RefLex Scheme – Annotation Guidelines

Arndt Riester & Stefan Baumann

1. Introduction

1.1 Short history of the semantics of definite descriptions (partly based on Elbourne 2013: chapters 1 and 3)

The analysis of referring expressions (e.g. Barack Obama, the red bicycle, she, one of the tigers) has a long history in linguistics and philosophy. Our contemporary picture is shaped, in particular, by Frege's view (Frege 1891, 1892) on (definite) descriptions, which has been widely adopted in current linguistic theory, with certain changes and enhancements. One of Frege's claims was that the successful use of definite descriptions, as well as of proper names, presupposes the existence of exactly one individual or entity to which the expression is referring. In other words, it makes no sense to ask whether a statement that contains a definite expression is true or false if the entity referred to by the definite either does not exist or is not unique. Russell (1905) presented a different, and very influential, approach according to which definites assert both existence and uniqueness. (According to him, the sentence 'The king of France is wise.' would explicitly express that there is currently exactly one king of France and that he is wise.) Russell’s – incorrect – view dominated the field until Strawson (1950) restored and refined the presuppositional view (without actually mentioning Frege). Current approaches to definite descriptions (e.g. Elbourne 2013) are still Fregean in the sense that they assume a definite to presuppose the uniqueness of their referent, but some (Neale 1990, Coppock and Beaver 2015) are questioning whether definites actually always presuppose existence. We will not go into the details of this latter point. As for uniqueness, however, there is an issue that we need to discuss. In a lot of cases, it seems clearly wrong that definite descriptions should indicate uniqueness in a strict sense, since many referents of definite descriptions that we encounter in everyday spoken or written discourse are by no means the only ones of their kind in the world; for instance, the table, the cup, the road etc. These expressions seem to behave differently than truly uniquely referring items like the sun, the present Pope, most proper names, or complex descriptions like the square root of 4. Nevertheless, contemporary semantic theory (e.g. Elbourne 2013, Kamp ms) has managed to maintain Frege's uniqueness assumption by relativizing it to smaller domains, contexts or situations. This means that the phrase the table is permissible, and indicates unique reference, if our context of discussion is confined, for instance, to a certain room, or if a unique table is already salient in the ongoing discourse.

A competitor to the “Frege-Strawson theory” (Elbourne 2013: 45) of definite descriptions is the familiarity theory as represented by Christophersen (1939), Heim (1982) or Roberts (2003). On this account, the use of a definite description is permissible if the entity referred to is at least weakly familiar, i.e. entailed by the interlocutors' common ground (Roberts 2003: 306), while indefinites are, by contrast, typically used to introduce new entities. There is a class of counterexamples against the familiarity theory. Hawkins (1978: 130ff.) has called them unfamiliar definites, e.g. expressions like the woman Max went out with last night. Definites of this kind are able to establish the uniqueness of their referent without calling upon the interlocutors' knowledge or the discourse context. Hence, they are able to truly add a new referent to the common ground.
1.2 Information status

In the last section, we gave a very rough overview on the theory of (definite) descriptions. This theory represents the backdrop against which the notion of information status has been developed as a data-oriented – rather than philosophical or semantic – classification of referring expressions (terms) in written and spoken corpora. The idea is to group terms that occur in natural texts into different types, in order to have a closer look at their linguistic properties or to use the classification for a variety of computational purposes. Apart from the different methodologies prevalent in philosophy of language on the one hand, and corpus annotation on the other hand, the two can also be characterized by opposing goals: while formal semantics strives to arrive at a detailed characterization of intricate linguistic phenomena (e.g. definiteness), the goal of linguistic annotation is, rather, to produce a robust classification, which should be reproducible with high reliability by non-experts. The history of information status annotation starts with Prince (1981), while the notion information status itself, to our knowledge, was first used in Prince (1992). The terminology and delineation of the different information status classes in the literature is non-standardized, if not to say chaotic. Major proposals for the classification of referring expressions are formulated in Gundel et al. (1993), Chafe (1994), Lambrech (1994), Poesio and Vieira (1998), Eckert and Strube (2000), Nissim et al. (2004), Götzte et al. (2007), and Riester et al. (2010). The present guidelines introduce the two-dimensional RefLex annotation scheme developed in Baumann and Riester (2012). This paper discusses and integrates diverse aspects of and previous ideas on information status, and also provides a comparison of terminology.

1.3 RefLex scheme

The main idea behind the RefLex annotation scheme is that information status should be analyzed at two levels or dimensions, namely a referential and a lexical (or conceptual) dimension. The roots of this idea lie in theories of information structure (focus-background). Among the approaches to focus some attach particular importance to the distinction between given and new information (e.g. Halliday 1967, Halliday and Hasan 1976, Schwarzschild 1999). The (over-simplified) idea in these approaches is that new information represents the focus, the main point of an utterance, while given information is backrounded. What is important for us at this point is that, as Schwarzschild (1999) points out, the givenness of a constituent must be defined differently for referring expressions¹ and for non-referring expressions, say, (predicate-denoting) nouns, verbs or adjectives, whose information status is restricted to the lexical dimension. (Note that for languages which lack determiners, nouns must simultaneously be analysed as lexical expressions and as referring entities.) As for referring expressions, these are defined as given if and only if they have a coreferential antecedent, i.e. an expression in the previous discourse that refers to the same entity. By contrast, non-referring expressions are defined as given if and only if the expression itself was used in the previous discourse. (Actually, Schwarzschild talks about entailment here: a noun is entailed by a previous occurrence of the same noun or, for instance, by the previous occurrence of a hyponym or synonym.) Since referring expressions (except for pronouns) are typically built from non-referring expressions (a definite description must contain at least a noun), this leads to informationally challenging constellations like the ones shown in (1) and (2). In the following, we

