Information Structure Annotation and Secondary Accents

Arndt Riester[1] *& Stefan Baumann*[2]
[1]IMS, University of Stuttgart
[2]IfL Phonetics, University of Cologne

Abstract
We present a proposal for an annotation system for information structure that combines contemporary corpus-oriented accounts of information status with insights from the recent theoretical debate (e.g. Selkirk, 2007; Beaver & Velleman, subm.) on the basic pragmatic sources which lead to primary and secondary accentuation; in particular, the combination of the given-new distinction with the classical triggers for F-marking by Rooth (1992). We comment on the yet undecided question whether one or several kinds of focus should be considered in the annotation task. A key property of our scheme is its distinction between a lexical and a referential level. This allows us to describe fine-grained properties of texts, e.g. the information structurally and prosodically relevant observation that a given discourse referent may be taken up by means of lexically new material. The annotation system is demonstrated for examples from transcripts of spoken corpora as well as sentences taken from the theoretically oriented literature. We report on the inter-annotator agreement reached, and show how the system can be used in the investigation of subtle prosodic phenomena like secondary accents, which have been claimed to mark second occurrence focus.

1 Contrastive focus vs. information focus

A longstanding issue in information structure theory is the differentiation between so-called *contrastive focus* and *information focus* (focus related to the novelty of a constituent). Both types of focus are commonly marked by primary pitch accents, i.e. by strong prosodic prominence. While the distinction is usually demonstrated on the basis of intuitive minimal pairs like (1), from Selkirk (2007), its fundamental semantic distinction has remained controversial.

(1) a. I gave one to SArah$_{CF}$, not to CAITlin$_{CF}$.
b. I gave one to SArah$_{IF}$.

What examples like (1) seem to suggest is that *contrastive focus* requires the overt availability of a pair of alternatives. One problem of contemporary focus literature is that, usually, cases like (1a) are grouped together with examples involving focus-sensitive particles like (2a) or question-answer sequences like (2b), following the paradigm of Rooth (1992).

(2) a. Semanticists only talk about ONly$_F$.
b. What did the semanticist talk about this time? She talked about ANSwers$_F$.

No overt alternatives are involved in these examples, which has led researchers to quite different reactions. An obvious move is to abandon the notion of *contrastive focus* altogether, and to try and find a uniform explanation for focusing in general. This is the path that is pursued by the "lumpers" camp[1], e.g. Büring (to appear); Rooth

[1]Quote: Beaver and Velleman (subm., Sect. 1)

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 111-127, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

(2010). By contrast, "splitters" like Selkirk (2007) and Beaver and Velleman (subm.) explicitly claim that it is necessary to distinguish between two *sources* that determine the assignment of *prominence* – while avoiding the question whether there are one or several *types* of focus. The two sources are, on the one hand, *novelty* (marked in logical form by means of an N feature or, indirectly, by lack of a *G(iven)* feature) and, on the other hand, a collection of factors which are varyingly pooled under the notions of either "contrastive focus" (Götze et al., 2007), "focus" (Rooth, 1992; Selkirk, 2007) or "importance" (Beaver and Velleman, subm.). They are usually assumed to carry an F feature in logical form and to share the common property of evoking a set of (implicit or explicit) *alternatives*.

Since the terminological situation is obviously complicated we try to be careful in using notions like "focus". We acknowledge the need for a two-factor account of identifying the information structural setup of sentences and discourses. However, we think that we would go too far if we reserved the *focus* notion for expressions whose prominence is due to an alternative-related property such as explicit contrast, the presence of exhaustive particles or a wh-question. If consequently applied, this would lead to the conclusion that some standard examples of focus such as (3a,b) no longer ARE focus examples since the only obvious reason for the accents at hand is the novelty of their host phrases.

(3) a. Mary went into a store. She [bought a book about BATS].
b. Let me tell you a secret about Sally and John. Sally is [in LOVE] with John.

This, however, is the situation that we permanently encounter in transcripts from monologues and many other kinds of texts. A differentiation between answers to overt questions and ordinary, unsolicited, information-conveying sentences is artificial also for the reason that there are theories explaining the structuring of discourse by use of (often implicit) *questions under discussion*, e.g. Roberts (1996).

The position we are mildly favouring is, therefore, to not exclude the use of the term *focus* in cases like (3a,b), which only involve given and new information, and to use the term *elicited alternatives* for direct sources of focusing (F-marking). The more fine-grained a classification system is, the less does the question matter whether the classes are reducible to one or two types of focus. Instead, we would like to find out whether these fine-grained differences that we can detect at the pragmatic level are related to subtle differences at the prosodic level. We are not only interested in the location and realisation of the nuclear accent but also in pre- and postnuclear secondary accents and other prosodic phenomena.

2 Annotation of information structure

In this paper, we make a proposal for the annotation of *information structure*. We use this rather underspecified term to avoid the intricacies, discussed in the previous

section, surrounding the notion of *focus*. Moreover, we take notice of the persisting terminological confusion in the field, an unfortunate matter which we do not expect to be overcome soon. However, we feel the need to clarify which aspects of information structure we are interested in. Our main interest lies in the focus-background distinction, which we are going to analyze in more detail than usually seen in contemporary accounts. Our basic coordinates are the formal accounts of focus as provided, on the one hand, by Rooth (1992), and on the other hand, by Schwarzschild (1999), which we take to describe complementary, yet compatible, aspects of the focus notion.[2]

According to other accounts, *topic* is taken to be the complement of focus (Hajičová et al., 1998). However, we follow e.g. Krifka (2007) in distinguishing the topic-comment dimension from the focus-background dimension. We will not be discussing the former. Neither shall we consider a theme-rheme distinction.

2.1 Previous approaches to annotating information structure

Annotations of focus and related information structural features are often said to be "difficult" as compared to, for instance, morphosyntactic annotations. This is likely due to the fact that informal definitions of focus are often remarkably vague whereas insights from the formal-semantic literature are not easily transfered to corpus data.

