A Recursive Annotation Scheme for Referential Information Status

Arndt Riester1, David Lorenz2, Nina Seemann1

1Institute for Natural Language Processing, University of Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany
2 English Department, University of Freiburg, Rempartstr. 15, 79085 Freiburg, Germany
arndt.riester@ims.uni-stuttgart.de, david.lorenz@frequenz.uni-freiburg.de, nina.seemann@googlemail.com

Abstract

We provide a robust and detailed annotation scheme for information status, which is easy to use, follows a semantic rather than cognitive motivation, and achieves reasonable inter-annotator scores. Our annotation scheme is based on two main assumptions: firstly, that information status strongly depends on (in)definiteness, and secondly, that it ought to be understood as a property of referents rather than words. Therefore, our scheme banks on overt (in)definiteness marking and provides different categories for each class. Definates are grouped according to the information source by which the referent is identified. A special aspect of the scheme is that non-anaphoric expressions (e.g. names) are classified as to whether their referents are likely to be known or unknown to an expected audience. The annotation scheme provides a solution for annotating complex nominal expressions which may recursively contain embedded expressions. In annotating a corpus of German radio news bulletins, a kappa score of .66 for the full scheme was achieved, a core scheme of six top-level categories yields $\kappa = .78$.

1. Introduction

We provide a robust and detailed annotation scheme for information status, which covers all – also embedded – DPs/PPs in a text (excluding idioms) and achieves reasonable inter-annotator scores. Other annotation schemes, e.g. Nissim et al. (2004) or Dipper et al. (2007) take information status to be an independent dimension from (in)definiteness. They adopt three-way distinctions, which – in varying shares – mix the partitions postulated by Prince (1981) (evoked / inerferable / new), Prince (1992) (discourse-old / discourse new but hearer-old (unused) / discourse-new) and Chafe (1994) (given / accessible / new). By contrast, our scheme is built on the conviction that information status is strongly dependent on (in)definiteness. At least for languages, in which morphological (in)definiteness marking is available, like English or German, annotations become more meaningful (and easier to accomplish) if such marking is not neglected.

Instead of assuming a cognitive hierarchy, we claim that definite expressions are grouped according to information sources in which they are identified: the discourse context (what has explicitly been uttered), the utterance context (entities which are perceptible or available by default to the discourse participants, cf. Kaplan (1989)), the encyclopaedic context (entities familiar to the recipient) as well as frames, which act as small contexts at the interface between lexical and world knowledge. Our classification subsumes annotations for coreference and bridging which have already been treated, for instance, in Poesio and Vieira (1998), Poesio (2004) or Nedoluzhko et al. (2009). We also propose a classification for indefinites.

Another important point is that the notion of information status (in particular, givenness) can be understood in two ways: as a property of words (suggesting a definition of givenness as recurrence or hyperonymy with regard to a preceding word), or as a property of the referents of determined phrases (on this view, givenness means coreference). While the two notions sometimes overlap, there are cases in which they contradict each other: identical words in a text may (and in fact often do) refer to different individuals, whereas lexically unrelated phrases may refer to one and the same entity. In the present scheme we have opted for the referential interpretation of information status,¹ which inevitably means that the markables we are targeting must be full (and possibly complex) prepositional phrases or determiner phrases (Montague’s (1974) “terms”, often referred to as “noun phrases” but here the DP terminology is crucial since only determined phrases are referential). Complex phrases, in turn, may recursively contain other DPs or PPs, which have their own information status. For instance, the phrase $[DP \text{ the red gem } [PP \text{ in } [DP \text{ the Queen’s crown}]]$ is built from linguistic material referring to three different individuals/entities (the Queen, the crown, the gem) which potentially possess different information status and need to be annotated recursively. In the following, we will provide an overview on the taxonomy using examples adapted from various issues of Time, The Guardian, and Nature.

2. The scheme

Availability in the discourse context: GIVEN

A referential expression is GIVEN² if and only if it is definite and has a coreferential antecedent in the discourse context. Our notions of coreference and discourse context are those adopted, for instance, in Discourse Representation Theory (Kamp and Reyle, 1993). GIVEN entities are further classified as shown in Table 1.

Several coreferential items taken together form a givenness chain. Subsequent GIVEN items must be classified with regard to the entire chain, not only with regard to its immediate predecessor. Occasionally, labels conflict with each other.\footnote{1}{cf. Baumann and Riester (2010) for a two-level scheme}

\footnote{2}{In earlier publications related to the scheme, e.g. Riester (2008b), this category is called D-GIVEN (discourse-given) to indicate that – contra e.g. Prince (1981) – our use of the term givenness is restricted to textually available or explicitly uttered entities. Since it appears that this is the view which is by now widely accepted within contemporary literature, it seems safe to use the term GIVEN without further specification.}


1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction

1. Introduction
In English, differences of case are ignored. For instance, the last item in the sequence *President Obama*. . . *Obama*. . . *Obama* counts as short with regard to the first and as repeated with regard to the second item. In such cases, as a rule, the label repeated wins.