¹ The syntactic domain of a referential expression is either called determiner phrase (DP, in generative linguistics since Abney 1987), or nominal phrase (NP, in wide parts of computational linguistics and many other linguistic areas).
adopt the convention to include the relevant expressions – also called *markables* or *mentions* – in square brackets. Antecedents are underlined.

(1) UN Special Envoy Ahtisaari is making the case for an independence of Kosovo under international control. This would be the only political and economic option for the future of [the Serbian province].

The referring expression 'the Serbian province' in (1) is *referentially given* (*r*-given) since it corefers with 'Kosovo'. At the same time, the word 'Serbian' is *lexically new* (*l*-new) since it is not entailed by the previous discourse.

(2) An earthquake has hit Central Japan. Also in the island state of Vanuatu in the Southern Pacific [two quakes] have been registered.

In contrast to (1), the referring expression [two quakes] in (2) is analyzed as *referentially new* (*r*-new), on the understanding that the two Pacific quakes are not coreferent with the one in Japan. (They are not the same entity.) The word 'quake', however, is either a synonym or hypernym of the previously mentioned noun 'earthquake', and is therefore classified as *lexically given* (*l*-given).

### 1.4 Relative uniqueness and referential information status

As indicated above, we adopt a qualified variant of the Fregean approach to definites, namely that they always refer to a unique entity *within a relevant domain or context*. Our fine-grained classification of referential information status is oriented precisely towards the question of which classes of contexts can be distinguished. This understanding of the use of definites is partly based on ideas developed by Hans Kamp (cf. Kamp, ms). We distinguish the following contexts:

- **The referents of expressions which are unique in the previous discourse context** – because they were mentioned earlier – are labelled as *r*-given. They are typically referred to by means of (third person) pronouns, repetitions or short forms of proper names, and short DPs like 'the man'. (It is important to emphasize that all definitions we give are *pragmatic* in nature. We explicitly avoid classification rules based on word class or syntactic and prosodic features.) Referential givenness describes a relation which is known in the literature as *coreference*, see e.g. BBN Technologies (2007), Pradhan et al. (2007), Krasavina and Chiarosc (2007), Rodriguez et al. (2010), Recasens and Martí (2010). Note, however, that if two expressions in a sequence can be said to be *coreferential*, it is only the second one that must be labelled as *r*-given while the first one might be discourse-new.

- **If an entity does not have a coreferential antecedent but can be understood as unique with respect to a previously introduced situation or scenario**, we will be using the label *r*-bridging. The notion derives from the term *bridging anaphor* (Clark 1977, Poesio and Vieira 1998, Asher and Lascarides 1998, Löbner 1998): a (typically definite) expression signals identifiability; the recipient, however, is unable to identify the referent of the expression itself. As a remedy, she builds a “bridge” in order to link the expression to previously mentioned material. Bridging anaphors are sometimes also called *associative anaphors*. Like *r*-given expressions, bridging anaphors cannot be interpreted – and therefore do not occur – in isolation.
Discourse-new expressions which refer to truly unique entities (in the global context) are labelled as $r$-unused. We distinguish between two subclasses: on the one hand, the label $r$-unused-unknown is assigned to referring expressions which come with a sufficient amount of descriptive material to enable the hearer to create a new discourse referent without any previous knowledge (Hawkins’s unfamiliar definites). On the other hand, $r$-unused-known is a label assigned to globally unique entities which are already known by the hearer. In annotation practice, it will often be difficult to draw a clear line between $r$-unused-unknown and $r$-unused-known because addressees may differ in the amount of their encyclopaedic knowledge, or because the annotator does not know to whom a text was originally addressed. Note also that the $r$-unused labels only apply when a globally unique entity is mentioned for the first time. On each subsequent mention it will count as $r$-given.

Expressions referring to uniquely identifiable entities in the context of a dialogue situation (e.g. visually) receive the label $r$-environment on their first mention. The discourse participants (I, you, we) are always classified as $r$-given-sit. More detailed information will be given below.

Expressions denoting a discourse-new and non-uniquely identifiable referent are labelled $r$-new. In West-Germanic languages, they are typically marked by indefiniteness.

