Universität Stuttgart

Arndt Riester and Stefan Baumann

The RefLex Scheme – Annotation Guidelines

SinSpeC

Working Papers of the SFB 732
"Incremental Specification in Context"

SinSpeC 14 (2017) ISSN 1867-3082

Universität Stuttgart

SinSpeC issues do not appear on a strict schedule.

Volume 14 (2017)

Series Editor: Jonas Kuhn
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
D-70569 Stuttgart

Published by **O**nline **P**ublikationsverbund der **U**niversität **S**tuttgart (OPUS)

Published 2017

ISSN **1867-3082**

About *SinSpeC*

SinSpeC are the Working Papers of the Sonderforschungsbereich (SFB) 732 "Incremental Specification in Context". The SFB 732 is a collaboratory research center at the University of Stuttgart and has been funded by the German Research Foundation (DFG) since July 1, 2006.

The SFB 732 brings together scientists from the areas of linguistics, computational linguistics and signal processing at the University of Stuttgart. Their common scientific goals are to achieve a better understanding of the mechanisms that lead to ambiguity control/disambiguation as well as the enrichment of missing/incomplete information and to develop methods that are able to fully describe these mechanisms.

For further information about the SFB please visit:

http://www.uni-stuttgart.de/linguistik/sfb732/

SinSpeC aims at publishing ongoing work within the SFB in a fast and uncomplicated way in order to make the results of our work here known to the scientific community and strengthen our international relationships. It publishes papers by the staff of the SFB as well as papers by visiting scholars or invited scholars.

SinSpeC is available online on the above website.

A 'Print on Demand' version can be ordered at the same address.

Contact Information:

Director of the SFB 732:

Prof. Dr. Jonas Kuhn
jonas@ims.uni-stuttgart.de

Coordinator of the SFB 732:

Dr. Sabine Mohr
sabine@ifla.uni-stuttgart.de

SFB 732
Universität Stuttgart
Keplerstr. 17
D-70174 Stuttgart

Phone: 0711/685-83115
Fax: 0711/685-83120

The RefLex Scheme – Annotation Guidelines

Arndt Riester (Universität Stuttgart)
arndt.riester@ims.uni-stuttgart.de

Stefan Baumann (Universität zu Köln)
stefan.baumann@uni-koeln.de

Abstract: *The purpose of the RefLex annotation scheme (Baumann and Riester 2012) is the two-dimensional analysis of textual or spoken corpus data with regard to referential information status (including coreference and bridging) as well as lexical information status (semantic relations). We provide some linguistic-philosophical background followed by detailed guidelines, which can be used in combination with various annotation tools.*

Keywords: anaphora, bridging, coreference, corpus annotation, information status, information structure, referring expressions, referential and lexical givenness, semantic relations

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 Freges'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 many 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 referents of the phrases *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* to both speaker and addressee – "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), Lambrecht (1994), Poesio and Vieira (1998), Eckert and Strube (2000), Nissim et al. (2004), Götze 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 central idea behind the RefLex annotation scheme is that information status should be analysed 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* (mostly, *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 *backgrounded.* 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 expressions[1] 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 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 analysed as *referentially new (r-new)*, on the understanding that the two Pacific quakes are not coreferential 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] The syntactic domain of a typical referring 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).

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), in particular his notion of the *articulated context*. 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, however, to emphasize that all definitions we give are *semantic-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 Chiarcos (2007), Rodríguez 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 called *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 (and non-referring terms) are classified according to the scheme in Table 1.

Table 1: Annotation tags of the r-level	
Tag	Contextual class
r-given-sit	Referents contained in text-external context (communicative situation)
r-environment	
r-given	Referents mentioned in previous discourse context
r-given-displaced	
r-cataphor	Discourse-new entities that depend on other expressions in the discourse context
r-bridging	
r-bridging-contained	Globally unique entities that are discourse-new and independent of the discourse context
r-unused-unknown	
r-unused-known	
r-new	Non-unique, discourse-new entities
r-expletive	Non-referring expressions
r-idiom	
+generic	Optional features
+predicative	

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).

d. Vocatives, e.g. at the beginning of a conversation:

(3) [Herr Maas], wie geht es [Ihnen]?

[Mr. Maas], how are [you] doing?

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.

Note, furthermore, that it makes sense to draw coreference links (Section 2.9.5) between recurring deictic expressions, given that they really denote the same entity or set. The annotator should be aware that there is temporal progression (relevant e.g. for *now*), and deictic pronouns may come with different radii. *Here* may refer to a tiny spot, the room, to the entire country, continent or even planet. Likewise, in a single text, there may be different simultaneous uses of *we*, which refer to different groups, as shown in Figure 1. These uses can only be identified from the context.

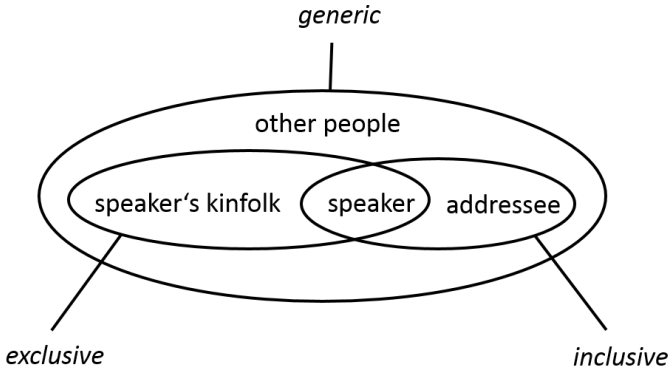

Figure 1: Different uses of 'we'

2.1.2 r-environment: The label applies to an expression whose referent is immediately present in the 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.

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

(5) [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

(6) I met <u>a man</u> yesterday. [The man] told me a story.

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

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

(8) 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

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

(10) Ghanas Präsident 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. (Lit. '[He] turned [himself] against (...)')

d. Repetition of the same referent with a different expression *(epithets)*

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

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

(13) 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*

(14) 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.

In (14), both *Iran* and *Teheran* are meant to refer to the Iranian government. This is why they are annotated as coreferential here, while under normal circumstances, of course, Teheran is a part of Iran.

