A recognition–based meta–scheme for dialogue acts annotation

Claudia Soria, Vito Pirrelli
CNR - Istituto di Linguistica Computazionale
Via della Faggiola 32, I-56126 Pisa, Italy
{soria,vito}@ilc.pi.cnr.it

Abstract

The paper describes a new formal framework for comparison, design and standardization of annotation schemes for dialogue acts. The framework takes a recognition–based approach to dialogue tagging and defines four independent taxonomies of tags, one for each orthogonal dimension of linguistic and contextual analysis assumed to have a bearing on identification of illocutionary acts. The advantages and limitations of this proposal over other previous attempts are discussed and concretely exemplified.

1 Introduction

Recent years have witnessed a growing concern with the provision of ***standardized*** formats for exchange, integration and use of ***shareable*** annotated dialogues, and the resulting development of formal frameworks intended to compare, standardize and customize annotation schemes for dialogue acts (see (Allen and Core, 1997; Core and Allen, 1997; Larsson, 1998; Ichikawa et al., 1998). Arguably, these efforts should be instrumental in speeding up progress in the field, meeting at the same time the rapidly increasing demands of dialogue system technology.

It is important to observe that any framework of this kind should be able to ***explicitly*** characterize both scope and nature of the dialogue phenomena covered by a given tag set, since they appear to vary considerably from scheme to scheme, as a function of i) the analytical standpoints adopted and ii) the dimensions of linguistic and contextual analysis taken into account. We hereafter introduce some key–ideas (namely, ***recognition–based*** vs ***generation–based annotation*** and ***annotation meta–scheme***) that have, in our view of things, the potential of making explicit in a principled and declarative way the relationship between tag definitions and underlying dimensions of analysis. Careful consideration of this relationship makes it possible to conceive of a dialogue tag as a point in an n–dimensional space, rather than as an undecomposable conceptual unit. As we will see, this offers a number of advantages over other existing approaches to scheme comparison and standardization.[1]

1.1 Recognition–based annotation

It is useful to recognize two complementary approaches to labeling utterances with dialogue acts, hereafter referred to for convenience as a *generation-based* and a *recognition-based* perspective. The generation perspective is chiefly concerned with the question "given a dialogue utterance, what underlying mental process might have produced it?". Such a mental process can be defined i) as a communicative "intention", or, alternatively, ii) in terms of a formal characterization of the reasoning process underlying dialogues, with specific emphasis on the effects of speech acts on the agents' ***mental states*** (or ***information states***) and, ultimately, on dialogue planning (Poesio and Traum, 1998; Poesio et al., 1999). The recognition perspective, on the other hand, addresses the question: "given a dialogue utterance, on the basis of what available linguistic or contextual clues can one recognize its underlying intention(s)?". By linguistic and contextual clues, we mean here a variety of more or less overtly available information, ranging from the surface linguistic realization of an utterance, to its propositional content and the pragmatic context where the dialogue is situated.

A generation-based approach lays emphasis on the (assumed) ***accessibility*** of the mental states/intentions of a speaker in a dialogue, either through an explicit representation of these states (as feature–based informational structures), or through a step of abductive inference on the annotator's part. In the recognition–based approach, attention is shifted to the ***interpretability*** of an utterance as conveying a certain intention, where interpretability is a function of the information available to the hearer/annotator at a certain point in time. Ideally, the two perspectives should lead to the same annotated dialogue. In practice, this is often not the case, due to the wide range of variation in the information accessible to the hearer/annotator.

[1] This work is carried out in the framework of the MATE project. In particular, we would like to acknowledge our debt to (Klein et al., 1998).

In a generation–based approach, an utterance can simultaneously *be intended* to respond, promise, request, inform etc. A recognition–based perspective makes use of a different notion of *multifunctionality* whereby several intentions can be *recognized on the basis of distinct dimensions of linguistic and extra–linguistic information*. For example, an utterance like `I want to go to Boston` can be i) a claim, if judged on its linguistic declarative form only, ii) an answer, relative to a previously uttered request, and iii) an order, if - say - addressed to a taxi–driver. In this perspective, it is relatively immaterial whether, e.g., the utterance was *ultimately* and *primarily* intended as an assert; rather, it is sufficient to observe that one *could* interpret `I want to go to Boston` as an assert, on the basis of a certain type of available linguistic or contextual information.