2. R-level

Referring expressions are classified according to the scheme in Table 1.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Contextual class</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$-given-sit</td>
<td>Referents contained in text-external context</td>
</tr>
<tr>
<td>$r$-environment</td>
<td>(communicative situation)</td>
</tr>
<tr>
<td>$r$-given</td>
<td>Referents mentioned in previous discourse context</td>
</tr>
<tr>
<td>$r$-given-displaced</td>
<td></td>
</tr>
<tr>
<td>$r$-cataphor</td>
<td>Discourse-new entities that depend on other concepts / frames in the discourse context</td>
</tr>
<tr>
<td>$r$-bridging</td>
<td></td>
</tr>
<tr>
<td>$r$-bridging-contained</td>
<td></td>
</tr>
<tr>
<td>$r$-unused-unknown</td>
<td>Globally unique entities that are discourse-new and independent of the discourse context</td>
</tr>
<tr>
<td>$r$-unused-known</td>
<td></td>
</tr>
<tr>
<td>$r$-new</td>
<td>Non-unique, discourse-new entities</td>
</tr>
<tr>
<td>+generic</td>
<td>Optional feature for generic entities</td>
</tr>
<tr>
<td>+predicative</td>
<td>Optional feature for predicative expressions</td>
</tr>
</tbody>
</table>

*Table 1: Annotation tags of the r-level*
2.1 Referents contained in the text-external context, the communicative situation or environment (deixis)

2.1.1 r-given-sit: This label applies to an expression whose referent is immediately present in the text-external context. The use of the expression is not accompanied by a pointing gesture (which is why we speak of *symbolic deixis*, cf. Levinson 1983: 65). The following cases can be distinguished:

a. Expression refers to participants in a conversation, i.e. first and second person pronouns.

b. Expression refers to the time of utterance, or time intervals relative to the time of utterance: e.g. *now, last week, 200 years ago*.

c. Expression refers to a unique entity in the visual context or to the location of the utterance itself: e.g. *here, in the fridge, the yellow triangle* (in a map task or visual world paradigm).

We do not annotate adverbial quantifiers like *always, often, usually, every Wednesday, mittwochs 'on Wednesdays', morgens 'every morning’ etc.* because they do not refer to a unique entity.

2.1.2 r-environment: The label applies to an expression whose referent is immediately present in text-external context and which needs to be accompanied by a pointing gesture or gaze (*gestural deixis*).

This category only applies in face-to-face communication. It is often used with demonstratives.

(1) [This chair] (*pointing*) is wobbly.

(2) [He] (*pointing*) is the person in charge.

2.2 Referents present in the previous discourse context (coreference)

2.2.1 r-given (referentially given): The expression is coreferential with an antecedent in the previous discourse. Examples:

a. Repetition of the same referent with the same content expression

(3) I met a man yesterday. [The man] told me a story.

b. Repetition in a reduced, abbreviated or otherwise modified form

(4) John owns a bicycle. He takes [the bike] with him wherever he goes.

(5) Putin hält ein neues Partnerschaftsabkommen mit der Europäischen Union für notwendig. In einem Gastbeitrag für die FAZ betont Putin die Bedeutung der Beziehungen seines Landes mit der EU. (*Putin considers a new partnership agreement with the European Union necessary. In a guest contribution for the Frankfurter Allgemeine Zeitung* Putin stressed the importance of his
country's relations [with the EU].')

c. Pronominal reference

(6) I met a man yesterday. [He] told me a story.

(7) Ghana's President Kufour betonte, dazu dürfe es nie wieder kommen. Zugleich wandte [er] [sich] gegen Forderungen nach Entschädigung. ('Ghana's president Kufour stressed that this must never happen again. At the same time, [he] opposed claims for compensations.')

d. Repetition of the same referent with a different expression

(8) I met a man yesterday. [The traveller] told me a story.

(9) Ole was a brilliant athlete. The local press had nothing but praise for [the tennis player].

(10) The pope's butler was questioned by Vatican investigators. [Paolo Gabriele] has been held under guard at the Vatican since his arrest.

e. Rhetorical devices expressing coreference, e.g. metonymy, synecdoche

(11) Der Westen verdächtigt den Iran, nach Kernwaffen zu streben. Der EU-Außenbeauftragte Solana betonte, die Tür zu Verhandlungen [mit Teheran] bleibe offen. ('The West suspects Iran to strive for nuclear weapons. EU High Representative Solana pointed out that the door to negotiations [with Teheran] remained open.')

d. Abstract anaphors (Asher 1993, Dipper and Zinsmeister 2012, Kolhatkar et al. 2013) referring to facts, propositions, properties, questions, issues, events or states


(13) Paul sings in the shower. John does [it], too.

(14) Is war necessary? [This question] divides people and political parties.

2.2.2 r-given-displaced: If the coreferential antecedent of an expression occurs earlier than the previous five clauses (in written texts) or intonation phrases (if prosodic information is available), the label r-given-displaced is used.

We assume that a referent is valid during the whole discourse, i.e. a referent that has been introduced will not become fully new again, cf. Yule (1981). Nevertheless, the choice of a distance of five units is arbitrary to a certain degree. In annotation tools which allow for an automatic processing of the distance between anaphoric links, the sub-label displaced may be unnecessary.
2.3 Discourse-new entities whose interpretation depends on context

2.3.1 \textit{r-bridging}: This label is used for non-coreferential anaphoric expressions which are dependent on and unique with respect to a previously introduced scenario.

A bridging anaphor (associative anaphor) can only be felicitously used due to the contextual availability of another (non-coreferential) item ("anchor"). The anchor typically establishes a context scenario or situation in which the bridging anaphor plays a unique and perhaps even prototypical role. In some cases, the anchor is not a specific word but rather a whole stretch of text.

