Evaluating Dialogue Act Tagging with Naive and Expert Annotators

Jeroen Geertzen[1,2], **Volha Petukhova**[1], **Harry Bunt**[1]

[1]Dept. of Communication & Information Sciences
Tilburg University

[2]Dept. of Industrial Design
Eindhoven University of Technology

{j.geertzen,v.petukhova,harry.bunt}@uvt.nl

Abstract

In this paper the dialogue act annotation of naive and expert annotators, both annotating the same data, are compared in order to characterise the insights annotations made by different kind of annotators may provide for evaluating dialogue act tagsets. It is argued that the agreement among naive annotators provides insight in the clarity of the tagset, whereas agreement among expert annotators provides an indication of how reliably the tagset can be applied when errors are ruled out that are due to deficiencies in understanding the concepts of the tagset, to a lack of experience in using the annotation tool, or to little experience in annotation more generally. An indication of the differences between the two groups in terms of inter-annotator agreement and tagging accuracy on task-oriented dialogue in different domains, annotated with the DIT^{++} dialogue act tagset is presented, and the annotations of both groups are assessed against a gold standard. Additionally, the effect of the reduction of the tagset's granularity on the performances of both groups is looked into. In general, it is concluded that the annotations of both groups provide complementary insights in reliability, clarity, and more fundamental conceptual issues.

1. Introduction

Dialogue act annotations with high reliability are a prerequisite for obtaining sound theoretical insights on dialogue or obtaining training data for automatic dialogue act tagging. A dialogue act scheme can be applied reliably if the assignment of the categories in the scheme does not depend on individual judgement, but on a shared understanding of what the categories mean and how they are to be used. Manual dialogue act classification is usually evaluated in terms of inter-annotator agreement. Agreement is sometimes measured as a percentage of the cases on which the annotators agree (percentage agreement), but more often expected agreement is taken into account by using for instance the kappa statistic (Cohen, 1960; Carletta, 1996). Inter-annotator agreement expresses the degree to which annotations that have been made by multiple annotators can be relied upon. An issue in determining inter-annotator agreement is what kind of annotators to use. Carletta (1996) argues that in annotating with schemes such as those in discourse and dialogue analysis there are no real experts, and that what counts is how totally naive annotators manage based on written instructions. When totally naive annotators are used, however, factors such as the clarity of the written instructions and the interface of the annotation tool have a bigger impact on performance than when annotators are used who are familiar with the tagset and have a good overview of the annotation concepts that can be used. Moreover, when the aim is to obtain annotations that are as accurate as possible and the dialogue act tagset is rather complex, the use of expert annotators seems more obvious. It can be argued that both evaluation based on naive annotators and evaluation based on expert annotators can provide indications of the usability of the tagset, but that evaluation based on naive annotators provides more insight in the clarity of the concepts in the tagset, whereas evaluation based on expert annotators provides an indication of how reliably the tagset can be applied when errors are ruled out that are due to deficiencies in conceptual understanding, to a lack of experience in using the annotation tool, or to little experience in annotation more generally.

When inter-annotator agreement scores for data annotated with a particular tagset indicate high reliability[1] it is not guaranteed that there is high agreement on the assignment of the *right* concept. Even though it is not likely to happen often, annotators could agree in assigning a certain concept but disagree with an expert on what would be the correct concept to assign. Therefore, to obtain a reliable evaluation, inter-annotator agreement scores should ideally be complemented with accuracy scores, i.e. scores that express how many of the annotations are actually correct according to a reference annotation (a *gold standard*).

In this paper a study is presented in which we compare the difference in inter-annotator agreement of naive and expert annotators on task-oriented dialogue for the DIT^{++} dialogue act tagset, and assess the accuracy of naive and expert annotation against a gold standard. In Section 2. we will discuss the dialogue act data, the dialogue act scheme that we used, and the annotator groups that have participated in the experiments. The results will be presented in Section 3. and Section 4. The effect of reducing the complexity of the tagset on the agreement scores is addressed in Section 5., which is followed by a discussion and conclusions in Section 6.

2. Experiment outline

2.1. Naive versus expert annotators

The aim of the annotation experiment is to contrast annotations performed by naive annotators with those performed by expert annotators and evaluate on both inter-annotator agreement and tagging accuracy. Naive annotators can be characterised as subjects that have not been linguistically trained but that have participated in an introductory session explaining the dialogue data, the dialogue act tagset,

[1]In the case of Cohen's kappa, this is often taken to be between 0.8 and 1.0. For a general discussion, see e.g. (Landis and Koch, 1977; Krippendorff, 1980).