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

(15) Paul sings in the shower. Mary finds [that] weird.
(*[that]:* the fact that John sings in the shower)

(16) Paul sings in the shower. John does [it], too.
(*[it]:* the property/activity of singing in the shower)

(17) 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 r-cataphor: A cataphor is an expression whose referent is established only later on in the text. Cataphoric expressions are coreferential with subsequent items *(postcedents)*.

(18) Nine days after [she] won the women's 800m world championship in Berlin, Caster Semenya returned home to the plains of Limpopo.

(19) In [its] ruling, the Supreme Court ordered the election commission to formally dismiss him.

(20) Gestern Abend haben sich die Staats- und Regierungschefs [darauf] verständigt, die Erklärung um einen Passus zu erweitern.

Last night the heads of state and government agreed on extending the declaration by another passage. (Lit. '(...) have agreed on [this]: to extend the declaration (...)')

2.3.2 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. Bridging anaphors can be thought of as expressions which carry their antecedent as a silent (elliptical) argument.

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

(22) If the construction starts soon, [the new townhall] will be finished already in 2020.

(23) The referee lost control over [the football match].

(24) 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 accounts (e.g. Asher 1998: 83), indefinite descriptions that are interpreted as parts of previously introduced entities, or as being involved in previously mentioned events, may also count as bridging anaphors, e.g. (25).

(25) A bird is sitting in the tree. It has just lost [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. Furthermore, as the above examples show, bridging relations are not restricted to whole-part combinations. 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 *accessibility* at the lexical level of annotation (see Section 3.2.2).

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.

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

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

(28) [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.

(29) [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.

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

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

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

(33) Bei einem Festakt [im ehemaligen Handelsposten 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.

(34) [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 or r-unused-known* (see below)*.*

If there is no obvious bridging relation between the outer and the inner concept, then this makes the label *r-unused* 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 one of the *r-unused labels.*

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

Permutation: [A new townhall] (will be built, and) ☺ [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: (They built) [a new townhall], (and) ??[the swimming pool] created discontent among the voters.

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

Example 3: John says that we should ask [his hairdresser].
(embedded possessive pronoun)

Permutation: [John/He] says that we should ask ??[the hairdresser].

Result: Assign the label *r-unused-unknown* to the phrase [his hairdresser].

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.

(35) [The Pope] wore orange socks.

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

Names that come without descriptions will typically fall in this category even if, strictly speaking, there might be other persons in the world which coincidentally bear the same name. In using such a name, the speaker ignores the existence of potential name twins.

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*.

(37) I'm looking for [a friend]. He owes me [money].

(38) Why do you spend so much time in Italy? I'm married to [a Neapolitan].

(39) [Party supporters] have said they have [enough support in parliament] to elect [a new prime minister].

(40) [Three people] walked across the street.

(41) [Ein Militärsprecher] bestätigte [Explosionen] und den Tod [von mindestens zwei Soldaten].

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

Note that we also count bare mass nouns and (wh-)question pronouns as indefinite.

(42) Elizabeth poured [sugar] in her coffee.

(43) There was [police] in front of the building.

(44) [Who] is going to carry these boxes upstairs?

2.6 Non-referring noun phrases

Not all terms that have the syntactic category of NP/DP/PP are actually referring to some discourse entity. We distinguish two classes of non-referring expressions: expletives and idiomatic expressions.

2.6.1 r-expletive: An *expletive* (also called *pleonastic*) pronoun occupies a syntactic position in a clause without actually referring to anything.

(45) [It] is snowing.

(46) [Es] hat Festnahmen gegeben.

[There] have been arrests.

Note that some pronouns which seem to behave syntactically like dummy elements should better be analysed as (abstract) cataphors; compare Section 2.3.1.

(47) [It] is great <u>that so many people are happy with this</u>. *r-cataphor*

(48) [Es] gefällt mir, <u>dass wir derselben Meinung sind</u>. *r-cataphor*

I like [it] <u>that we are of the same mind</u>.

2.6.2 r-idiom: Idioms are fixed expressions which have a figurative meaning and, therefore, typically do not introduce a proper discourse referent. Although idioms can occur in various syntactic categories, we are only interested in NPs/DPs/PPs because these could potentially be mistaken as referring expressions. It is a characteristic of idiomatic expressions that they can be replaced by non-figurative ones, which might not even contain any NP, e.g. *to go back [to the drawing board]*$_{r\text{-}idiom}$ *= to start all over.*

(49) [The early bird] catches [the worm].

(50) Sie hat [Schwein] gehabt.

She has had a stroke of luck. (Lit. 'She has had [a pig]'.)

2.7 Additional features

2.7.1 *+generic:* This additional feature/tag is assigned to referring expressions denoting a class, or a non-specific or hypothetical entity (cf. Krifka et al. 1995, Mari et al. 2013, Friedrich et al. 2015).[2]

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, in principle, combine with all r-categories. Generic entities (and only these) can recur in the indefinite form after having been previously mentioned. This is the only case in which an indefinite expression may be labelled as anaphoric *(r-given+generic).* In the following, we list different types of generic expressions.

a. Class

(51) [A cat] is a mammal. *r-new+generic*

(52) [The lion] is a huge animal. *r-unused-known+generic*

(53) [Lions] are huge animals. *r-new+generic*

(54) 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. Non-specific or hypothetical entities

Indefinite generic phrases may express non-specificity rather than class reference.

(55) a. Kanzlerin Merkel hat [vor einem Scheitern der Reformbemühungen] gewarnt.

Chancellor Merkel warned [of a failure of the reform efforts.] *r-new+generic*

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

[A failure] would be a historic lapse, she pointed out. *r-given+generic*

[2] Note that in the original formulation of the RefLex scheme (Baumann and Riester 2012), we assumed a separate category *r-generic*. This, however, resulted in a rather inhomogeneous classification, which is why we think that using a combinable feature is a better solution.

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

[Pressure and intimidation] would have no effect, Foreign Minister Mottaki declared. *r-new+generic*

(57) a. Der hessische Ministerpräsident Koch hat [vor Mindestlöhnen] gewarnt.

