus>					

(44)

<WORDS>	Is	this	book	in	German	?
<NFocus>	nf					
<CFocus>						

<WORDS>	No	,	it	is	not	.
<NFocus>	nf-sol					
<CFocus>						

- VP-focus: extended over the whole VP of the answer:

(45)

<WORDS>	What	is	Mary	doing	?
<NFocus>	nf				
<CFocus>					

<WORDS>	She	is	reading	a	book	.
<NFocus>		nf-sol				
<CFocus>						

- narrow (XP-) focus: extended over one constituent or on a part of a constituent only

(46)

<WORDS>	Who	is	reading	a	book	?
<NFocus>	nf					
<CFocus>						

<WORDS>	Mary	is	reading	a	book	.
<NFocus>	nf-sol					
<CFocus>						

(47)

<WORDS>	What	is	Mary	reading	?
<NFocus>	nf				
<CFocus>					

<WORDS>	She	is	reading	a	book	.
<NFocus>				nf-sol		
<CFocus>						

(48)

<WORDS>	What	sort	of	books	does	Mary	read	?
<NFocus>	nf							
<CFocus>								

<WORDS>	She	reads	books	on	linguistics	.
<NFocus>				nf-sol		
<CFocus>						

- discontinuous focus domain: instances of discontinuous focus domains are given when a question is so explicit that it asks for two or more non-adjacent parts of an utterance. The index shows that the parts annotated for focus belong to one and the same focus domain that is interrupted by discourse-given material. This is useful to distinguish cases of discontinuous focus domains from those of multiple foci (cf. 4.4).

(49)

<WORDS>	What	did	Paul	do	with	the	book	?
<NFocus>	nf							
<CFocus>								

<WORDS>	He	gave	it	to	Mary	.
<NFocus>		nf_1		nf_1		
<CFocus>						

Unsolicited new-information focus (*nf-unsol*)

In running texts, for example in a narrative, report etc., the domain of *unsolicited new-information focus* extends over that part of the information that carries forward the discourse. It applies, for instance, to newly added discourse referents, i.e. new individuals like persons, events, facts, states/qualities, time intervals and locations which can be refered to by pronouns in the following discourse. *Nf-unsol* further applies to new relations between given discourse referents, i.e. to all sorts of predicates: verbal and nominal predicates, quantificational determiners (*every, all, each, always, often* etc.).

In order to determine the domain of *nf-unsol,* we adopt a strategy already used for the identification of the focus domain in cases of question-answer sequences. We assume that for each sentence in a running text a preceding implicit question exists. That part of the sentence that supplies the new or missing information according to the implicit question is the information that carries the discourse further and has therefore to be annotated for *nf-unsol.*

Comment: Note that the domain of *nf-unsol* can also vary and be discontinuous as described for *nf-sol* above.

Text-initial sentences are usually all-focus sentences (also called presentational sentences which introduce new discourse referents). The entire initial sentence is annotated for focus.

With *non-initial sentences*, pay attention to the relation between given and newly established information, the latter being the domain of *nf-unsol*. In order to determine *nf-unsol*, try to formulate the *most general* question for each sentence on the basis of the given material, according to specific discourse types and the (probable) intention of the speaker to highlight that information which is able to develop the discourse.

The following is a sample annotation of *nf-unsol* in a narrative sequence:

(50) [1] Once upon a time, there was a wizard. [2] He lived in a beautiful castle. [3] All around the castle, there were green fields full of precious flowers. [4] One day, the wizard decided to leave his castle.

<WORDS>	Once	upon	a	time	there	was	a	wizard	.
<NFocus>	nf-unsol								
<CFocus>									
<FOCUS QUEST.>	no focus question possible / Who/What is the story going to be about?								

(51)

<WORDS>	He	lived	in	a	beautiful	castle	.
<NFocus>		nf-unsol					
<CFocus>							
<FOCUS QUEST.>	What about the wizard?						

In (51), questions like *"Where did he live?"* as well as *"What about his dwelling?"* are possible, too, but nevertheless they do not fit as a proper continuation of the discourse as established so far.

(52)

<WORDS>	All	around	the	castle	,	…
<NFocus>						…
<CFocus>						…
<FOCUS QUEST.>	What about the castle?					

<WORDS>	…	there	were	green	fields	full	of	precious	flowers	.
<NFocus>	…	nf-unsol								
<CFocus>	…									
<FOCUS QUEST.>	What about the castle?									

(53)

<WORDS>	One	day	,	the	wizard	decided	to	leave	his	castle	.
<NFocus>				nf-unsol							
<CFocus>											
<FOCUS QUEST.>	What happened then?										

Note that in (53), the role of the sentence in discourse structure plays a crucial role in formulating the focus question and assigning the domain of *nf-unsol*. As the sentence in (53) opens a new paragraph, its function is similar to that of the text-initial sentence in (50). Consequently, "*the wizard"* – though mentioned before – belongs to the information necessary to complete the implicit question and is therefore part of *nf-unsol.*

5.3 Contrastive Focus (*cf*)

5.3.1 Core Annotation Scheme

We understand contrastive focus (*cf*) as that element of the sentence that evokes a notion of contrast to (an element of) another utterance.