Typical participants in a scenario: Bridging

The class of bridging entities is defined as those definites which neither possess a corefential antecedent nor can be understood in the absence of a preceding expression which defines a scenario or frame in which they play a typical role. We do not attempt a subclassification according to this role, e.g. part-whole, participant in a situation etc., as proposed in Nissim et al. (2004). Such roles have been treated in frame semantics (Fillmore, 1985) and, as we think, are not relevant to information status. Note, further, that the notion of bridging has originally been defined in a broader sense than we are using it here (Clark, 1977; Asher and Lascarides, 1998); nevertheless our use explicitly focuses on the prevalent cases in the contemporary literature.

We use the simple label bridging if and only if a salient bridging antecedent can be identified. In our annotations, we draw an anchoring link from the bridging anaphor to that antecedent. Sometimes a frame is evoked by a stretch of text rather than by a single word. In these cases, we use the label bridging-text, without any specific anchor. A special case are complex DPs in which the bridging “antecedent” is the genitive or prepositional adjunct contained in the phrase itself. We assign the label bridging-contained to these DPs. 

### Table 1: Subcategories of given

<table>
<thead>
<tr>
<th>GIVEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example sentences</td>
</tr>
</tbody>
</table>

- **-PRONOUN**: “Russia is ruled by the same people who own [it].”
- **-REFLEXIVE**: Cancer hides itself within normal structures like blood vessels.
- **-SHORT**: Both had the blessings of Dr. Richard Klausner. But even Klausner had to be persuaded at first.
- **-REPEATED**: It is hard to think of any industrial society that in its essentials is less like the U.S. than Japan. [...] Japan is a nation with very deep cultural roots and habits.
- **-EPITHET**: Before the European Union’s ban on incandescent lightbulbs went into effect on Sept. 1, consumers across Europe raided stores to stockpile the \[familiar\] bulbs.

### Table 2: SITUATIVE

Typical participants in a scenario: Bridging

The class of bridging entities is defined as those definites which neither possess a corefential antecedent nor can be understood in the absence of a preceding expression which defines a scenario or frame in which they play a typical role. We do not attempt a subclassification according to this role, e.g. part-whole, participant in a situation etc., as proposed in Nissim et al. (2004). Such roles have been treated in frame semantics (Fillmore, 1985) and, as we think, are not relevant to information status. Note, further, that the notion of bridging has originally been defined in a broader sense than we are using it here (Clark, 1977; Asher and Lascarides, 1998); nevertheless our use explicitly focuses on the prevalent cases in the contemporary literature.

We use the simple label bridging if and only if a salient bridging antecedent can be identified. In our annotations, we draw an anchoring link from the bridging anaphor to that antecedent. Sometimes a frame is evoked by a stretch of text rather than by a single word. In these cases, we use the label bridging-text, without any specific anchor. A special case are complex DPs in which the bridging “antecedent” is the genitive or prepositional adjunct contained in the phrase itself. We assign the label bridging-contained to these DPs. 

#### Table 3: Subtypes of bridging

<table>
<thead>
<tr>
<th>BRIDGING</th>
<th>Examples sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0</td>
<td>Their two-day visit is the second step in the beginning of a dialogue with Burma. [...] For years, the US had isolated the junta with political and economic sanctions.</td>
</tr>
<tr>
<td>-TEXT</td>
<td>United were trailing 3-1 when Fletcher was felled in the area by Aleksei Berezutski. The Scotland midfielder then yellow-carded by the referee.</td>
</tr>
<tr>
<td>-CONTAINED</td>
<td>The Republicans won the ... governorship of Virginia.</td>
</tr>
</tbody>
</table>

### Table 3: Subtypes of bridging

(Non-)Availability in the recipients’ encyclopaedic context: ACCESSIBLE / UNUSED

According to Prince (1981), textual annotations of the sort described here should reflect a speaker or writer’s assumptions concerning the hearer’s knowledge state (“assumed familiarity”), which is a simplification with regard to Clark and Haviland’s (1977) notion shared knowledge. However,

---

3. Under such a view instances of gestural deixis find their referents in the so-called environment context (Kamp, to appear).

4. Inspired by Prince’s (1981) class inferrable-contained
we think that annotating the writer’s assumptions is an unreachable goal, if it is to mean more than distinguishing certain morphosyntactic patterns – particularly when it comes to classifying non-anaphoric, discourse-new entities.