For instance, in her study of focus and topic in a corpus of spoken Danish, Paggio (2006) defines focus, quoting Lambrecht (1994), as "non-presupposed information". This, in combination with a number of heuristics and general principles (such as "all sentences have a focus") is used as a guideline for the annotators. Paggio reports a kappa score between 0.7 and 0.8 on controlled monologue and dialogue data such as descriptions and map tasks. In her setting, however, annotators made use of prosodic information, which makes the annotation task simpler but also semantically intransparent.

In the *LISA* (Linguistic Information Structure Annotation) guidelines (Götze et al., 2007), information structure is annotated on three layers: information status (*given / accessible / new*, restricted to referring expressions), topic and focus. Focus is defined as "[t]hat part of an expression which provides the most relevant information in a particular context" (p.170). "New-information focus" is distinguished from "contrastive focus". *New information* may come as *solicited* (in response to a question) or *unsolicited*.[3] The guidelines additionally contain a useful list of triggering constellations for contrastive focus (see Sect. 3.3 below). As for the focus layer, Ritz et al. (2008) report an inter-annotator agreement of 0.41 to 0.62 for different types of texts based on predefined markables (but no prosodic information). Telling from these rather low scores, annotating focus has not yet reached a satisfactory level.

[2] It has been noted in Beaver and Clark (2008, Sect. 2.4), though, that the usage of F-features is not the same on the two accounts.

[3] Note that this stands in contrast to e.g. Rooth (1992); Selkirk (2007), who would count solicited, but not unsolicited, information as a trigger for F-marking akin to contrastive focus.

We specifically want to point out what we see as an unfortunate decision in the LISA guidelines: the choice to separate the annotations of, on the one hand, information status, and, on the other hand, new information focus. Both describe the given-new distinction, thus, the same kind of information. They differ in that information status allows for a more differentiated classification but is limited to referential expressions. It is our explicit goal to overcome this separation and, thereby, generalise the notion of information status to all expression types.

2.2 A new labeling system for information status: the RefLex scheme

In the following, we will briefly introduce a labeling system for information status, which distinguishes between a referential level and a lexical level (and which we therefore call the *RefLex* scheme). We will clarify why it is desirable to use such a fine-grained system rather than just distinguishing between "given" and "new" constituents. Note that we are not claiming that the annotation labels presented below represent syntactic features of some kind, in the way as, for instance, Selkirk (2007) treats her F and G markings. We will make no predictions as regards the precise functioning of the syntax-phonology interface. Nevertheless, a crucial point of the whole procedure is the assumption that the information status labels have an impact on prosody.

The category descriptions below are kept very short, since we have introduced them in great detail elsewhere (Baumann and Riester, subm.). By use of the following choice of R-categories it is possible to classify *referential* determiner phrases and prepositional phrases occurring in natural discourse; by use of the L-categories we can classify the information status of content words and non-referential phrases.

2.2.1 R-GIVEN and L-GIVEN

Givenness, loosely following Schwarzschild (1999, 151), can be interpreted as either synonymy / hyponymy of lexemes or as identity between referring expressions. Likewise, Halliday and colleagues[4] distinguish between *lexical cohesion* and various referential relations. We call the two notions *L-givenness* and *R-givenness*, respectively. Interesting constellations can be observed if the two notions are simultaneously applied, as shown below.

R-labels apply at the DP or PP level. For instance, in examples (4), (5) and (7) we find various kinds of coreferential expressions. Lexical givenness, on the other hand, applies in (5) and (7) on the repeated words, and in (6) on the hypernym "guy".

(4)

A colleague came in.	The	idiot	dropped a vase.
	R-GIVEN		

(5)

A student came in.	Another	student	greeted	him.
		L-GIVEN		
				R-GIVEN

[4] e.g. Halliday and Hasan (1976, 288); Halliday and Matthiessen (2004)

(6)

A policeman came in.	Another	guy	left.
		L-GIVEN	

(7)

A man came in.	The	man	coughed.
		L-GIVEN	
	R-GIVEN		

The most important take-home message is that neither is referential givenness a prerequisite for lexical givenness, as shown in (4), nor the other way round, see (5) and (6), although the two sometimes combine, as in (7).

2.2.2 R-NEW, L-NEW, R-UNUSED

Novelty is, on most treatments of information structure and discussions of the *given/new* distinction, understood as "novelty in the discourse". Remarkably, however, Prince (1992) additionally distinguishes between *discourse novelty* and *hearer novelty*, the latter representing a stronger notion since unmentioned (i.e. discourse-new) entities may nevertheless be familiar to the addressee (i.e. hearer-old). In her earlier paper, Prince (1981) uses the labels *unused* (discourse-new, hearer-old) and *brand-new* (discourse-new, hearer-new) for the same opposition. The labels R-NEW and R-UNUSED that are employed on our account are defined in a slightly different way: both describe discourse-new referential expressions but, while R-NEW is reserved for indefinites, R-UNUSED stands for uniquely identifiable, definite, but not necessarily *known*, entities used on the first occasion in a text. This decision, on the one hand, does justice to the long-standing semantic tradition to keep indefinites and definites (for instance, proper names) apart, and, on the other hand, accounts for the difficulty to decide with certainty whether, for instance, a *named entity* is hearer-known or not[5].

Independently of what has just been said, it is furthermore possible to separately describe the discourse novelty of *lexemes* (L-NEW) and of the *discourse referents* (R-NEW, R-UNUSED) which they introduce. Examples of the three categories in combination are given in (8) to (10).

(8)

A	man	came in.	Another	man	left.
	L-NEW			L-GIVEN	
R-NEW			R-NEW		

(9)

George	came in.	Mary	likes	George.
L-NEW		L-NEW		L-GIVEN
R-UNUSED		R-UNUSED		R-GIVEN

(10)

The	man	who stole	my	wallet	is very tall.
	L-NEW			L-NEW	
			R-UNUSED		
R-UNUSED					

[5] Although the respective subclassifications can be made with a reasonable degree of agreement, cf. Riester et al. (2010).