It is important to emphasize at this stage that virtually no existing annotation scheme for dialogue acts *can be said to instantiate either perspective only.* In fact, the vast majority of tag sets exhibit, to different degrees, a combination of the two approaches. In the remainder of this paper, we will elaborate the recognition–based perspective as a basis for annotation scheme comparability, standardization and customization.

1.2 The notion of meta–scheme

We call an *annotation meta–scheme* a formal framework for comparing annotation schemes, which can also be used as a practical blue–print to scheme design and customization. A crucial feature of the annotation meta–scheme illustrated here is that it is intended to *make explicit the type of linguistic and contextual information relied upon in the process of tagging dialogue utterances with illocutionary acts.* In this respect, the meta–scheme is chiefly recognition–based.

In practice, this is achieved by defining one independent taxonomy of utterance tags for each of the orthogonal dimensions of linguistic or contextual analysis which have a bearing on the definition of dialogue acts. For example, in some cases dialogue acts are identified on the basis of the *linguistic form* of an utterance only. We thus find it convenient to define an autonomous typology of tags based on purely grammatical facts such as, e.g., subject–auxiliary inversion, *wh*–words, a rise of intonation etc. Surely, tags defined along this dimension will often fail to convey the *primary intention* of a given utterance: for example, an interrogative sentence may conceal an order, and an explicit performative may turn an assert into a request. Yet this should not worry us, as long as the relation between a tag and its supporting dimension of analysis is explicitly stated.

It should be appreciated that, in existing annotation schemes, the relationship between linguistic and contextual clues on the one hand and tag definitions on the other hand is characterized only implicitly. Linguistic and contextual dimensions of analysis are simultaneously drawn upon in tag definitions in a complex way, so that the relationship of these dimensions with each tag is often only indirect. This will be illustrated in more detail in the following sections. Suffice it to point out here that, far from being a methodological flaw, this practice responds to the practical need of annotating utterances in a maximally economic way, i.e. with the sparsest possible set of tags. Clearly, requirements of economy and ease of annotation are appropriate for labeling a dialogue text with a specific application or a specific theoretical framework in mind. However, they may get in the way when it comes to *comparing* different annotation schemes, or *exporting* the annotation scheme developed for a given application to another domain. In these latter cases, perspicuity of the linguistic and contextual content of tags should be given priority over other more practical concerns.

2 .Previous standardization efforts

In this section we will sketchily overview two of the most important attempts at providing standardized dialogue–act tags for general annotation, namely DAMSL (Allen and Core, 1997; Core and Allen, 1997) and Larsson's (Larsson, 1998), with particular emphasis on the assumptions underlying their methodological approach.

DAMSL is certainly the most influential effort in the provision of standards for dialogue annotation to date (Allen and Core, 1997; Core and Allen, 1997). It is designed to offer a general, underspecified scheme, potentially usable in different domains, and susceptible of further specification into finer grained domain–specific categories. DAMSL is credited for taking the issue of utterance multifunctionality most seriously: an utterance can be tagged at the same time along several orthogonal dimensions of annotation, each of them defining an independent layer of communicative intention. Accordingly, the same utterance can be interpreted, e.g., as giving information, making a request, making a promise etc. It is important to emphasize here that, in DAMSL, multiple dimensions serve the purpose of capturing different facets of an illocutionary act and are not intended to *directly* reflect the different linguistic and contextual dimensions on the basis of which these facets are recognized. In this sense, DAMSL multi–dimensionality is predominantly generation–based. Nonetheless, tag definitions are a mixed bag of generation and recognition–based criteria.

At the core of the DAMSL taxonomy lies a bipartition between the so–called forward– and backward–looking dialogue functions, a fairly faithful rendering of Searlian speech act categories (Searle, 1969).

The assumed orthogonality of all dimensions makes virtually *any* combination of DAMSL dimensions admissible for annotation, in a potentially combinatorial explosion of multiple tags. Finally, although originally conceived as a meta-scheme, DAMSL has been used and circulated since its conception as yet another independent scheme in its own right, often proving too general to be of practical use. More importantly, the fact that it provides non-exclusive categories seems to have a negative impact on its reliability (Core and Allen, 1997).

