Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects

Stefan Baumann[1], Arndt Riester[2]

[1] I*f*L Phonetik, University of Cologne, Germany
[2] IMS, University of Stuttgart, Germany

stefan.baumann@uni-koeln.de, arndt.riester@ims.uni-stuttgart.de

Abstract

The main objective of the paper is to show that for an adequate analysis of an item's information status in spoken language two levels of givenness have to be investigated: a *referential* and a *lexical* level. This separation is a crucial step towards our goal to arrive at the best possible classification of nominal expressions occurring in natural discourse which reflects our understanding of all important aspects related to an item's givenness or novelty ("information status"). For that purpose, we first introduce our motivation for this division which stems from the observation that both levels of information status have an influence on an item's prosodic marking. Section 2 presents basic concepts including the cognitive dimensions which have been discussed in the literature on givenness, in particular the role of knowledge, consciousness and (un)importance. In Section 3, we give an overview of influential proposals to classify and order the various concepts. In doing so, we explain and relate the most important concepts that play a role in classifying nominal expressions: coreference, bridging, inference, hearer knowledge, indexicality, embeddedness. A number of paradoxes and inconsistencies of the presented approaches are discussed, which can be resolved by the insight that there are two different notions of givenness which apply to expressions of different syntactic-semantic type. We refer to them as *referential givenness (r-givenness)* and *lexical givenness (l-givenness),* and provide the semantic basis for these two concepts. In Section 4, we propose a fine-grained two-level annotation scheme for the analysis of an item's information status, called the *RefLex scheme*. Section 5 takes a closer look at the prosody of the proposed labels in combination and develops a couple of hypotheses which are tested in two small exemplary corpora of spontaneous and read German speech.

1. Introduction

It has often been claimed for West Germanic languages such as English or German that *new* – as well as *contrastive* – information is marked by accentuation whereas *given* information gets deaccented, i.e. it does not receive a pitch accent although it would be expected to be accented in an unmarked case (see e.g. Halliday 1967; Allerton 1978; Ladd 1980). These claims have to a certain extent been confirmed in experimental work (e.g. Cruttenden 2006) or critically assessed (e.g. Terken and Hirschberg 1994). However, in order to verify these and similar claims in natural spoken language, speech corpora have to be built and annotated in terms of given and new information. In the last three decades, various proposals have been

brought forward for labelling schemes designed to enable annotations of the given-new distinction, also known as *information status*. However, none of these schemes have proven detailed enough to capture and distinguish all sorts of informational distinctions which are necessary to explain even the most elementary intonational patterns. For instance, two patterns which annotators have struggled with to the present day are shown in the two sequences in (1) and (2) (both taken from Büring 2007).

(1) A: *Did you see Dr. Cremer to get your root canal?*
B: *Don't remind me. I'd like to STRANgle the butcher.* [1]

(2) A: *Why do you study Italian?*
B: *I'm MArried to an Italian.*

It is obvious that in (1) and (2) there are different types of givenness involved. Nevertheless, both seem to have an influence on the prosodic realisation. More specifically, the example in (1), which Büring adapted from Ladd (1980), shows that the determiner phrase (DP) *the butcher* is deaccented because it is interpreted as *coreferential* with the previously mentioned *Dr. Cremer*. On the other hand, the mention of *Italian* in (2B) is treated as given, although it is not coreferential with the first mention: *Italian* in the question denotes the language, *Italian* in the answer denotes a person. Thus, *Italian* in (2B) simply is *lexically given*, also leading to deaccentuation.

In this paper we take the basic insight derived from these examples as a motivation for the development of a new, two-layered, type of annotation system for information status, which, moreover, not only identifies different sorts of "given" expressions but also "new" and "intermediary" ones. The paper is structured as follows: after some remarks about the main uses of the term *givenness* in the recent linguistic history in Section 2, we will turn to a discussion of existing proposals to information status (Section 3). In Section 4, we present our own scheme which is divided into a *referential* and a *lexical* level. Section 5 reports on some empirical findings from two small annotated corpora of spontaneous and read speech. After the conclusion, there is an appendix, which provides an overview of the most important label combinations as well as a comparison of the proposed categories with previous labelling schemes for information status.

2. Concepts of givenness

People convey information by expressing properties of and relations between individuals in specific situations. The expressions which denote these individuals, typically determiner phrases (traditionally: "noun phrases"), pronouns or prepositional phrases, do not only possess a certain (lexical and/or referential) *meaning* but can also be regarded to have a specific cognitive *information status* or *degree of givenness* relative to various contextual factors.

In the literature, the notion of *givenness* has been understood in at least three different senses, which are related to different cognitive dimensions: the first one has to do with the *knowledge* assumed to be shared by speaker and listener, the second with what the speaker assumes to be in the listener's *consciousness* at the time of utterance, and the third with how a speaker presents a piece of information with regard to what s/he considers *important* or not, cf. the overview presented by Ellen Prince (1981: 225-232).

2.1. Knowledge

Influential representatives of an approach based on the first dimension are Clark and Haviland (1974). They define given information as known information, considered by the speaker as an entry in the listener's memory structure (both long-term and working memory) representing an individual. Importantly, the listener is not required to be thinking of this element at the time of utterance. The aspect of assumed mutual or "shared" knowledge and beliefs is also a central part in Stalnaker's (1974) widely used notion of *common ground*. Similar to Clark and Haviland and to some extent also to Stalnaker, Prince (1981) considers the dimension of knowledge as basic to an understanding of givenness. However, she discards the term "shared knowledge", since it suggests the point of view of an omniscient observer who knows what goes on in the listener's mind, and replaces it by "assumed familiarity", thus taking the perspective of the speaker. Other linguists have referred to known entities as being present "in the permanent registry" (Kuno 1972) or "culturally copresent" (Clark and Marshall 1981).

2.2. Consciousness

The second definition of givenness implies that the speaker assumes that the listener has or could appropriately have some particular entity in his/her consciousness at the time of hearing the utterance. The main representative of this view on givenness is Chafe (e.g. 1976, 1994), who claims that "it is ultimately impossible to understand the distinction between given and new information without taking consciousness into account" (Chafe 1994: 72). Chafe repeatedly showed that knowing something and thinking of something are different mental states, which have to be neatly kept apart.

The notion of *activation* is central to Chafe's cognitive approach and applies to both speaker and listener. He defines givenness in terms of the activation cost a speaker has to invest in order to transfer an idea from a previous state of consciousness in the listener's mind into an active state. If a referent is already active in the listener's mind at the time of the utterance (i.e. it has been mentioned recently), it is *given*, and if a referent becomes activated from a previously inactive state, it is *new*. Chafe additionally assumes a third state, semi-active, meaning that the respective entities have not been mentioned themselves but are directly linked to consciously present ones. He calls information in this state *accessible*.

Note that there is also a stronger notion of givenness in terms of consciousness, which is referred to as *predictability* (e.g. by Kuno 1972). Literally, predictability means the hearer's ability to recover a previously mentioned linguistic item prior to its subsequent mention. It should be kept in mind that not all consciously present items count as predictable.

2.3. Unimportance

The third way of defining givenness is concerned with the intentions of a speaker and is thus essentially pragmatic in nature. Important representatives of this approach are Halliday and colleagues (e.g. Halliday 1967) as well as Kuno (1972). This approach implies that a speaker

may present an item as given not only because it has been mentioned in the preceding discourse or because it is known but also for the purpose of marking its unimportance. That is, as Halliday and Matthiessen (2004: 91) put it: "The meaning is: this is not news". Conversely, new information is presented as newsworthy or important information, irrespective of whether it has been mentioned before. It is from these usages of givenness and newness that Halliday (1967) developed his theory of focus.

There are two problems with this approach. First, defining givenness in terms of how an item is "presented" necessarily leads to a circular argumentation if we want to look at the relation between an item's information status and its linguistic coding. Second, the dependency on a speaker's intentions and assumptions in determining an item's information status involves a large amount of subjectivity.

In the present paper, we regard the discourse context as a cognitive dimension shared by the interlocutors at the time and place of utterance. It consists of a set of discourse referents as well as the predicative and relational information attributed to them which are a direct consequence of the ongoing communication. In doing so, we adopt a view which is widespread in semantic theory (e.g. Kamp and Reyle 1993). Among the three dimensions of givenness, the dimension of consciousness is determined first and foremost by the dynamic discourse context. It is this level of cognitive activation which may be considered central for the analysis of an item's givenness (or information status), even more so than the dimension of knowledge which is not subject to immediate contextual changes in the same way. The dimension of unimportance should be kept apart from the other two dimensions, since it does not describe an item's actual information status but rather the speaker's concern to play an item's relevance up or down.

2.4. Other models and combined approaches

Most of the older approaches to givenness do not clearly differentiate between the dimensions discussed above. Stalnaker's (1974) *common ground*, for instance, is defined as the set of propositions the participants in a conversation assume to be true for the sake of communication, thus representing a mixture of knowledge and conscious awareness (see also Krifka 2007: 15-17).

In a more recent approach, Prince (1992) deliberately combines these two dimensions of givenness. In her analysis of a short written text (a fund-raising letter), she distinguishes between *hearer-old* and *hearer-new* entities on the one hand and *discourse-old* and *discourse-new* entities on the other. Hearer-status is equivalent to the "knowledge dimension of givenness". Prince's second notion of givenness, the one of discourse-old versus discourse-new, is equivalent to the "consciousness dimension of givenness". This means that an entity may be given or new with respect to the current discourse. While discourse-old entails hearer-old, discourse-new entities can be either hearer-new or hearer-old.

Another combined approach is the one of Lambrecht (1994). To Chafe's three-fold distinction of given (active), accessible (semi-active) and new (inactive) information, he adds the category of *identifiable* information, which "has to do with a speaker's assessment of whether a discourse referent is already stored in the hearer's mind or not" (Lambrecht 1994: 76). While *activation* is related to consciousness, *identifiability* is in some sense related to *knowledge*. More specifically, identifiable items can be either in the long-term memory of the hearer (inactive) but also be salient within a scenario under current discussion, available

in the text-external setting or represent a generic expression. Of course, previously mentioned (active) items are identifiable as well. Usually, non-generic indefinites are unidentifiable. There is some degree of uncertainty whether *definite* expressions can count as unidentifiable as well. Lambrecht (1994: 109) illustrates the relation between identifiability and activation in a diagram, shown in Figure 1.