2.2.3 R-BRIDGING, L-ACCESSIBLE

Prince (1981) and also Chafe (1994) have pointed out that it is desirable to not only distinguish between *given* and *new* information but to take into account at least a third, intermediate, class: expressions which have not been mentioned explicitly but are *inferrable* from material in the discourse. Chafe (1994) uses the term *accessible* for such information but he does not distinguish between different levels, as we would like to do. As far as referents are concerned, a closely related phenomenon has been discussed under the notion of *bridging* (Clark, 1977; Asher and Lascarides, 1998), shown in example (11).

(11)

Bill	discovered	a romantic	house.	The	door	was open.
L-NEW			L-NEW		L-ACCESSIBLE	
R-UNUSED		R-NEW		R-BRIDGING		

The label L-ACCESSIBLE is defined for words which are hyponyms or meronyms (part expressions) of other words in the recent discourse context (i.e. not further away than 5 clauses). The label R-BRIDGING, on the other hand, is defined quite differently as a definite expression whose licensing depends on a previously introduced scenario or frame. So, while in (11), "door" and "house" stand in a part-whole relation ("door" is lexically accessible), no such relation exists between "murdered" and "harpoon" in (12). Since the harpoon is an unusual murder instrument, it is labeled L-NEW. Nevertheless, we would still like to say that this is a case of bridging, since the second sentence could not be uttered felicitously at the beginning of a discourse.

(12)

John	was murdered yesterday.	The	harpoon	was lying nearby.
L-NEW			L-NEW	
R-UNUSED		R-BRIDGING		

Other than R-UNUSED expressions, the interpretation of items labeled R-BRIDGING is context-dependent. In contrast to the label R-GIVEN, R-BRIDGING implies non-coreference. Indefinites never receive the label R-BRIDGING in the present system. In (13), lexical accessibility combines with referential novelty.

(13)

John	lives	in	Italy	and is married	to a	Neapolitan.
L-NEW			L-NEW			L-ACCESSIBLE
R-UNUSED		R-UNUSED			R-NEW	

Note that identifying R-BRIDGING as a separate referential class of information status in between given and new expressions does not only derive from purely theoretical considerations but can be shown to have a significant influence on the realisation of nuclear accents; an issue which is highly relevant for information structure theory. Röhr and Baumann (2010) demonstrate in experiments that *inferrable* information is significantly more often produced with low or falling accents as compared to new information, which is predominantly realised with (perceptually more prominent) high accents. An

example is shown in (14), whose prosodic realisation (with an early peak accent that is low on the accented syllable) can be seen in Figure 1.[6]

(14) Thomas darf heute im Zoo seinen Lieblingsaffen füttern. [...] Er steckt sich [R-BRIDGING die [L-NEW Banane]] ein.
'Today, Thomas has got the permission to feed his favourite monkey at the zoo. He pockets the banana.'

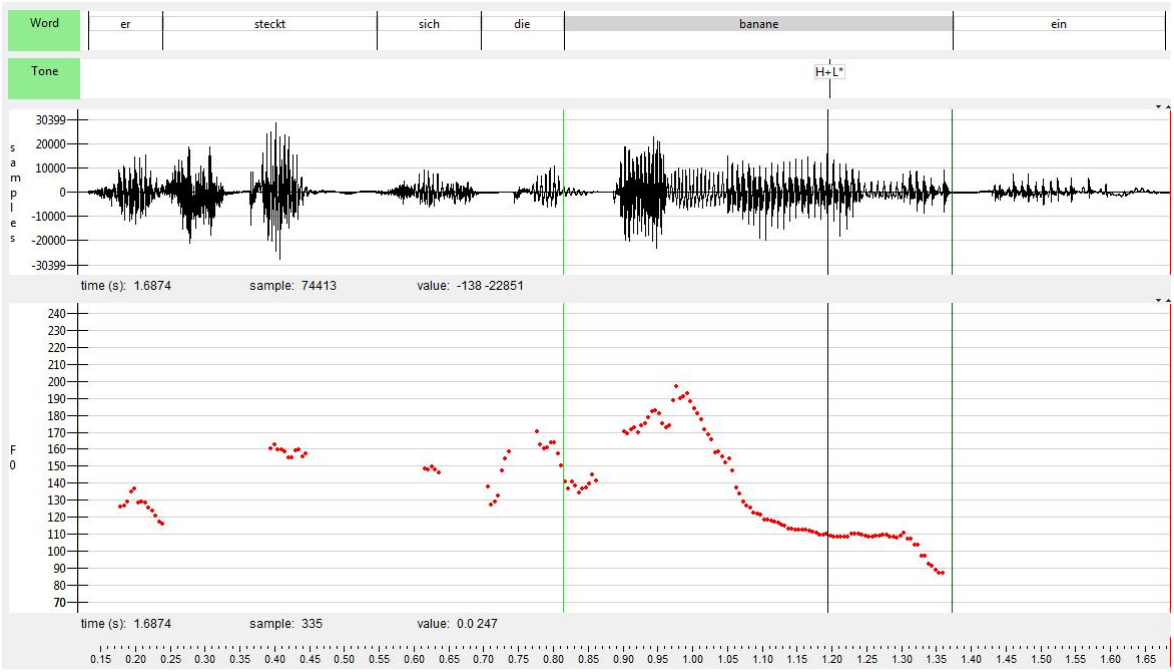

Figure 1: Possible realisation of an L-NEW, R-BRIDGING expression ("die Banane")

2.2.4 R-GENERIC

Definite or indefinite expressions which refer to a kind, see (15) and (16), receive the label R-GENERIC.

