Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme

Jeroen Geertzen and Harry Bunt
Language and Information Science
Tilburg University, P.O. Box 90153
NL-5000 LE Tilburg, The Netherlands
{j.geertzen,h.bunt}@uvt.nl

Abstract

We present a first analysis of inter-annotator agreement for the DIT^{++} tagset of dialogue acts, a comprehensive, layered, multidimensional set of 86 tags. Within a dimension or a layer, subsets of tags are often hierarchically organised. We argue that especially for such highly structured annotation schemes the well-known kappa statistic is not an adequate measure of inter-annotator agreement. Instead, we propose a statistic that takes the structural properties of the tagset into account, and we discuss the application of this statistic in an annotation experiment. The experiment shows promising agreement scores for most dimensions in the tagset and provides useful insights into the usability of the annotation scheme, but also indicates that several additional factors influence annotator agreement. We finally suggest that the proposed approach for measuring agreement per dimension can be a good basis for measuring annotator agreement over the dimensions of a multidimensional annotation scheme.

1 Introduction

The DIT^{++} tagset (Bunt, 2005) was designed to combine in one comprehensive annotation scheme the communicative functions of dialogue acts distinguished in Dynamic Interpretation Theory (DIT, (Bunt, 2000; Bunt and Girard, 2005)), and many of those in DAMSL (Allen and Core, 1997) and in other annotation schemes. An important difference between the DIT^{++} and DAMSL schemes is the more elaborate and fine-grained set of functions for feedback and other aspects of dialogue control that is available in DIT, partly inspired by the work of Allwood (Allwood et al., 1993). As it is often thought that more elaborate and fine-grained annotation schemes are difficult for annotators to apply consistently, we decided to address this issue in an annotation experiment on which we report in this paper. A frequently used way of evaluating human dialogue act classification is inter-annotator agreement. Agreement is sometimes measured as percentage of the cases on which the annotators agree, but more often expected agreement is taken into account in using the kappa statistic (Cohen, 1960; Carletta, 1996), which is given by:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where p_o is the observed proportion of agreement and p_e is the proportion of agreement expected by chance. Ever since its introduction in general (Cohen, 1960) and in computational linguistics (Carletta, 1996), many researchers have pointed out that there are quite some problems in using κ (e.g. (Di Eugenio and Glass, 2004)), one of which is the discrepancy between p_0 and κ for skewed class distribution.

Another is that the degree of disagreement is not taken into account, which is relevant for any non-nominal scale. To address this problem, a weighted κ has been proposed (Cohen, 1968) that penalizes disagreement according to their degree rather than treating all disagreements equally. It would be arguable that in a similar way, characteristics of dialogue acts in a particular taxonomy and possible pragmatic relatedness between them should be taken into account to express annotator agreement. For dialogue act taxonomies which are structured in a meaningful way, such as those that

express hierarchical relations between concepts in the taxonomy, the taxonomic structure can be exploited to express how much annotators disagree when they choose different concepts that are directly or indirectly related. Recent work that accounts for some of these aspects is a metric for automatic dialogue act classification (Lesch et al., 2005) that uses distance in a hierarchical structure of multidimensional labels.

In the following sections of this paper, we will first briefly consider the dimensions in the DIT^{++} scheme and highlight the taxonomic characteristics that will turn out to be relevant in later stage. We will then introduce a variant of weighted κ for inter-annotator agreement called κ_{tw} that adopts a taxonomy-dependent weighting, and discuss its use.

2 Annotation using DIT

DIT is a context-change (or information-state update) approach to the analysis of dialogue, which describes utterance meaning in terms of context update operations called 'dialogue acts'. A dialogue act in DIT has two components: (1) the semantic content, being the objects, events, properties, relations, etc. that are considered; and (2) the communicative function, that describes how the addressee is intended to use the semantic content for updating his context model when he understands the utterance correctly. DIT takes a multidimensional view on dialogue in the sense that speakers may use utterances to address several aspects of the communication simultaneously, as reflected in the multifunctionality of utterances. One such aspect is the performance of the task or activity for which the dialogue takes place; another is the monitoring of each other's attention, understanding and uptake through feedback acts; others include for instance the turn-taking process and the timing of communicative actions, and finally yet another aspect is formed by the social obligations that may arise such as greeting, apologising, or thanking. The various aspects of communication that can be addressed independently are called *dimensions* (Bunt and Girard, 2005; Bunt, 2006). The DIT^{++} tagset distinguishes 11 dimensions, which all contain a number of communicative functions that are specific to that dimension, such as `TURN GIVING`, `PAUSING`, and `APOLOGY`.