(15) The city is planning a new townhall, and [the construction] will start next year.

(16) If \textit{the construction} starts soon, [the new townhall] will be finished already in 2020.

(17) \textit{The referee} lost control over [the football match].

(18) In Ägypten hat [die Regierung] Sicherheitsvorkehrungen getroffen, um Proteste [der Opposition] [gegen das Verfassungsreferendum] zu verhindern. ('In Egypt [the government] has taken safety precautions to prevent protests by [the opposition] [against the constitutional referendum].')

Note that in some other annotation schemes, \textit{partitive} indefinites, i.e. indefinite descriptions that are interpreted as parts of previously introduced entities, may also count as bridging anaphors, e.g. (19).

(19) \textit{A bird is sitting in the tree. It has just lost \textit{a feather}}.

This, however, introduces a considerable degree of uncertainty in the annotation system, since under such a treatment each indefinite expression would have to be considered as a potential bridging anaphor. In the RefLex scheme, only entities which are unique within their scenario (i.e. definites in languages that provide definite articles) qualify as bridging anaphors. The semantic contribution of the whole-part relation is expressed under the notion of \textit{accessibility} at the lexical level of annotation (see Section 3.2.2).

2.3.2 \textit{r-cataphor}: A cataphor is an expression whose referent is established only later on in the text. Cataphoric expressions refer to subsequent items (postcedents).

Nine days after [she] won the women's 800m world championship in Berlin, \textit{Caster Semenya} returned home to the plains of Limpopo.

In [its] ruling, \textit{the supreme court} ordered the election commission to formally dismiss him.

Gestern Abend haben sich die Staats- und Regierungschefs [darauf] verständigt, \textit{die Erklärung um einen Passus zu erweitern}. ('Last night the heads of state and government agreed on [the solution] to extend the declaration by another passage.') (od. agreed on extending... – dann ist es aber nicht mehr kataphorisch...)
2.4 Globally unique descriptions – context-free expressions

2.4.1 r-bridging-contained: This label applies to a non-coreferential anaphoric expression that is anchored to an embedded phrase.

If the anchor is realised as a syntactic argument within a complex bridging anaphor, the entire phrase is marked as r-bridging-contained, as in the following examples.

[The construction of the new townhall] will start next year.

[The opening day of the G20 summit] was threatening to deteriorate.

[Die Staats- und Regierungschefs der 27 EU-Staaten] kommen heute in Berlin zu einem Festakt zusammen. ('The heads of state and government of the 27 EU countries] get together in Berlin today for a ceremonial act.')

[the highest mountain of the Himalayan], [the oldest brother of my office mate] etc.

2.4.2 r-unused-unknown: This label describes a discourse-new expression which is identifiable from its own linguistic description, but which is not generally known.

Put differently, the label is used for an item that the speaker does not expect to be known by the hearer but which the speaker presents in a form that guarantees the uniqueness of its referent.

[The swimming pool of the new townhall] created discontent among the voters.

[The woman Max went out with last night] wore orange socks.

[Martti Ahtisaari, United Nations Special Envoy], is making the case for an independence of Kosovo under international control.

Bei einem Festakt [in the former trading post Elmina in Ghana] wurde an über zehn Millionen Afrikaner erinnert, die als Sklaven verschifft wurden. ('A ceremonial act [in the former trading post Elmina in Ghana] was reminiscent of more than ten million Africans who were shipped as slaves.')

[The pope's butler] was questioned by Vatican investigators.

Caution: The category r-bridging-contained can easily be mixed up with the category r-unused-unknown.

If there is no obvious bridging relation between the outer and the inner concept, then this makes the
label $r$-*unused-known* appropriate. In contrast, the category $r$-*bridging-contained* often describes prototypical relations between the nominal head of a complex phrase and its possessor or nominal argument (e.g. each summit has an opening day, (nearly) each state has a government and so on).

**Permutation test:** Try to dislocate the embedded argument of a complex definite description to the left. If the remaining "anaphor" is still interpretable in relation to the dislocated "antecedent", assign the label $r$-*bridging-contained*. If not, assign the label $r$-*unused-known*.

Example 1:  
[The construction of the new townhall] will start next year.

Permutation:  
(there is) [a new townhall], (and) $\odot$ [the construction] will start next year.

Result:  
Assign the label $r$-*bridging-contained* to the phrase [The construction of the new townhall].

Example 2:  
[The swimming pool of the new townhall] created discontent among the voters.

Permutation:  
(there is) [a new townhall], (and) ??[the swimming pool] created discontent among the voters.

Result:  
Assign the label $r$-*unused-known* to the phrase [The swimming pool of the new townhall].

2.4.3 **$r$-*unused-known*:** This label applies to unique discourse-new expressions which are generally known, i.e. to items which the speaker assumes the hearer (or the expected audience) to be familiar with. The item is neither derivable from the current discourse, nor is it visible.

[The Pope] wore orange socks.

[Der Iran] will an seinem Atomprogramm festhalten.  
('[Iran] intends to hold on to its nuclear programme.')

2.5 **Discourse-new expression with non-unique description**

2.5.1 **$r$-new:** Expressions introducing a new, non-unique referent are labelled $r$-*new*.