(15)

The	fox	is	a	predator.
	L-NEW			L-NEW
R-GENERIC			R-GENERIC	

(16)

Mary	only likes	vegetables.
L-NEW		L-NEW
R-UNUSED		R-GENERIC

The examples of R and L labels presented in the sections above only show a small number of combinations that are possible in the annotation system, which allows for very detailed information structural investigations of discourses. For a comprehensive list of possible combinations consult Baumann and Riester (subm.). In the following section we will turn to a number of practical issues which arise when we apply the

[6]Screen shot of the target sentence using the speech analysis tool EMU (Cassidy and Harrington, 2001), displaying labeling tiers for words and intonation (accents annotated according to GToBI, following Grice et al. (2005), as well as the oscillogram and pitch contour)

annotation scheme to corpus data. Finally, we present some proposals for using the system in the task of identifying and describing regions of texts which are particularly interesting as far as prosody is concerned. Some of these have received wide attention in the semantic literature, such as so-called *second occurrence focus*.

3 Annotating corpus data

3.1 Annotation of syntactic phrases

In previous literature on information status (Prince, 1981, 1992; Nissim et al., 2004; Götze et al., 2007; Riester et al., 2010) usually only referential expressions (syntactically: DPs, PPs) are considered as the units for annotation. However, ever since in the development of information *structure* theory, givenness and novelty have been defined for all syntactic categories.

It is our claim that, in defining information status at the L-level, we are providing the foundations for a comprehensive information structural analysis of sentences and discourses. Of course, the question what counts as a unit for annotation is influenced by the choice of syntactic theory underlying the analysis. For the time being, we shall be classifying projections of *content words* like verbs, nouns, adjectives and adverbs, i.e. non-referential phrases of category VP, NP,[7] AP, AdvP and S. A basic overview of what counts as an R-level or an L-level unit is shown in Fig. 2.

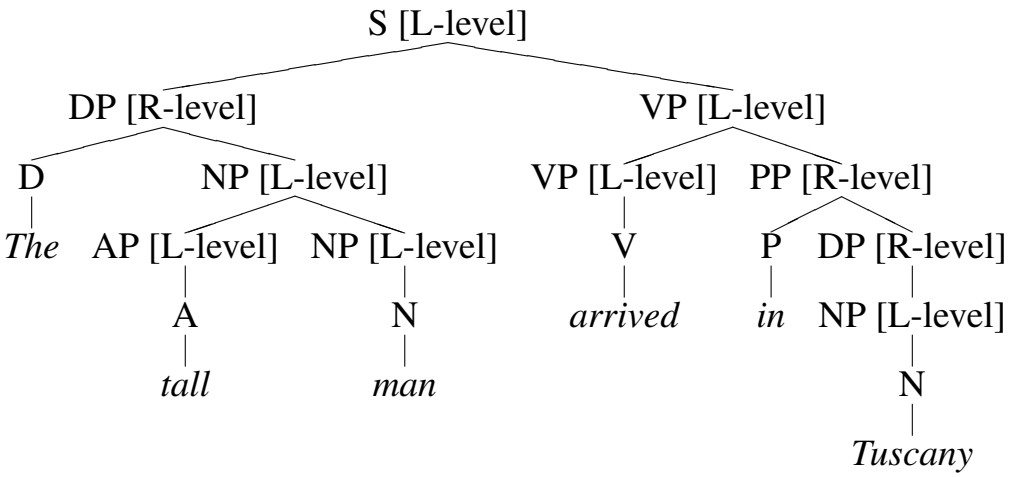

Figure 2: Basic target units for *RefLex* annotations

We would like to point out that what we are proposing amounts to a practical explication – and further development – of the approach taken by Schwarzschild (1999, 151), who distinguishes between categories of type e (R-level) and of type $\langle \alpha, \beta \rangle$ (L-level). Our definition of the L-level, however, is much simpler than Schwarzschild's since we completely relinquish his notion of *Existential F-closure*. However, we make use of

[7] We are assuming the DP hypothesis. Accordingly, we take NPs to denote properties, i.e. sets of individuals, whereas DPs denote (or refer to) a single individual or group entity.

his idea to generalise lexical relations to a notion of entailment.[8] In corpus annotation practice, the linguistic scheme shown in Fig. 2 will have to be adapted to various constraining factors, such as the properties of the chosen parser with its specific syntactic tagset, as well as features of the annotation tool. Fig. 3 shows the annotation of a German sentence, which was parsed using XLE and the German LFG grammar by Rohrer and Forst (2006), and converted to be used with the SALTO tool (Burchardt et al., 2006), which produces output in TIGER/SALSA-XML. In the rest of the paper, we shall abstract over such individual choices, since it is our goal to provide the general annotation procedure and not one that is tied to a specific annotation tool, format or syntactic theory.