Besides dimension-specific communicative functions, DIT also distinguishes a layer of communicative functions that are not specific to any particular dimension but that can be used to address any aspect of communication. These functions, which include questions, answers, statements, and commissive as well as directive acts, are called *general purpose functions*. A dialogue act falls within a specific dimension if it has a communicative function specific for that dimension or if it has a general-purpose function and a semantic content relating to that dimension. Dialogue utterances can in principle have a function (but never more than one) in each of the dimensions, so annotators using the DIT^{++} scheme can assign at most one tag for each of the 11 dimensions to any given utterance.

Both within the set of general-purpose communicative function tags and within the sets of dimension-specific tags, tags can be hierarchically related in such a way that a label lower in a hierarchy is more specific than a label higher in the same hierarchy. Tag F_1 is more specific than tag F_2 if F_1 defines a context update operation that includes the update operation corresponding to F_2. For instance, consider a part of the taxonomy for general purpose functions (Figure 1).

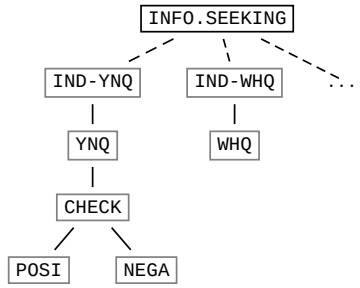

Figure 1: Two hierarchies in the information seeking general purpose functions.

For an utterance to be assigned a `YN-QUESTION`, we assume the speaker believes that the addressee knows the truth value of the proposition presented. For an utterance to be assigned a `CHECK`, we assume the speaker *additionally* has a weak belief that the proposition that forms the semantic content is true. And for a `POSI-CHECK`, there is the additional assumption that the speaker believes (weakly) that the hearer also believes that the proposition is true.[1]

Similar to the hierarchical relations between `YN-Question`, `CHECK`, and `POSI-CHECK`, other parts

[1]For a formal description of each function in the DIT^{++} tagset see `http://ls0143.uvt.nl/dit/`

of the annotation scheme contain hierarchically related functions.

The following example illustrates the use of DIT^{++} communicative functions for a very simple translated) dialogue fragment[2].

1 S at what time do you want to travel today?
`TASK = WH-Q, TURN-MANAGEMENT = GIVE`

2 U at ten.
`TASK = WH-A, TURN-MANAGEMENT = GIVE`

3 S so you want to leave at ten in the morning?
`TASK = POSI-CHECK, TURN-MANAGEMENT = GIVE`

4 U yes that is right.
`TASK = CONFIRM, TURN-MANAGEMENT = GIVE`

3 Agreement using κ

3.1 Related work

Inter-annotator agreements have been calculated with the purpose of qualitatively evaluating tagsets and individual tags. For DAMSL, the first agreement results were presented in (Core and Allen, 1997), based on the analysis of TRAINS 91-93 dialogues (Gross et al., 1993; Heeman and Allen, 1995). In this analysis, 604 utterances were tagged by mostly two annotators. Following the suggestions in (Carletta, 1996), Core et al. consider kappa scores above 0.67 to indicate significant agreement and scores above 0.8 reliable agreement. Another more recent analysis was performed for 8 dialogues of the MONROE corpus (Stent, 2000), counting 2897 utterances in total, processed by two annotators for 13 DAMSL dimensions. Other analyses apply DAMSL derived schemes (such as SWITCHBOARD-DAMSL) to various corpora (e.g. (Di Eugenio et al., 1998; Shriberg et al., 2004)). For the comprehensive DIT^{++} taxonomy, the work reported here represents the first investigation of annotator agreement.