A different approach to standardization is taken in Larsson (Larsson, 1998), who suggests to model the comparison of two different encoding schemes as a *mapping function* between the two corresponding hierarchies of tags (taxonomies). The correspondence induced by the mapping function can be *one-to-one*, *one-to-many* and *one-to-none*. Two tags which are in a one-to-one relationship are taken to be *synonymous*. A one-to-many relationship is interpreted as suggesting that one tag in a taxonomy *subsumes* more than one tag in another taxonomy, as illustrated in figure 1 for the relationship between Info-request in DAMSL and the tags Check, Align, Query-yn and Query-w in the HCRC MAP TASK annotation scheme (Carletta et al., 1996). One-to-many mappings (and many-to-one) hold between those branches in two taxonomies which are specified at different levels of granularity. Finally, a one-to-none correspondence signifies that a particular taxonomy is silent on a range of phenomena which happen to be overtly marked in another taxonomy. For instance, since MAP TASK provides no tag for the category of *commissives*, this phenomenon is understood to be covered by tags provided in DAMSL only. Eventually, a more general and comprehensive hierarchy subsuming the two compared schemes is built by a) taking the intersection set of synonymous tags, b) taking one-to-none tags from either taxonomy only, c) representing a one-to-many tag relationship as a mother-daughters hierarchy of the corresponding nodes. For reasons that will be made clear in the following section, this approach ends up considerably *re-defining* scope and applicability of the tags considered. For example, when a Reply-y of MAP TASK is classified as a daughter node of DAMSL Answer, one is in fact ignoring that, in MAP TASK, Reply-y has a rather broader scope than the one entailed by this correspondence.

To sum up, the standardization efforts reviewed in this section are not concerned with drawing a principled line between a generation-based and a recognition-based perspective. As a result, tags of different schemes are typically related to one another through functional synonymy, subsumption or generation-based multifunctionality. As we will see in the following section, this may in some cases obscure the precise nature of these relations.

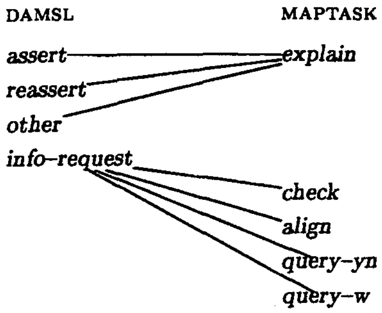

Figure 1: Many-to-one/one-to-many mapping

3 Scheme Comparison

As already pointed out above, Larsson's approach to developing more comprehensive tag hierarchies by mapping comparable tag sets logically presupposes three types of correspondence being at work: one-to-one, one-to-many and one-to-none. This is pictorially illustrated in figure 1, which summarizes Larsson's (Larsson, 1998) mapping function between DAMSL and MAP TASK, in the area of asserts and requests. However, the assumption that different tag sets tend to partition the *same range of phenomena at different levels of granularity*, in much the same way two taxonomies may mutually differ at the level of depth at which (some of) their branches are specified, is unwarranted. In fact, different annotation schemes take different analytical perspectives on dialogue phenomena, and end up with carving them up into different categories. This situation typically produces *many-to-many* tag correspondences.

In a pilot experiment, we used four different dialogue-act schemes[2] to annotate a small corpus of five English task-oriented dialogues.[3] All dialogues were manually tagged by two different annotators with all annotation schemes. We then counted, for any pair of tags t_A and t_B in the tag sets A and B, how many times they are found to mark the same

[2] The annotation schemes considered are: DAMSL (Allen and Core, 1997), DAMSL-*Switchboard* (Jurafsky, Shriberg, and Biasca, 1997), VERBMOBIL 2 (Alexandersson et al., 1998), and the HCRC MAP TASK annotation scheme (Carletta et al., 1996).

[3] Sources: a human-human dialogue on room furnishing, from the COCONUT corpus (di Eugenio, Jordan, and Pylkkaenen, 1997); a human-human dialogue of appointment scheduling, from the VERBMOBIL corpus (Alexandersson et al., 1998); one human-human dialogue of instruction giving, from the MAP TASK corpus (Carletta et al., 1996); one human-machine dialogue containing travel information, from the TOOT corpus (see `http://www.cs.umd.edu/users/traum/DSD/hw1.html`); one WOZ dialogue on interactive problem solving, from the TRAINS corpus (see `http://www.cs.rochester.edu:80/re-search/trains/annotation`).