Therefore, in order to make the annotations tractable, we introduce a further simplification and propose to label whether the annotator himself expects a certain expression to be known or unknown by the intended audience of a text, cf. Rießer (2008a), Rießer (2008b). Moreover, this does justice to the fact that, for instance, news texts normally do not only have one addressee but are designed for a large number of recipients, which of course never possess exactly the same knowledge.

Entities which are expected to be widely known are labeled ACCESSIBLE-GENERAL-TOKEN, those which refer to a (generic) class or type are labeled ACCESSIBLE-GENERAL-TYPE.

Entities which are expected to be unknown to the audience are labeled ACCESSIBLE-DESCRIPTION.

From a semantic point of view, only ACCESSIBLE-GENERAL entities form part of the recipients’ encyclopaedic context (Kamp, to appear), whereas entities labeled ACCESSIBLE-DESCRIPTION are unknown and have to be accommodated. This important distinction is not sufficiently accounted for in any of the previous annotation schemes for information status.

<table>
<thead>
<tr>
<th>ACCESSIBLE</th>
<th>Example sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>-GENERAL-TYPE (UNUSED-TYPE)</td>
<td>Dispersal in the Eurasian badger is believed to be very limited.</td>
</tr>
<tr>
<td>-GENERAL-TOKEN (UNUSED-KNOWN)</td>
<td>Secretary of state Hillary Clinton jettisoned the respect earned on a tough diplomatic mission to Pakistan by claiming that [...].</td>
</tr>
<tr>
<td>-DESCRIPTION (UNUSED-UNKNOWN)</td>
<td>The hoard was found in January near Harrogate by David and Andrew Whelan, father and son hobby metal detectorists, in a bare field due to be ploughed for spring sowing.</td>
</tr>
</tbody>
</table>

Note that the use of ACCESSIBLE on these occasions differs from the one in Chafe (1994), who uses it to describe information which is derived from previously mentioned material. In Table 4 we therefore also include an alternative notion (UNUSED). The distinction between ACCESSIBLE-GENERAL-TOKEN / UNUSED-KNOWN and ACCESSIBLE-DESCRIPTION / UNUSED-UNKNOWN entities also corresponds to the difference between familiar and non-familiar but uniquely identifiable discussed by Gundel et al. (1993).

**ACCESSIBLE-DESCRIPTION vs. BRIDGING-CONTAINED**

An important rule concerns the demarcation between the categories ACCESSIBLE-DESCRIPTION (when assigned to complex definite descriptions) and BRIDGING-CONTAINED, which both apply to context-free expressions: BRIDGING-CONTAINED items may occur in autonomy because they possess a subconstituent representing the anchor for the entire expression. ACCESSIBLE-DESCRIPTION items generally consist of sufficient linguistic material to enable the addressee to accommodate (Lewis, 1979) an adequate discourse entity. The two types are distinguished by the following rule: an item labeled as BRIDGING-CONTAINED refers to an object or person that stands in a prototypical relation to its genitive or PP-argument. By contrast, the relation holding between the referent of an expression labeled ACCESSIBLE-DESCRIPTION and its argument phrase is non-prototypical; compare Table 5.

| BRIDGING-CONTAINED | The gunman was on the roof of the checkpoint when he opened fire. |
| ACCESSIBLE-DESCRIPTION | [he was] convicted of helping to organise the seizure of Osama Moustafa Hassan Nasr from a Milan street in February 2003. |

Table 5: BRIDGING-CONTAINED vs. ACCESSIBLE-DESCRIPTION

A further test is that it is always possible to separate the parts of BRIDGING-CONTAINED phrases (the checkpoint ... the roof), while it is considerably marked (till outright impossible) to split expressions labeled ACCESSIBLE-DESCRIPTION (Osama Moustafa Hassan Nasr ... the seizure).

**Indefinites: INDEF**

Indefinites are grouped into specific ones (called NEW because they introduce a novel discourse referent that may be taken up later) and GENERIC ones (those referring to a class or a non-specific concept under consideration). There are, furthermore, indefinites standing in a characteristic relation to the context: they may pick up a concept mentioned before (literally or paraphrastically: RESumptive), or they may represent a subset of an aforementioned group: PARTITIVE. Similar to the BRIDGING cases, that group may occasionally occur embedded in a complex phrase: PARTITIVE-CONTAINED. See Table 6.

**Other categories**

The remaining categories are: CATAPHOR (cataphoric definite or indefinite expressions which are not reflexives); EXPLETIVE (pronouns lacking a thematic role); NULL (nothing, nobody etc.); and RELATIVE (a label for non-restrictive relative clauses. By contrast, restrictive relative clauses are interective modifiers forming part of the DP and, thus, are included under the information-status label assigned to this entire DP.)