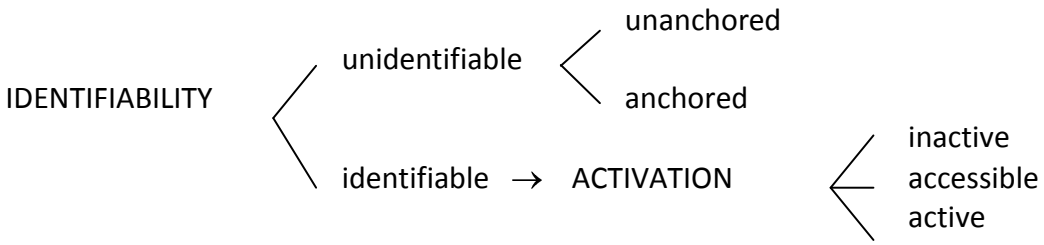


Figure 1: Cognitive categories of identifiability and activation in Lambrecht (1994)

We will not discuss Lambrecht's proposal in detail. However, in the appendix to this article there is an overview table which relates his approach to other ones discussed in the paper.

3. Paradoxes and inconsistencies of existing approaches to information status: Preliminaries for a new arrangement

In this section we give an overview of the concepts proposed in the influential paper by Prince (1981) and explore their properties and weaknesses in comparing them with similar concepts proposed in other approaches. Our goal is develop a new annotation system for information status, a better classification than the existing ones that adheres to the following conditions: (i) we want to be able to classify all nominal expressions occurring in natural discourse as reliably as possible, (ii) the classification should reflect our understanding of all important aspects related to the givenness or novelty (the *information status*) of natural language expressions (iii) the classification should allow for at least a partial hierarchical ordering of the classes proposed and (iv) a corpus annotated according to the new scheme should be beneficial for research on prosody.

Prince (1981) proposes to integrate various conceptions of givenness according to a hierarchy of "assumed familiarity", shown in Figure 2. This term describes the perspective of the speaker who has certain assumptions about the mental state of the hearer.

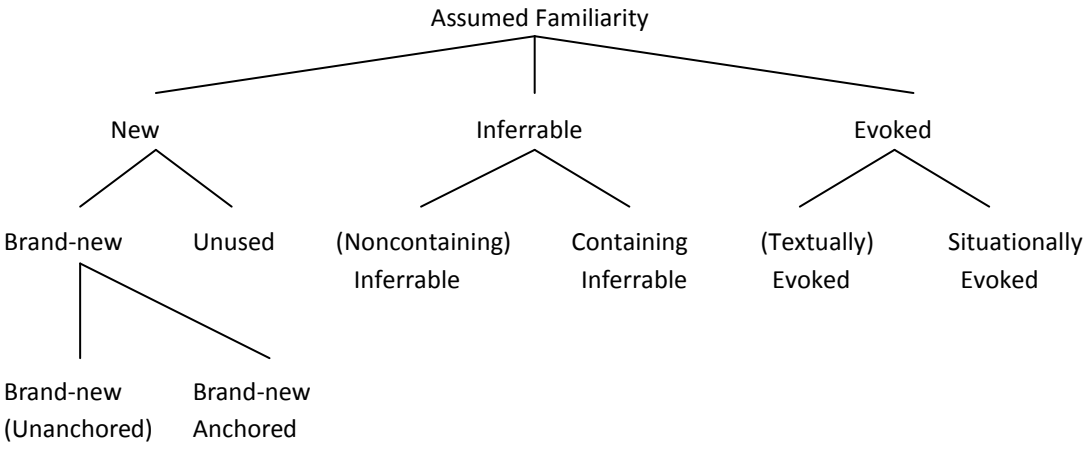


Figure 2: Prince's (1981: 237) Assumed Familiarity Scale

3.1. New

First, we will have a look at the class of *new* items. Prince defines this class on the basis of text – or discourse – alone. A necessary condition for an item[2] to be called *new* is that it may not have occurred in the previous text. Depending on whether the addressee knew the item or not, it will be subclassified as *unused* or as *brand-new*. This is shown in the list of examples in (3).

(3) *We were sitting in the lobby.*

a. *A strange-looking fellow came up to us.* } brand-new } unanchored
b. *Harry Smith came up to us.*
c. *The man who had stood at the bar came up to us.* anchored
d. *George Clooney came up to us.* unused

There are several reasons for us to feel uneasy about the classification in (3). First of all, we notice that indefinites (3a), proper names (3b,d) as well as definite descriptions (3c) are all grouped together (they are all *new*), which is rather peculiar from a semantic perspective.

Second, what is remarkable is that the proper names in (3b) and (3d) are grouped into quite different subclasses. Certainly, we would like to say that, at the time this article is written, George Clooney is a widely known person, while nobody we know has heard of Harry Smith. But does this justify treating the latter on a par with indefinite expressions? In more contemporary approaches to information status like Nissim et al. (2004) and Götze et al. (2007), the gap between *known* and *unknown* expressions widens even further since, on these accounts, only *unknown* (*brand-new*) names would count as *new*, while known ones receive some intermediate (*non-new*) status.[3] If we assume that some text or spoken corpus is analysed in terms of information status following one or the other scheme – perhaps in order to investigate correlations between information status classes and certain intonational parameters – we must be aware that classificational discrepancies of the kind described may have drastic consequences for the empirical results obtained and, therefore, should be considered with careful thought.

A point that is raised and discussed in Riester (2008a, 2008b) is that deciding whether a definite expression is *known* or *unknown* can be much more difficult than for the clear cases in (3b) and (3d). Consider, for instance, the underlined expression in (4).

(4) *Sven Akalaj, the Foreign Minister of Bosnia and Herzegovina, said in an interview*...

Following Prince (1981), we would probably select the label *brand-new anchored* for the item in (4). However, this choice rather derives from a mixture of semantic and morphosyntactic cues. On the one hand, we would not want to say that the referent is a known person, which is a semantic-pragmatic criterion. On the other hand, it contains an anchoring apposition. In any case, the classification is not motivated by the writer's "assumed familiarity" with the item.

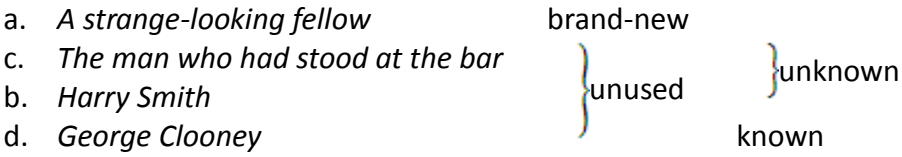
There are several problems here. First, we have no access to what goes on in some speaker's or writer's mind, which renders Prince's task of labelling the speaker's assumed

familiarity virtually impossible. Secondly, persons, places or other entities are rarely ever "objectively" known[4] or unknown but only with respect to some intended recipient (Riester 2008a, 2008b). There is likewise no access to the recipient's mind. If a text is addressed to a larger crowd – news texts are a typical example for this – knowledge and ignorance are unevenly distributed among the recipients.

For the task of labelling information status it seems to be best to assume some prototypical recipient and estimate his or her encyclopaedic knowledge. This may sound like a rather uncertain task but it is better than to indulge in a potentially circular argumentation which uses formal criteria for speculating about what the writer may have thought when producing his text.

However, there is one move which we should make in order to soften the consequences of potential misclassifications. We propose to rearrange Prince's (1981) classification of *new* entities as shown in (5).

(5)



Expressions which cannot be clearly classified as known or unknown, like the one in example (4), do not need any such subclassification. A further welcome consequence of this new arrangement of *discourse-new* entities is that it keeps indefinites and definites apart, which have been treated differently in semantic theory. Only definites ("unused") trigger an identificational presupposition (Karttunen 1974; Heim 1982; van der Sandt 1992), which in some cases may be resolved in the addressee's knowledge context (Riester 2009) ("unused-known") and sometimes may need to be accommodated ("unused-unknown").

3.2. Inferable / accessible / bridging

The next class assumed by Prince (1981) are *inferable* items. They take an intermediate position between *new* and *evoked* entities. Similarly, in the work of Chafe (1994), there is an intermediate class, called *accessible* or *semi-active*, localised between *given* and *new* items. The idea behind both approaches is that such *inferable* / *accessible* items are in some sense *discourse-new* but at the same time inferentially related to some expression introduced earlier. As examples, consider the underlined expressions in (6) and (7), adapted from Clark (1977).

(6) *I walked into my hotel room.*
a. *The ceiling was very high.*
b. *The windows looked out to the bay.*

(7) *John was murdered yesterday*.
a. *The murderer got away.*
b. *The weapon was lying nearby.*

In the examples in (6) and (7), the hearer encounters certain definite expressions, e.g. *the*

ceiling in (6a), which signal identifiability. However, since no ceiling has been mentioned before and the expression is not unique from a global perspective, the hearer is urged to restrict the search space in such a way that uniqueness is guaranteed. The previously mentioned *hotel room* is a suitable restrictor and *the ceiling* is thus interpreted as "the ceiling of the hotel room". This procedure has been discussed in terms of the so-called *Given-New Contract* (Clark and Haviland 1974; Clark 1977). The hearer's mental process of linking a non-coreferential but nevertheless context-dependent expression to previous material is also known as *bridging*.[5] The respective definite expressions are therefore called *bridging anaphora* or *associative anaphora* (cf. Poesio and Vieira 1998; Vieira and Poesio 2000; Poesio 2004).

The fact that hearers are able to infer or construct a *bridge* between some contextual antecedent and the bridging anaphor, however, should not be confused with the claim, which is sometimes made, that the anaphor itself must be *inferable* or *accessible*, although this is the case in (6) and (7). If there is a room, we may usually infer that it has a ceiling. However, it is also very common to have bridging anaphors which are unexpected, non-inferable and therefore more informative than the ones just mentioned; compare, for instance, (8) and (9).

(8) *I walked into my hotel room. <u>The chandeliers</u> sparked brightly.*

(9) *John was murdered yesterday. <u>The Japanese kitchen knife</u> lay nearby.*

The annotation scheme that we will develop in Section 4 of this paper is able to distinguish between these cases. Finally, consider the special case in (10), where what used to be the *antecedent* in (6a) has shifted and become an embedded part of the anaphor.

(10) *<u>The ceiling of the hotel room</u> was very high.*

Prince (1981) uses the label *containing inferable* for such cases (although they are not clearly distinguished from her class *brand-new anchored*). We will refer to them in the following sections as *bridging-contained.*

3.3. Situationally evoked / indexicals / deixis

The final main class defined in Prince (1981) is called *evoked.* It consists of the two strands *textually evoked* and *situationally evoked*. Halliday and Hasan (1976) refer to these two types as *endophoric* and *exophoric* reference. It is obvious that these correspond to, on the one hand, explicitly mentioned entities and, on the other hand, entities present in the conversational situation. We start with the latter ones, which are also known as *indexicals* or *deictic expressions.*

Kaplan (1989) defines two classes of indexicals: pure indexicals and demonstratives.[6] Indexicals are characterized by the fact that, as Kaplan puts it, "the referent is dependent on the context of use" (1989: 490). By this, he meant to separate such expressions from bound uses of definite descriptions and pronouns as in (11).