The Hessian governor Koch warned [of minimum wages]. *r-new+generic*

b. Baden-Württembergs Ministerpräsident Oettinger wandte sich ebenfalls [gegen Mindestlöhne]. *r-given+generic*

Governor Oettinger of Baden Wuerttemberg also opposed [minimum wages].

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

(58) a. I'm looking for [a doctor]. (any doctor: non-specific) *r-new+generic*

b. I'm looking for [a doctor]. He owes me money.

(a certain doctor: specific) *r-new*

(59) They have enough support in parliament to elect [a new prime minister.] *r-new+generic*

c. Negation

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

(60) a. I don't have [a car]. *r-new+generic*

b. I have [no car]. *r-new+generic*

Attention:

(61) My neighbour has a cat. I haven't seen [it] yet. *r-given*
(a specific cat)

2.7.2 +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*.

(62) a. In this company, Mary is [the boss]. *r-bridging+predicative*

b. They elected him [President of France]. *r-unused-known+predicative*

c. The price rose [to seven dollars]. *r-new+predicative*

d. [As a child], Peter had lived in Brandenburg. *r-new+predicative*

e. Er ist [gelernter Friseur]. *r-new+predicative*

He is [a trained hairdresser].

f. I consider her [a genius]. *r-new+predicative*

2.8 Decision tree for the r-level

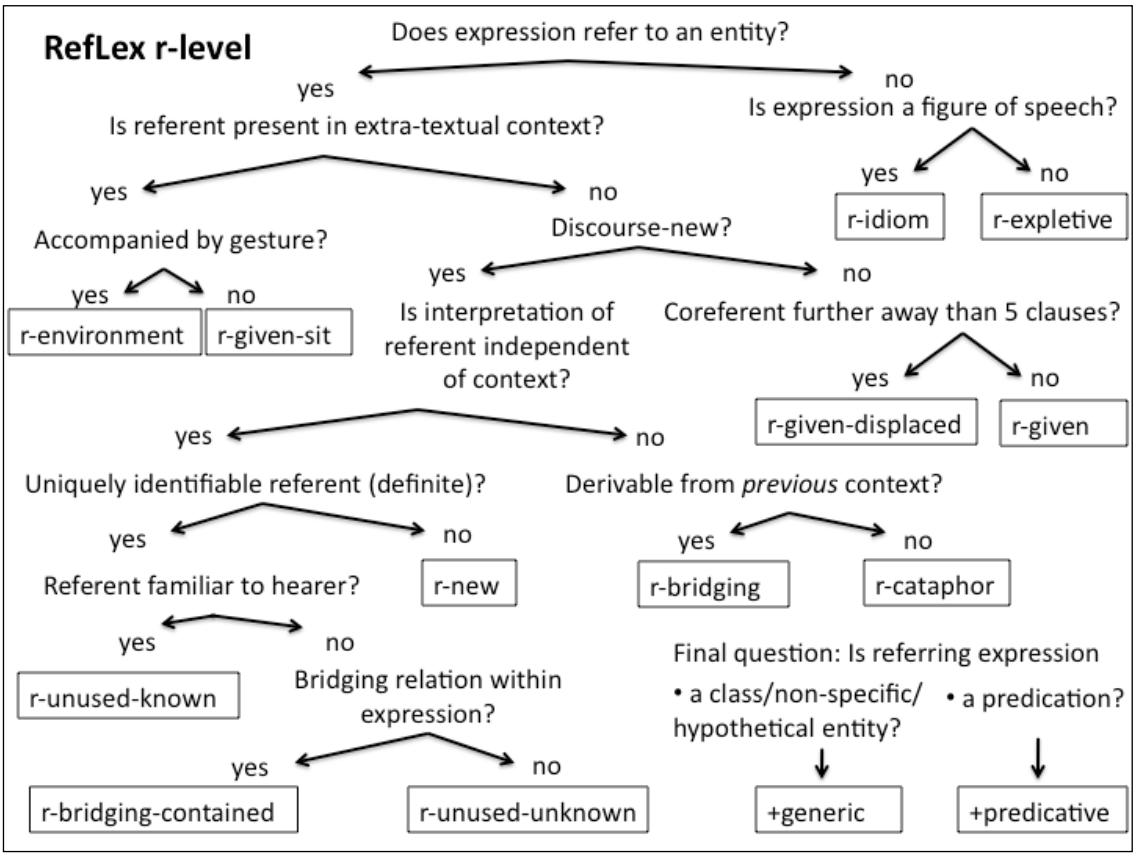

Figure 2: Decision tree for r-level

2.9 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 pre-processed using different parsers. We intend to raise the reader's awareness that tool and pre-processing choices may have an influence on the actual annotation process, which is independent of the theoretical-linguistic properties of the RefLex system and which might require a mild degree of adaptation on behalf of the annotator.

2.9.1 Annotation units

R-labels are assigned to referring expressions, in particular phrases that occur as verbal arguments.[3] Depending on the syntactic framework chosen, such phrases are analysed as DPs (determiner phrases, see Figure 3a) or NPs (noun phrases, see Figure 3b). In these figures *(SALTO)*, data have been pre-processed with different parsers.

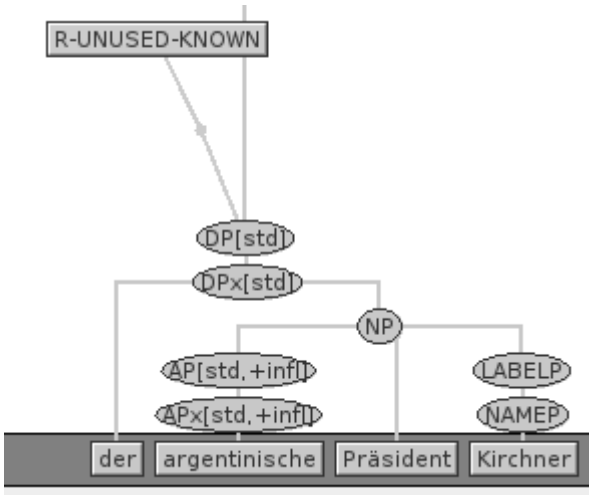

Figure 3a: Deep DP analysis of the phrase [Argentinian President Kirchner]

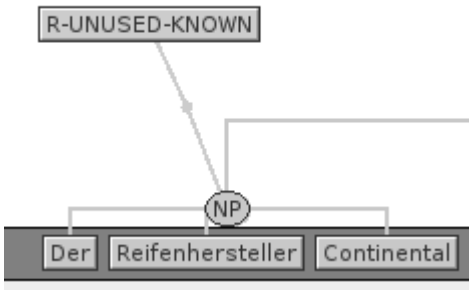

Figure 3b: Flat NP analysis of the phrase [the tyre manufacturer Continental]

Discourse particles (e.g. *even, only, also;* for German: *sogar, nur, auch, schon, noch, ja* etc.) 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.

