An annotation scheme for information status in dialogue [*]

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman

Institute for Communicating and Collaborative Systems
{mnissim|sdingar1|jeanc|steedman}@inf.ed.ac.uk
University of Edinburgh, United Kingdom

Abstract

We present an annotation scheme for information status (IS) in dialogue, and validate it on three Switchboard dialogues. We show that our scheme has good reproducibility, and compare it with previous attempts to code IS and related features. We eventually apply the scheme to 147 dialogues, thus producing a corpus that contains nearly 70,000 NPs annotated for IS and over 15,000 coreference links.

1 Introduction

In this paper we define the Information Status (IS) of an entity as reflecting the speaker's assumptions about the hearer's knowledge/beliefs, and we express it by the well-known old/new distinction. This distinction can also be thought of as indicating how much a discourse entity contributes to changing or updating the discourse model.[1] It has long been recognised that the IS of entities influences syntactic form (e.g. (Ariel, 1990; Prince, 1992; Gundel et al., 1993; Birner and Ward, 1998; Wasow, 2002)). Thus, a corpus annotated for IS would allow detailed studies of syntactic choice (such as active vs. passive and dative placement), as well as provide useful information for other NLP tasks such as anaphora resolution, parsing, and text classification. This paper presents the first scheme purely developed for IS annotation for all NP types, and figures for its reliability obtained from a study on a portion of the Switchboard corpus (Godfrey et al., 1992). The annotation scheme is described in Section 2, together with some theoretical background on IS. In Section 3 we present a study aimed at validating the scheme, and we discuss the results and the distribution of entities in the final corpus. Finally, we compare this work with related annotation efforts and conclude (Section 4). The companion paper (Carletta et al., 2004) describes the processing techniques used to achieve the results presented here.

2 Annotation Scheme

Our annotation scheme mainly builds on (Prince, 1992) and (Eckert and Strube, 2001), as well as on related work on annotation of anaphoric links (Passonneau, 1996; Hirschman and Chinchor, 1997; Davies et al., 1998; Poesio, 2000). In defining "old" and "new", Prince uses two cross-cutting dichotomies: the *hearer*'s point of view and the *discourse model*. We consider here three of the four possible combinations that arise: if an entity is both known to the hearer and has already been mentioned in the conversation it is defined as *old*; if it is unknown to the hearer and has not been previously referred to is *new*. If it is newly mentioned in the dialogue but the hearer can *infer* it from the previous conversation it is *mediated*.[2] The latter is the

case of generally known entities (such as "the sun", or "the Pope" (Löbner, 1985)), and *bridging* (Clark, 1975), where an entity is related to a previously introduced one.[3]

We use such a three-way classification because finer-grained distinctions for IS (e.g. (Prince, 1981; Lambrecht, 1994)) have proved hard to distinguish reliably in practice (see Section 4). However, we organised our scheme *hierarchically*, so that finer-grained categories can be specified as subtypes for the main classes. The benefit of this approach lies in the fact that it preserves a high-level, more reliable distinction while allowing a finer-grained classification that can be exploited for specific tasks.

In addition to the three main categories, we introduced two classes to deal with errors in the markable extraction procedure and with problems in text comprehension. The automatic extraction of markables is not flawless, also due to mistakes and/or inconsistencies in the original Treebank annotation, so that some markables should actually be excluded (Carletta et al., 2004). For example, it would be meaningless to assign an IS value to "course" in the phrase "of course", or "there" in "there is/are". A category non-applicable is used for such cases, for idiomatic occurrences, and expletive uses of "it". Traces are automatically extracted as markables, but are left unannotated (no tag is assigned). Rarely, the annotators find some fragments difficult to understand, thus making it impossible for them to make a decision on the IS of a specific entity. When this happens, a category not-understood can be assigned. Entities marked as non-applicable or not-understood are excluded from any further annotation. For all other markables, the annotators must choose between old, mediated, and new. For the first two, subtypes *can* also be specified: subtype assignment is encouraged but not compulsory.

Old An entity is old if it has been previously mentioned, i.e. if it is *coreferential* with an already introduced entity, if it is a generic pronoun, or if it is a personal pronoun referring to the dialogue participants. Six different subtypes are available for old entities: identity, event, general, generic, ident_generic, relative. In Example 1, for instance, "us" would be marked as old because it corefers with "we", and a subtype identity would also be assigned.[4]

---

[1] We follow (Prince, 1992) in using " old" rather than "given" to refer to "not-new" information, but regard the two as identical. For the sake of space we do not discuss terminology. For an overview see (Vallduví, 1992; Steedman, 2000).