(54) from OHG Tatian 229, 28 – 230, 01 (John 11, 9-10):

oba uuer gengit In tage / ni bispurnit. […] /[…] *oba her get In naht / bispurnit.* […] (If anyone walks in the day, he does not stumble […]. But if he walks in the night, he stumbles.)

<WORDS>	oba	uuer	Gengit	In	tage
<GLOSS>	if	anyone	Walks	in	day
<NFocus>					
<CFocus>				cf	
<TRANS>	If anyone walks in the day, …				

<WORDS>	oba	her	get	In	naht
<GLOSS>	if	he	walks	in	night
<NFocus>					
<CFocus>				cf	
<TRANS>	But if he walks in the night, …				

Contrastive focus may also extend over different domains of an utterance. In alternative questions and the answers to them it covers the whole CP, cf. (55).

(55)

<WORDS>	Is	it	raining	or	not	?
<NFocus>						
<CFocus>	cf				cf	

<WORDS>	Yes	,	it	is	.
<NFocus>			nf		
<CFocus>			cf		

In other cases, it will cover only a part of a lexical constituent, for example prefixes, the auxiliary part of analytical tense forms etc., cf. (56).

(56)

<WORDS>	We	do	not	export	but	import	goods	.
<MOPRH>	We	do	not	ex- port	but	im- port	goods	.
<NFocus>								
<CFocus>				cf		cf		

In case there is more than only one contrast in a sentence, an index is used to identify the contrasted pairs, cf. (57).

(57)

<WORDS>	Mary	likes	apples	but	Bill	prefers	strawberries	.
<NFocus>								
<CFocus>	cf_1		cf_2		cf_1		cf_2	

5.3.2 Extended Annotation Scheme: Subcategories of Contrastive Focus (cf)

Contrastive subtype replacing (*cf-repl*)

This subtype of contrastive focus corrects the contextually given information by replacing parts of it for suppletive information.

(58)

<WORDS>	I	heard	that	Mary	is	growing	vegetables	now	?
<NFocus>	nf-unsol								
<CFocus>									

<WORDS>	No	,	she	is	growing	bananas	.
<NFocus>							
<CFocus>						cf-repl	

Contrastive subtype selection (*cf-sel*)

An element out of a given set of explicitly expressed alternatives is selected. The classic instance of a selective focus is found in answers to alternative questions with *or*, as in the following example.

(59)

<WORDS>	Do you want to go	to	the	red	or	to	the	blue	house	?
<NFocus>										
<CFocus>				cf				cf		

<WORDS>	I	want	to	go	to	the	red	one	.
<NFocus>									
<CFocus>							cf-sel		

Contrastive subtype partiality (*cf-part*)

The *cf* introduces a (new) part or subset of a previously mentioned entity.

(60)

<WORDS>	What	are	your	sisters	doing	?
<NFocus>	nf					
<CFocus>						

(61)

<WORDS>	My	older	sister	works	as	a	secretary	,
<NFocus>				nf-sol				
<CFocus>		cf-part_1		cf_2				

<WORDS>	but	my	younger	sister	is	still	going	to	school	.
<NFocus>					nf-sol					
<CFocus>			cf-part_1		cf_2					

Contrastive subtype implication (*cf-impl*)

An utterance with this subtype of contrastive focus implies that the requested information holds true not for the information provided explicitly in the answer but for other alternatives that are accessible in the context.

(62)

<WORDS>	Where	is	the	weather-cock	?
<NFocus>	nf				
<CFocus>					

<WORDS>	Well	,	on	the	red	roof	,	there is no weather-cock	.
<NFocus>									
<CFocus>					cf-impl				

Here, the speaker implies that the weather-cock is on a roof other than the red one. Difference to *cf-part* is difficult. Pay attention that in *cf-part* the set of alternatives is explicitly given. For example, a question like "*Where on the roofs is the weather-cock?"* allows for *cf-part* in the answer because the set of alternatives, "*the roofs"*, is explicitly given.

Contrastive subtype: truth-value (verum) (*cf-ver*)

This subtype of contrastive focus emphasizes the truth-value of the proposition. The annotation domain for truth-value focus is the whole proposition. (Note: In the literature, it is common to mark only the focus exponent [here: did].)

(63) context:

A: The exam was difficult, nevertheless lots of students passed.

B: Yes, that's true. Lots of students did pass.

<WORDS>	Lots	of	students	did	pass	.
<NFocus>						
<CFocus>	cf-ver					

Comment: There are cases in which the truth-value of the proposition is set and emphasized at the same time.

(64)

<WORDS>	Nobody believed that	,	but	Mary	did	go	to	Berlin	.
<NFocus>				nf					
<CFocus>				cf-ver					

In this case the truth-value of the proposition that *Mary went to Berlin* which is open in the context is being specified and emphasized at the same time.