and the use of an annotation tool. Expert annotators can be characterised as linguistically trained subjects that have experience in annotating dialogue and are thoroughly familiar with the tagset.
In the role of naive annotators, six undergraduate students annotated the selected dialogue material. They had been introduced to the annotation scheme and the underlying theory as part of a course in pragmatics. During this course they had approximately four hours of lecturing and a few small annotation exercises. Two PhD students annotated as experts[2]. They have been actively working with the annotation scheme for more than two years and have annotated substantial parts of dialogue corpora. In order to calculate accuracy scores, i.e. to asses to what extent the annotators in both groups have annotated correctly, a gold standard is required. To obtain such a gold standard annotation, the authors[3] have analysed and discussed the available annotations and have established full agreement. The few cases for which fundamental disagreement or unclarity remained were kept out of the gold standard.
For all dialogues, the audio recordings were transcribed and the annotators annotated pre-segmented utterances for which full agreement had been established on segmentation 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[4]. The task was facilitated by the use of an annotation tool that had been built for this occasion (Geertzen, 2007). This tool allowed the subjects to assign each utterance one tag for each dimension without any further constraints. Both the naive and expert annotators could provide comments with each utterance for indicating problems, explaining the decision to choose a particular tag, or indicating that none of the available dimensions was addressed. The last mentioned case did not happen for the expert annotators and happened two times for the naive annotators.

2.2. Corpus data

The dialogues that were annotated are task-oriented and are all in Dutch. To account for different complexities in the interaction, both human-machine and human-human dialogues are considered. The dialogues analysed are drawn from different corpora: OVIS (Strik et al., 1997), DIAMOND (Geertzen and Bunt, 2006), and a collection of Map Task dialogues (Caspers, 2000). The number of utterances that are drawn from each corpus are specified in Table 1.
On average, naive annotators needed 23.2 seconds to annotate each utterance where expert annotators needed 11.8 seconds.

[2]Two of the authors participated as expert annotators.

[3]One of which is actively involved in the definition and refinement of the dialogue acts.

[4]Both the definitions and guidelines have been used and tested in earlier annotation sessions and have been improved over time as a result of feedback and analysis of disagreement. The dialogue act definitions and guidelines can be found at `http://dit.uvt.nl/` and `http://dit.uvt.nl/guide/`, respectively.

corpus	domain	type	#utt
OVIS	train connections	H-M	193
DIAMOND	operation of a fax machine	H-M	131
		H-H	114
DUTCH MAPTASK	map task	H-H	120
			558

Table 1: Characteristics of the utterances considered.

2.3. Dialogue act tagset

The DIT^{++} tagset was designed to combine in one comprehensive annotation scheme the communicative functions of dialogue acts distinguished in Dynamic Interpretation Theory (DIT, (Bunt, 2000)), and many of those in DAMSL (Allen and Core, 1997) and in other annotation schemes. Important differences between the DIT^{++} and DAMSL schemes are the more clearly defined notion of dimension (Bunt, 2006) and 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 (see: Allwood et al. (1993)).
The DIT^{++} taxonomy distinguishes 11 dimensions, addressing information about the task domain (*Task*); providing communicative feedback (*Auto-* and *Allo-feedback*); managing difficulties in speaking (*Own Communication Management* and *Partner Communication Management*), dealing with *Turn Management*, *Contact Management* and *Time Management*, addressing the structure of the dialogue (*Dialogue Structuring* and *Topic Management*), and dealing with social conventions (*Social Obligations Management*). For each dimension, at most one communicative function can be assigned. The taxonomy contains two types of communicative functions: those linked to a particular dimension ('dimension-specific functions') an those which can be applied in any dimension ('general-purpose functions').

3. Quantitative comparative results

Table 2 shows the inter-annotator agreement statistics for each dimension, averaged over all annotation pairs. With *annotation pair* is meant a pair of assignments an utterance received from two annotators for a particular dimension. The kappa figures in the table are based on those cases in which both annotators assigned a function to a specific utterance for a specific dimension. For each annotator group, scores for observed agreement (p_o), expected agreement (p_e), and Kappa (κ_{tw}) are listed in the first, second, and third column, respectively. These statistics are taxonomically weighted (see: Geertzen and Bunt (2006)) and as such take into account semantic and pragmatic relatedness of concepts. This means that when there is disagreement on two dialogue acts that have much in common, disagreement is considered partial instead of full (as is the case with Cohen's standard kappa) with the result that the disagreement is more accurately quantified. Table 3 is included to have an idea how the disagreement scores are when standard kappa instead of κ_{tw} is used, .
The column *#pairs* indicates on how many annotation pairs the statistics are based. The last column shows the ap*-ratio*.