(63) There was a flock of sheep grazing on the meadow.
a. [Every sheep] had a small bell around [its] neck.
(every sheep from the mentioned flock) *r-given*
b. [Some sheep] were white with a black face.
(subgroup mentioned for the first time) *r-new*

2.9.2 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. Note that in *Slate* different labels are color-coded. An example is shown in Figure 4.

[3] Note that elliptical constructions or zero anaphora are not labelled, at least not for German and English, which are considered non-pro-drop languages.

Figure 4: Annotating embedded phrases in Slate

Do not forget to label possessive pronouns!

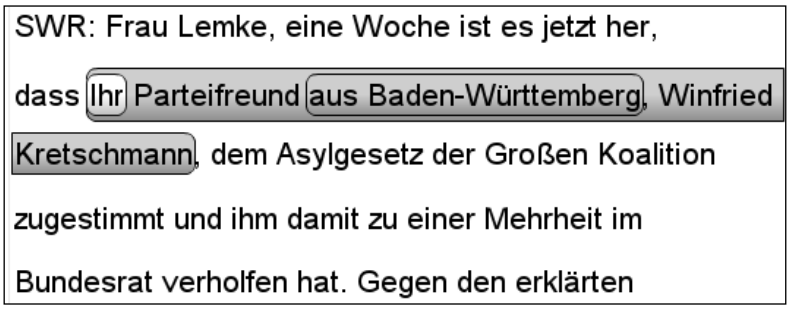

Figure 5: Embedded possessive pronoun: [[your] fellow party member [from Baden Wuerttemberg], Winfried Kretschmann]

In the absence of syntactic analyses or the possibility of label embedding, the nesting of phrases can be modelled by creating several annotation layers, as e.g. in the tool *EXMARaLDA* (Schmidt & Wörner 2014) (see Figure 6). Here, no anaphoric links are annotated but the RefLex domains and labels are aligned with the words in the speech signal (this is done similarly in the speech analysis tool *Praat*; Boersma & Weenink 2012).

	[00:0	13 [00:0	14 [00:0	15 [00:0	16 [00:04	17 [00:04.5]	18 [00:05.0]	19 [00:0	20 [00:05.4]	21 [00:0	22 [00:0	23 [00:0	24 [00:0
SN [txt]	n	and	point	to	policy	agreements	between	the	members	of	the	Five	Eyes
SN [r-level 1]				r-unused-unknown									
SN [r-level 2]							r-bridging-contained						
SN [r-level 3]										r-given			

Figure 6: Annotation of nested phrases without syntactic analysis in EXMARaLDA.

2.9.3 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: (i) The preposition linguistically "belongs" to the referring expression rather than to the embedding verb. (ii) 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. (iii) Other languages (e.g. Finnish) make ample use of case endings rather than prepositions (e.g. *rakennukse-ssa 'in the building',* lit. *'building-in'*), 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:

(64) She asked [$_{PP}$ for [$_{DP}$ the bill]$_i$]$_j$. (Phrases *i* and *j* have the same referent.)

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.

(65) There was a tree in the garden. Paul sat [behind [the tree]$_{i:\ r\text{-}given}$]$_{j:\ r\text{-}new}$
[It]$_{i:\ r\text{-}given}$ was a maple tree.
It was shady [there]$_{j:\ r\text{-}given}$.

2.9.4 Partitives vs. quantified expressions

Partitives, like in Example (66b), typically represent discourse-new subgroups of previously introduced entities, and are treated as nested expressions.

(66) a. [Ten ducks]$_{r\text{-}new}$ were sitting beside the pond.
b. [Four [of the ducks]$_{r\text{-}given}$]$_{r\text{-}new}$ were male.

Note that certain quantified expressions, like the one in (67), look syntactically similar to partitives. Yet, they are analyzed as a single referring expression.

(67) [Hundreds of ducks]$_{r\text{-}new}$ were sitting beside the pond.

2.9.5 Anaphoric links

R-given and *r-bridging* anaphors typically have antecedents. However, only *r-given* (as well as *r-given-sit*) expressions may form coreference chains that consist of more than two markables. The link originating from an anaphor should always be drawn to the immediately preceding antecedent (in annotation tools like *Slate* or *SALTO*, which provide such a functionality).

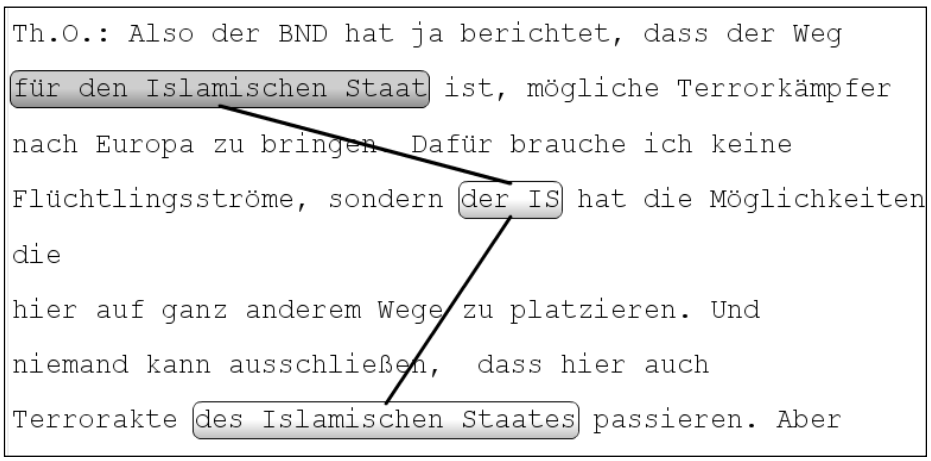

Figure 7: Coreference chain: [for the Islamic State] … [the IS] … [of the Islamic State]

2.9.6 Appositions and relative clauses

Appositions (Figure 8) and relative clauses are grouped together with the expressions they modify, i.e. they are not annotated separately. This also holds for relative clauses containing a coreferential phrase which is properly embedded, like in Figure 9.