[2] This type corresponds to Prince's (1981; 1992) *inferrables*.

[3] The fourth combination (an entity already introduced in the dialogue but unknown to the hearer) is theoretically plausible but too rare to be useful to code.

[4] All examples in this paper are from the Switchboard Corpus. The markable in question is typed in boldface; antecedents or trigger entities, where present, are typed in italics.

(1)   [. . .] *we* camped in a tent, and uh there were two other couples with **us**.

In addition, a coreference link is marked up between anaphor and antecedent, thus creating anaphoric chains (see also (Carletta et al., 2004)). The subtype event applies whenever the antecedent is a verb phrase (VP) rather than an NP. In Example 2, "it" is old/event, as its antecedent is the VP "educate three". As we do not consider VPs as markables, no link can be marked up.

(2)   I most certainly couldn't *educate three*. I don't know how my parents did **it**.

Also classified as old are personal pronouns referring to the dialogue participants as well as generic pronouns. In the first case, a subtype general is specified, whereas the subtype for the second case is generic. An instance of old/generic is "you" in Example 3.

(3)   up here **you** got to wait until Aug- August until the water warms up.

In a chain of generic references, the subtype ident_generic is assigned, and a coreference link is marked up. Coreference is also marked up for relative pronouns: they receive a subtype relative and are linked back to their head.

Mediated   Mediated entities have not yet been directly introduced in the dialogue, but are inferrable from previously mentioned ones, or generally known to the hearer. We specify nine subtypes: general, bound, part, situation, event, set, poss, func_value, aggregation.[5] Generally known entities such as "the moon" or "Italy" are assigned a subtype general. Most proper nouns fall into this subclass, but it is up to the annotator to decide between a mediated/general or a new category, depending on the context.[6]

Also mediated are bound pronouns, such as "them" in Example 4, which are assigned a subtype bound.

(4)   [. . .] it's hard to raise *one child* without **them** thinking they're the pivot point of the universe.

A subtype poss is used to mark all kinds of intra-phrasal possessive relations (prenominal as well as postnominal).

Four subtypes (part, situation, event, and set) are specifically used to mark instances of bridging, i.e. entities that are inferrable because a related entity has been previously introduced in the dialogue. The subtype part is used to mark part-whole relations for physical objects, both as intra- and inter-phrasal relations. (This category is to be preferred to poss whenever applicable.) The occurrence of "the door" in Example 5, for instance, is annotated as mediated/part.

(5)   When I come *home* in the evenings my dog greets me at **the door**.

For similar relations that do not involve physical objects, i.e. if an entity is part of a situation set up by a previously introduced entity, we use the subtype situation.[7] This applies, e.g., to the NP "the specifications" in Example 6.

(6)   I guess I don-, don't really have a problem with *capital punishment*. I'm not really sure what [breathing] **the exact specifications** are for Texas.

Like in the category old, we introduce a subtype event. This is applied whenever an entity is related to a previously mentioned VP. In Example 7, e.g., "the bus" is triggered by *travelling around Yucatan*.

(7)   We were *travelling around Yucatan*, and **the bus** was really full.

Whenever an entity referred to is a subset of, a superset of, or a member of the same set as a previously mentioned entity, the subtype set is applied.

Rarely, an entity refers to a value of a previously mentioned function, as "zero" and "ten" in Example 8. In such cases a subtype func-value is assigned.

(8)   I had kind of gotten used to *centigrade temperature* you know – if it's between **zero** and **ten** it's cold.

Lastly, a subtype aggregation is used to classify coordinated NPs. Two old entities do not give rise to an old coordinated NP, unless it has been previously introduced as such. A mediated/aggregation tag is assigned instead.

New   The category new is assigned to entities that have not yet been introduced in the dialogue and that the hearer cannot infer from previously mentioned entities. No subtypes are specified for this category.

The guidelines contain a decision tree the annotators use to establish priority in case more than one class is appropriate for a given entity. For example, if a mediated/general entity is also old/identity the latter is to be preferred to the former. Similar precedence relations hold among subtypes.

To provide more robust and reliable clues in annotating bridging types (e.g. for distinguishing between poss and part), we provided replacement tests specified for each type and referred to relations encoded in knowledge bases such as WordNet (Fellbaum, 1998) (for part) and FrameNet (Baker et al., 1998) (for situation).