---

5 ACCESSIBLE-GENERAL is used in Dipper et al. (2007).
6 Compare Prince (1981) and Baumann and Rießer (2010).
### Example sentences

- **INDEF**
  - In a hilltop suburb south of Jerusalem called Efrat, Sharon Katz serves a neat plate of sliced cake.

- **NEW**
  - Serious beer drinkers should head straight to this 550-year-old institution.

- **GENERIC**
  - That's close to how a cancer vaccine works, but not precisely. Most experts see cancer vaccines as a hybrid of treatment and prevention.

- **RESUMPTIVE**
  - That's close to how a cancer vaccine works, but not precisely. Most experts see cancer vaccines as a hybrid of treatment and prevention.

- **PARTITIVE**
  - At violent clashes between the police and demonstrating Kurds, 52 police officers and three demonstrators were injured.

- **PARTITIVE**
  - In one of the eggs containing four haploid nuclei, the nuclei were dividing.

### Table 6: Subtypes of INDEF

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDEF</td>
<td>Example sentences</td>
</tr>
<tr>
<td>NEW</td>
<td>In a hilltop suburb south of . . . Jerusalem called Efrat</td>
</tr>
<tr>
<td>GENERIC</td>
<td>Serious beer drinkers should head straight to this 550-year-old institution.</td>
</tr>
<tr>
<td>RESUMPTIVE</td>
<td>That's close to how a cancer vaccine works, but not precisely. Most experts see cancer vaccines as a hybrid of treatment and prevention.</td>
</tr>
<tr>
<td>PARTITIVE</td>
<td>At violent clashes between the police and demonstrating Kurds, 52 police officers and three demonstrators were injured.</td>
</tr>
<tr>
<td>CONTAINED</td>
<td>In one of the eggs containing . . . four haploid nuclei, the nuclei were dividing.</td>
</tr>
</tbody>
</table>

### 3. Evaluation

The present evaluation has been conducted for German. We have annotated a corpus of transcripts from German radio news bulletins. The annotations were done by two students of linguistics, well-trained in the annotation task, using the SALTO/SALSA tool (Burchardt et al., 2006). For this evaluation, we chose a section comprising at least 1149 DPs/PPs. Since the coders themselves had to identify the markables while annotating, a few markables were accidentally overlooked by either one of the coders. We are ignoring these. The number of markables annotated by both coders is 1100. Of those, 757 are labeled identically. Table 7 shows the distribution of annotations we obtain.

According to the metric of Cohen (1960) and Artstein and Poesio (2008), we achieve a $\kappa$ score of .66 using the entire scheme. Computing the inter-annotator agreement for a core variant of our scheme using six top-level categories GIVEN, SITUATIVE, BRIDGING, ACCESSIBLE, INDEF and OTHER yields $\kappa = .78$. This is lower than the scores reported in Nissim et al. (2004) (.788 (ext.) / .845 (core)) for their annotations on dialogue data (Switchboard). However, their system distinguishes fewer categories (16) and does not classify all DPs/PPs. Nissim et al. (2004) exclude locatives and directionals and moreover use a category not-understood to exclude some controversial cases. Ritz et al. (2008) report $\kappa = .55$ performance for the scheme by Dipper et al. (2007) on newspaper commentaries (the most closely related to our radio news transcripts), going up to .73 for question/answer pairs. Hempelmann and others (2005) report .72 for Prince’s (1981) classification using seven categories on narratives from 4th grade textbooks.

### 4. Conclusion

We think that the way our system proposes annotate information status recursively is convincing (cf. Figure 1), because information status is defined at the level of determined phrases. In other words, it is the discourse referents involved, not the expressions used, which are classified according to information status.

![Figure 1: Nested Annotation](image)

The scheme straightforwardly connects to semantic theory, since the categories we use are definable in DRT, compare e.g. Riester (2009). We integrate, in a transparent way, insights from the semantic-philosophical tradition on (in)definiteness, anaphora and accommodation. In related work, we could show that the classes we assume tend to be asymmetrically ordered when cooccurring in the same clause (Cahill and Riester, 2009). Furthermore, some of them are statistically correlated with distinct pitch accent distributions (Schweitzer et al., 2009).

### 5. Acknowledgements

This work was funded by the Collaborative Research Center Incremental Specification in Context (SFB 732) at the University of Stuttgart. We particularly would like to thank Stefan Baumann, Dag Haug, Hans Kamp and Massimo Poesio for discussion, as well as two anonymous reviewers for their comments.
said Kirchner in Córdoba . . . emphasised the Argentinian head of state

Figure 2: Anaphoric link: GIVEN-EPITHET

Table 7: Annotation results for two coders
6. References