In West-Germanic languages, new referents are typically introduced by indefinite expressions. In languages without morphosyntactic marking of (in-)definiteness, all discourse-new referring expressions that are not uniquely identifiable are labelled $r$-*new*.

I’m looking for [a friend]. He owes me [money].

Why do you spend so much time in Italy? I’m married to [a Neapolitan].
[Party supporters] have said they have [enough support in parliament] to elect [a new prime minister].

('A military spokesman] confirmed [explosions] and the death [of at least two soldiers].')

2.6 Additional Feature +generic

+generic: This additional feature/tag is assigned to referring expressions denoting a class, an abstract entity, or a non-specific or hypothetical entity (cf. Krifka et al 1995, Mari et al. 2013, Friedrich et al. 2015).

In Germanic languages, generic expressions may be marked by articles or take the form of (singular or plural) bare nouns. In order to determine whether a bare noun is uniquely or non-uniquely referring, provisionally insert the definite and the indefinite article. Depending on which one preserves the meaning of the bare noun more appropriately, choose the label r-unused (unique reference) or r-new (non-unique reference).

The feature +generic may combine with all r-categories. Generic entities (and only these) can recur in the indefinite form after having been previously mentioned. In the following, we list different types of generic expressions.

a. Class

[A cat] is a mammal.  r-new+generic
[The lion] is a huge animal.  r-unused-known+generic
[Lions] are huge animals.  r-new+generic

As a fan [of fantasy fiction] it's been entertaining watching mainstream cultural critics' baffled responses to Game of Thrones.  r-unused-known+generic

b. Abstract entities

Nach zwei Jahren erhielt er endlich [Asyl].  r-unused-known+generic
('After two years he was finally given [asylum].')

Der hessische Ministerpräsident Koch hat [vor Mindestlöhnen] gewarnt.  r-new+generic
('The Hessian governor Koch warned [of minimum wages].')

Baden-Württembergs Ministerpräsident Oettinger wandte sich ebenfalls [gegen Mindestlöhne].  r-given+generic
('The Baden-Wuerttemberg governor Oettinger also opposed [minimum wages].')
c. Non-specific or hypothetical entities

Indefinite generic phrases may express non-specificity instead of class reference or abstraction.

Kanzlerin Merkel hat [vor einem Scheitern der Reformbemühungen] gewarnt.

('Chancellor Merkel warned [of a failure of the reform efforts.]')

[Ein Scheitern] wäre ein historisches Versäumnis, betonte sie.

('A failure] would be a historic lapse, she pointed out.')

[Druck und Einschüchterung] würden nichts bewirken, erklärte Außenminister Mottaki.

('Pressure and intimidation] would have no effect, Foreign Minister Mottaki declared.')

Often, expressions indicate hypothetical referents, i.e. no concrete referent is introduced in the discourse.

I'm looking for [a doctor]. (any doctor: non-specific) vs. I'm looking for [a doctor]. He owes me money. (specific)

They have enough support in parliament to elect [a new prime minister.]

d. Negation

Entities in the scope of a negation operator are usually not instantiated. We treat them like generic entities.

I don’t have [a car].

I have [no car].

2.7 Additional feature +predicative

The label +predicative is assigned to markables which express properties of other referring expressions that are often but not exclusively indicated by the presence of a copula verb. The label potentially combines with any other label category. However, we refrain from combining the labels +predicative and +generic.

In this company, Peter is [the boss].

They elected him [President of the United States].

The price rises [to seven dollars].

[As a child], Peter had lived in Brandenburg.

Er ist [gelernter Friseur].

('He is [a trained hairdresser].')
2.8 Decision tree for the r-level

![Decision tree for r-level](image)

Figure 1: Decision tree for r-level

2.8 Annotation conventions for the r-level

In the following, we present examples from different written and spoken registers that were analysed using different annotation tools (SALTO, Burchardt et al. 2006; Slate, Kaplan et al. 2012; EXMARaLDA, Schmidt & Wörner 2014). The SALTO data were syntactically preprocessed using different parsers. By this we intend to raise the reader's awareness that tool and preprocessing choices may have an influence on the actual annotation process, which is unrelated to the theoretical-linguistic properties of the RefLEx system.

2.8.1 Annotation units

R-labels are assigned to referring expressions, in particular phrases that occur as verbal arguments. Depending on the syntactic framework chosen, such phrases are analysed as DPs (determiner phrases) or NPs (noun phrases).
Discourse particles (e.g. even, only, also) do not belong to the referring expression and are therefore not part of the markable. By contrast, quantifiers/determiners do belong to the markable. Note that there are definite quantifiers such as every and all, and indefinite quantifiers like some, many, most, a few, one, two, three etc.

2.8.2 Complex phrases
Referring expressions, especially in formal, written language, are often nested inside each other. In such cases, we follow the convention to assign one r-label to each referring expression.

Do not forget to label possessive pronouns!