3.2 Experiment outline

As noted, existing work on annotator agreement analysis has mostly involved only two annotators. It may be argued that especially for annotation of concepts that are rather complex, an odd number of annotators is desirable. First, it allows having majority agreement unless all annotators choose entirely different. Second, it allows to deal better with the undesirable situation that one annotator chooses quite differently from the others. The agreement scores reported in this paper are all calculated on the basis of the annotations of three annotators, using the method proposed in (Davies and Fleiss, 1982).

The dialogues that were annotated are task-oriented and are all in Dutch. To account for different complexities of interaction, both human-machine and human-human dialogues are considered. Moreover, the dialogues analyzed are drawn from different corpora: OVIS (Strik et al., 1997), DIAMOND (Geertzen et al., 2004), and a collection of Map Task dialogues (Caspers, 2000); see Table 1, where the number of annotated utterances is also indicated.

corpus	domain	type	#utt
OVIS	TRAINS like interactions on train connections	H-M	193
DIAMOND1	interactions on how to operate a fax device	H-M	131
DIAMOND2	interactions on how to operate a fax device	H-H	114
MAPTASK	HCRC Map Task like interaction	H-H	120
			558

Table 1: Characteristics of the utterances considered

Six undergraduate students annotated the selected dialogue material. They had been introduced to the DIT^{++} annotation scheme and the underlying theory while participating in a course on pragmatics. During this course they were exposed to approximately four hours of lecturing and few small annotation exercises. For all dialogues, the audio recordings were transcribed and the annotators annotated presegmented utterances for which full agreement was established on segmentation level beforehand. During the annotation sessions the annotators had — apart from the transcribed speech — access to the audio recordings, to the on-line definitions of the communicative functions in the scheme and to a very brief, 1-page set of annotation guidelines[3]. The task was facilitated by the use of an annotation tool that had been built for this occasion; this tool allowed the subjects to assign each utterance one DIT^{++} tag for each dimension without any further constraints. In total 1,674 utterances were annotated.

3.3 Problems with standard κ

If we were to apply the standard κ statistic to DIT^{++} annotations, we would not do justice to an important aspect of the annotation scheme concerning the differences between alternative tags,

[2]Drawn from the OVIS corpus (Strik et al., 1997): OVIS2:104/001/001:008-011

[3]See `http://ls0143.uvt.nl/dit`

and hence the possible differences in the disagreement between annotators using alternative tags. An aspect in which the DIT^{++} scheme differs from other taxonomies for dialogue acts is that, as noted in Section 2, communicative functions (CFs) within a dimension as well as general-purpose CFs are often structured into hierarchies in which a difference in level represents a relation of specificity. When annotators differ in that they assign tags which both belong to the same hierarchy, they may differ in the degree of specificity that they want to express, but they agree to the extent that these tags inherit the same elements from tags higher in the hierarchy. Inter-annotator disagreement is in such a case much less than if they would choose two unrelated tags. This is for instance obvious in the following example of the annotations of two utterances by two annotators:

1	S	what do you want to know?	WHQ	YNQ
2	U	can I print now?	YNQ	CHECK

With utterance 1, the annotators should be said simply to disagree (in fact, annotator 2 incorrectly assigns a YNQ function). Concerning utterance 2 the annotators also disagree, but Figure 1 and the definitions given in Section 2 tell us that the disagreement in this case is quite small, as a CHECK inherits the properties of a YNQ. We therefore should not use a black-and-white measure of agreement, like the standard κ, but we should have a measure for *partial annotator agreement*.

In order to measure partial (dis-)agreement between annotators in an adequate way, we should not just take into account whether two tags are hierarchically related or not, but also how far they are apart in the hierarchy, to reflect that two tags which are only one level apart are semantically more closely related than tags that are several levels apart. We will take this additional requirement into account when designing a weighted disagreement statistic in the next section.

4 Agreement based on structural taxonomic properties

The agreement coefficient we are looking for should in the first place be *weighted* in the sense that it takes into account the magnitude of disagreement. Two such coefficients are weighted kappa (κ_w, (Cohen, 1968)) and alpha (Krippendorff, 1980). For our purposes, we adopt κ_w for its property to take into account a probability distribution typical for each annotator, generalize it to the case for multiple annotators by taking the average over the scores of annotator pairs, and define a function to be used as distance metric.