(11) *A farmer who owns a donkey feeds <u>it</u>.*

A definition like this must nowadays be considered insufficient, since it leaves aside dynamic theories of meaning. There is by now a well-researched class of expressions which are also context-dependent – most prominently, various sorts of anaphors – which are not amongst those expressions which Kaplan had in mind when defining his class of indexicals.[7]

The subclass of demonstratives describes expressions which are necessarily accompanied by a pointing gesture ("demonstration"). Among such expressions are personal pronouns (*he, she, it, they*), on their "gesturally deictic" reading, demonstrative pronouns (*this, that*), expressions like *here* (pointing) and *there*, as well as definite and demonstrative descriptions (*the man, this man* – again only in their pointing use). Since pointing requires visual identification on behalf of the addressee, it can only occur in particular communicative situations like face-to-face dialogue. By contrast, pure (symbolic) indexicals (*I, you, now, here* (meaning: location of the speaker)*, tomorrow, yesterday*) need not (and usually do not) come with a pointing gesture.

3.4. Textually evoked / discourse old / given / active

We now turn to the class of *textually evoked* or, henceforth, simply *given* items. Examples are shown in (12).

(12) *On my way home a dog barked at me.*
a. *<u>The dog</u> belongs to my neighbour's oldest son.*
b. *<u>The animal</u> belongs to my neighbour's oldest son.*
c. *<u>It</u> belongs to my neighbour's oldest son.*

The cases in (12) are fairly uncontroversial. In all of them, the previously mentioned *dog* is taken up by the phrases indicated in (12a-c), which signal their respective identifiability by means of (a) repeated, (b) entailed or (c) pronominal linguistic material. The *anaphoric* expressions are said to be *coreferential* with their antecedent. Note, however, that entailment or pronominalisation are not necessary conditions for calling an expression *given*, as shown below.

(13) *On my way home, a dog barked at me.*
<u>The fierce German Shepherd</u> seemed to be quite aggressive.

(14) *Ole was a brilliant athlete.*
The local press had nothing but praise for <u>the tennis player</u>.

Examples (13) and (14) are meant to demonstrate that the antecedents in a *coreference chain* need not be any more specific or informative than their subsequent anaphor. In the cases shown, the anaphors even entail their antecedents as regards content (e.g. if something is a fierce German Shepherd then it is also a dog). Clark (1977) refers to such postponed informative definites as *epithets*.[8] It may even be the case that the relation between the antecedent and the anaphoric expression is not one of lexical entailment but is just an informative attribution on behalf of the speaker. For instance, instead of "German Shepherd" the speaker might have used "stupid beast", regardless of the fact that not all dogs are stupid beasts and not all stupid beasts are dogs. Likewise, instead of telling us that Ole is a tennis player, our conversation partner could have let us know that Ole can also be

referred to – in newspaper style – as "the 23-year old father of twins*",* which does not stand in an entailment relation to the phrase "a brilliant athlete".

In general, however, the coreferential (or *given*) reading in all of the cases mentioned can only be obtained if the anaphoric expression does not receive the nuclear pitch accent (see footnote 1). In (13), the long subject DP *the fierce German Shepherd* is likely to be marked by at least one *prenuclear* accent but the *nucleus* has to fall on *agGRESsive*. Similarly, in example (14) the main (nuclear) accent has to fall on *PRAISE*, whereas the anaphor *the tennis player* has to be deaccented.[9] If *TENnis player* is accented instead (which would be the default intonation in an all-new sentence), it cannot be understood as *given* and the hearer is led to wonder which tennis player the speaker was talking about (as in the *butcher* example in (1) above).

3.5. Two levels

Conversely, entailment or repetition is not a guarantee for *coreference*. Consider the sequence in (15).

(15) *On my way home, a dog barked at me. It made me think of ANna's dog.*

There are two instances of the word *dog* but the dogs mentioned are not the same creature. Nevertheless, although the two expressions are not coreferential, there is a strong correspondence between the two mentions of *dog.* One hint for this is the fact that it is impossible to place an accent on the second instance of the word *dog*, which would be possible if it were replaced by any other unrelated word; see (16).

(16) *On my way home, a dog barked at me. It made me think of Anna's BOYfriend.*

We conclude that the phenomenon in (16) and many similar examples cannot be explained in terms of a *givenness notion* which is based on coreference. The same observation is made implicitly in Schwarzschild's (1999) article on givenness and focus. Schwarzschild (1999: 151) provides us with the following definition of givenness:

Definition I: *given*
An utterance U counts as *given* if it has a salient antecedent A and

a) If U is type *e,* then A and U corefer;
b) Otherwise: modulo existential type-shifting, A entails the existential F-closure of U.

Let us investigate the most important aspects of this definition. First we note that *utterances* or, better, *expressions* are grouped into two kinds: type *e* and type "non-*e*". This requires some specification. In the following, we shall treat all "term" expressions (syntactically: DPs) as being of type *e*, including indefinites.[10] On the other hand, content words like nouns, verbs or adjectives but also modified NPs like *green apple* are treated as type "non-*e*"; particularly, as properties. In other words, for the time being only DPs will be treated as *referential expressions*, to which Schwarzschild's definition (b) is applicable. Only for these expressions, or *at this particular syntactic-semantic level,* givenness is equal to coreference.

The second part of the definition contains a technical twist, which we shall simplify to some extent. According to Schwarzschild (1999), a "non-referential" expression, e.g. an

expression *U* of type <e,t>, is *given* if the following procedure is successful: (i) contextually identify an antecedent expression *A* of the same type, (ii) replace any potential focus marked subparts of ⟦*U*⟧ by existentially bound variables (*existential F-closure*), (iii) replace the predicate-initial λ by an existential quantifier within both *A* and *U (existential type-shift)*, (iv) validate whether the resulting meaning ⟦*A'*⟧ entails ⟦*U'*⟧.

Since we are only interested in a technical *givenness* definition but not in Schwarzschild's Optimality-Theoretic machinery for predicting focus, we ignore all aspects dealing with focus marking (step (iii)).[11] The result is a more conservative *givenness* notion, which we demonstrate on the basis of example (17).

(17) *On my way home, a big German Shepherd barked at me.*
It reminded me of Anna's dog.

As we already noticed with example (15), the two mentioned dogs are not identical. Therefore, the DPs [a big German Shepherd] and [Anna's dog] are not coreferential, the latter is not *given*. On the other hand, the modified NP [big German Shepherd] and the word [dog] may be translated as the predicate expressions in (18a,b).

(18) a. $\lambda x.[big(x) \wedge German\_Shepherd(x)]$
b. $\lambda x.dog(x)$

In the next step, the lambdas are replaced by existential quantifiers and the resulting formulas are verified as to whether the first entails the second (see (19)).

(19) $\exists x.[big(x) \wedge German\_Shepherd(x)] \models \exists x.dog(x)$

Since the entailment in (19) is valid, the expression *dog* counts as *given* according to the simplified (b) part of Definition I. Of course, what is at the heart of this procedure is simply a verification of *synonymy* or (right-hand-side) *hypernymy* of the expressions involved. However, it is a little more than that, since these notions are generalised from words to phrases, like the sequence [*big German Shepherd*] in the first sentence of (17).

As a summary we state that there are two different notions of *givenness* which apply to expressions of different semantic type. Referential expressions (type *e*) are *given* if they possess a coreferential antecedent (another discourse referent). Non-referential expressions (mostly, type <e,t>) are *given* if they have an "antecedent" which is either an identical expression, a synonym or a hyponym. We refer to the former type as *referential givenness (r-givenness),* to the latter as *lexical givenness (l-givenness).*

A distinction between the two levels can also be found in the system of *cohesion* within the framework of Systemic-Functional Linguistics (e.g. Halliday and Hasan 1976; Halliday and Matthiessen 2004). Cohesion describes the lexicogrammatical links between elements in a discourse. While reference operates at phrase level and creates links between elements from the situation (exophoric) or from the text (endophoric), lexical cohesion operates at word level and is achieved through the choice of lexical items (Halliday and Matthiessen 2004: 535). The kinds of lexical relations playing a role in lexical cohesion are claimed to be repetition, synonymy, hyponymy and meronymy. As Halliday and Matthiessen (2004: 572) point out, coreference is not a necessary factor in cases of lexical cohesion, as already exemplified in the *Italian* example (2) above.

The different sources of givenness and in particular their effect on a constituent's

prosodic realisation have also been thoroughly discussed by van Deemter (1994, 1999). He argues that an item may either be deaccented due to *object-givenness*, if it stands in a coreference (or "identity-anaphoric") relation to an antecedent, or due to *concept-givenness*, if the item constitutes a non-identity anaphor of a subsumed, i.e. extensionally included, word.

According to van Deemter, concept-givenness comprises cases of identical words, as in (20), where the generic expression *a car* in the second sentence subsumes the exemplar mentioned earlier.

(20) *If Susan owns a car, she must be rich.*
Well anyway, you don't NEED a car in New York City. (Van Deemter 1994: 19)

It also applies to cases of "subsuming anaphors".[12] Van Deemter (1999: 7) gives the following examples. The anaphoric hypernym *string instruments* (the superordinate term) can be deaccented, as in (21) (see also example (17) above):

(21) *Bach wrote many pieces for viola. He must have LOVED string instruments.*

In contrast, the anaphoric hyponym *viola* (the subsumed term) cannot be deaccented, resulting in (22):

(22) *Bach wrote many pieces for string instruments. He must have loved the viOla.*

but # *He must have LOVED the viola.*

Subordinate anaphors like *viola* in (22), would not be concept-given, but new, and thus have to be accented. Allerton (1978: 142) explains this pattern

> "by the fact that while the hyponym frequently implies the superordinate (Lyons 1968: 455), and the part frequently implies the whole, the reverse applies only rarely in either case. In other words, the use of the hyponym or 'part' word involves adding extra information in a way the reverse sequence does not, and this extra 'new' information requires some degree of stress".

Van Deemter postulates "memory limitations" for concept-givenness (Van Deemter 1994: 20), since he claims that recency is a much stronger factor in concept-givenness than in object-givenness. This difference may actually stem from a difference in the affected levels, suggesting that *words* (i.e. *referring expressions*) are more transient than *referents* (i.e. their *denotations*). Such an interpretation would support the idea of lexical cohesion as being crucial for cases of concept-givenness, and coreference being crucial for object-givenness. This idea will be reflected in our proposed annotation scheme presented in the next section.