SWR: Frau Lemke, eine Woche ist es jetzt her, dass
Ihr Parteifreund aus Baden-Württemberg, Winfried Kretschmann, dem Asylgesetz der Großen Koalition zugestimmt und ihm damit zu einer Mehrheit im Bundesrat verholfen hat. Gegen den erklärten Willen

Figure 4: Embedded possessive pronoun
In the absence of syntactic analyses, the nesting of phrases can be modelled by creating several annotation layers as shown in the following figure (EXMARaLDA annotation tool; also possible e.g. in the speech analysis tool Praat (Boersma & Weenink 2012).

Figure 5: Annotation of nested phrases without syntactic analysis

### 2.8.2.1 Prepositional phrases

If a referring expression starts with a preposition, we assign an r-label to the entire PP. There are several arguments in favour of this decision: (1) The preposition linguistically 'belongs' to the referring expression rather than to the embedding verb. (2) Some languages (notably German) display cases of conflated preposition-determiners (e.g. zum, im, ins), which leave no other choice than to label the PP. (3) Some languages (e.g. Finnish) make ample use of case endings rather than prepositions (e.g. rakenne-sa ‘in the building’), again forcing label assignment to the entire locative unit. Generally, this means that r-labels should be assigned at the highest node of the referring expression (no matter whether it is analysed as a PP, DP or NP).

Many prepositions are meaningless because they are subcategorised by the embedding verb, which means that there is no semantic difference between the PP and the DP/NP:

She asked [pp for [dp the bill]], (identical “referents” for i and j)

However, for instance with locative or temporal PPs, it makes sense to distinguish different levels of embedding (and thus to potentially assign more than one label), as in the following example where i refers to a tree, and j refers to a location.

There was a tree in the garden. Paul sat [behind [the tree]], r-given

a) [It]: r-given was a maple tree.
b) It was shady [there], r-given.

### 2.8.2.2 Appositions and relative clauses

Appositions and relative clauses are grouped together with the expressions they modify, i.e. they are not annotated separately.

What are the Panama Papers? The Panama Papers are 11.5 million documents – or 2.6 terabytes of data – provided by an unnamed source to a German newspaper, **Süddeutsche Zeitung**, more than one year ago. They were

Figure 6: Referring expression with apposition
2.8.2.3 Anaphoric links

R-given and r-bridging anaphors typically have antecedents. But only r-given expressions may form coreference chains. The link originating from an anaphor should always be drawn to the immediately preceding antecedent (in annotation tools which provide such a functionality).

Th.0.: Also der BND hat ja berichtet, dass der Weg für den Islamischen Staat ist, mögliche Terrorkämpfer nach Europa zu bringen. Dafür brauche ich keine Flüchtlingsströme, sondern der IS hat die Möglichkeiten die hier auf ganz anderem Wege zu platzieren. Und niemand kann ausschließen, dass hier auch Terrorakte des Islamischen Staates passieren. Aber

Figure 7: Coreference chain: [for the Islamic State] ... [the IS] ... [the Islamic State]

2.8.2.4 Discontinuous referring expressions

Depending on the syntactic framework provided, there may be crossing edges and therefore syntactic nodes for discontinuous phrases. In Figure 8, we see a verbal complex (verschrieben werden 'be prescribed') that separates a PP in two halves.²

Figure 8: Discontinuous phrase, analysed with crossing edges: 'may not be prescribed to women under 30 who want to go on the pill for the first time'

If the syntactic framework does not provide such a node (or if there is no syntactic analysis available), the two parts of the phrase receive the same label, plus the respective flags +left and +right.

² The little flag is used to represent the +generic feature.
SWR: Und der Bremer Weg? Also der hat ja, ist ja auf enormen Widerstand natürlich bei den Fußballvereinen gestoßen, ist ja völlig klar.

*Figure 9*: Discontinuous phrase [auf enormen Widerstand bei den Fußballvereinen] / [on fierce resistance on behalf of the football clubs], marked with a special (non-coreferential) link (Slate tool)

### 2.8.2.5 Coordination

Coordinations of referring expressions have to be annotated on (at least) two tiers: Each of the conjuncts receives a label on one tier, and the coordination itself receives another one on an extra tier (Figure 10 on the left).

*Figure 10*: Anaphoric retrieval (on the right: Koalition 'coalition') of a coordinated expression (on the left: Union und SPD 'Conservative Party and Social Democrats')

Ecke. Habe die Recherche damals gemacht, bin natürlich auch, habe immer die SPD dafür kritisiert und dann mit Kurt Beck begonnen, die Anträge zu stellen bei der Kommission, um ein Übergleiten und eine Privatisierung vorzunehmen. Die dann in der, in dem ersten

*Figure 11*: Annotating a coordinated expression as well as the two conjuncts
The reason why such a rather complicated measure is necessary is that both the coordination itself (Figure 10 on the right: Union und SPD = Koalition) or one of its conjuncts (Figure 11 on the right: Bund) can be picked up anaphorically.

![Figure 11: Anaphoric retrieval (on the right: Bund 'state') of a conjunct (on the left: Bund und Länder 'state and provinces')](image)

SWR: Das ist eine lange Geschichte. Sie haben es erwähnt, mit dem Nürburgring, auch Haus und Zweibrücken, drei Behilfeverfahren. Was muss Ihre Regierung tun, was müssen Sie als Grüne tun, um bei den Bürgern wieder Vertrauen herzustellen?