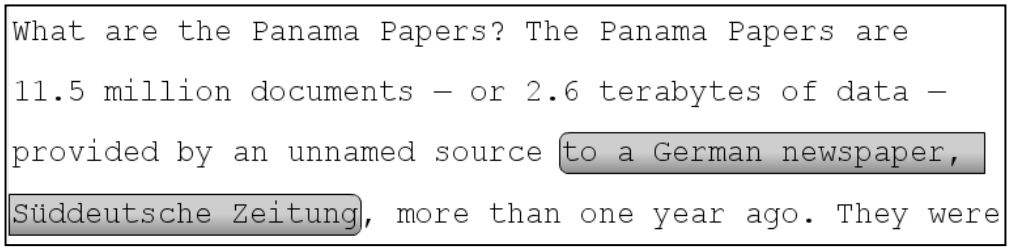

Figure 8: Referring expression with apposition

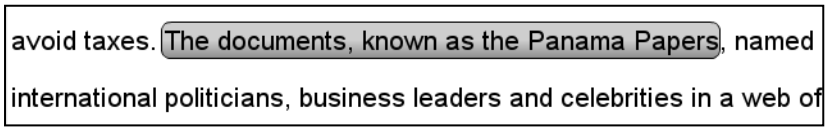

Figure 9: Referring expression with (reduced) relative clause and an embedded phrase ('the Panama Papers') that corefers with the entire expression

2.9.7 Discontinuous markables

For a variety of (syntactic or processing) reasons, a referring expression may be broken in two parts, although the entire expression should receive one common label.

(68) [Wir] haben [alle] davon gehört.

[We] have [all] heard of it. *([we... all]: r-given-sit)*

Depending on the annotation tool used, different strategies may be used to annotate such expressions. If there is a syntactic analysis available that provides crossing edges there may be a syntactic node available for the discontinuous phrase. In Figure 10 *(SALTO)*, we see a verbal complex *(verschrieben werden 'be prescribed')* that separates a PP into two halves.[4]

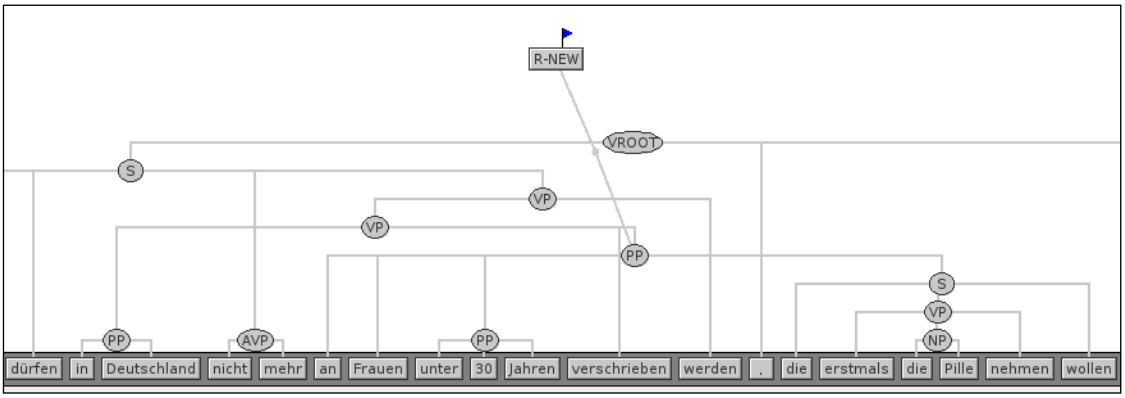

Figure 10: Discontinuous PP: [an Frauen unter 30 Jahren, (...) die erstmals die Pille nehmen] / [to women under 30 (...) who want to go on the pill for the first time], analysed with crossing edges

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*. Yet another alternative, shown below in *Slate*, is a special type of link that does not indicate coreference but discontinuity.

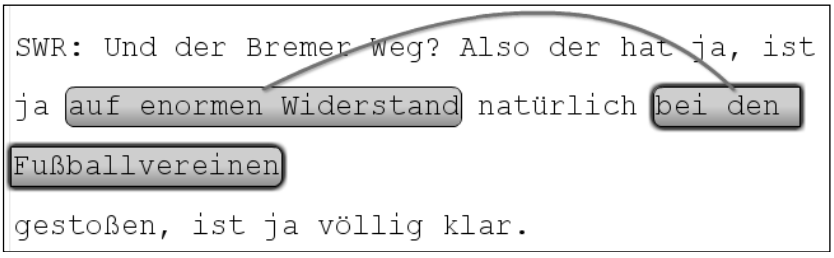

Figure 11: Discontinuous phrase [auf enormen Widerstand (...) bei den Fußballvereinen] / [on fierce resistance (...) on behalf of the football clubs], marked with a special (non-coreferential) link (Slate)

2.9.8 Coordination and aggregation

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, see Figure 12 *(SALTO)* and Figure 13 *(Slate)*. If the information status of the conjuncts differ, then the coordination as a whole receives the "newer" label.

[4] The little flag is used to represent the *+generic* feature in *SALTO*.