Concluding this section, let us consider an example by van Deemter (1994: 5):

(23) *Clinton visited many towns; when he finally arrived in CLINton, he was late.*

Van Deemter argues for an obligatory accent on the "anaphor" *Clinton*, denoting a town, although it displays the same referring expression as the antecedent *Clinton*, denoting a person. At first sight, this sentence serves as counter-evidence for the relevance of lexical

givenness for an item's prosody and, more specifically, Büring's *Italian* example (2), where the non-coreferential anaphor had to be deaccented. However, it seems that the accent on *Clinton* in (23) simply indicates a new or unpredictable piece of information, namely the existence of *a TOWN called Clinton*, apart from the already known fact that there is *a PREsident called Clinton*. In cases like these, a referent's newness or unpredictability may override its lexical givenness, just as a contrastive focus may override any kind of givenness. Thus, an example like (23) does not undermine the general relevance of givenness at a lexical level as shown in the *Italian* example.[13]

A note on the relation between givenness and focus: we are well aware that not only an item's information status has an influence on its prosodic realisation but also a number of features that are linked to the level of contrast, focus or (shifted) topic. Thus, a comprehensive annotation scheme will include labels for configurations which *elicit alternatives*, such as focus-sensitive particles or overtly contrastive expressions (see Riester and Baumann 2011) and also take shifted syntactic positions of given expressions into account. A detailed discussion of these features would be beyond the scope of this paper, however.

4. The RefLex scheme

In the present paper we suggest that the concept of information status should be applied to two distinct levels, namely a *referential level* and a *lexical level*. The annotation scheme that we will introduce below is therefore called the *RefLex scheme*. We focus attention on the information status of nominal expressions.[14]

A (possibly complex) phrase is assigned both referential and lexical information status. However, the two apply at different syntactic levels. Referential information status is assigned at the level of PP and DP, whereas lexical information status applies at the word level, or possibly the modified NP level (see Figure 3).

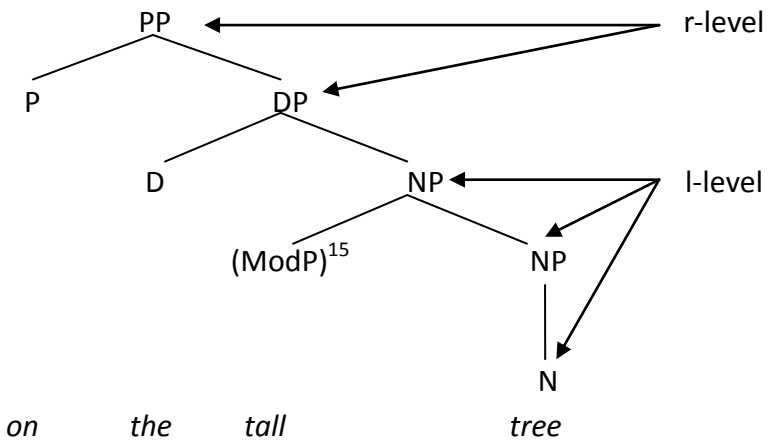


Figure 3: Syntactic domains of the referential (*r*) and lexical (*l*) levels for the example phrase *on the tall tree*

4.1. RefLex scheme: The referential level

Table 1 contains the labels we propose to be relevant at the referential level, accounting for the source of a referent's givenness, the distance from its last mention as well as its realisation as a definite or an indefinite expression. We will give examples of the proposed categories below.

Table 1: Labels for the annotation of discourse referents; "r-" indicates the referential level

Units: definite DPs	
r-given	anaphor corefers with antecedent in previous discourse
r-given-sit	referent is immediately present in text-external context (in particular discourse participants) – symbolic deixis
r-given-displaced	coreferring antecedent does not occur in previous 5 intonation phrases or clauses
r-environment	refers to item in text-external context (conversational environment) – gestural deixis / demonstratives
r-bridging	non-coreferring anaphor, dependent on previously introduced scenario
r-bridging-contained	bridging anaphor which is anchored to an embedded phrase
r-unused-known	discourse-new item which is generally known
r-unused-unknown	discourse-new item which is identifiable from its own linguistic description but not generally known
Units: definite or indefinite DPs	
r-cataphor	item whose referent is established later on in the text
r-generic	abstract or generic item
Units: indefinite DPs	
r-new	specific or existential indefinite introducing a new referent

Referential givenness *(r-given)* is defined as a coreference relation between an antecedent and an anaphor, as illustrated in (24):[16]

(24) *I met <u>a man</u> yesterday. <u>The man</u> told me a story*.
r-given

A special case of referential givenness is represented by the participants in a conversation and by references to the time and place of the speech setting. These referents can be thought of as being available by default. They are encoded by pure indexicals like *you, I, now, tomorrow* or *here*, and are labelled *r-given-sit* (cf. Section 3.3). We do not annotate adverbial quantifiers like e.g. *always, often, usually, every Wednesday,* because they do not refer to unique entities.

If the antecedent has not been mentioned within the last five information units[17], the anaphor is labelled *r-given-displaced*. We borrow the term *displaced* from Yule's (1981) notion *displaced non-new*, which applies to entities denoting referents which have been

established previously in the discourse. Yule uses *displaced non-new* as a counterpart to *current non-new* entities which are the most recently introduced referents in the discourse. Although we assume a non-linear deactivation process for discourse referents, we propose that they are "valid" for a whole discourse, i.e. once a referent is introduced it will not become "new" again within the same discourse. We are aware, however, that any way of fixing the number of intervening units between discourse items, may they be referents or words, is necessarily arbitrary as long as it is not based on psycholinguistic or corpus evidence.[18] Nevertheless, the rather fine-grained annotation system proposed here may be a useful instrument for this field of research.

The label used for the visible objects in the communicative environment which are not available in the speech setting by default (as e.g. the interlocutors) is *r-environment*. In (25), *this chair* is supposed to refer to a visible chair not mentioned before in the discourse. As we already discussed in Section 3.3, such expressions only occur in special genres in which gesturing is an option. It does not occur, for instance, in news texts or narratives lacking elements of direct speech.

(25) *This chair (pointing) is wobbly*.
r-environment

If a definite expression can only be used felicitously by virtue of the contextual availability of a previous non-coreferential item (functioning as an "anchor"), we use the label *r-bridging*. In some cases, the anchor is not a specific referent but rather a whole stretch of text. In (26), there is a bridging relation between *the football match* and *the referee*, who is prototypically present in a football match.

(26) *Recently, at the football match, the referee was wearing orange socks.*
r-bridging

If the anchor surfaces as a syntactic argument within a complex bridging anaphor, we propose to mark the entire phrase (which in this case refers to the person who is a referee) as *r-bridging-contained*, as in (27):

(27) *The referee of the football match was wearing orange socks.*
r-bridging-contained

By the label *r-unused* we indicate a definite description whose referent has not been previously introduced into the current discourse and which is marked as identifiable even in the absence of context. This category is divided into two subcategories. The label *r-unused-known* stands for an item which the annotator assumes the hearer/reader or the expected audience to have knowledge of, as is the case in (28):

(28) *The Pope was wearing orange socks.*
r-unused-known

By contrast, an item is *r-unused-unknown* if the annotator does not expect the denoted referent to be known by the audience. Examples for this type are given in (29) and (30).

(29) *The woman Max went out with last night was wearing orange socks.*
r-unused-unknown

(30) *The 20-minute interruption of the football match was caused by a man*
r-unused-unknown
wearing only orange socks.

At this point, recall our discussion in Section 3.1 on the difficulties of labelling hearer knowledge. Very often, especially in texts addressed to a larger audience, discourse-new items cannot be easily categorized as known like the expressions in (28) or unknown like the ones in (29), (30). The present system can handle this problem elegantly by omitting any subclassification, as shown in the example discussed in (4) above and repeated here as (31).

(31) *Sven Akalaj, the Foreign Minister of Bosnia and Herzegovina, said in an interview...*
r-unused

In some cases, it may be difficult to choose between the labels *r-unused-unknown* and *r-bridging-contained*. The difference between them lies in the *prototypicality* of the relation between the head of the expression and the argument serving as the adjunct: *r-bridging-contained* describes prototypical relations, as in (27), where *the referee* is a necessary part of *the football match*, whereas (30) is a case of *r-unused-unknown*, since a 20-minute interruption is unusual for a football match. In other words, a referee is highly predictable or unsurprising in the context of a football match whereas a 20-minute interruption is surprising and therefore more informative.

Cataphoric expressions refer to subsequent items. They can be marked morphosyntactically either by definite or indefinite noun phrases and are labelled *r-cataphor*, as in (32):

(32) *Nine days after she won the women's 800m world championship in Berlin,*
r-cataphor
Caster Semenya returned home to the plains of Limpopo.

The label *r-generic* applies to abstract or generic referring expressions. They may be morphologically definite or indefinite. Example (33) displays genuine generic terms

(33) *The lion / A lion is a big animal.*
r-generic

while (34) illustrates that an abstract referent may surface as an indefinite expression even if it is contextually given in some way. The characteristic property of the expression *a collapse* in this case seems to lie in its hypothetical nature.

(34) *The president warned of a collapse of the banking system.*
A collapse could seriously damage the economy.
r-generic

Singular bare nouns receive the label *r-generic* if they denote an abstract referent, which seems to be the typical case in English or German. However, languages may differ

considerably in their morphological marking, which is why we prefer to employ a semantic rather than a morphological definition. (35) gives an example in which an indefinite article would misleadingly signal that there are several instances of the concept involved, neither would a definite article be appropriate here.

(35) *After two years, he was finally granted <u>asylum</u>.*
r-generic

Summing up loosely, the label *r-generic* is applied to kind-referring, abstract or hypothetical entities, often embedded under propositional attitude descriptions as in (34), which signal a preference not to be taken up by a personal pronoun in the following discourse.

Finally, referring expressions which are new in the discourse, not kind-referring, non-hypothetical but specific or existential, and which are marked by an indefinite article, are labelled *r-new*. Sometimes it is difficult to differentiate between indefinite generic and new referents, however. Generic expressions go hand in hand with non-specificity, as in (36) below. In (37), on the other hand, the term *a friend* refers to a specific person the speaker has in mind. Thus, it is marked as *r-new*.

(36) *I'm looking for <u>a friend</u>. I'm in serious trouble.* (non-specific)
r-generic

(37) *I'm looking for <u>a friend</u>. He owes me money.* (specific)
r-new

Note that "non-specific" should not be equated with "unknown"; it rather means that the introduction of an individual-type referent into the discourse is typically suppressed.