E.L.: Also ich glaube erstens, wenn dieser, diese Bescheide da sind und wir alle wissen, wie die Richtung ist, dann haben wir da mal Ruhe drin. Weil wir haben jetzt sehr lange gewartet und viel, viel Unsicherheit gehabt. Und wenn Unsicherheit da ist und eine Regierung nicht klar sagen kann, wie das jetzt weitergeht

mit ihren Vorschlägen, weil sie immer darauf wartet, dass die Kommission zusammentritt, dass dann können Sie nicht gerade Sicherheit vermitteln. Jetzt haben wir Konzepte für die entsprechenden Regionen. Wir arbeiten daran schon sehr lange. Aber es wird noch mal deutlicher, wie diese dann auch

![Figure 12: Aggregation, using a special link](image)

2.8.3 Direct speech

Elements which occur in direct speech are not coreferential with elements that have occurred before the direct speech section. Thus, direct speech is treated as separate, embedded, discourse.
3. L-level

Lexical expressions are classified according to the scheme in Table 2.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>l-given-same</td>
<td>active, i.e. salient concepts</td>
</tr>
<tr>
<td>l-given-syn</td>
<td></td>
</tr>
<tr>
<td>l-given-super</td>
<td></td>
</tr>
<tr>
<td>l-given-whole</td>
<td></td>
</tr>
<tr>
<td>l-accessible-sub</td>
<td>semi-active, i.e. derivable concepts</td>
</tr>
<tr>
<td>l-accessible-part</td>
<td></td>
</tr>
<tr>
<td>l-accessible-stem</td>
<td></td>
</tr>
<tr>
<td>l-new</td>
<td>inactive concepts</td>
</tr>
</tbody>
</table>

Table 2: Annotation tags of the l-level

The lexical level applies to the word domain, more specifically to content words such as nouns, adjectives and verbs. Pronouns and other functional categories are not annotated at the l-level. 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.

3.1 l-given: The label expresses that the markable is identical (-same), synonymous (-syn), hypernymic (-super), or holonymic (-whole) with/to an expression in the discourse context.

3.1.1 l-given-same: Recurrence of same (content) word

Look at the funny dog over there! I like that [dog].

Look at the funny dog over there! It makes me think of Anna's [dog].

Der Iran will an seinem Atomprogramm festhalten. Der Westen verdächtigt den [Iran], nach Kernwaffen zu streben.
('Iran intends to hold on to its nuclear programme. The West suspects [Iran] to strive for nuclear weapons.')

Barack Obama was expected to press Merkel on the pooling of liability for single currency countries' debt. But there is no chance of [Merkel] agreeing to underwrite the debt of other European countries for the foreseeable future.

3.1.2 l-given-syn: Relation between words at the same hierarchical level (synonyms)

John owns a bicycle. You absolutely need a [bike] if you work at Stanford.

(van Deemter 1999)
Der Iran will an seinem Atomprogramm festhalten. Außenminister Mottaki sagte, auch die schärfsten Strafmaßnahmen seien zu schwach, um die iranische Nation zu einem Verzicht auf ihre [Nuklear-Politik] zu zwingen. ('Iran intends to hold on to its nuclear programme. Foreign Minister Mottaki said that even the most severe sanctions were too weak to force the Iranian nation to abandon its [nuclear policy].')

Union und SPD haben eine Teileinigung zur Neuregelung des Niedriglohnsektors erreicht. Man habe Einigkeit über ein Kombilohnmodell für junge Arbeitslose erzielt. ('The Conservative Party and the Social Democrats reached a sub-agreement on a revision of the low-pay sector. They said they [achieved] a consensus on a combination wage model for the young unemployed.')

Putin hält ein neues Partnerschaftsabkommen mit der Europäischen Union für notwendig. In einem Gastbeitrag für die FAZ betont Putin die Bedeutung der Beziehungen seines Landes mit der [EU]. ('Putin considers a new partnership agreement with the European Union necessary. In a guest contribution for the Frankfurter Allgemeine Zeitung Putin stressed the importance of his country's relations with the [EU].')

The PC is ready to obtain data and [receive] alarms from an external system.

3.1.3 l-given-super: Word is lexically superordinate to previous word (markable is a hypernym)

Do you like dogs? I like all [animals].

Why do you study Italian? I always wanted to learn a Romance [language].

John owns a bicycle. You absolutely need a [vehicle] if you work at Stanford.

Die 27 Staats- und Regierungschefs der EU wollen die "Berliner Erklärung" unterzeichen. In dem [Dokument] legt sich die EU auf Reformen bis Frühjahr 2009 fest. ('The 27 heads of state and government of the EU want to sign the Berlin Declaration. In the [document] the EU commits itself to reforms until spring 2009.')

The outcomes of the Rio+20 Earth Summit will be very different to those of the past but that doesn't mean the [summit] will fail.

3.1.4 l-given-whole: Word is lexically superordinate to previous word (markable is a holonym)

Why do you spend so much time in Naples? It's my favourite city in [Italy].

Britain is building alliances to block a legally binding charter of fundamental rights. With the Tories on the attack over alleged government acquiescence in an embryonic
"constitution" for the [EU], it emerged yesterday that there is a wide opposition to the maximalist version of the project.