MAPTASK DAMSL	explain	check	query-yn	query-w
assert	0.43	0.02	0.01	-
reassert	0.5	0.5	-	-
open–option	0.2	-	0.17	0.42
offer	0.42	0.1	0.2	0.26
conv–opening	0.5	-	0.5	-
info–request	-	0.12	0.34	0.54

Table 1: Many–to–many mapping I

token utterance. This measure is proportional to the degree of translatability between tag sets, and provides a firmer ground for assessing their level of correspondence than sheer inspection of tag definitions does. Results of the experiment show that the prevalent pattern of correspondence is, in fact, *many–to–many*. Table 1 illustrates this point, showing the actual correspondences between DAMSL and MAP TASK, in the common area of asserts and requests. For each slot of table 1 at the crossing of DAMSL tag t_D and MAP TASK tag t_M, we report the averaged number of times an utterance labeled as t_D is also assigned t_M, divided by the total number of utterances tagged as t_D. These figures show two things. First, Larsson's mappings reflect prevalent patterns of tag correspondence only partially. Secondly, such patterns are far from being exhaustive of the range of possible use of the tags involved. To give but one example, out of 10 utterances tagged as MAP TASK Explain in one of our test dialogues, 9 are tagged as DAMSL Assert, 6 as DAMSL Offer, and 3 as DAMSL Open-option. We conclude that Larsson's approach is useful to uncover degrees of correspondence between tag sets, but is still too shallow to shed light on the nature of this correspondence.

Let us now compare MAP TASK and VERBMOBIL. Both schemes are mono–dimensional, meaning that they assign only one tag per utterance. Yet, this does not seem to simplify their pattern of correspondence, which turns out to be, once more, many–to–many, as illustrated in table 2. Consider, for example, the relationship between MAP TASK Reply-y and VERBMOBIL Accept and Feedback-positive. Reply-y is almost exclusively concerned with the *linguistic form* of an utterance, while VERBMOBIL Accept and Feedback-positive are mainly based on the relationship between a reply and the propositional content of the utterance being replied to. This important difference is levelled out when one tries to represent it as a mapping function from the MAP TASK tag set onto the tag set of VERBMOBIL. A more promising key to an understanding of the intricate relationship between MAP TASK and VERBMOBIL can be found when things are looked at from a purely recognition–based perspective. It turns out that the dimensions of information *implicitly* called upon in the definition of most existing dialogue tag sets are considerably varied. To limit ourselves to some of the tags in table 2, such dimensions range from syntax (Reply-y) to propositional content (Feedback-positive) and co–text (Accept). Many–to–many mapping can thus be viewed as the result of the following situation: i) for each tag set, tags are defined in relation to their relevance to an intended goal (be it practical or theoretical); ii) the definition calls upon a number of relatively independent classificatory dimensions; iii) neither all tags in the same tag set nor tags belonging to different schemes consistently share the same dimensions. This situation is illustrated in more detail in the following sections.

MAPTASK VERBMOBIL	acknldg	reply-y	ready
accept	0.77	0.23	-
feedbck–positive	0.34	0.46	0.2
backchannel	0.45	-	0.45

Table 2: Many–to–many mapping II

4 Recognition–based comparability

The classificatory dimensions selected in this section for a recognition–based comparison are simply those more consistently (however implicitly) assumed for tag definition by the dialogue–acts community. In particular, each dimension in the list below covers a specific level of information taken as criterial for tag–assignment in the tag definitions overviewed in our pilot experiment:

- **D1, Grammatical information**: tag-assignment presupposes availability of morphosyntactic, syntactic, prosodic and lexical information (limited to grammatical words only): see, for example, wh–questions and yes-/no–questions in SWITCHBOARD
- **D2, Information about lexical and semantic content**: tag-assignment presupposes knowledge about the propositional content of an utterance, e.g. in terms of its logical structure, topic representation, inter-clausal dependencies within the utterance and occurrence of semantically full words (as opposed to grammatical words): see, for example, the category Assert

in DAMSL, defined as a truth-conditional claim about the world

- **D3, Co-textual information**: tag-assignment presupposes knowledge of the previous/following utterance(s) (see all "backward-looking" or responsive categories)
- **D4, Pragmatic information**: tag-assignment requires knowledge of the context of the dialogue: e.g. the social relationship of speaker/hearer, the physical setting of the interaction, the specific domain talked about etc.: this is the case of indirect speech acts, such as `I'm cold`, tagged as an order when used to mean `Close the window`.