5.4 Multiple foci and joint occurrence of nf and cf

Multiple foci can be found in various contexts, like in multiple questions and their answers. In some cases, *nf* and *cf* co-occur in one and the same utterance. Typically, a *cf* is embedded or nested within an *nf*.

- answer to multiple questions:

(65)

<WORDS>	Who	met	whom	?
<NFocus>	nf		nf	
<CFocus>				

<WORDS>	An	American	farmer	met	a	Canadian	farmer	.
<NFocus>	nf				nf			
< cfocus >		cf				cf		

- contrast within a sentence with a single *nf* focus domain:

(66)

<WORDS>	What	happened	?
<NFocus>	nf		
<CFocus>			

<WORDS>	An	American	farmer	met	a	Canadian	farmer	.
<NFocus>	nf							
<CFocus>		cf				cf		

- *cf* and *nf* can also completely fall together:

(67)

<WORDS>	Which	brother	helped	which	brother	?
<NFocus>	nf			nf		
<CFocus>						

<WORDS>	The	oldest	brother	helped	the	youngest	brother	.
<NFocus>		nf				nf		
<CFocus>		cf				cf		

- *cf* and *nf* can completely diverge from each other:

(68') (An adapted example from Jacobs 1991: 201f.)

The children left the remainings of their meals everywhere in the apartment. Mary is responsible for the dirt in the bedroom and John for that in the bathroom.

(68)

<WORDS>	And	who	has	eaten	in	the	living	room	?
<NFocus>		nf							
<CFocus>							cf		

5.5 Operator-bound focus (…+*op*)

All kinds of foci given above can occur as bound by focus operators like the particles *only, even, also etc.* as well as negation operators. Different focus association is also possible. In the cases given below, the focus operator *only* triggers two different foci.

(69a) (Rooth 1985)

<WORDS>	Mary	only	introduced	Bill	to	Sue	.
<CLASS>		foc-prt					
<NFocus>				nf+op			
<CFocus>							

(69b) (Rooth 1985)

<WORDS>	Mary	only	introduced	Bill	to	Sue	.
<CLASS>		foc-prt					
<NFocus>						nf+op	
<CFocus>							

5.6 Annotation Procedure

Please complete the following steps:

Q1: Is the sentence a declarative or a non-declarative one?

- if non-declarative (imperative, question): go to *Q3*
- if declarative: go to *Q2*

Q2: Does the utterance complete an explicit wh-question?

- Yes: the constituent which is congruent to the wh-word is to be annotated "*nf-sol*"
- No: go to *Q3*

Q3: Does a constituent of the utterance (or the utterance as a whole) evoke the notion of contrast to another constituent in previous context?

- Yes: annotate it for "*cf*" – for further annotation go to *Q4*
- No: go to *Q5*

Q4: Does the context enable you to further specify the contrastive relation according to the inventory given in 4.3.2?

- Yes: annotate according to the inventory given in 4.3.2.
- No: restrict the annotation to "*cf*"

Q5: Which part of the utterance reveals the new and most important information in discourse? Try to identify the domain by asking implicit questions as done in the example in 4.2.2!

- annotate the identified costituent or domain as "*nf-unsol*"

Q6: Is it possible to add to the utterance a formula like "It is true / It is not true ...", "Is it true / Is it not true ...?" to the respective proposition without changing its meaning/function within the discourse?

- Yes: annotate it as "*cf-ver*" according to 4.3.2.5
- No: no additional specification is necessary

Q7: Does the sentence contain a focus operator?

- Yes: annotate the constituent that is bound by it for "+*op*"
- No: no additional specification is necessary

6 Recommended Annotation Procedure

(1) Preparation of the Data

Make sure that the data is prepared for the annotation with information structure. In particular, check for the annotation of sentences and NPs and PPs according to the Syntax Annotation Guidelines.

If the data is not annotated accordingly, do this annotation first!

(2) Annotation step 1: Information Status and Topic

Start from the beginning of the discourse.

For every sentence:

(a) Check for the referentiality of each NP and PP in the sentence (cf. Section 2.2.1).

(b) Specify the Information Status of every referring NP- and PP-marked constituent. Follow the instructions in 2.3.!

(c) Test for the Topic status of each NP and PP in the sentence. Follow the guidelines in Section 3!

(3) Annotation step 2: Focus

Start from the beginning of the discourse. For every sentence:

- Apply the annotation procedure for the Focus Annotation Scheme in Section 4.6.

(4) Check for Completeness

Check for the completeness of the Annotation:

(a) Check for the complete annotation of Information Status for all referring NPs and PPs.

(b) Check for the complete annotation of new-information focus: for each sentence a new-information focus should be assigned.

(5) Finishing the Annotation

Don't forget to save the annotation!

7 References

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. 2004. An Annotation Scheme for Information Structure in Dialogue. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal, May.

Jacobs, Joachim. 1991. *Informationsstruktur und Grammatik*. Opladen: Westdeutscher Verlag.

Rooth, Mats. 1985. *Association with Focus*. Ph.D. dissertation, University of Massachusetts.