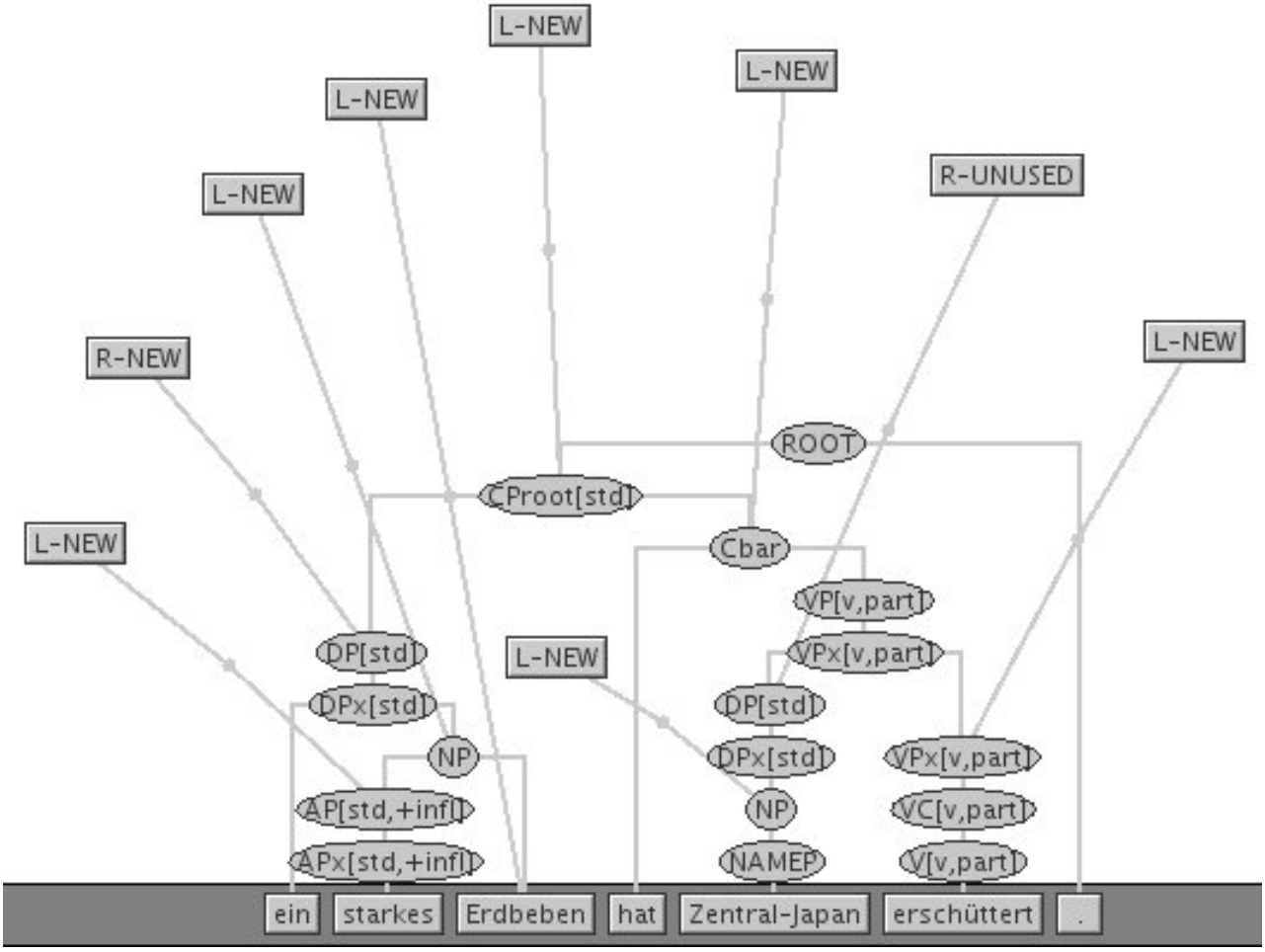

Figure 3: Sentence annotated in SALTO: *A strong earthquake has hit central Japan.*

3.2 Example annotation of a radio news feature

In the following, we will briefly show how the extended annotations can be applied to an example from a German radio news bulletin before turning to the reanalysis of some theoretically more advanced examples. The news example is (17), which will be

[8] According to this approach, the previous mention of "chihuahua" **entails** the successively mentioned hypernym "dog", as well as a successive mention of "small dog", although we wouldn't normally want to say that the latter phrase is a "hypernym", cf. Baumann and Riester (subm., Sect. 3).

annotated as in (18-20). We use a simplified table notation and additionally provide the GToBI labels (Grice et al., 2005) for the corresponding speech data.[9] Note, however, that in our envisaged annotation process, the labelers will have no access to prosodic information, since it is the correspondence between prosody and information structure which we are ultimately intending to investigate.

(17) a. Ein starkes Erdbeben hat Zentral-Japan erschüttert.
'An strong earthquake has hit central Japan.'

b. Die Behörden gaben eine Tsunami-Warnung für den Südwesten heraus.
'The authorities have issued a tsunami warning for the Southwest.'

c. Auch im Inselstaat Vanuatu im Südpazifik wurden zwei Beben registriert.
'Also in the island state of Vanuatu in the Southern Pacific two earthquakes have been registered.'

(18)

	H*	L+H*		H*	H*	H+!H* L-%
Ein	starkes	Erdbeben	hat	Zentral-	Japan	erschüttert.
	L-NEW	L-NEW		L-NEW		L-NEW
	L-NEW			R-UNUSED		
R-NEW			L-NEW			
L-NEW						

(19)[10]

	H*			L+H*		L+H*	L-%
Die	Behörden	gaben	eine	Tsunami-Warnung	für den	Südwesten	heraus.
	L-NEW			L-NEW		L-NEW	L-NEW
R-BRIDGING					R-BRIDGING		
			R-NEW				
	L-NEW						
L-NEW							

(20)

H*		H*	L+H*		L+H*		L*	L*+H	H+L* L-%
Auch	im	Inselstaat	Vanuatu	im	Südpazifik	wurden	zwei	Beben	registriert.
		L-NEW	L-NEW		L-NEW			L-GIVEN	L-NEW
				R-UNUSED			R-NEW		
	R-UNUSED								
						L-ACCESSIBLE			
L-ACCESSIBLE									

Next to the general assignment of L-labels to verbal and adjectival phrases and clauses, there are a few important observations which relate to complex phrases like [$_{\text{R-NEW}}$ eine Tsunami-Warnung [$_{\text{R-BRIDGING}}$ für den Südwesten]] in (19), or [$_{\text{R-UNUSED}}$ im Inselstaat Vanuatu [$_{\text{R-UNUSED}}$ im Südpazifik]] in (20). In each, one referential phrase has another one embedded in it. Since the two possess different referents, two R-labels are nested inside each other.

3.3 Elicited alternatives

In Sect. 1, we already discussed the need to consider two main sources that may have an influence on the prosodic realisation of a sentence: besides *information status* we

[9] One of the anonymous reviewers requested that we include the respective prosodic information. Unfortunately, at the current stage, it is impossible to provide a satisfactory discussion of the discourse-prosody interface of this example, especially since prosodic correlates of information structure usually require a broad-scale statistical analysis.