3   Validation of the Scheme

We describe here the experiment we carried out to test the reliability of the annotation scheme and the distribution of the categories in a corpus of 147 dialogues.

3.1   Corpus Collection and Preparation

The corpus we use for our study is Switchboard, a collection of spontaneous telephone conversations, averaging 6 minutes in length, between speakers of American English on predetermined topics. A set of approximately 650 dialogues is parsed, as part of the Penn Treebank. We used a portion of this set for our study. Three dialogues were used to assess the annotation scheme and 147 in total were eventually annotated. All dialogues were converted into XML

---

[5]Some of the subtypes are inspired by categories developed for bridging markup in annotation schemes for anaphoric phenomena, especially in (Passonneau, 1996; Davies et al., 1998).

[6]Ariel (1990) argues for a difference between first, last and full proper nouns. This insight could be taken into account by treating full names as likely to be new and first names mediated.

---

[7]This includes elements of the thematic grid of an already introduced entity. It subsumes Passonneau's (1996) class "arg".

(Carletta et al., 2004). We exploited the pre-existing morphological and syntactic markup to automatically select and filter NPs to be annotated. Locative, directional, and adverbial NPs were excluded. Disfluencies were also omitted. Possessive pronouns were added to the set of markables.

3.2 Method

The reliability of the scheme was assessed on three Switchboard dialogues, containing a total of 1738 NPs. The annotators were the first author (AnnM) and a paid annotator with a background in AI/linguistics (AnnV). The annotators followed specific guidelines containing instructions, examples, and a decision tree. The annotation was performed using the NITE XML Toolkit, specifically customized for this task (Carletta et al., 2004).

A demonstration dialogue was separately annotated by AnnM and the second author in a first pass, and then together in a consultation and disagreement reconciliation phase to obtain a gold standard. AnnV used this demonstration file to get acquainted with the annotation scheme. All questions and comments arising during this phase were discussed with AnnM, and the guidelines were amended accordingly. After this exercise, AnnM and AnnV separately annotated a dialogue (DiaA) on the same topic as the demonstration dialogue ("family life") for further training. The annotation was performed independently, and afterward divergences were discussed and guidelines amended. Two more dialogues (DiaB and DiaC, on different topics) were then annotated to further assess reliability.

3.3 Results and Discussion

We evaluated annotation reliability by using the Kappa statistic (Carletta, 1996). Good quality annotation of discourse phenomena normally yields a kappa (K) of about .80. We assessed the validity of the scheme on the four-way classification into the three main categories (old, mediated and new) and the non-applicable category. The latter was included as the exclusion of certain markables is not always straightforward and annotators might disagree upon some cases. We also evaluated the annotation including the subtypes. All cases where at least one annotator assigned a not-understood tag were excluded from the agreement evaluation (14 markables). Also excluded were all traces (222 markables), which the annotators left unmarked. The total markables considered for evaluation over the three dialogues was therefore 1502.

The annotation of the three dialogues yielded $K = .845$ for the high-level categories, and $K = .788$ when including subtypes ($N = 1502$; $k = 2$).[8] These results show that overall the annotation is reliable and that therefore the scheme has good reproducibility. When including subtypes agreement decreases, but backing-off to the high-level categories is always possible, thus showing the virtues of a hierarchically organised scheme (see Section 4).

Theoretical issues and the annotators's experience suggested that some categories are more difficult to distinguish than others. We therefore carried out reliability tests for single categories. Indeed, it emerged that mediated and new were more difficult to apply than old, for which agreement

---

[8] N stands for the number of instances annotated and k for the number of annotators. Unless otherwise specified, $N = 1502$ and $k = 2$ hold for all K scores reported in this paper.

was measured at $K = .902$. Both mediated and new are quite reliable anyway, with $K = .800$ and $K = .794$ respectively. Agreement for non-applicable was $K = .846$.

Table 1: Reliability for old and mediated subtypes. T indicates how many times a given category was chosen by the annotators. T=11, for instance, might indicate that it was chosen 4 times by one annotator and 7 by the other.