4.1 Cohen's weighted κ

Assuming the case of two annotators, let p_{ij} denote the proportion of utterances for which the first and second annotator assigned categories i and j, respectively. Then Cohen defines κ_w in terms of *disagreement* rather than *agreement* where $q_o = 1 - p_o$ and $q_e = 1 - p_e$ such that Equation 1 can be rewritten to:

$$\kappa = 1 - \frac{q_o}{q_e} \tag{2}$$

To arrive at κ_w, the proportions q_o and q_e in Equation 2 are replaced by weighted functions over all possible category pairs:

$$\kappa_w = 1 - \frac{\sum v_{ij} \cdot p_{oij}}{\sum v_{ij} \cdot p_{eij}} \tag{3}$$

where v_{ij} denotes the disagreement weight. To calculate this weight we need to specify a distance function as metric.

4.2 A taxonomic metric

The task of defining a function in order to calculate the difference between a pair of categories requires us to determine semantic-pragmatic relatedness between the CFs in the taxonomy. For any annotation scheme, whether it is hierarchically structured or not, we could assign for each possible pair of categories a value that expresses the semantic-pragmatic relatedness between the two categories compared to all other possible pairs. However, it seems quite difficult to find universal characteristics for CFs to be used to express relatedness on a rational scale. When we consider a taxonomy that is structured in a meaningful way, in this case one that expresses hierarchical relations between CF based on their effect on information states, the taxonomic structure can be exploited to express in a systematic fashion how much annotators disagree when they choose different concepts that are directly or indirectly related.

The assignment of different CFs to a specific utterance by two annotators represents full disagreement in the following cases:

1. the two CFs belong to different dimensions;

2. one of the two CFs is general-purpose; the other is dimension-specific;[4]

3. the two CFs belong to the same dimension but not to the same hierarchy;

4. the two CFs belong to the same hierarchy but are not located in the same branch. Two CFs are said to be located in the same branch when one of the two CFs is an ancestor of the other.

If, by contrast, the two CFs take part in a ancestor-offspring relation within a hierarchy (either within a dimension or among the general-purpose CFs), then the CFs are related and this assignment represents partial disagreement. A distance metric that measures this disagreement, which we denote as δ, should have the following properties:

1. δ should be a real number normalized in the range $[0 \ldots 1]$;

2. Let C be the (unordered) set of CFs.[5] For every two CFs $c_1, c_2 \in C$, $\delta(c_1, c_2) = 0$ when c_1 and c_2 are not related;

3. Let C be the (unordered) set of CFs. For every communicative function $c \in C$, $\delta(c, c) = 1$;

4. Let C be the (unordered) set of CFs. For every two CFs $c_1, c_2 \in C$, $\delta(c_1, c_2) = \delta(c_2, c_1)$.

Furthermore, when c_1 and c_2 are related, we should specify how distance between them in the hierarchy should be expressed in terms of partial disagreement. For this, we should take the following aspects into account:

1. The distance in levels between c_1 and c_2 in the hierarchy is proportional to the magnitude of the disagreement;

[4]This is in fact a simplification. For instance, an INFORM act of which the semantic content conveys that the speaker did not understand the previous utterance forms an act in the Auto-Feedback dimension (see Note 6), and a tagging to this effect should perhaps not be considered to express full disagreement with the assignment of the dimension-specific tag AUTO-FEEDBACK-Int$^-$. See also the next footnote.

[5]Strictly speaking, in DIT a dialogue act annotation tag is either (a) the name of a dimension-specific function, or (b) a pair consisting of the name of a general-purpose function and the name of a dimension. However, in view of the simplification mentioned in the previous note, for the sake of this paper we may as well consider tags containing a general-purpose function as simply consisting of that function.

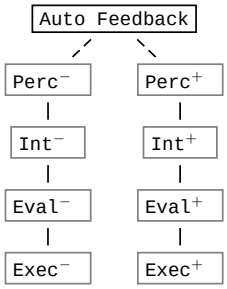

Figure 2: Hierarchical structures in the auto feedback dimension.