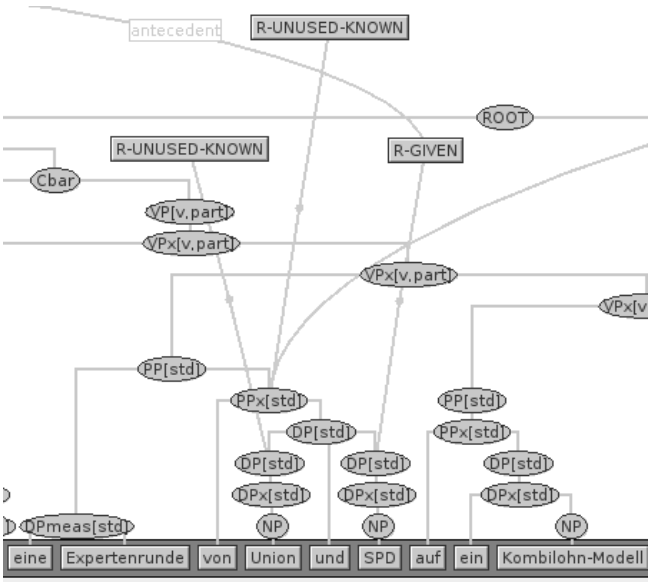

Figure 12: Annotation of a coordinated expression [von [Union] und [SPD]] / [of [Conservative Party] and [Social Democrats]] and its conjuncts in SALTO

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 13: Annotating a coordinated expression [[a transition] and [a privatisation]] as well as its two conjuncts in Slate

The reason why nested labels are necessary is that both the coordination itself or any of its conjuncts can be taken up anaphorically. An overview of different possibilities is given below.

(69) [[The Conservatives]$_i$ and [the Social Democrats]$_j$]$_k$ have found an agreement.

a. [They]$_k$ decided not to raise taxes.

b. It was [the Conservative Party]$_i$ who had promised this to [their]$_i$ voters.

In the case that a plural pronoun is grouping together two or more referents which have not occurred as conjuncts of a coordination we speak of *aggregation* or *summation* cf. Kamp & Reyle (1993: Chapter 4). An example in *Slate* is shown below in Figure 14. Note that aggregation, like discontinuity mentioned above, makes use of yet another special link type, which is not to be confused with coreference.

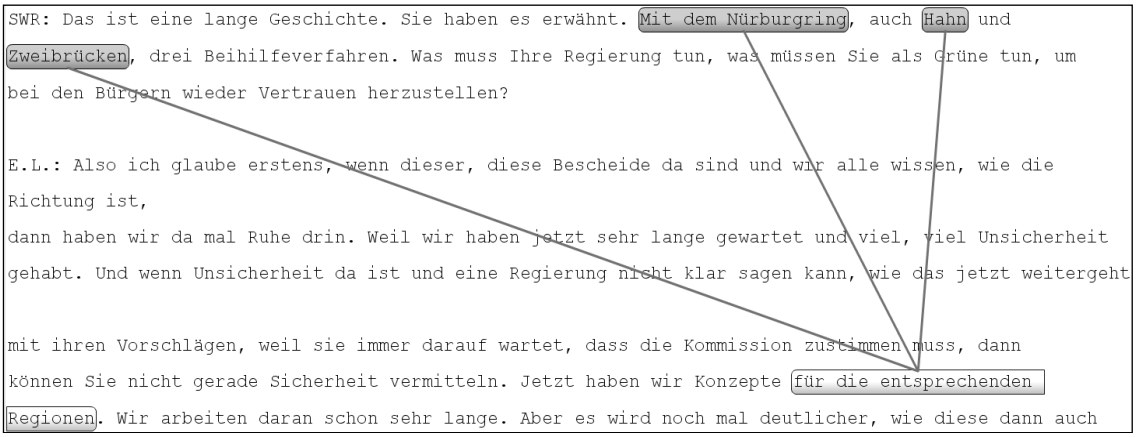

Figure 14: Aggregation, using a special link type: 'With Nürburgring, also Hahn and Zweibrücken'. The three place names are grouped together under the expression [for the respective regions].

2.9.9 Direct speech

Elements which occur in direct/quoted speech are not coreferential with elements that have occurred before the direct speech section. Thus, direct speech is treated as a separate – embedded – discourse.

(70) Once upon a time there was a dear little girl who was loved by everyone who looked at her, but most of all by her grandmother (...) One day her mother said to her: "Come, Little Red Riding Hood, here is a piece of cake and a bottle of wine; take them to [your grandmother], (...)" *r-unused-known,* not *r-given*

In the same way, the text body is assumed to be separated from headlines, abstracts etc.

3 L-level

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

Table 2: Annotation tags of the l-level	
Tag	Salience class
l-given-same	active, i.e. salient concepts
l-given-syn	
l-given-super	
l-given-whole	
l-accessible-sub	semi-active, i.e. derivable concepts
l-accessible-part	
l-acessible-stem	
l-new	inactive concepts

The lexical level applies to the word domain, more specifically to content words such as nouns, adjectives, (content) adverbs 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

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

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

(73) 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.

(74) 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)

(75) John owns a bicycle. You absolutely need a [bike] if you work at Stanford.
(van Deemter 1999)

(76) 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].

(77) 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.

(78) 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].

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

3.1.3 l-given-super: A word is lexically superordinate to previous word in the sense that the markable is a hypernym, i.e. a superset of the antecedent expression.

(80) Do you like dogs? I like all [animals].

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

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

(83) Die 27 Staats- und Regierungschefs der EU wollen die "Berliner Erklärung" unterzeichnen. 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.

(84) 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: A word is lexically superordinate to previous word in the sense that the markable is a holonym of the antecedent.

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

(86) 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.[5]

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).*

[5] This example will probably soon find itself among the victims of the Brexit.

3.2.1 l-accessible-sub: A word is lexically subordinate to previous word in the sense that the markable is a hyponym.

(87) Do you like animals? I like all [dogs].

(88) John does own a vehicle. But you absolutely need a [bicycle] if you work at Stanford.

(89) 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: A word is lexically subordinate to previous word in the sense that the markable is a meronym (an expression denoting a part).