[10] The particle verb "gaben…heraus" ("issued") is annotated on "heraus".

have to identify features that are linked to Alternative Semantics. Götze et al. (2007, 178ff.) provide a number of such features under the heading of *contrastive focus*. Since we think that neither "focus" nor "contrast" are ideal labels for this class of features, for reasons discussed above, we will simply use the label ALT. A minimal list of important triggers of alternatives is shown in Table 1.

Sublabel of ALT	Description
FS	Item is associated with a focus-sensitive particle.
OC	Item is an element of a pair or list of overtly contrastive expressions (sentence-internally or across sentences); this subsumes e.g. *corrections* and *coordinated* expressions.
SE	Item selects one element from a pair or list of *previously* introduced alternatives.
VF	Verum focus

Table 1: Configurations which elicit alternatives

We think that Table 1 summarizes the relevant alternative-eliciting features. Note that, for instance, the prosodic prominence of an answer to an overt question is already adequately described by means of novelty at the R- or L-level, or the feature ALT-SE.[11]

When we apply this additional set of features to example (20), we obtain the following additional tier of *elicited alternatives* shown in (21).

(21)

<table>
<tr><td>H*</td><td></td><td>H*</td><td>L+H*</td><td></td><td>L+H*</td><td></td><td>L*</td><td>L*+H</td><td>H+L* L-%</td></tr>
<tr><td>Auch</td><td>im</td><td>Inselstaat</td><td>Vanuatu</td><td>im</td><td>Südpazifik</td><td>wurden</td><td>zwei</td><td>Beben</td><td>registriert.</td></tr>
<tr><td rowspan="4"></td><td></td><td>L-NEW</td><td>L-NEW</td><td></td><td>L-NEW</td><td>L-NEW</td><td></td><td>L-GIVEN</td><td>L-NEW</td></tr>
<tr><td colspan="2"></td><td colspan="2">R-UNUSED</td><td rowspan="2"></td><td rowspan="2"></td><td colspan="2">R-NEW</td><td rowspan="2"></td></tr>
<tr><td colspan="4">R-UNUSED</td><td colspan="2"></td></tr>
<tr><td colspan="7">L-ACCESSIBLE</td></tr>
<tr><td colspan="10">L-NEW</td></tr>
<tr><td></td><td colspan="5">ALT-FS / -OC</td><td></td><td></td><td></td><td></td></tr>
</table>

We observe that the phrase "im Inselstaat Vanuatu im Südpazifik" is associated with the additive particle "auch". It furthermore contrasts with "Zentral-Japan".

4 Inter-annotator agreement

We are now briefly going to discuss the inter-annotator agreement that we achieved for the proposed scheme, in particular for the two levels of information status. In a small experiment the two authors of this article independently annotated a text consisting of a transcript from spontaneous speech, comprising 65 sentences. Beforehand, we agreed on the set of markables to be annotated. In total, R-annotations were assigned to

[11] Likewise, we think that we do not need the feature like *implication* (Götze et al., 2007, 181), which again can be captured with our label L-NEW.

133 markables, L-annotations were assigned to 275 markables, following the schemes summarized in Table 2.[12]

R-Level		L-Level	
Units: DP, PP, *that*-CP		Units: AP, AdvP, NP, VP, S	
Label	Description	Label	Description
R-GIVEN	corefential anaphor	L-GIVEN	word identity / synonym / hypernym / holonym / superset
R-BRIDGING	non-coreferential context-dependent expression	L-ACCESSIBLE	hyponym / meronym / subset / otherwise related
R-UNUSED	definite discourse-new expression	L-NEW	unrelated expression (within last five clauses)
R-NEW	specific indefinite		
R-GENERIC	generic definite or indefinite		
OTHER	e.g. cataphors		

Table 2: Overview basic *RefLex* scheme

We achieve a κ score (Cohen, 1960) of 0.70 for the R-level and 0.78 for the L-level. We were not able to provide results for the annotation of elicited alternatives since the text chosen contained only 9 markables for ALT-labels.

5 Second occurrence focus and secondary accents

In the remaining part of this article we turn to an issue which has received much attention in both the theoretical and experimental literature: second occurrence focus, see example (22) from Partee (1999). We discuss this phenomenon in order to show how our annotation scheme for information structure ultimately might pave the way to a corpus analysis of second occurrence focus and other phenomena involving secondary (i.e. weaker) accents.

(22) A: Everyone already knew that Mary only eats VEgetables$_F$.
B: If even PAUL knew that Mary only eats VEgetables$_{SOF}$ then he should have suggested a different REStaurant.

Describing the precise conditions which license second occurrence focus (SOF) is not straightforward. Selkirk (2007) characterizes a *SOF* as a given constituent (since it has been mentioned before) which is at the same time focused (in (22) due to association

[12] See Baumann and Riester (subm., Sect. 4) for an extended scheme.

with "only") and whose antecedent is also focused. Beaver and Velleman (subm.) avoid reference to focusing by saying that a *SOF* must be "important" (F-marked, see above) as well as "predictable" (roughly: part of a larger constituent which is also given). Following our proposed annotation scheme, example (22) will receive the analysis given in (23-24).