2. The magnitude of disagreement between c_1 and c_2 being located in two different levels of depths n and $n+1$ *might* be considered to be more different than that between to levels of depth $n+1$ and $n+2$. If this would be the case, the deeper two levels are located in the tree, the smaller the differences between the nodes on those levels. For the hierarchies in DIT, we keep the magnitude of disagreement linear with the difference in levels, and independent of level depth;

Given the considerations above, we propose the following metric:

$$\delta(c_i, c_j) = a^{\Delta(c_i, c_j)} \cdot b^{\Gamma(c_i, c_j)} \qquad (4)$$

where:

- a is a constant for which $0 < a < 1$, expressing how much distance there is between two adjacent levels in the hierarchy; a plausible value for a could be 0.75;

- Δ is a function that returns the difference in depth between the levels of c_i and c_j;

- b is a constant for which $0 < b \leq 1$, expressing in what rate differences should become smaller when the depth in the hierarchy gets larger. If there is no reason to assume that differences on a higher depth in the hierarchy are of less magnitude than differences on a lower depth, then $b = 1$;

- $\Gamma(c_i, c_j)$ is a function that returns the minimal depth of c_i and c_j.

To provide some examples of how δ would be calculated, let us consider the general purpose functions in Figure 1. Consider also Figure 2, that represents two hierarchies of CFs in the auto

feedback dimension[6], and let us assume the values of the various parameters those that are suggested above. We then get the following calculations:

$$\begin{array}{ll} \delta(IND-YNQ, CHECK) & = 0.75^2 \cdot 1 \quad = 0.563 \\ \delta(YNQ, CHECK) & = 0.75^1 \cdot 1 \quad = 0.75 \\ \delta(Perc^+, Perc^+) & = 0.75^0 \cdot 1 \quad = 1 \\ \delta(Perc^+, Eval^+) & = 0.75^2 \cdot 1 \quad = 0.563 \\ \delta(Int^-, Int^+) & \quad = 0 \\ \delta(POSI, NEGA) & \quad = 0 \end{array}$$

To conclude, we can simply take δ to be the weighting in Cohen's κ_w and come to a coefficient which we will call *taxonomically weighted kappa*, denoted by κ_{tw}:

$$\kappa_{tw} = 1 - \frac{\sum(1-\delta(i,j)) \cdot p_{oij}}{\sum(1-\delta(i,j)) \cdot p_{eij}} \qquad (5)$$

4.3 κ_{tw} statistics for DIT

Considering the DIT^{++} taxonomy, it may be argued that due to the many hierarchies in the topology of the general-purpose functions, this is the part where most is to be gained by employing κ_{tw}.

Table 2 shows the statistics for each dimension, averaged over all annotation pairs. With *annotation pair* is understood the pair of assignments an utterance received by two annotators for a particular dimension. The figures in the table are based on those cases in which both annotators assigned a function to a specific utterance for a specific dimension. Cases where either one annotator does not assign a function while the other does, or where both annotators do not assign a function, are not considered. Scores for standard κ and κ_{tw} can be found in the first two columns. The column *#pairs* indicates on how many annotation pairs the statistics are based. The last column shows the ap*-ratio*. This figure indicates which fraction of all annotated functions in that dimension are represented by annotation pairs. When $\#ap$ denotes the number of annotation pairs and $\#pa$ denotes the number of partial annotations (annotations in which one annotator assigned a function and the other did not), then the ap-ratio is calculated as $\#ap/(\#pa + \#ap)$. We can observe that due to the use of the taxonomic weighting both *feedback* dimensions and the *task* dimension gained substantially in annotator agreement.

[6]Auto-feedback: feedback on the processing (perception, understanding, evaluation,..) of previous utterances by the speaker. DIT also distinguishes allo-feedback, where the speaker provides or elicits information about the addressee's processing.

Dimension	κ	κ_{tw}	#pairs	ap-ratio
task	0.47	0.71	848	0.87
task:action discussion	0.61	0.61	91	0.37
auto feedback	0.21	0.57	127	0.34
allo feedback	0.42	0.58	17	0.14
turn management	0.82	0.82	115	0.18
time management	0.58	0.58	68	0.72
contact management	1.00	1.00	8	0.17
topic management	nav	nav	2	0.08
own com. management	1.00	1.00	2	0.08
partner com. management	nav	nav	1	0.07
dialogue struct. management	0.74	0.74	15	0.31
social obl. management	1.00	1.00	61	0.80

Table 2: Scores for corrected κ and κ_{tw} per DIT dimension.