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

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

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

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

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

(94) 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

(95) Why do you study Italian? I'm married to an [Italian].[6] (Büring 2007)

[6] Since we are dealing with two different concepts in this example, we decided not to label the second occurrence of *Italian* as *l-given-same*, contrary to what we say in Baumann and Riester (2012).

(96) Eine Erhebung für die Zeit von Juni 2006 bis Februar dieses Jahres habe ergeben, dass Flugzeuge aus EU-Ländern fünf Mal verbotene Sperrzonen [überflogen] hätten.

According to an inquiry for the time between June 2006 and February this year, aircrafts (lit. 'flight tools') from EU countries had [flown] across restricted zones five times.

(97) Die Picknickdecke ist kariert. Ich hasse [Karomuster].

The picnic 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)

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

(99) Der [Iran] will an seinem [Atomprogramm] [festhalten].

[Iran] intends to [hold on] to its [nuclear programme].

(100) 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 (content) 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 (e.g. *football league*) 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*, a link marking discontinuity etc.; comparable to solutions proposed in Section 2.9.7.

3.4.3 Hierarchies

If several labels are possible at the same time, the following order of preference 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.

(101) Germany has a [population] of 80 million people. *l-new*, not *l-accessible-part*

(102) Klose is the oldest player in his [team]. *l-new*, not *l-given-whole*

However, we do annotate hyponymic relations *(l-given-super* and *l-accessible-sub)* between a proper noun/name and a common noun, where appropriate.

(103) Germany and other [countries] will return to Central European Time on October 29. (Germany is a country, therefore: *[countries]*$_{l\text{-}given\text{-}super}$)

3.4.6 Cross-categorical relations

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

Acknowledgement

The financial support of our research projects (Sonderforschungsbereich 732, Project A6, and Project BA 4734/1-2) by Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

References

Abney, Steven (1987), 'The English noun phrase in its sentential aspect', (Massachusetts Institute of Technology).

Asher, Nicholas (1993). Reference to abstract objects in discourse. Kluwer: Dordrecht.

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

Baumann, Stefan and Riester, Arndt (2012), 'Referential and lexical givenness: Semantic, prosodic and cognitive aspects', in Gorka Elordieta and Pilar Prieto (eds.), *Prosody and Meaning* (25; Berlin: De Gruyter Mouton), 119-61.

BBN Technologies (2007), 'Co-reference guidelines for English OntoNotes Version 7.0', https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-coreference-guidelines.pdf

Boersma, Paul and Weenink, David (2012). Praat: doing phonetics by computer, http://www.praat.org/

Büring, Daniel (2007). Intonation, Semantics and Information Structure. In: Ramchand, G. Reiss, C. (Eds.), The Oxford Handbook of Linguistic Interfaces.

Burchardt, Aljoscha, Erk, Katrin, Frank, Anette, Kowalski, Andrea and Padó, Sebastian (2006). SALTO - A Versatile Multi-Level Annotation Tool. Proceedings of LREC, Genoa, Italy.

Carlson, Gregory and Pelletier, Francis (1995), *The Generic Book* (University of Chicago Press).

Chafe, Wallace (1994), *Discourse, consciousness, and time* (University of Chicago Press).

Christophersen, Paul (1939), *The articles: A study of their theory and use in English* (Copenhagen: Munksgaard).

Clark, Herbert (1977), 'Bridging', in Philip Johnson-Laird and Peter Wason (eds.), *Thinking: Readings in cognitive science* (Cambridge University Press), 411-20.

Coppock, Elizabeth and Beaver, David (2015), 'Definiteness and determinacy', *Linguistics and Philosophy*, 38, 377-435.

Dipper, Stefanie and Zinsmeister, Heike (2012), 'Annotating abstract anaphora', *Language Resources and Evaluation,* 46, 37-52.

Eckert, Miriam and Strube, Michael (2000), 'Dialogue acts, synchronizing units, and anaphora resolution', *Journal of Semantics,* 17 (1), 51-89.

Elbourne, Paul (2013), *Definite descriptions* (Oxford University Press).

Frege, Gottlob (1891), *Function und Begriff: Vortrag gehalten in der Sitzung vom 9. Januar, 1891 der Jenaischen Gesellschaft für Medicin und Naturwissenschaft* (Jena: Hermann Pohle).

--- (1892), 'Über Sinn und Bedeutung', *Zeitschrift für Philosophie und Philosophische Kritik,* 100, 25-50.

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.

Götze, Michael, et al. (2007), 'Information structure', in Stefanie Dipper, Michael Götze and Stavros Skopeteas (eds.), *Information structure in cross-linguistic corpora* (ISIS 7: Universitätsverlag Potsdam), 147-87.

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

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).

Hawkins, John (1978), *Definiteness and indefiniteness: a study in reference and grammaticality prediction* (London: Croom Helm).

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

Kamp, Hans (manuscript), 'Entity Representations and Articulated Context', (University of Stuttgart / University of Texas at Austin).