Dimension	naive annotators					expert annotators				
	p_o	p_e	κ_{tw}	#pairs	ap-ratio	p_o	p_e	κ_{tw}	#pairs	ap-ratio
task	0.63	0.17	0.56	3000	0.81	0.85	0.16	0.82	298	0.78
auto feedback	0.67	0.48	0.36	615	0.53	0.92	0.57	0.82	85	0.64
allo feedback	0.53	0.29	0.33	91	0.02	0.85	0.24	0.81	23	0.38
turn	0.67	0.44	0.40	6	0.10	0.84	0.68	0.48	86	0.68
time	0.87	0.84	0.20	169	0.51	0.98	0.87	0.88	65	0.89
contact	0.80	0.66	0.41	10	0.19	0.75	0.38	0.60	8	0.50
topic	nav	nav	nav	2	0.06	nav	nav	nav	nav	nav
own communication	1.00	0.50	1.00	2	0.06	1.00	0.38	1.00	4	0.17
partner communication	1.00	1.00	nav	3	1.00	1.00	1.00	nav	2	1.00
dialogue structuring	0.80	0.30	0.71	83	0.32	0.92	0.38	0.88	14	0.65
social obligations	0.95	0.28	0.93	369	0.72	0.93	0.24	0.91	30	0.86

Table 2: Inter-annotator agreement for naive and expert annotators, per dimension, drawn from the set of all annotation pairs.

This figure indicates which fraction of all annotated functions in that dimension are present in annotation pairs. If $\#ap$ denotes the number of annotation pairs and $\#pa$ 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 $\#ap/(\#pa + \#ap)$.

From Table 2, it is obvious that for almost all dimensions, expert annotators obtain substantially higher agreement, as was to be expected. Considering the ap-ratio's for both annotator groups, it can be observed that for most dimensions expert annotators agree more on whether or not to assign a communicative function.

The scores for tagging accuracy are found in Table 4. Accuracy was calculated for both groups of annotators in two ways: by *taxonomically weighted Kappa* scores (column κ_{tw}), and by means of taxonomically weighted *percentage agreement* with the gold standard (column p_o). For each annotator a taxonomically weighted kappa score is calculated with the gold standard. The resulting scores are averaged to obtain a single score for each group. This is done for each dimension in the tagset. Second, for each annotator group the percentage agreement is calculated by similarly averaging individual percentage agreements. Note that both accuracy scores are slightly higher than the corresponding average scores for inter-annotator comparison. When we generalise over all dimensions and calculate a single accuracy score for each group, naive annotators score 0.67 and experts score 0.92. The considerably higher score for experts is not a surprise considering the per-dimension scores, but it is particularly interesting to see if there are annotators that deviate substantially in accuracy from the others in the group. For if this is the case this tells us more if the tagging accuracy per dimension is positively or negatively biased. The accuracy scores of individual annotators are visualised in Figure 1.

From this figure, we see that for the naive annotators (N1 until N6), there is more deviation from the group mean than for experts (E1 and E2). More importantly, annotator N6 deviates considerably from the other annotators in the group, causing the performance of the naive annotators to be biased positively. For the two expert annotators, having high tagging accuracy, there is only little deviation from

Dimension	naive annotators			expert annotators		
	p_o	p_e	κ	p_o	p_e	κ
task	0.45	0.09	0.40	0.83	0.16	0.90
auto feedback	0.31	0.14	0.20	0.87	0.45	0.77
allo feedback	0.26	0.10	0.18	0.74	0.17	0.69

Table 3: κ scores for dimensions where κ and κ_{tw} differ.

Dimension	naive annotators			expert annotators		
	p_o	p_e	κ_{tw}	p_o	p_e	κ_{tw}
task	0.64	0.16	0.58	0.91	0.16	0.90
auto feedback	0.74	0.46	0.52	0.94	0.48	0.88
allo feedback	0.58	0.19	0.48	0.95	0.22	0.94
turn	0.67	0.52	0.31	0.92	0.67	0.76
time	0.92	0.81	0.57	0.99	0.88	0.94
contact	1.00	0.60	1.00	0.91	0.48	0.83
topic	nav	nav	nav	nav	nav	nav
own comm.	1.00	0.52	1.00	1.00	0.38	1.00
partner comm.	1.00	1.00	nav	1.00	1.00	nav
dialogue struct.	0.89	0.36	0.82	0.87	0.34	0.81
social obl.	0.96	0.26	0.94	0.95	0.23	0.94

Table 4: Tagging accuracy for naive and expert annotators, per dimension, drawn from the set of all annotation pairs.