4.2. Referential information status and the Givenness Hierarchy

In Section 2.2, we briefly discussed Chafe's (1976, 1994) categorisation of information in the form of a mini-hierarchy of *given, accessible* and *new,* according to whether the information was already present in the discourse (active / given), immediately related to active information (semi-active / accessible) or unrelated to the previous discourse (inactive / new). From the extensive discussion in the previous sections it has become plain that such a simple ternary distinction is not ideal to account for the variety of types of referential information status found in natural language. For instance, as we already showed in Section 3.1, it is not obvious why indefinites (*new,* according to Chafe) should be treated as on a par with those definite expressions which, following our scheme, are clearly distinguishable as *r-unused*. We will make use of Chafe's classification when discussing lexical information status below. The problem seems to be that there is confusion concerning the question of what dimension such a hierarchy of expressions should express. Another well-known cognitive scale, which seems better suited to organise classes of referential information status, is the *Givenness Hierarchy* proposed by Gundel et al. (1993), shown in (38).

(38) in focus > *activated > familiar > uniquely identifiable > referential* > type-identifiable

We will only discuss the middle section marked in italics. According to Gundel et al., each property subsumes all other properties to its right. A *referential* expression is one which introduces or picks up a discourse referent. *Unique identifiability* is a property usually attributed to definites. *Familiarity* is a stronger notion, which requires an expression to be hearer-known, while *activation* additionally requires conscious awareness. Taking this simplified characterization as a basis, we can order our r-level-expressions as shown in Table 2.

Table 2: Referential information status and the Givenness Hierarchy

<table>
<tr><th>referential (but not uniquely identifiable)</th><th>uniquely identifiable (but not familiar)</th><th>familiar (but not activated)</th><th>activated</th></tr>
<tr><td colspan="3"></td><td>r-given, r-given-sit</td></tr>
<tr><td colspan="2"></td><td>r-unused-known, r-given-displaced</td><td></td></tr>
<tr><td></td><td colspan="3">r-bridging, r-environment</td></tr>
<tr><td></td><td colspan="2">r-bridging-contained</td><td></td></tr>
<tr><td colspan="4">r-generic</td></tr>
<tr><td></td><td>r-unused-unknown</td><td colspan="2"></td></tr>
<tr><td>r-new, r-cataphor</td><td colspan="3"></td></tr>
</table>

As we can see from the table, there is some meaningful correspondence between the perspective taken by Gundel et al. (1993) and our approach, although the two systems do not map directly onto each other. Some of our categories have no clear cut-off points, as for instance expressions carrying the RefLex label *r-bridging*. There are cases in which they would be classified as *activated*, and others in which they would only count as *uniquely identifiable* or *familiar*.

4.3. RefLex scheme – the lexical level

The lexical level applies to the word domain, more specifically to nouns or NPs (*green car*) and other set-denoting expressions. Pronouns and other functional categories are not annotated at the l-level. The proposed labels are shown in Table 3 and are explained in detail below. Note that at this level Chafe's (1994) terminology *given / accessible / new* is employed. However, we use it to classify words rather than their referents, as Chafe did. Nevertheless, our classification is "Chafean" in spirit in the sense that *given* describes an *active* word, *accessible* characterizes a *semi-active* word and *new* describes an *inactive* word.

Table 3: Labels for the annotation of nouns and NPs; "l-" indicates the lexical level

Units: nouns, modifiers, NPs	
l-given-same	recurrence of same expression
l-given-syn	relation between nouns at the same hierarchical level (synonyms)

l-given-super	noun is lexically superordinate to previous noun (markable is a hypernym or holonym, or generally a superset)
l-accessible-sub	noun is lexically subordinate to previous noun (markable is a hyponym or meronym, or generally a subset)
l-accessible-other	two related nouns, whose hierarchical lexical relation cannot be clearly determined (e.g. within a scenario)
l-new	noun not related to another noun within last 5 intonation phrases or clauses

The label *l-given-same* is used if the same nominal expression recurs, as in (39), which is a repetition of example (24):

(39) *I met a man yesterday. The man told me a story.*
l-given-same

Note that the two instances of the same expression are required to share the same semantic sense, i.e. homonyms such as *ball* (a. event, b. object) do not count as *l-given-same*. If two nouns can be regarded to be at the same semantic level, as e.g. in a synonymy relation like *lift – elevator*, we propose to mark the "anaphor" as *l-given-syn*.

In cases where the anaphor is superordinate to its antecedent, e.g. in hyponym-hypernym or part-whole relations, the label *l-given-super* is used. An example is (40), already discussed — as (21) — in Section 3.5:

(40) *Bach wrote many pieces for viola. He must have loved string instruments.*
l-given-super

The opposite case occurs if an anaphor is subordinate to its antecedent, e.g. in hypernym-hyponym or whole-part relations. An example, also discussed in Section 3.5. as example (22), of this type is shown in (41). Here, the anaphor cannot be considered *l-given* since it is more informative, and is thus labelled *l-accessible-sub (compare Section 2.2.)*:

(41) *Bach wrote many pieces for string instruments.*
He must have loved the viola.
l-accessible-sub

If there is a succession of two related nouns whose hierarchical relation is not clear-cut in lexico-semantic terms (e.g. *lawyer* in a *court room* scenario), we propose to use the label *l-accessible-other*. The label suggests that the word in question is not as easily available from the discourse context as a repeated word, a synonym, a hypernym or a holonym but still related.

Finally, a noun that is lexico-semantically unrelated to another noun is annotated as *l-new*. As mentioned above, we assume a reset after five information units, i.e. a word can regain the label *l-new*. This procedure implies a shorter period of "cognitive givenness" of lexical items compared to referents (cf. the discussion of concept-givenness versus object-givenness in Section 3.5), which is supported by studies on lexical cohesion. Morris and Hirst (1991), e.g.,

showed that there can be up to three intermediary sentences (which may include about five units of information, i.e. intonation phrases or clauses) between a word and the preceding element of a lexical chain before the words should be considered unrelated.

5. The prosody of referential and lexical givenness

In what follows we will give selected examples in order to formulate some basic hypotheses about the prosodic marking of various kinds of nominal expressions, classified according to both levels of our annotation system presented in Section 4. A wide range of the most important label combinations are given in Appendix A. Furthermore, we will present a short empirical evaluation of the hypotheses on data from two small annotated corpora of spontaneous and read speech (for a more detailed analysis of these data see Baumann and Riester, submitted). We are well aware that the hypotheses, in the way they are formulated, can represent no more than a common baseline that is usually assumed in theoretical studies. Older experimental work, e.g. Terken and Hirschberg (1994), has already provided evidence that we should not expect them to be fully verified. Nevertheless, postulating hypotheses like in the following enables us to compare the degrees of their fulfilment, isolate interesting sets of data, and finally draw conclusions about other factors which may lead to (de-)accentuation.

If an expression denotes a coreference relation with an antecedent, it is usually assumed to be deaccented, irrespective of its classification at the lexical level. A particularly clear case should be a combination of referential and lexical givenness, as in (42).

(42) *Look at <u>the funny dog</u> over there! I LIKE <u>that dog</u>.*
r-given
l-given-same

We can thus formulate our first hypothesis as follows:

Hypothesis I: Given referents *(r-given)* encoded by given lexical items *(l-given-same)* are deaccented.

The examples in the literature suggest that even “lexical novelty” is overwritten by referential givenness, as in (43) (discussed as (1) above), leading to hypothesis II.

(43) A: *Did you see <u>Dr. Cremer</u> to get your root canal?*
B: *Don’t remind me. I’d like to STRANgle <u>the butcher</u>.*
r-given
l-new

Hypothesis II: Given referents *(r-given)* encoded by new lexical items *(l-new)* are deaccented.

New referents are generally assumed to receive accents, especially if their lexical expression is new as well. However, if they are realized with a given noun phrase, they are predicted to be deaccented (hypothesis III), as proposed in the Introduction when we discussed Büring's *Italian* example (2), repeated here as (44).

(44) A: *Why do you study <u>Italian</u>?*
B: *I'm MArried to <u>an Italian</u>.*
r-new
l-given-same

Hypothesis III: New referents *(r-new)* encoded by given lexical items *(l-given-same)* are deaccented.

5.1. Structural reasons for accentuation

The ways in which the accentuation of a (nominal) expression may be realized prosodically in view of its information status may be influenced by various other factors. Apart from mainly pragmatic conditions such as (contrastive or non-contrastive) focussing – compare e.g. Selkirk (2007) – or downplaying (see Section 2.3) there are several structural aspects which may influence an item's accentability. If a nominal expression is part of a predicative construction, e.g., it is more likely to receive an accent. Recall e.g. (21), discussed in Section 3.5 and repeated here as (45), in which the noun *string instrument* is supposed to be deaccented since it is lexically superordinate (l-given-super) to its antecedent *viola*. In (46), on the other hand, *string instrument* occurs in a predication and is assigned a nuclear pitch accent. We propose to annotate a construction as *predicative* in order to account for its special status in terms of prosodic marking.

(45) *Bach wrote many pieces for <u>viola</u>. He must have LOVED <u>string instruments</u>.*
r-generic
l-given-super

(46) *Bach wrote many pieces for <u>viola</u>.*
It is commonly known that [a viola is <u>a STRING instrument</u>.]
r-generic
l-given-super
predicative

The main reason for this accent pattern is that in these cases the semantic relation or referential identity is explicitly asserted rather than taken for granted and is therefore treated as new information.

It is also common that the sheer length of sentences (or utterances, respectively) affects the distribution of accents. Thus, in short phrases a referentially or lexically given item may be accented due to the lack of an alternative. Conversely, long phrases may require accentuation of items which would normally be deaccented. Consider example (47), in which a complete deaccentuation after the anticipated nuclear accent on *praise* (in capital letters) appears to be difficult to produce for rhythmic reasons. In fact, there are several anchor points (marked by small capitals) for potential postnuclear secondary accents (also called *phrase accents*; see Grice et al. 2000) or even nuclear pitch accents.

(47) *Ole was a brilliant tennis player. The local press had nothing but PRAISE for the 27-YEAR-old FAther of a BEAUtiful DAUGHter.*

Moreover, nuclear accents have a special status, both structurally and functionally: since a nuclear accent represents the only obligatory prominence in a phrase (in British School approaches, e.g. Crystal 1969, as well as in Autosegmental-Metrical Phonology, cf. Ladd 2008), speakers and listeners are sensitive to its position. This implies that a prenuclear prominence is often harder to classify as an "accent", since its phonological status is less clearly determined – also because its prominence value is judged *relative* to the surrounding prominences (including the nucleus; see e.g. Jagdfeld and Baumann 2011) and/or according to the listener's expectations about metrical structure (see e.g. Calhoun 2010). That is, prenuclear accents are often placed for rhythmic reasons, irrespective of an item's information status (see Baumann et al. (2007), and Büring's (2006, 2007) "ornamental" accents). Since the accented-deaccented dichotomy can thus be more clearly shown in (potentially) nuclear position, we mostly present examples in which the item in question occurs sentence finally.