By way of illustration, table 3 below provides a recognition–based interpretation of tags in DAMSL, SWITCHBOARD, MAP TASK and VERBMOBIL, related to Searle's class of Representatives.

Category & Scheme	D1	D2	D3	D4
Assert (DAMSL)	+	+	+	
Statement (SWBD)	+	+		
Explain (MAPTASK)	+		+	+
Inform (VERBMOBIL)	+	+	+	+

Table 3: Assert Categories vs Dimensions

An *Assert* in DAMSL is an utterance "whose primary intention is to make claims about the world, also in the weaker form of hypothesizing or suggesting that something might be true" (Allen and Core, 1997). A typical Assert, thus, will be realized with a declarative clause type and a specific prosodic contour (D1 in table 3); moreover, an Assert is defined as an utterance whose propositional content is truth-conditional (D2) and has new informational status (D3).

The general category Statement in SWITCHBOARD (Jurafsky, Shriberg, and Biasca, 1997) is mainly identified on the basis of lexical and grammatical information, more or less of the kind required for Assert in DAMSL. In particular, a Statement-non-opinion requires co-occurrence of first–person personal pronouns (D1), and of a personal story as the content of the utterance (D2). Similarly, a Statement-opinion presupposes verbs expressing opinion such as "think" and "believe" (D1) and a personal opinion as the content of the utterance (D2). The Explain category in MAP TASK is defined as an utterance "stating information which has not been elicited by the partner" (Carletta et al., 1996). Thus, recognition of an instance of Explain involves, besides lexico-grammatical clues about the linguistic form of an utterance (D1), also consideration of adjacency-pairs constraints (D3). D4 is also indirectly invoked to disambiguate between a true Explain and a declarative utterance used as an order (Instruct). Finally, Inform in VERBMOBIL (Alexandersson et al., 1998) is defined as a default tag, to be used when other tags fail to apply. This makes it reasonable to ground Inform on all available dimensions of analysis at the same time.

Analytical dimensions are also called upon differently within the same tag set. This is illustrated in Table 4 for the MAP TASK tags.

	D1	D2	D3	D4
Explain	+		+	+
Instruct	+			+
Query–yn	+		+	
Query–w	+		+	
Check	+		+	+
Align	+	+	+	
Reply–y	+	+	+	
Reply–n	+	+	+	
Reply–w		+	+	
Acknowledge	+		+	
Clarify	+		+	

Table 4: Dimensions in MAPTASK

Recognition of an Instruct move is predominantly based on grammatical factors; however, pragmatic knowledge is also invoked in case of indirect requests. Query-yn and Query-w moves are mainly defined in terms of their grammatical form, together with knowledge of the following response (hence D3). To apply a Check tag to an utterance, an annotator must look for an interrogative form (D1), an initiative value and an old informational status (D3); finally, an inference about the mental state of the speaker (D4) is also required. Recognition of an Align move relies on the following clues: surface indicators of the utterance being a request (generally prosodic factors), a limited set of words such as "okay", "right" etc. (D2), the fact that the utterance closes a sequence of turns whereby some information has been exchanged (D3). All the five responsive categories presuppose knowledge of the previous move(s) in a dialogue (D3). Furthermore, identification of Replies-y, Replies-n, and Replies-w is based both on the occurrence of specific prosodic contours (e.g. a non-rising one) and on the intended propositional content of the utterance (D2). The same holds for Acknowledge and Clarify which, in addition, are more strictly defined in relation to specific lexical items (D2) and to the content of the utterance these moves respond to (D3).

To sum up, we find the projection plots of tables 3 and 4 an insightful way of making explicit the range of analytical variability among tags i) of different schemes and ii) within the same scheme. Two tags lying close along one dimension of analysis can easily turn out to be diametrically opposed along another dimension. Only by teasing out the multiple recognition–based dimensions called upon in the definition of each tag, we can gain some insights into the pattern of their correspondence, and eventually sharpen up scheme comparability considerably. A multidimensional recognition–based meta–scheme was designed to achieve this purpose, as detailed in the following section.