3.2 l-accessible: The markable is hyponymic (-sub) or meronymic (-part) to an expression in the discourse context, or a recurring stem or element in a compound (-stem).

3.2.1 l-accessible-sub: Word is lexically subordinate to previous word (markable is a hyponym)

Do you like animals? I like all [dogs].

John owns a vehicle. But you absolutely need a [bicycle] if you work at Stanford.

Akademiker in Deutschland zahlen nach einer Untersuchung über Steuern weniger an das Hochschulsystem zurück, als sie an Ausbildungsleistungen erhalten haben. Besonders deutlich sei dies bei den [Ärzten].

('According to a study on taxes, academics in Germany refund less money to the higher education system than they have received as training aid. This was particularly obvious in the case of [physicians].')

3.2.2 l-accessible-part: Word is lexically subordinate to previous word (markable is a meronym)

Why do you spend so much time in Italy? I have a friend in [Naples].

I walked into my hotel room. The [ceiling] was very high.

Germany's chancellor is under pressure to soften her hardline stance on the austerity measures Europe imposed on indebted [eurozone] members.

In a number of cases, l-accessible-part can also be used to describe prototypical occurrences within a frame or scenario.

We went to a restaurant last night. John argued with a [waiter].

A press conference took place at the US Congress. I spoke to a [journalist] afterwards.

3.2.3 l-accessible-stem: A recurring stem or element in a compound

Why do you study Italian? I'm married to an [Italian]. (Büring 2007)


('According to an inquiry for the time between June 2006 and February this year, aircrafts from EU countries had [flown] across restricted zones five times.')
Die Picknickdecke ist kariert. Ich hasse [Karomuster].
('The picknick blanket is checked, I hate [chequers].')

3.3 l-new: All expressions that are unrelated to the existing discourse receive the label l-new.

3.3.1 l-new: Word is not related to another word within the last five intonation phrases (if prosodic information is available) or clauses (in written texts)

[Pakistan's] [highest] [court] has [declared] that the country's [prime minister] is [disqualified] from [office].

Der [Iran] will an seinem [Atomprogramm] [festhalten].
('[Iran] intends to [hold on] to its [nuclear programme].')

I walked into my hotel room. The [chandeliers] sparked brightly.
(= no prototypical part of a hotel room)

3.4 Annotation conventions for the l-level

3.4.1 Annotation units
The basic annotation units at the l-level are content words such as nouns, full verbs, adjectives and adverbs. In contrast to the r-level (which includes determiners and prepositions), the l-labels are attached as low as possible, i.e. at word level. However, compounds are treated as single units.

3.4.2 Particle verbs
The verb and its particle receive the same label, plus a respective feature +left or +right.

3.4.3 Hierarchies
If several labels are possible, the following order applies:
l-given-same > l-given-syn > l-given-super > l-given-whole > l-accessible-stem > l-accessible-sub > l-accessible-part

3.4.4 Displacement
We assume a decay of cognitive activation of elements at the l-level after five intonation phrases or clauses (in contrast to discourse entities at the r-level, cf. section 2.2.2). After this threshold an element will count as l-new again.

3.4.5 Proper nouns (names) and common nouns
Meronymic relations (l-given-whole, l-accessible-part) are not annotated between a proper noun/name and a common noun.

Germany has a [population] of 80 million people.  (l-new, not l-accessible-part)
Klose is the oldest player in his [team].

(l-new, not l-given-whole)

However, we do annotate hyponymic relations (only l-given-super) between a proper noun/name and a common noun.

Germany and other [countries] will return to Central European Time on October 29.

(l-given-super)

3.4.6 Cross-categorial relations

We do not assume any lexical relations across word classes, except for the label l-accessible-stem (see 3.2.3).

Bibliography

Abney, Steven (1987), 'The English noun phrase in its sentential aspect', (Massachusetts Institute of Technology).
Christoffersen, Paul (1939), The articles: A study of their theory and use in English (Copenhagen: Munksgaard).
Coppock, Elizabeth and Beaver, David (2015), 'Definiteness and determinacy', Linguistics and Philosophy, 38, 377-435.
Elbourne, Paul (2013), Definite descriptions (Oxford University Press).
Friedrich, Annemarie et al. (2015), 'Annotating genericity: a survey, a scheme, and a corpus',
Proceedings of the 9th Linguistic Annotation Workshop (LAW), Denver, 21-30.


Halliday, Michael (1967), 'Notes on transitivity and theme in English (part 2)', Journal of linguistics, 3 (02), 199-244.

Halliday, Michael and Hasan, Ruqaiya (1976), Cohesion in English (London: Longman).


Heim, Irene (1982), 'The semantics of definite and indefinite noun phrases', (University of Massachusetts).


Mari, Alda, Beyssade, Claire and Del Prete, Fabio (2013), Genericity (Oxford University Press).


Russell, Bertrand (1905), 'On denoting', Mind, 14, 479-93.
Schwarzschild, Roger (1999), 'GIVENness, AvoidF and other constraints on the placement of accent', Natural Language Semantics, 7 (2), 141-77.
Strawson, Peter (1950), 'On referring', Mind, 59, 320-44.

Appendix: Differences between coreference annotation in Reflex and in OntoNotes