5.2. Empirical evidence from two corpora of German speech

Let us examine our hypotheses by looking at two small spoken corpora in German (see Baumann and Riester 2010; Baumann and Riester, submitted). The first corpus consists of six spontaneous monologues, which were labelled for information status (RefLex scheme) and prosody (pitch accents, break indices and boundary tones following GToBI; see Grice et al. 2005) by two independent annotators. The monologues were produced by six native speakers (three female, three male) aged between 27 and 42, whose only instruction was to tell a story of their choice for no longer than five minutes. The stories, which were digitally recorded in a quiet room, consist of 374 intonation phrases comprising 3008 words.

The second corpus comprises ten stories in which ten different target words occurred in various combinations of information status levels. The texts were read by ten native speakers of German (seven female, three male), aged between 22 and 31. All of them originated from the area around Cologne and Düsseldorf. The subjects' task was to read out the texts in a contextually appropriate manner. The semantic and prosodic labelling was done with Praat (Boersma and Weenink 1996) by three independent annotators. So far, we annotated a selection of 23 sentences produced by five different speakers, adding up to 115 sentences in total. They consist of 134 intonation phrases and 1058 words.

Hypothesis I was confirmed for read speech, since 71% of referentially and lexically given non-pronominal items (n=34) did not receive a pitch accent, and only 5% were assigned a nuclear accent. In spontaneous speech, however, only 17% of fully given items (n=41) were deaccented, i.e. the hypothesis was not confirmed. An example from the read speech corpus is shown in Figure 4. Here, the name *Nina* is an immediate repetition of the same referent (context: *Tom findet die Nina überaus hübsch.* 'Tom thinks that Nina is extremely pretty.'). According to hypothesis I, it is deaccented.

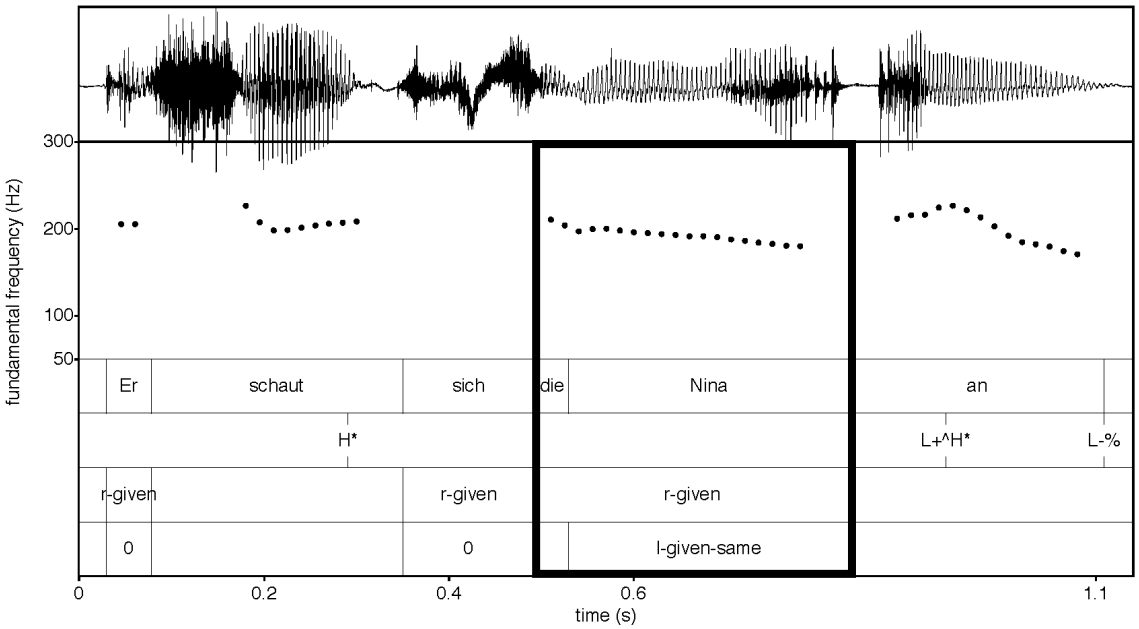


Figure 4: Praat screen shot of the read utterance *Er schaut sich die Nina an* ('He has a look at Nina') with label tiers and pitch contour; the framed phrase *die Nina* (lit. 'the Nina'), labelled for information status as *r-given* and *l-given-same*, is deaccented.

Hypothesis II could not be confirmed for either of the two data types, because referentially given but lexically new nominal expressions (as in the *butcher* example) were deaccented in only 18% of the cases in spontaneous speech (n=17) and 24% of the cases in read speech (n=25).

While it could be shown that referentially *plus* lexically new items were nearly always marked by (nuclear and prenuclear) pitch accents (spontaneous data: 96%; read data: 100%), the combination of *r-new* and *l-given-same* was deaccented in only 31% (spontaneous data, n=13) and 10% (read data, n=10) of the cases. Thus, hypothesis III was not confirmed. Figure 5 shows a typical example of a referentially new but lexically given item in our corpus of read speech. The word in question, *Preis* ('price'), has been mentioned before but denoted a different referent (context: *Über den Preis müssen sie allerdings erst noch verhandeln. Sie sprechen die Dame an.* 'However, they still have to negotiate the price. They approach the lady.'). The second mention of *Preis* is marked by a pitch accent, which contradicts the prediction stated in hypothesis III. Note, however, that the accent is (only) prenuclear and represents an early peak accent, which are two factors that reduce the degree of prominence of the accented item – and consequently its degree of (perceived) newness (see Röhr and Baumann (2011) for a recent investigation of the role accent type and position play in marking and decoding information status in German).

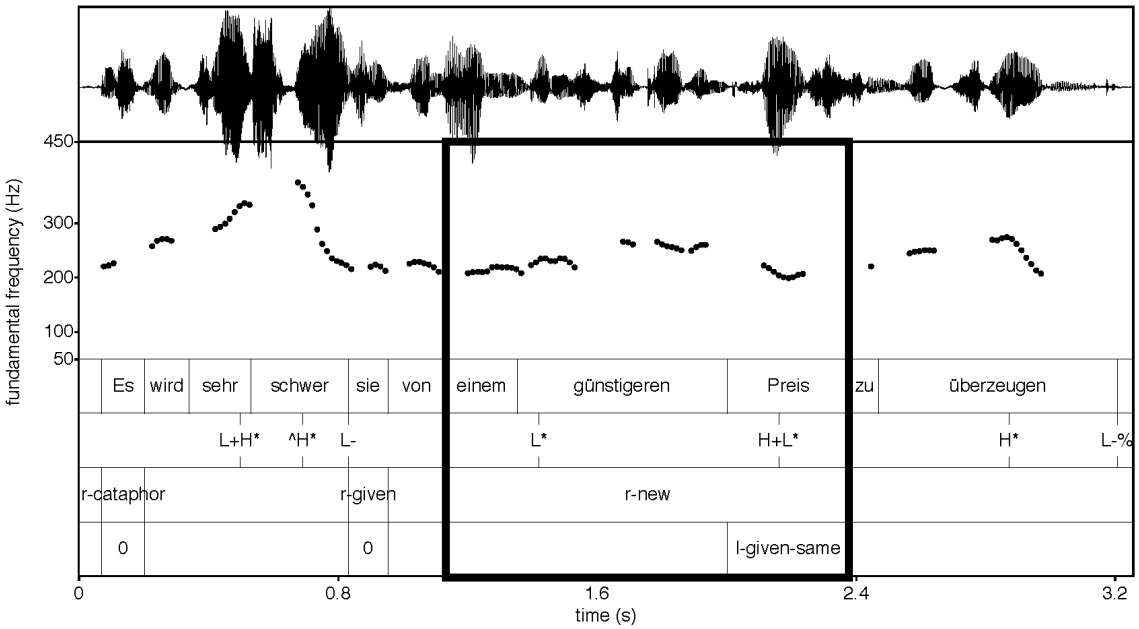


Figure 5: Praat screen shot of the read utterance *Es wird sehr schwer, sie von einem günstigeren Preis zu überzeugen* ('It will be very difficult to convince her of a cheaper price') with label tiers and pitch contour; the head of the framed phrase *einem günstigeren Preis* ('a cheaper price'), labelled for information status as *r-new* and *l-given-same*, is marked by a prenuclear pitch accent.

The analysis of the two corpora confirms the usefulness of both a referential and a lexical level for an investigation of information status in spoken language. It is particularly striking that only some of the hypotheses, which were derived from constructed examples discussed in the theoretical literature, could be confirmed. Nevertheless, there were clear differences between spontaneous and read speech. In read speech, we found more agreement with the hypotheses and a general tendency of a stepwise increase in prosodic prominence from given to new items, which could not be observed in the spontaneous data.

Interestingly, we found more accents than assumed in both kinds of data, particularly in spontaneous speech (similar results were found by Ito et al. (2004) for American English). The most obvious reason is that some of the subjects spoke rather slowly, leading to a large number of short phrases, which in turn contain at least one (nuclear) accent. Furthermore, many referring expressions in the spontaneous data occurred in predications, 89% of which were accented, and in contrastive structures, which received an accent in 79% of the cases. Certainly, however, we need further evidence by larger, statistically evaluated corpora from various sources and with varying degrees of spontaneity.

6. Conclusion and outlook

We have shown that for an adequate analysis of an item's information status in spoken language two levels of givenness have to be investigated: a *referential* and a *lexical* level. This separation is a crucial step towards an adequate classification (and annotation) of nominal expressions occurring in natural discourse. This classification reflects our understanding of all important aspects related to an item's givenness or novelty.

Two main problems in the analysis of discourse structure, information structure and

cognition are, on the one hand, non-standardized vocabulary and, on the other hand, a large gap between theory and actual data. It has been our goal in this paper to address both problems and propose some solutions. In fact, we think that defining criteria for the analysis of abstract notions like *givenness* and various (sub-)classes of information status enables annotators to identify them in naturally occurring language, which is at the same time a step towards a more standardised terminology.

Our research draws on earlier approaches to givenness and information status. We hope we have been able to do justice to their original insights in Sections 1 to 3 but we also hope to have shown that the theoretical backgrounds need to be retold and developed further as we have done in Section 4, in the form of our two-layer RefLex annotation scheme.

In Section 5, we took a closer look at the prosody of the two proposed levels in cases where they interact. As a point of departure, we looked at constructed examples (partly taken from the theoretical literature) in order to arrive at a set of simple hypotheses on the relation between different types and combinations of information status levels and their prosodic realization. A hypothesis thus derived is that coreference leads to deaccentuation, irrespective of the choice of words at the lexical level. If an expression is *not* referentially given, the question of the expression's accentability should depend on the question whether the expression enters in some lexical relation to previously mentioned material or not.