CATEGORY	K (T)	CATEGORY	K (T)
o/identity	.904 (601)	m/part	.594 (10)
o/event	.837 (92)	m/situation	.719 (46)
o/general	.937 (365)	m/event	.794 (20)
o/generic	.845 (112)	m/set	.696 (244)
o/id_generic	.876 (95)	m/poss	.907 (87)
o/relative	.982 (59)	m/func_value	n/a (0)
m/general	.862 (104)	m/aggreg.	1.00 (16)
m/bound	.961 (27)		

Table 1 summarises reliability scores for the subtypes. As expected, old subtypes are easier to assign than mediated ones. Within the latter, subtypes for which syntactic clues are of help (such as poss and aggregation) are more reliable than those for which semantics and pragmatics play a stronger role, such as set, situation, and part. Grouping some classes (such as poss and part, or part and situation) and/or developing better identification clues are directions we will explore to improve reliability. Although no instances of func_value were found in the assessment dialogues, further annotation revealed that this subtype is quite easy to assign, although generally rare (see Figure 1).

Training of annotators and amendments of the guidelines with additional examples and instructions seem to have had a significant influence on the reliability of the scheme. The annotation of the training dialogue DiaA yielded $K = .794$ ($N = 371$; $k = 2$), whereas K on DiaB and DiaC was measured at .860 ($N = 579$; $k = 2$), and .857 ($N = 552$; $k = 2$), respectively. Training also significantly influenced the reliability of the category non-applicable, as annotation of DiaA yielded an unsatisfactory $K = .614$, but for DiaB and DiaC it yielded $K = .916$ and $K = .888$, respectively.

The annotators found the decision tree very useful when having to choose between more than one applicable subtypes, and we believe it had a significant impact on the reliability of the scheme.

The scheme was then applied for the annotation of more Switchboard dialogues. Currently, our corpus is composed of 147 dialogues for a total of 43358 sentences with 69004 marked up valid NPs, 35299 of which are old, 23816 mediated and 9889 new (8127 were excluded as non-applicable, and 160 were not understood), and 16324 coreference links. Figure 1 shows the distribution of old and mediated subtypes in the resulting corpus.

4 Related Work and Conclusions

To our knowledge, (Eckert and Strube, 2001) is the only other work that explicitly refers to IS annotation. They also use a Prince's (1992)-based old/mediated/new distinction for annotating Switchboard dialogues. However, their IS annotation is specifically designed for salience ranking of candidate antecedents for anaphora resolution, and not described in detail. They do not report figures on inter-annotator agreement so that a proper comparison with our
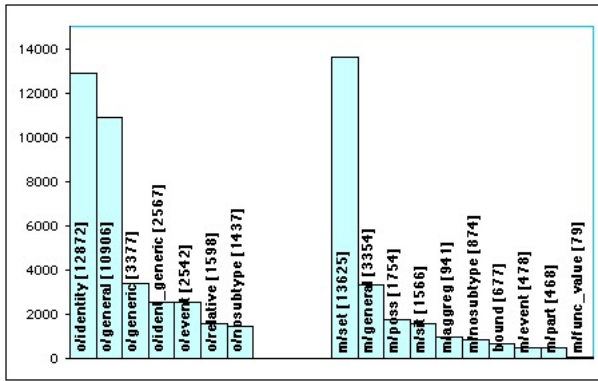
Figure 1: Distribution of old and mediated subtypes

experiment is not feasible. Among the schemes that deal with annotation of anaphoric NPs, our scheme is especially comparable with DRAMA (Passonneau, 1996) and MATE (Davies et al., 1998). Both schemes have a hierarchical structure. In DRAMA, types of *inferrables* can be specified, within a division into conceptual (mainly pragmatically determined) vs. linguistic (mainly based on argument structure) inference. No annotation experiment with interannotator agreement figures is however reported. MATE provides subtypes for bridging relations, but they were not applied in any annotation excercise, so that reliability and distribution of categories were only tested on the "core scheme" that only concerns true coreference.

Our experiment aimed at validating the scheme is partially comparable with work on the annotation of anaphoric relations, especially where bridging annotation is involved, as it relates to our mediated entities. However, the comparison is not straightforward. First, specific syntactic classes (e.g. definite NPs and demonstratives) are usually preselected for annotation, whereas we mark up all NPs. Second, we annotate IS as such, whereas they try to classify different *uses* of specific NP types. Third, we do not try and identify a single antecedent for mediated (bridging) NPs.