5 The meta–scheme

To construct our meta–scheme, we took the classificatory dimensions D1–D4 introduced in the previous section as a basis for the definition of four independent taxonomies of utterance tags, some of which consist, in their turn, of further sub-dimensions, as detailed in the following paragraphs.

D1: Grammatical Information This includes the set of morpho–syntactic, prosodic and lexical clues, traditionally referred to as "illocutionary force indicating devices" (Searle, 1969). They range from verb mood (indicative vs. imperative) and word order (e.g., subject inversion) to prosodic tone (rising vs. falling) and lexico-grammatical markers (do–auxiliaries, wh–words, etc.).

The tag values specified along this dimension indicate the illocutionary intention of an utterance as a function of grammatical information only:

- Assert
- Request
  - Request–Imperative
  - Request–Interrogative
    - Request–wh
    - Request–y/n
    - Request–or
- Exclamation

Tag values are defined as follows.

Assert: if an utterance is of a declarative clause type (with a final falling tone and an unmarked SVO order), then it should be tagged as an *Assert*, whose recognizable illocutionary force can be paraphrased as a "claim about the world (where the world includes the speaker). According to our definition, the following utterances should be tagged as D1 *Asserts* (real examples): `I lost a chair`; `Not a problem with the time`; `the lamp and table sound good`; `so I think we're done`; `This is the AT&T Amtrak train schedule system`; `Yes`, `No`.

Request: if an utterance instantiates an imperative or interrogative clause type, then it should be tagged as a Request, whose typical illocutionary force is an attempt by the speaker to get the hearer to do something (classical Directives). The following utterances should thus be tagged as Requests at D1 (real examples): `Do you know the time?`; `Tell me the time`; `Go to Corning`; `Turn right`; `Could you pass me the salt?`.

Exclamation: if an utterance instantiates an exclamative clause type, then it should be tagged as an Exclamation, whose typical illocutionary force is the expression of a particular state of mind of the speaker, as in the following examples: `Hi!`; `Sorry`; `Right!` (uttered with the appropriate intonation); `Of course!`.

D2: Semantic Information This dimension serves the purpose of characterizing an utterance in terms of its propositional and lexical content. We can further specify three classificatory subdimensions, reflecting three independent aspects of semantic information at the utterance level.

- **D2.1: Truth–conditionality** The following values of this attribute label an utterance as having a truth-conditional propositional content or not:
  - truth–cond
  - ntruth–cond
- **D2.2: Polarity**
  - **Positive**: the speaker asserts something, as in `Yes`, or `I think so`.
  - **Negative**: the speaker denies something, as in `No`, or `I don't think so`.
- **D2.3 Performative**: this tag says that an utterance contains an explicit performative, as in `I promise...`, `I suggest...` etc.

D3: Co-textual Information Co–textual information has to do with the relationship of an utterance with previous or following utterances in a discourse. This dimension is criterial for, e.g., tagging an utterance as a reply. Also distinctions referring to the informational status of an utterance, i.e. whether it conveys new or old information, are to be encoded along this dimension. This dimension also includes information about the degree of compliance of a reply with its corresponding initiative.

- **D3.1: Adjacency Pairs**
  - **Initiative**: the utterance prompts an expectation
  - **Reply**: the utterance fulfills an expectation
- **D3.2: Compliance**
  - **Compliant**: the utterance fulfills the expectation set up by a previous utterance in the expected way

- **Non-Compliant**: the utterance fulfills the expectation set up by a previous utterance in an unexpected/dispreferred way

- **D3.3: Presupposition**
  - **New**: the utterance provides information which is new to the hearer
  - **Old**: the utterance provides information which is old to the hearer

D4: Pragmatic Information This dimension characterizes an utterance on the basis of pragmatic information, i.e. knowledge of the social relationship between speaker/hearer, the physical setting of the interaction, the topic of the dialogue etc. Two sub–dimensions are identified here:

- **D4.1: Illocutionary Force**
  - Representative
  - Directive
  - Commissive
  - Expressive

  These represent the classical top categories of Searle's typology of speech acts (Searle, 1969). The possibility of further specify them is left open.