(23)

<table>
<tr><td>Everyone</td><td>already</td><td>knew</td><td>that</td><td>Mary</td><td>only</td><td>eats</td><td>vegetables.</td></tr>
<tr><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4">L-NEW</td><td rowspan="4"></td><td>L-NEW</td><td rowspan="3"></td><td rowspan="2">L-NEW</td><td>L-NEW</td></tr>
<tr><td>R-UNUSED</td><td>R-GENERIC</td></tr>
<tr><td colspan="2">L-NEW</td></tr>
<tr><td colspan="5">L-NEW</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>ALT-FS</td></tr>
</table>

(24)

<table>
<tr><td>If</td><td>even</td><td>Paul</td><td>knew</td><td>that</td><td>Mary</td><td>only</td><td>eats</td><td>vegetables</td><td>then ...</td></tr>
<tr><td rowspan="5"></td><td rowspan="5"></td><td>L-NEW</td><td rowspan="5">L-GIVEN</td><td rowspan="5"></td><td>L-GIVEN</td><td rowspan="4"></td><td rowspan="2">L-GIVEN</td><td>L-GIVEN</td><td rowspan="5"></td></tr>
<tr><td rowspan="4">R-UNUSED</td><td>R-GIVEN</td><td>R-GENERIC</td></tr>
<tr><td colspan="2">L-GIVEN</td></tr>
<tr><td colspan="3">L-GIVEN</td></tr>
<tr><td colspan="4">R-GIVEN</td></tr>
<tr><td></td><td></td><td>ALT-FS</td><td></td><td></td><td></td><td></td><td></td><td>ALT-FS</td><td></td></tr>
</table>

Beaver et al. (2007) showed that the word "vegetables" in (24) is realised with a secondary accent which is not marked by pitch movement but rather by means of increased duration of the focused word in comparison with a deaccented version.

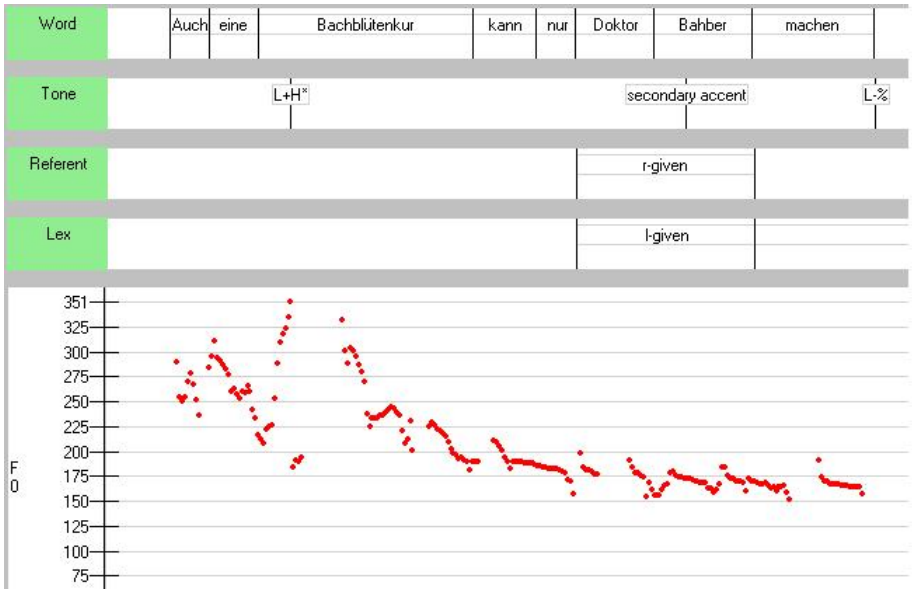

Figure 4: Realisation of second occurrence focus (R-GIVEN, ALT-FS) in German ("Dr. Bahber")

Similar results were found for German by Ishihara and Féry (2006) as well as Baumann et al. (2010). Figure 4 shows an example of second occurrence focus in German taken from the discourse in (25). The nuclear accent in (25b) is clearly placed on "Bachblütenkur", whereas "Bahber" only receives a secondary prominence.

(25) a. Eine Akupunktur kann nur Dr. Bahber machen.
'An acupuncture can only be done by Dr Bahber.'
b. Auch eine Bachblütenkur kann nur Dr. Bahber machen.'
'Also a cure with Bach flowers can only be done by Dr Bahber.' (Baumann et al., 2010, 63)

While second occurrence focus has received much attention in the literature on information structure, it is not easy to find good corresponding examples in corpus data. Nevertheless, secondary accents occur quite frequently, and it is instructive to investigate what other instances of secondary prominence have in common with examples like (24) or (25b). A good candidate is the phrase "mein afrikanischer Freund" *(my African friend)* in (26), found in our corpus of spontaneous monologues (see also Figure 5).

(26) [. . .] der junge Mann [. . .] Das Visum musste leider abgelehnt werden, weil Herr Nwahiri – so heißt [R-GIVEN mein [L-NEW afrikanischer [L-NEW Freund]]] – . . .
'[. . .] the young man [. . .] The visa unfortunately had to be dismissed because Mr. Nwahiri – that's the name of my African friend – . . . '

Figure 5: Realisation of an epithet (R-GIVEN, L-NEW) – "mein afrikanischer Freund"

This expression is an example of what is called an *epithet* (Clark, 1977; Schlenker, 2005). Such expressions can usually be characterized as coreferential expressions (R-GIVEN) which consist of new lexical material (L-NEW). (In this case "my African friend" refers back to "the young man"). They are not identical with cases of second occurrence focus (which, as we said, are defined as combinations of given or predictable and contrastive features) but exhibit a similar combination of boosting and inhibiting factors. Epithets typically cannot be produced with a nuclear accent because this would block the interpretation that the intended referent has been mentioned before, but they may receive a secondary prominence (cf. Figure 5).

Finally, we tentatively assume that the realisation of the secondary accent in example 24, can be described by assuming a joint effect of the ALT-FS feature and the referential givenness of the *fact* to which the *that*-clause refers. But this surely is worth of closer examination.

6 Summary

We have presented an annotation system for information structure which combines the advantages of a detailed classification of information status with the categorial freedom necessary to determine the givenness, accessibility or novelty of all parts of a clause and, therefore, focus-background information.

An important improvement is the differentiation between lexical relations like synonymy and hyponymy which hold between lexemes or set-denoting categories, and anaphora-related notions such as coreference or bridging which target referential expressions. Rather than saying that an expression is "given" or "new" we are now able to express that, for instance, a given individual is referred to by means of new lexical material. We also support the use of a further information structural level, which we call *elicitated alternatives* and which captures contrastive and other alternative-related properties of focus that do not belong to the domain of information status.