When we look at the agreement statistics and consider κ scores above 0.67 to be significant and scores above 0.8 considerably reliable, as is usual for κ statistics, we can find the dimensions `TURN-MANAGEMENT`, `CONTACT MANAGEMENT`, and `SOCIAL-OBLIGATIONS-MANAGEMENT` to be reliable and `DIALOGUE STRUCT. MANAGEMENT` to be significant. For some dimensions, the occurences of functions in these dimensions in the annotated dialogue material were too few to draw conclusions. When we also take the ap-ratio into account, only the dimensions `TASK`, `TIME MANAGEMENT`, and `SOCIAL-OBLIGATIONS-MANAGEMENT` combine a fair agreement on functions with fair agreement on whether or not to annotate in these dimensions. Especially for the other dimensions, the question should be raised for which cases and for what reasons the ap-ratio is low. This question asks for further qualitative analysis, which is beyond the scope of this paper[7].

5 Discussion

In the previous sections, we showed how the taxonomically weighted κ_{tw} that we proposed can be more suitable for taxonomies that contain hierarchical structures, like the DIT$^+$+) taxonomy. However, there are some specific and general issues that deserve more attention.

A question that might be raised in using κ_{tw} as opposed to ordinary κ, is if the assumption that the interpretations of κ proposed in literature in terms of reliability is also valid for κ_{tw} statistics. This is ultimately an empirical issue, to be decided by which κ_{tw} scores researchers find to correspond to fair or near agreement between annotators.

Another point of discussion is the arbitrariness of the values of the parameters that can be chosen in δ. In this paper we proposed $a = 0.75$ and $\beta = 0.5$. Choosing different values may change

[7]See (Geertzen, 2006) for more details.

the disagreement of two distinct CFs located in the same hierarchy considerably. Still, we think that by interpolating smoothly between the intuitively clear cases at the two extreme ends of the scale, it is possible to choose reasonable values for the parameters that scale well, given the average hierarchy depth.

A more general problem, inherent in almost any (dialogue act) annotation activity is that when we consider the possible factors that influence the agreement scores, we find that they can be numerous. Starting with the tagset, unclear definitions and vague concepts are a major source of disagreement. Other factors are the quality and extensiveness of annotation instructions, and the experience of the annotators. These were kept constant throughout the experiment reported in this paper, but clearly the use of more experienced or better trained annotators could have a great influence. Then there is the influence that the use of an annotation tool can have. Does the tool gives hints on annotation consistency (e.g. an `ANSWER` should be preceded by a `QUESTION`), does it enforce consistency, or does it not consider annotation consistency at all? Are the possible choices for annotators presented in such a way that each choice is equally well visible and accessible? Clearly, when we do not control these factors sufficiently, we run the risk that what we measure does not express what we try to quantify: (dis)agreement among annotators about the description of what happens in a dialogue.

6 Conclusion and future work

In this paper we have presented agreement scores for Cohen's unweighted κ and claimed that for annotation schemes with hierarchically related tags, a weighted κ gives a better indication of (dis)agreement than unweighted κ. The κ scores for some dimensions seem not particularly spectacular but become more interesting when looking at semantic-pragmatic differences between dialogue acts or CFs. Even though there are somewhat arbitrary aspects in weighting, when parameters are carefully chosen a weighted metric gives a better representation of the inter-annotator agreements. More generally, we propose that semantic-pragmatic relatedness between taxonomic concepts should be taken into account when calculating inter-annotator (dis)agreement. While we used DIT^{++} as tagset, the weighting function we proposed can be employed in any taxonomy containing hierarchically related concepts, since we only used *structural* properties of the taxonomy.