- **D4.2: task vs communication**
  - Task
  - Communication

This sub–dimension is intended to capture the traditional distinction between utterances used to perform a task, and utterances whose main function is smoothing and ensuring the communication process as such. Thus, for instance, utterances such as `Is there a train at Avon?` or `I want to go to Boston` are clearly task-related, while utterances such as `Can you hear me?` or `I don't understand you` are communication-based.

5.1 The meta–scheme at work

How do tags in the meta–scheme relate to the tags in DAMSL, SWITCHBOARD, MAP TASK and VERBMOBIL? What does this relationship tell us about the degree of similarity between the annotation schemes? An objective way of addressing these questions is to use the meta–scheme itself for labeling all five dialogues in the pilot experiment of section 3, to then assess the degree of scheme correspondence in terms of the number of utterances which are found to be marked up with the same tags, similarly to what was done in section 3. Note that the use of a meta–scheme to tag a dialogue should not suggest that the meta–scheme is, as such, an adequate tool for annotation. First, tags are largely under–specified. Moreover, the focus of annotation is shifted here from the identification of primary illocutionary acts to the recognizable linguistic and contextual clues for their identification. We will return to this important point in the following section. Table 5 exemplifies the annotation of a dialogue excerpt (two turns, three utterances) with the categories in the meta–scheme.

D1:req-wh D2:req-info D3.1:I D4:direct	u:	what time would engine two and three leave Elmira?
D1:assert D2.1:truthcond D3.1:R D3.2:ncomp D4:represent	s:	well they're not scheduled yet
D1:assert D2.1:ntruthcond D3.1:R D3.2:comp D4:represent	s:	but we can send them at any time we want

Table 5: Sample annotation

Table 6 reports the degree of multidimensional similarity between MAP TASK Explain, on the one hand, and DAMSL Assert, Re-Assert, Open-Option, Offer and Info-Request on the other hand. In the table, each tag is represented as a point in the n-dimensional space staked out by the meta–scheme. The first column gives the invariant meta–scheme tags which are shared by all utterances tagged as Explain. A dash ('-') in the column signifies that tags vary along the corresponding dimension: this means that the dimension is not criterial for the definition of *Explain*. This is the case of D2.2 (polarity), D3.2 (compliance) and D4.1 (pragmatic illocutionary force). In the remaining columns, we put '=' to signify dimensional equivalence, i.e. identity of invariant meta–scheme tags, and '≠' to express diversity. Once more, a dash is used to indicate that the corresponding dimension is orthogonal to the information conveyed by the tag. Intuitively, the tags more similar to Explain are those with more '=' and fewer '≠' in the corresponding column.

Note that Assert turns out to be the tag with the highest number of matching dimensions ('='), and the lowest number of mismatches ('≠'). This explains why MAP TASK Explain is the most natural candidate for replacing DAMSL Assert, as suggested by Larsson. We can now give reasons for that: Assert differs from Explain in that the former, unlike the latter, conveys no stable initiative force. Note further, however, that Explain is not defined along dimension D4.1, which, in turn, defines tags such as Open-option, Offer and Info-Request. This suggests that Explain is also likely to replace these tags when they are assigned to assertive and truth conditional

	explain	assert	reassert	openopt	offer	inforeq
D1	assert	=	=	=	-	-
D2.1	truth–cond	=	=	=	-	-
D2.2	-	-	-	-	-	-
D3.1	Init	-	=	-	=	=
D3.2	-	-	-	-	-	-
D3.3	new	=	≠	=	=	=
D4.1	-	-	≠	≠	≠	≠
D4.2	task	=	=	=	=	=

Table 6: Multidimensional tag correspondences

utterances, that is when these utterances happen to meet the criteria for identification of Explain. Incidentally, it should be noted that the evidence of table 6 provides a justification of the figures reported in table 1, which would otherwise remain counterintuitive in the light of tag definitions.