Poesio and Vieira (1998) describe two experiments for the classification of definite descriptions. Their annotation schemes are mainly built on Prince's (1981; 1992) and Hawkins' (1978) taxonomies. A first experiment with four classes (coreference, bridging, discourse new generally known entities, and idioms) yields $K = .68$ ($N = 1040$; $k = 3$). K rises to $.73$ if idioms (comparable to our non-applicable class) are excluded. There is no specific class for new entities (arguably because they are only concerned with definites). In a second experiment they classify types of definites into four classes (coreference, bridging, larger situation, and unfamiliar) using semantic rather than syntactic criteria. The annotation, performed by naïve annotators, yields $K = .63$ ($N = 430$; $k = 3$). By grouping the classes *after* the annotation into only two categories (coreferential vs. discourse new) they obtain $K = .76$.

Spenader (2001) also carried out a study on definite NPs by using a flat classification into eight classes, ranging from coreference to new. The annotation experiment yielded $K = .45$ ($N = 406$; $k = 2$). Spenader ascribes the main responsibility for the low agreement to the fact that many entities are related to the previous context in different ways, rather than a single one. We believe the large number of classes also contributes to the low performance.

Salmon-Alt and Vieira (2002) annotated NPs introduced by the definite article and demonstrative determiners in French and Portuguese, allowing classification into four categories (pronominal coreference, full NP coreference, other types of anaphora, new). They obtained $K = .52$ ($N = 461$; $k = 2$) for French definite NPs, and $K = .48$ ($N = 541$; $k = 2$) for Portuguese. For demonstratives, K was calculated only on three classes (other anaphora types were merged with new *after* the annotation), and was better than for definites ($K = .79$ ($N = 291$; $k = 2$) for French and $K = .65$ ($N = 243$; $k = 2$) for Portuguese).

These studies show that an improvement in results can be obtained only when conflating different classes. This was done *after* the annotation, though, since the original schemes were flat. Our hierarchically organised scheme neatly circumvents this problem without giving up further specification. By allowing a higher-level classification: mediated, it also reflects more naturally the fact that some entities are mediated via more than a single specific relation in the context. The same effect can be noted in the difference in agreement when including or excluding subtypes.

To our knowledge, the corpus we have annotated is the largest available with this kind of annotation, in addition to other word- and sentence-level markup. Among the several applications such a resource lends itself to are discourse analysis, text classification, and language generation.

5 References

M. Ariel. 1990. *Accessing Noun Phrase Antecedents*. Routledge, New York.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of COLING-ACL*, pages 86–90.

B. Birner and G. Ward. 1998. *Information Status and Noncanonical Word Order in English*. Cambridge University Press.

J. Carletta, S. Dingare, M. Nissim, and T. Nikitina. 2004. Using the NITE XML Toolkit on the Switchboard Corpus to study syntactic choice: a case study. In *Proc. of LREC2004*.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

H. H. Clark. 1975. Bridging. In R. Schank and B. Nash-Webber, editors, *Theoretical Issues in Natural Language Processing*. The MIT Press, Cambridge, MA.

S. Davies, M. Poesio, F. Bruneseaux, and L. Romary. 1998. Annotating coreference in dialogues: Proposal for a scheme for MATE, July. http://www.hcrc.ed.ac.uk/~poesio/anno_manual.html.

M. Eckert and M. Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, pages 517–520.

J. Gundel, N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

J. A. Hawkins. 1978. *Definiteness and Indefiniteness*. Croom Helm, London.

L. Hirschman and N. Chinchor. 1997. MUC-7 coreference task definition. In *Proc. of MUC-7*.

K. Lambrecht. 1994. *Information structure and sentence form. Topic, focus, and the mental representation of discourse referents*. CUP, Cambridge.

S. Löbner. 1985. Definites. *Journal of Semantics*, 4:279–326.

R. Passonneau. 1996. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript, December.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite descriptions use. *Computational Linguistics*, 24(2):183–216.

M. Poesio. 2000. The GNOME annotation scheme manual. http://www.hcrc.ed.ac.uk/~gnome/anno_manual.html.

E. F. Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*. Academic Press, New York.

E. Prince. 1992. The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins, Philadelphia/Amsterdam.

S. Salmon-Alt and R. Vieira. 2002. Nominal expressions in multilingual corpora: Definites and demonstratives. In *Proc. of LREC2002, 2002*, pages 1627–1634.

J. Spenader. 2001. Between binding and accommodation. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Proceedings of Bi-Dialog 2001*, pages 162–173.

M. Steedman. 2000. *The Syntactic Process*. The MIT Press, Cambridge, MA.

E. Valldduví. 1992. *The Informational Component*. Garland, New York.

T. Wasow. 2002. *Postverbal Behavior*. CSLI Publications.