Simple as these hypotheses were formulated, it is not surprising that they could only partly be confirmed in our two small exemplary corpora of read and spontaneous speech since they do not address the issue of other known factors which have an impact on prosodic realization, such as (contrastive) focus, topic shift or predication. Still RefLex annotations enable a much better identification of relevant cases, which can then be further analysed and classified. Interestingly though, the accent patterns found in read speech proved to be more in accordance with the hypotheses than the accent patterns in spontaneous speech.

A crucial next step will be to investigate the *types* of pitch accent used in the described cases since it is likely that relevant information about the degree of an item's activation is encoded in the actual phonological choice of accent, which, in turn, evokes different levels (or degrees) of prosodic prominence. Several studies have shown that a simple dichotomy of accentuation versus deaccentuation is inappropriate for an account of information status in general. Pierrehumbert and Hirschberg (1990) for American English and Kohler (1991) for German were the first who proved that the accent type or, respectively, the tonal configuration, are important cues for inferring a referent's information status as well as higher-level semantic-pragmatic relations. In future work, it is our goal to reinterpret these and more recent results from the literature in terms of the RefLex scheme.

Acknowledgements

This research was partly enabled by the Stuttgart Collaborative Research Centre (SFB) 732 "Incremental Specification in Context", Project A1, and by the Cologne Project GR 1610/5 "Degrees of Activation and Focus-Background Structure in Spontaneous Speech – The Relation between Prosodic Marking and Syntactic and Semantic Structure Building". We acknowledge the kind support of the German Research Foundation (DFG). For their assistance during various steps in the analysis of the spoken data, we thank Christine Röhr, Marlene Biehl, Simon Ritter and Rainer Schmitz. We would also like to thank Hans Kamp as well as three anonymous reviewers for valuable comments on an earlier manuscript of this article.

References

Allerton, David J. 1978 The notion of 'Givenness' and its relation to presupposition and theme. *Lingua* 44: 133-168.

Arnold, Jennifer 1998 Reference form and discourse patterns. Ph.D. dissertation, Stanford University.

Asher, Nicholas and Alex Lascarides 1998 Bridging. *Journal of Semantics* 15 (1): 83-113.

Barwise, John and Robin Cooper 1981 Generalized quantifiers and natural language. *Linguistics and Philosophy* 4: 159-219.

Baumann, Stefan, Johannes Becker, Martine Grice and Doris Mücke 2007 Tonal and articulatory marking of focus in German. In: *Proceedings 16th ICPhS*, Saarbrücken. 1029-1032.

Baumann, Stefan and Arndt Riester 2010 Annotating information status in spontaneous speech. In: *Proceedings Speech Prosody 2010*, Chicago. 100092:1-4.

Baumann, Stefan and Arndt Riester submitted Coreference, lexical givenness and prosody in German. In: Susanne Winkler and Jutta Hartmann (eds.), Information Structure Triggers: Effects on (De)accentuation, Dislocation and Deletion. *Lingua*, Special Issue.

Boersma, Paul and David Weenink 1996 PRAAT, a system for doing phonetics by computer (version 3.4.). Report 132, Institute of Phonetic Sciences of the University of Amsterdam. URL: http://www.fon.hum.uva.nl/praat/

Bos, Johan, Anne-Marie Mineur and Paul Buitelaar 1995 Bridging as coercive accommodation. In: Ewan Klein, Suresh Manandhar, Werner Nutt and Jorg Siekmann (eds.), *Working Notes of the Conference on Computational Logic and Natural Language Processing*, Edinburgh: Human Communications Research Centre.

Büring, Daniel 2006 Focus projection and default prominence. In: Valéria Molnár and Susanne Winkler (eds.), *The Architecture of Focus*. Berlin/New York: Mouton De Gruyter.

Büring, Daniel 2007 Intonation, semantics and information structure. In: Gillian Ramchand and Charles Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces*, 445-474. Oxford: Oxford University Press.

Calhoun, Sasha 2010 the centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language* 86 (1): 1-42.

Chafe, Wallace 1976 Givenness, contrastiveness, definiteness, subjects, topics and point of view. In: Charles Li (ed.), *Subject and Topic*, 25-56. New York: Academic Press.

Chafe, Wallace 1987 Cognitive constraints on information flow. In: Russell Tomlin (ed.), *Coherence and Grounding in Discourse*, 21-52. Amsterdam: John Benjamins.

Chafe, Wallace 1994 *Discourse, Consciousness, and Time*. Chicago/London: University of Chicago Press.

Clark, Herbert H. 1977 Bridging. In: Philip N. Johnson-Laird and Peter C. Wason (eds.), *Thinking: Readings in Cognitive Science*, 411-420. Cambridge: Cambridge University Press.

Clark, Herbert H. and Susan E. Haviland 1974 Comprehension and the given-new contract. In: Roy Freedle (ed.), *Discourse Production and Comprehension*, 1-40. New Jersey: Ablex.

Clark, Herbert H. and Catherine R. Marshall 1981 Definite reference and mutual knowledge. In: Aravind Joshi, Bonnie Webber and Ivan Sag (eds.), *Elements of Discourse Understanding*, 10-63. Cambridge: Cambridge University Press.

Clark, Herbert H. and C.J. Sengul 1979 In search of referents for nouns and pronouns. *Memory and Cognition* 7: 35-41.

Cruttenden, Alan 2006 The de-accenting of given information: A cognitive universal? In: Giuliano Bernini and Marcia L. Schwartz (eds.), *Pragmatic Organization of Discourse in the Languages of Europe*, 311-356. The Hague: Mouton de Gruyter.

Crystal, David 1969 *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Fillmore, Charles J. 1975 Santa Cruz lectures on deixis. Technical Report, Indiana University Linguistics Club, Bloomington.

Götze, Michael, Cornelia Endriss, Stefan Hinterwimmer, Ines Fiedler, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, Ruben Stoel and Thomas Weskott 2007 Information structure. In: Stefanie Dipper, Michael Götze and Stavros Skopeteas (eds.), *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, 147-187. (Working Papers of the CRC 632, Interdisciplinary Studies on Information Structure 7.) Potsdam: Universitätsverlag Potsdam.

Grice, Martine 2006 Intonation. In: Keith Brown (ed.), *Encyclopedia of Language & Linguistics,* Second Edition, Volume 5, 778-788. Oxford: Elsevier.

Grice, Martine, Stefan Baumann and Ralf Benzmüller 2005 German intonation in autosegmental-metrical phonology. In: Sun-Ah Jun (ed.), *Prosodic Typology. The Phonology of Intonation and Phrasing*, 55-83. Oxford: Oxford University Press.

Grice, Martine, D. Robert Ladd and Amalia Arvaniti 2000 On the place of phrase accents in intonational phonology. *Phonology* 17 (2): 143-185.

Gundel, Jeanette, Nancy Hedberg and Ron Zacharski 1993 Cognitive status and the form of referring expressions in discourse. *Language* 69: 274-307.

Halliday, M.A.K. 1967 Notes on transitivity and theme in English, Part 2. *Journal of Linguistics* 3: 199-244.

Halliday, M.A.K. and Ruqaiya Hasan 1976 *Cohesion in English*. London: Longman.

Halliday, M.A.K. and Christian Matthiessen 2004 *An Introduction to Functional Grammar*. 3rd, revised edition. London: Edward Arnold.

Heim, Irene 1982 The semantics of definite and indefinite noun phrases. Ph.D. dissertation,

University of Massachusetts, Amherst.

Ito, Kiwako, Shari Speer and Mary Beckman 2004 Informational status and pitch accent distribution in spontaneous dialogues in English. In: *Proceedings Speech Prosody 2004,* Nara. 279-282.

Jagdfeld, Nils and Stefan Baumann 2011 Order effects on the perception of relative prominence. In: *Proceedings 17th ICPhS*, 958-961. Hong Kong.

Kamp, Hans and Uwe Reyle 1993 *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Dordrecht: Kluwer.

Kaplan, David 1989 Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In: Joseph Almog, John Perry and Howard Wettstein (eds.), *Themes from Kaplan*, 481-563. Oxford: Oxford University Press.

Karttunen, Lauri 1974 Presuppositions and linguistic context. *Theoretical Linguistics* 1: 181-194.

Keenan, Ed and Dag Westerståhl 1997 Generalized quantifiers in linguistics and logic. In: Johan van Benthem and Alice ter Meulen (eds.), *Handbook of Logic and Language*, 837-893. Amsterdam: Elsevier.

Kohler, Klaus 1991 Terminal intonation patterns in single-accent utterances of German: Phonetics, phonology and semantics. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel* 25: 115-185.

Krifka, Manfred 2007 Basic notions of information structure. In: Caroline Féry, Gisbert Fanselow and Manfred Krifka (eds.), *The Notions of Information Structure*, 13-55. (Working Papers of the CRC 632, Interdisciplinary Studies on Information Structure 6.) Potsdam: Universitätsverlag Potsdam.

Kuno, Susumo 1972 Functional sentence perspective: A case study from Japanese and English. *Linguistic Inquiry* 3: 269-320.

Ladd, D. Robert 1980 *The Structure of Intonational Meaning: Evidence from English*. Bloomington: Indiana University Press.

Ladd, D. Robert 2008 *Intonational Phonology*. Cambridge: Cambridge University Press.

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

Levinson, Stephen 2004 Deixis. In: Laurence R. Horn and Gregory Ward (eds.), *The Handbook of Pragmatics*, 97-121. Oxford: Blackwell.

Lyons, John 1968 *Introduction to Theoretical Linguistics*. London: Cambridge University Press.

Morris, Jane, and Graeme Hirst 1991 Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1): 21–48.

Nissim Malvina, Shipra Dingare, Jean Carletta and Mark Steedman 2004 An annotation scheme for information status in dialogue. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbon.

Pierrehumbert, Janet and Julia Hirschberg 1990 The Meaning of Intonational Contours in the Interpretation of Discourse. In: Philip R. Cohen, Jerry Morgan and Martha E. Pollack (eds.), *Intentions in Communication*, 271-311. Cambridge: MIT Press.

Poesio, Massimo 2004 The MATE/GNOME proposals for anaphoric annotation, revisited. In: *Proceedings of the Fifth SIGdial Workshop on Discourse and Dialogue at HLT-NAACL*. Boston.

Poesio, Massimo and Renata Vieira 1998 A corpus-based investigation of definite description use. *Computational Linguistics* 24(2): 183-216.

Prince, Ellen F. 1981 Toward a taxonomy of given-new information. In: Peter Cole (ed.), *Radical Pragmatics*, 223-256. New York: Academic Press.