Kamp, Hans and Reyle, Uwe (1993). *From Discourse to Logic* (Dordrecht: Kluwer).
Kaplan, Dain, et al. (2012), 'Slate -- A tool for maintaining and creating annotated corpora', *Journal for Language Technology and Computational Linguistics,* 26(2), 91-103.
Kolhatkar, Varada, Zinsmeister, Heike and Hirst, Greame (2013). 'Annotating anaphoric shell nouns with their antecedents', *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 112-121. Sofia, Bulgaria.
Krasavina, Olga and Chiarcos, Christian (2007), 'PoCoS: Potsdam coreference scheme', *Linguistic Annotation Workshop* (Prague: Association for Computational Linguistics), 156-63.
Krifka, Manfred, Pelletier, Francis, Carlson, Gregory, ter Meulen, Alice, Chierchia, Gennaro and Link, Godehard (1995), 'Introduction to Genericity', in Gregory Carlson and Francis Pelletier (eds.), *The Generic Book* (Chicago: University of Chicago Press), 1-124.
Levinson, Stephen (1983), *Pragmatics* (Cambridge University Press).
Lambrecht, Knud (1994), 'Information structure and sentence form: A theory of topic, focus, and the mental representations of discourse referents', (Cambridge University Press).
Löbner, Sebastian (1998), 'Definite associative anaphora', in Simon Botley (ed.), *Approaches to discourse anaphora. Proceedings of the Discourse Anaphora and Resolution Colloquium (DAARC-1996)* (Lancaster).
Mari, Alda, Beyssade, Claire and Del Prete, Fabio (2013), *Genericity* (Oxford University Press).
Neale, Stephen (1990), *Descriptions* (Cambridge, Mass.: MIT Press).
Nissim, Malvina, et al. (2004), 'An Annotation Scheme for Information Status in Dialogue', *4th Language Resources and Evaluation Conference (LREC)* (Lisbon).
Poesio, Massimo and Vieira, Renata (1998), 'A corpus-based investigation of definite description use', *Computational linguistics,* 24 (2), 183-216.
Pradhan, Sameer S, et al. (2007), 'Unrestricted coreference: Identifying entities and events in OntoNotes', *International Conference on Semantic Computing* (Irvine, CA), 446-53.
Prince, Ellen (1981), 'Toward a taxonomy of given-new information', in Peter Cole (ed.), *Radical pragmatics* (New York: Academic Press), 223-56.
--- (1992), 'The ZPG letter: Subjects, definiteness, and information-status', in Sandra and Mann Thompson, William (ed.), *Discourse description: Diverse analyses of a fund raising text* (Amsterdam: Benjamins), 295-325.
Recasens, Marta and Martí, M Antònia (2010), 'AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan', *Language resources and evaluation,* 44 (4), 315-45.
Riester, Arndt, Lorenz, David, and Seemann, Nina (2010), 'A Recursive Annotation Scheme for Referential Information Status', *Seventh Language Resources and Evaluation Conference (LREC)* (Valletta, Malta), 717-22.
Roberts, Craige (2003), 'Uniqueness in definite noun phrases', *Linguistics and philosophy,* 26 (3), 287-350.
Rodríguez, Kepa Joseba, et al. (2010), 'Anaphoric annotation of wikipedia and blogs in the live memories corpus', *Seventh Language Resources and Evaluation Conference* (Valletta, Malta), 157-63.
Russell, Bertrand (1905), 'On denoting', *Mind,* 14, 479-93.
Schmidt, Thomas & Wörner, Kai (2014). EXMARaLDA. In: Gut, U. et al. (eds.), Handbook on Corpus Phonology, Oxford University Press. 402-419.
Schwarzschild, Roger (1999), 'GIVENness, AvoidF and other constraints on the placement of accent', *Natural Language Semantics,* 7 (2), 141-77.
Simons, Mandy, et al. (2010), 'What projects and why', *Proceedings of Semantics and Linguistic Theory (SALT 20)* (20), 309-27.
Strawson, Peter (1950), 'On referring', *Mind,* 59, 320-44.
van Deemter, Kees (1999). Contrastive Stress, Contrariety, and Focus. In: Bosch, P., van der Sandt, R. (Eds.), Focus - Linguistic, Cognitive, and Computational Perspectives (Studies in Natural Language Processing). Cambridge University Press, Cambridge, 3-17.

Weischedel, Ralph et al. (2013). OntoNotes Release 5.0 LDC2013T19. Linguistic Data Consortium, Philadelphia.
Yule, George (1981), 'New, current and displaced entity reference', *Lingua,* 55 (1), 41-52.

Appendix: Differences between coreference annotation in *RefLex* and in *OntoNotes*

Information status annotation subsumes the annotation of coreference, and corpora annotated with coreference information play a crucial role as training data for the computational linguistic task of automatic coreference resolution. The *OntoNotes* corpus (Weischedel et al. 2013) has figured as a gold standard for many shared tasks in coreference resolution within the past decade. Although the main focus of the RefLex scheme is on the *classification* of textual mentions according to referential and lexical information status, the identification of *anaphoric links* (see Section 2.9.5) is nevertheless an important side aspect. Coreference annotation in RefLex is mostly compatible with the OntoNotes scheme (BBN Technologies 2007) as long as a number of systematic deviations are taken into account, which we list in the following:

RefLex	*OntoNotes*
Appositions that have the same referent as their head are grouped together with the head phrase, forming a single markable. (i) a. [John, a linguist]$_{\text{r-unused-unknown}}$ is coming for dinner.	There is a special coreference link (APPOS) between a head phrase and an apposition. (i) b. [John]$_{\text{x-HEAD}}$,[a linguist]$_{\text{x-ATTRIB}}$, is coming for dinner.
Prepositions, if present, are part of the markable. (ii) a. She goes [to Bruges].	Prepositions are kept outside the markable. (ii) b. She goes to [Bruges].
Modifiers of nouns do not represent separate markables. (iii) a. [the FBI spokesman]	Proper noun premodifiers (but no other modifiers like, for instance, common nouns or adjectives) are annotated as separate markables. Nationality acronyms like *U.S.* are not annotated separately. (iii) b. [the [FBI] spokesman]
Abstract anaphors may refer back to a full clause or verb phrase, depending on what is identified as their referent. (iv) a. Sales of passenger cars grew 22%. [The strong growth]$_{\text{r-given}}$ followed year-to-year increases.	Abstract anaphors are coreferenced with the verbal head of the assumed abstract antecedent. (iv) b. Sales of passenger cars [grew]$_{\text{x}}$ 22%. [The strong growth]$_{\text{x}}$ followed a year-to-year increase.

Several indefinite **generic expressions**, as well as generic *you*, can form coreference chains. (v) a. [Parents]$_{\text{r-unused-known+generic}}$ should be involved with [their]$_{\text{r-given+generic}}$ children's education, and [parents]$_{\text{r-given+generic}}$ should not blame schools all the time. (vi) a. Sometimes [you]$_{\text{r-given-sit+generic}}$ know [you]$_{\text{r-given-sit+generic}}$ simply have to help. (instances are linked)	Generic expressions can only be linked to pronouns or definite mentions of the same entity. (v) b. [Parents]$_{x}$ should be involved with [their]$_{x}$ children's education, and [parents]$_{y}$ should not blame schools all the time. (vi) b. Sometimes [you] know [you] simply have to help.