We have also quantitatively[8] evaluated the DIT^{++} tagset per dimension, and obtained an indication of its usability. We focussed on agreement per dimension, but when we desire a global indication of the difference in semantic-pragmatic interpretation of a complete utterance it requires us to consider other aspects. A truly multidimensional study of inter-annotator agreement should not only take intra-dimensional aspects into account but also relate the dimensions to each other. In (Bunt and Girard, 2005; Bunt, 2006) it is argued that dimensions should be *orthogonal*, meaning that an utterance can have a function in one dimension independent of functions in other dimensions. This is a somewhat utopical condition, since there are some functions that show correlations and dependencies with across dimensions. For this reason it makes sense to try to express the effect of the presence of strong correlations, dependencies and possible entailments in a multidimensional notion of (dis)agreement. Additionally, it may be desirable to take into account the importance that a CF can have. It is widely acknowledged that utterances are often multifunctional, but it could be argued that in many cases an utterance has a *primary* function and *secondary functions*; for instance, if an utterance has both a task-related function and one or more other functions, the task-related function is typically felt to be more important than the other functions, and disagreement about the task-related function is therefore felt to be more serious than disagreement about one of the other functions. This might be taken into account by adding a weighting function when combining agreement measures over multiple dimensions.

Other future work we plan is more methodological in nature, quantifying the relative effect of the factors that may have influenced the scores that we have found. This would create a situation in which there is more insight in *what* exactly is evaluated. As for evaluating the tagset, we for instance plan to further analyze co-occurence matrices to identify frequent misannotations, and to have annotators thinking aloud while performing the annotation task.

[8] Kappa statistics are indicative. To get a full understanding of what the figures represent, qualitative analysis by using e.g. co-occurence matrices is required, which is beyond the scope of this paper.

Acknowledgements

The authors thank three anonymous reviewers for their helpful comments on an earlier version of this paper.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript.

J. Allwood, J. Nivre, and E. Ahlsén. 1993. Manual for coding interaction management. Technical report, Göteborg University. Project report: Semantik och talspråk.

Harry C. Bunt and Yann Girard. 2005. Designing an open, multidimensional dialogue act taxonomy. In *Proceedings of the 9th Workshop on the Semantics and Pragmatics of Dialogue (DIALOR 2005)*, pages 37–44, Nancy, France, June.

Harry C. Bunt. 2000. Dialogue pragmatics and context specification. In Harry C. Bunt and William Black, editors, *Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics*, pages 81–150. John Benjamins, Amsterdam, The Netherlands.

Harry C. Bunt. 2005. A framework for dialogue act specification. In *Joint ISO-ACL Workshop on the Representation and Annotation of Semantic Information*, Tilburg, The Netherlands, January.

Harry C. Bunt. 2006. Dimensions in dialogue annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy, May.

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

Johanneke Caspers. 2000. Pitch accents, boundary tones and turn-taking in dutch map task dialogues. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, volume 1, pages 565–568, Beijing, China.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Mark G. Core and James F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In David Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Mark Davies and J.L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38:1047–1051.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

Barbara Di Eugenio, Pamela W. Jordan, Johanna D. Moore, and Richmond H. Thomason. 1998. An empirical investigation of proposals in collaborative dialogues. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 1998)*, pages 325–329, Montreal, Canada.

Jeroen Geertzen, Yann Girard, Roser Morante, Ielka van der Sluis, Hans Van Dam, Barbara Suijkerbuijk, Rintse van der Werf, and Harry Bunt. 2004. The diamond project (poster,project description). In *The 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)*. Barcelona, Spain.

Jeroen Geertzen. 2006. Inter-annotator agreement within dit^{++} dimensions. Technical report, Tilburg University, Tilburg, The Netherlands.

Derek Gross, James F. Allen, and David R. Traum. 1993. The TRAINS 91 dialogues. Technical Report TN92-1, University of Rochester, Rochester, NY, USA.

Peter A. Heeman and James F. Allen. 1995. The TRAINS 93 dialogues. Technical Report TN94-2, University of Rochester, Rochester, NY, USA.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, USA.

Stephan Lesch, Thomas Kleinbauer, and Jan Alexandersson. 2005. A new metric for the evaluation of dialog act classification. In *Proceedings of the 9th Workshop on the Semantics and Pragmatics of Dialogue (DIALOR 2005)*, pages 143–146, Nancy, France, june.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Boston, USA, April-May.

Amanda J. Stent. 2000. The monroe corpus. Technical Report TR728/TN99-2, University of Rochester, Rochester, UK.

Helmer Strik, Albert Russel, Henk van den Heuvel, Catia Cucchiarini, and Lou Boves. 1997. A spoken dialog system for the dutch public transport information service. *International Journal of Speech Technology*, 2(2):119–129.