Prince, Ellen F. 1992 The ZPG letter: Subjects, definiteness, and information-status. In: Sandra A. Thompson and William C. Mann (eds.), *Discourse Description: Diverse Analyses of a Fund Raising Text*, 295-325. Amsterdam: John Benjamins.

Riester, Arndt 2008a A semantic explication of 'information status' and the underspecification of the recipients' knowledge. In: Atle Grønn (ed.), *Proceedings of Sinn und Bedeutung* 12, 508-522. Oslo.

Riester, Arndt 2008b the components of focus and their use in annotating information structure. Ph.D. dissertation, Universität Stuttgart. (*Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung* (AIMS) 14 (2).)

Riester, Arndt 2009 partial accommodation and activation in definites. In: *Proceedings of the 18th International Congress of Linguists*, 134-152. Seoul: Linguistic Society of Korea.

Riester, Arndt and Stefan Baumann 2011 Information structure annotation and secondary accents. In: Stefanie Dipper and Heike Zinsmeister (eds.), *Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*, 111-127. (Bochumer Linguistische Arbeiten 3.)

Riester, Arndt, David Lorenz and Nina Seemann 2010 A recursive annotation scheme for referential information status. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, 717-722. La Valletta.

Röhr, Christine and Stefan Baumann 2011 Decoding information status by type and position of accent in German. In: *Proceedings 17th ICPhS*, 1706-1709. Hong Kong.

Schwarzschild, Roger 1999 GIVENness, AvoidF, and other constraints on the placement of accent. *Natural Language Semantics* 7(2): 141-177.

Selkirk, Elisabeth 2007 Contrastive focus, givenness and the unmarked status of "discourse-new". In: Caroline Féry, Gisbert Fanselow and Manfred Krifka (eds.), *The Notions of Information Structure*, 125-145. (Working Papers of the CRC 632, Interdisciplinary Studies on Information Structure 6.) Potsdam: Universitätsverlag Potsdam.

Stalnaker, Robert 1974 Pragmatic presuppositions. In: Milton K. Munitz and Peter K. Unger (eds.), *Semantics and Philosophy*, 197-214. New York: New York University Press.

Strawson, Peter F. 1950 On referring. *Mind* 59: 320-344.

Terken, Jacques and Julia Hirschberg 1994 Deaccentuation of words representing 'given'

information: Effects of persistence of grammatical function and surface position. *Language and Speech* 37(2): 125-145.

van Deemter, Kees 1994 What's new? A semantic perspective on sentence accent. *Journal of Semantics* 11: 1-31.

van Deemter, Kees 1999 Contrastive stress, contrariety, and focus. In: Peter Bosch and Rob van der Sandt (eds.), *Focus - Linguistic, Cognitive, and Computational Perspectives*, 3-17. (Studies in Natural Language Processing). Cambridge: Cambridge University Press.

van der Sandt, Rob 1992 Presupposition projection as anaphora resolution. *Journal of Semantics* 9: 333-377.

Vieira, Renata and Massimo Poesio 2000 Processing definite descriptions in corpora. In: Simon Botley and Tony McEnery (eds.), *Corpus-based and Computational Approaches to Anaphora*. London: UCL Press.

Yule, George 1981 New, current and displaced entity reference. *Lingua* 55: 41-52.

Appendix A

Combinations of r-level and l-level categories and their assumed (de-)accentuation.

A.1. Combinations with *r-given*

(48) *Look at <u>the funny dog</u> over there! I LIKE <u>that dog</u>.*
r-given
l-given-same

(49) *Ole was a brilliant <u>tennis player</u>.*
The local press had nothing but PRAISE for <u>the athlete</u>.
r-given
l-given-super

(50) *Ole was a brilliant athlete.*
The local press had nothing but PRAISE for <u>the tennis player</u>.
r-given
l-accessible-sub

(51) A: *Did you see <u>Dr. Cremer</u> to get your root canal?*
B: *Don't remind me. I'd like to STRANgle <u>the butcher</u>.*
r-given
l-new

A.2. Combinations with *r-bridging*

(52) *<u>The referee</u> lost conTROL over <u>the football match</u>.*
r-bridging
l-given-super

(53) *Recently at the football match, someone shouted at the refeREE.*
r-bridging
l-accessible-sub

(54) *I walked into my hotel room. The CEILING was very HIGH.*
r-bridging
l-accessible-sub

(55) *Recently, at the football match, they sold the SAUsages for FIVE EUros.*
r-bridging
l-new

(56) *I walked into my hotel room. The chandeLIERS sparked BRIGHTly.*
r-bridging
l-new

A.3. Combinations with *r-unused*

(57) *On my way home, a dog barked at me.*
It made me think of ANna's dog.
r-unused-(un)known[19]
l-given-same

(58) *On my way home, a big German Shepherd barked at me.*
It made me think of ANna's dog.
r-unused-(un)known
l-given-super

(59) *On my way home, a dog barked at me.*
a. *It made me think of Anna's German SHEPherd.*
r-unused-(un)known
l-accessible-sub

b. *It made me think of Anna's BOYfriend.*
r-unused-(un)known
l-new

A.4. Combinations with *r-generic*

(60) *Bach wrote many pieces for viola.*
a. *He must have LOVED the viola.*
r-generic
l-given-same

b. *He must have LOVED string instruments.*
r-generic
l-given-super

(61) *Bach wrote many pieces for string instruments.*
a. He must have loved the viOla.
r-generic
l-accessible-sub

b. *He also liked ANimals.*
r-generic
l-new

A.5. Combinations with *r-new*

(62) *Why do you study Italian?*
a. *I'm MArried to an Italian.*
r-new
l-given-same

b. *I always WANted to learn a Romance language.*
r-new
l-given-super

(63) *Why do you spend so much time in Italy? I'm MArried to a NeaPOlitan.*
r-new
l-accessible-sub

(64) *I'm looking for a FRIEND. He owes me money.*
r-new
l-new

Appendix B

Table: Comparison of Labelling Schemes for Information Status

RefLex (r-level)	**Prince (1981)**	**Lambrecht (1994)**	**Nissim et al. (2004)**	**Götze et al. (2007)**	**Riester et al. (2010)**
r-given	textually evoked	active / given	old	given-active	given
r-given-sit	situationally evoked	situationally accessible	old-general	accessible-situation	situative
r-given-displaced	-- (textually evoked)	textually accessible	-- (old)	given-inactive	-- (given)
r-environment	situationally evoked	situationally accessible	--	accessible-situation	situative
r-bridging	inferable	inferentially accessible	mediated	accessible-inferable	bridging
r-bridging-contained	containing inferable	--	mediated	--	bridging-contained
r-unused-known	new unused	inactive / unused	mediated-general	accessible-general	unused-known
r-unused-unknown	brand-new anchored, containing	unidentifiable anchored, inferentially	new	new	unused-unknown

	inferable	accessible			
r-cataphor	--	unidentifiable unanchored (?)	--	--	cataphor
r-generic	brand-new unanchored	active / given	old-generic	accessible-general	indef-generic, unused-type
r-new	brand-new	unidentifiable / brand-new	new	new	indef-new

[1] Capital letters indicate nuclear accents (defined as the final and structurally strongest pitch accent in an intonation phrase; see Grice 2006), small capitals pre- or postnuclear accents. The constituents in question are underlined.

[2] At this point we will not address the question whether an "item" is the word itself or the thing it refers to. This question will become crucial later on, though.

[3] In Nissim et al. (2004), familiar names and definites are labelled *mediated-general*, in Götze et al. (2007), the corresponding label is *accessible-general.* We cannot go into the details of these taxonomies here.

[4] Among the few prototypical examples are *the moon* and perhaps *President Barack Obama*.

[5] *Bridging* is a more general phenomenon, which can be used for explaining many other processes in which implicit meaning (e.g. rhetorical structure) is reconstructed, cf. Bos et al. 1995; Asher and Lascarides 1998.

[6] In the linguistic rather than philosophical tradition these are called instances of *symbolic deixis* and *gestural deixis* (Fillmore 1975; Levinson 2004).

[7] The difference between indexicals and anaphors is that the former are dependent on the so-called *utterance context*, while the latter depend on the *discourse context*.

[8] In Clark (1977), epithets are explained in terms of bridging. We explicitly treat them as coreferential and not as bridging anaphors. This is not necessarily a contradiction since our class of bridging anaphors does not cover all cases of bridging that have been discussed in the literature. The semantics of epithets is discussed in Riester (2009).

[9] Note that pitch accents cannot occur in postnuclear position.

[10] In Generalized Quantifier Theory (Barwise and Cooper 1981; Keenan and Westerståhl 1997) indefinites and other quantified expressions are translated as functions from predicates to truth values, type <<e,t>,t>. By contrast, we treat indefinites as referential expressions (compare also Strawson 1950: 341). We do not intend to challenge the quantificational aspects of existentials since we think there is room for both a semantic (quantificational) and pragmatic (referential) interpretation of indefinites; a discussion would obviously lead too far here. A practical argument for a referential treatment of indefinites is simply that they can occur as antecedents in a coreference chain. But in order for an expression to be *coreferential* it has to be *referential* in the first place.

[11] For the justification of this, consult Riester (2008a: Ch.3).

[12] Note, however, that these instances of "indefinite bridging" are of a different kind than the definites introduced in Section 3.

[13] Note that the non-coreferential anaphor gets deaccented (just as in the *Italian* example), if the aspect of "extra newness" is missing: *Clinton shares his name with a town; when he finally arRIVED in Clinton, he was late.*

[14] See Appendix B for a comparison between the categories of the RefLex scheme and the categories proposed by previous labelling schemes for information status.

[15] Modifiers can occur prenominally or postnominally and comprise adjectives, restrictive relative clauses as well as prepositional phrases.

[16] In the following examples, antecedents and anaphors are underlined. Labels indicating information status are only attributed to anaphors.

[17] If prosodic information is available, the *intonation phrase* serves as information unit; if not, we regard the *clause* as information unit.

[18] It is a matter of debate how far away a referent's last mention may be to still count as given, and to what extent recency effects can be thought of as a linear function of distance. According to Chafe (1987), a referent's deactivation from an earlier state may have at least two reasons: Apart from a simply gradual (or linear) decay mechanism, a referent may become less accessible due to interference from competing referents mentioned in the intervening discourse (cf. Arnold 1998: 22). The aspect of non-linearity is emphasized by Clark and Sengul (1979) who found a significantly higher availability of a referent mentioned in the previous clause compared to a referent mentioned two clauses back, while there was no significant effect between referents from two clauses back and referents from three clauses back.

[19] Here and in the following examples, the choice of *known* or *unknown* depends on the knowledge of the expected addressee.