6 Annotation and meta–scheme

As already pointed out above, the meta–scheme proposed here does not *per se* fulfill some important prerequisites for an annotation scheme. It is useful, at this stage, to elaborate this point. First, multidimensionality and orthogonality of the assumed multiple dimensions seem to be operationally cumbersome and, in general, detract from reliability in actual tagging practice. Furthermore, in the meta–scheme all classificatory dimensions are conceived of as being *on a par*. This means that we deliberately make no assumption as to what dimension of annotation ultimately provides information about the primary intended illocutionary act of an utterance, and how information along one dimension relates to information encoded at another dimension. This is not very informative from the point of view of annotation, but represents a very useful feature for scheme customization, as it makes it possible to modify/adapt an existing annotation scheme by collapsing some analytical dimensions in a controlled way.

Finally, it should be appreciated that the list of dimensions provided here is not meant to be either *exhaustive* or *minimal*, in the sense that every tag *should* be classified along each dimension. Other possible dimensions of analysis can include, for example, kinesic information, to account for dialogue acts performed through non-verbal communicative behavior, such as nodding, smiling and pointing. As long as dimensions are rigorously defined, this should clarify the intended use of a scheme considerably.

7 Conclusion

Tag sets are typically developed to respond to specific applications and practical usages, without bothering too much about how the tags themselves relate to the nature of information needed for their assignment in context. This is fine as long as tag sets are assessed in relation to the use they were originally intended for, but much less so if one wants to evaluate the extent to which one tag set translates into another tag set, or to assess the usability of a given tag set for other purposes/applications.

The multi–dimensional recognition–based meta–scheme described in these pages makes it explicit how intentions relate to the linguistic and contextual information needed for their identification. We showed that this is extremely helpful for scheme comparison, as it sheds light on the precise nature of tag correspondences, well beyond the intuitive grasp provided by tag definitions.

Preliminary experiments show that a translation of a dialogue tagged with an existing scheme into our meta–scheme is also a useful exercise to assess the internal consistency of the annotated material. If this is confirmed, then use of the meta–scheme should improve scheme design considerably, and should be able to provide procedural and testable guide–lines for dialogue annotators.

References

Alexandersson, J., B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. 1998. *Dialogue Acts in Verbmobil-2, Second Edition.* Verbmobil Report 226, DFKI Saarbruecken, Universitaet Stuttgart, TU Berlin, Universitaet des Saarlandes.

Allen, J. and M. Core. 1997. *Draft of DAMSL: Dialog Act Markup in Several Layers.* Technical report, Rochester.

Carletta, J., A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1996. *HCRC Dialogue Structure Coding Manual.* Technical Report HCRC TR-82, Human Communication Research Centre, University of Edinburgh, Edinburgh, Scotland.

Core, M. and J. Allen. 1997. "Coding Dialogs with the DAMSL Annotation Scheme". In *Proceedings of the AAAI Fall 1997 Symposium.*

di Eugenio, B., P. W. Jordan, and L. Pylkkaenen. 1997. *The COCONUT project: dialogue annotation manual.* Technical report.

Ichikawa, A., M. Araki, M. Ishizaki, S. Itabashi, T. Itoh, H. Kashioka, K. Kato, H. Kikuchi, T. Kumagai, A. Kurematsu, H. Koiso, M. Tamoto, S. Tutiya, S. Nakazato, Y. Horiuchi, K. Maekawa,

Y. Yamashita, and T. Yoshimura. 1998. "Standardising Annotation Schemes for Japanese Discourse". In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, pp. 731-736.

Jurafsky, D., L. Shriberg, and D. Biasca. 1997. *Switchboard SWBD*-DAMSL, *Shallow-Discourse-Function Annotation; Coders Manual, Draft 13.*

Klein, M., N. O. Bernsen, S. Davies, L. Dybkjaer, J. Garrido, H. Kasch, A. Mengel, V. Pirrelli, M. Poesio, S. Quazza, and C. Soria. 1998. *Supported Coding Schemes.* Technical Report D1.1, MATE.

Larsson, S. 1998. *Coding schemas for dialogue moves.* Technical report, Department of Linguistics, Goeteborg University.

Poesio, M. and D. Traum. 1998. "Towards an Axiomatization of Dialogue Acts". In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*, Enschede, The Netherlands, pp. 207-222.

Poesio, M., Cooper, R., Larsson, S., Traum, D. and C. Matheson. 1999. "Annotating Conversations for Information State Updates". Paper presented at Amstelogue99.

Searle, J. 1969. *Speech Acts.* Cambridge University Press