We have applied the annotation system to experimental and corpus data, as well as to theoretical examples that are taken from the literature on second occurrence focus. We also have sketched in what manner the detailed annotations which our system allows can be used for investigating phenomena which are prosodically marked by secondary accents.

In general, the labeling scheme serves to facilitate empirical investigations of subtle information structural and prosodic phenomena whose details by and large evade people's introspective abilities.

References

Nicholas Asher and Alex Lascarides. Bridging. *Journal of Semantics*, 15:83–113, 1998.

Stefan Baumann and Arndt Riester. Lexical and Referential Givenness: Semantic, Prosodic and Cognitive Aspects. In G. Elordieta and P. Prieto, editors, *Prosody and Meaning*, Trends in Linguistics. Mouton de Gruyter, Berlin, subm.

Stefan Baumann, Doris Mücke, and Johannes Becker. Expression of Second Occurrence Focus in German. *Linguistische Berichte*, (221):61–78, 2010.

David Beaver and Brady Clark. *Sense and Sensitivity. How Focus Determines Meaning*. Wiley & Sons, Chichester, 2008.

David Beaver and Dan Velleman. The Communicative Significance of Primary and Secondary Accents. submitted to *Lingua*, subm.

David Beaver et al. When Semantics Meets Phonetics: Acoustical Studies of Second Occurrence Focus. *Language*, 83(2), 2007.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Padó. SALTO: A Versatile Multi-Level Annotation Tool. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

Daniel Büring. Been There, Marked That – A Theory of Second Occurrence Focus. in a volume edited by Makoto Kanazawa and Christopher Tancredi, to appear.

Steve Cassidy and Jonathan Harrington. Multi-Level Annotation in the EMU Speech Database Management System. *Speech Communication*, 1-2(33):61–78, 2001.

Wallace L. Chafe. *Discourse, Consciousness, and Time*. University of Chicago Press, 1994.

Herbert H. Clark. Bridging. In P. Johnson-Laird and P. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 169–174. Cambridge University Press, 1977.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 1(20):37–46, 1960.

Caroline Féry, Gisbert Fanselow, and Manfred Krifka, editors. *The Notions of Information Structure*, volume 6 of *Interdisciplinary Studies on Information Structure*. Universitätsverlag Potsdam, 2007.

Michael Götze, Cornelia Endriss, Stefan Hinterwimmer, Ines Fiedler, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas adn Ruben Stoel, and Thomas Weskott. Information structure. In Stefanie Dipper, Michael Götze, and Stavros Skopeteas, editors, *Information Structure in Crosslinguistic Corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure*, number 7 in Working Papers of the CRC 632, Interdisciplinary Studies on Information Structure (ISIS), pages 147–187. 2007.

Martine Grice, Stefan Baumann, and Ralf Benzmüller. German Intonation in Autosegmental-Metrical Phonology. In Sun-Ah Jun, editor, *Prosodic Typology. The Phonology of Intonation and Phrasing*, pages 55–83. Oxford University Press, 2005.

Eva Hajičová, Barbara H. Partee, and Petr Sgall. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer, Dordrecht, 1998.

Michael Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.

Michael A. K. Halliday and C. Matthiessen. *An Introduction to Functional Grammar*. Edward Arnold, London, 2004.

Shinshiro Ishihara and Caroline Féry. Phonetic Correlates of Second Occurrence Focus. In *Proceedings of the 36th Meeting of the North Eastern Linguistics Society*, 2006.

Manfred Krifka. Basic Notions of Information Structure. In Féry et al. (2007), pages 13–56.

Knud Lambrecht. *Information Structure and Sentence Form*. Cambridge University Press, 1994.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. An Annotation Scheme for Information Status in Dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbon, 2004.

Patrizia Paggio. Annotating Information Structure in a Corpus of Spoken Danish. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1606–1609, Genoa, Italy, 2006.

Barbara H. Partee. Focus, Quantification and Semantics-Pragmatics Issues. In P. Bosch and R. van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 213–231. Cambridge University Press, 1999.

Ellen F. Prince. Toward a Taxonomy of Given-New Information. In P. Cole, editor, *Radical Pragmatics*, pages 233–255. Academic Press, New York, 1981.

Ellen F. Prince. The ZPG Letter: Subjects, Definiteness and Information Status. In W. C. Mann and S. A. Thompson, editors, *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pages 295–325. Benjamins, Amsterdam, 1992.

Arndt Riester, David Lorenz, and Nina Seemann. A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 717–722, Valletta, Malta, 2010.

Julia Ritz, Stefanie Dipper, and Michael Götze. Annotation of Information Structure: An Evaluation Across Different Types of Texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2137–2142, Marrakech, Morocco, 2008.

Craige Roberts. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. *OSU Working Papers in Linguistics*, 49, 1996.

Christine Röhr and Stefan Baumann. Prosodic Marking of Information Status in German. In *Proceedings of Speech Prosody*, Chicago, 2010.

Christian Rohrer and Martin Forst. Improving Coverage and Parsing Quality of a Large-Scale LFG for German. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genova, 2006.

Mats Rooth. A Theory of Focus Interpretation. *Natural Language Semantics*, 1(1):75–116, 1992.

Mats Rooth. Second Occurrence Focus and Relativized Stress F. In C. Féry and M. Zimmermann, editors, *Information Structure: Theoretical, Typological, and Experimental Approaches*. Oxford University Press, 2010.

Phillippe Schlenker. Minimize Restrictors! (Notes on Definite Descriptions, Condition C and Epithets). In E. Maier, C. Bary, and J. Huitink, editors, *Proceedings of Sinn und Bedeutung IX*, pages 385–416, Nijmegen, 2005.

Roger Schwarzschild. GIVENness, AvoidF, and Other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177, 1999.

Elisabeth Selkirk. Contrastive Focus, Givenness and the Unmarked Status of 'Discourse-New'. In Féry et al. (2007), pages 125–145.