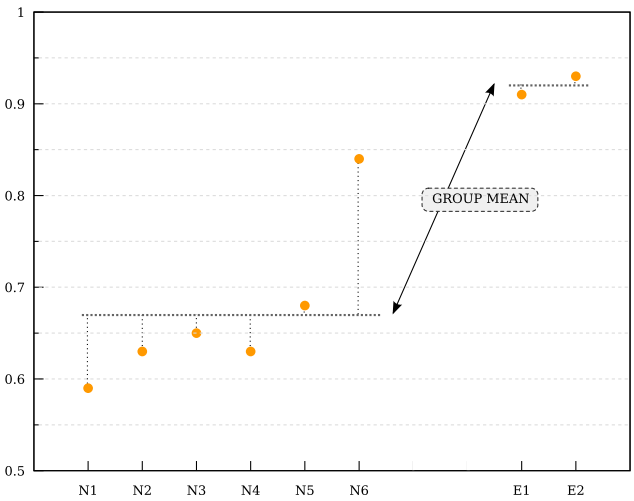

Figure 1: Tagging accuracy for naive and expert annotators.

the group mean.

4. Qualitative comparative results

To get a better understanding of the differences between naive and expert annotators as indicated by the statistics presented in the previous section, we can consider the co-occurrence matrices of dialogue acts and the actual annotations[5].

When the task and feedback dimensions are considered, which are relatively rich in dialogue acts, the intuition that naive annotators show more diversity in the dialogue act pairs that are involved in disagreements is confirmed. There are some cases in which both naive annotators and expert annotators show disagreement, with the difference that the magnitude of disagreement is less for the expert annotators. For instance, typical co-occurrences of dialogue acts of disagreements in the dimension Task are INFORM with ELABORATE and INFORM with WH-ANSWER, which occur for naive annotators 8.6 and 4.2 percent, respectively, and for expert annotators 1.7 and 1.3 percent, respectively, of all annotation pairs. Even though the experts do better than the naive annotators, this kind of pattern motivates action to be taken in improving the tagset with respect to the concept definitions involved.

Then, there are co-occurrences for which the naive annotators show considerable disagreement, and the experts do (almost) not. An example in the Task dimension is the co-occurrence of the communicative function INFORM with EXPLAIN. Sometimes, it occurred that naive annotators show, relatively to the number of annotation pairs, less disagreement than the experts. For instance, for naive annotators 0.7 percent of all annotation pairs involved the co-occurrence WH-ANSWER with INSTRUCT whereas for experts this was 2.0 percent. The reason why this happened becomes apparent when we take a look at the annotations that have been made in this context, for which the following dialogue excerpt[6], annotated for the Task dimension, is illustrative:

	utterance	expert 1	expert 2
S_1	do you want an overview of the codes?	YN-Q	YN-Q
U_1	yes	YN-A	YN-A
S_2	press function	INSTRUCT	WH-A
S_3	press key 13	INSTRUCT	WH-A
S_4	a list is being printed	INFORM	WH-A

Where naive annotators stayed close to question-answer adjacency pair patterns, the two experts generally disagreed on the specificity, in that expert 1 almost consistently annotated responses that were instructions as an INSTRUCT where expert 2 annotated them as a WH-ANSWER.

Analysis of the co-occurrence matrices showed a few other systematic differences between naive and expert annotators, most notably in Turn Management. As can be seen in Table 4, both naive and experts annotators failed to reach substantial agreement on assigning turn management functions. In dialogue, especially in multi-party interaction, interlocutors often signal eagerness to obtain the turn by interrupting the partner (TURN GRAB), to take the turn if available (TURN TAKE), to accept the turn when it was assigned to them (TURN ACCEPT), after finishing the contribution to explicitly assign the speaker role to an addressee (TURN ASSIGN), to drop the speaker role without putting any pressure on the addressee to take the turn (TURN RELEASE), or decide to continue as a speaker (TURN KEEP). Very often, interlocutors just start to speak if they want to say something and stop speaking if they are finished with their contributions. In these cases it is the question whether to annotate every first utterance in a turn as having a TURN TAKE function and every last utterance in the turn as having a TURN RELEASE function. The DIT^{++} annotation guidelines state[7] that there is no turn management when the speaker does not signal an intention to address the turn allocation explicitly and when the annotator does not have sufficient evidence in terms of utterance features (such as intonational cues). The lack of agreement was caused by a lack of such evidence. For example, to signal the intention to keep the turn the speaker may use, besides fillers such as *um* or *uh*, pauses, rising intonation, and the slowing down of speech rate. In particular the latter may be expressed subtly, which makes the annotator's decision rather subjective. Nevertheless, the experts annotators showed a more reliable intuition by reaching an agreement of 76.7 percent where naive annotators reached 66.7 percent. An example where prosodic rather than lexical cues address turn management is the following[8]:

	utterance	naive	expert
S_1	from which station to which station do you want to travel?	TAS:WH-Q	TAS:WH-Q
U_1	from...	TIM:STALL	TIM:STALL TUM:KEEP

Another source of disagreement on turn management originates from dealing with multifunctionality. For instance, discourse markers such as *and*, *or*, or *but* are known to have multiple functions in dialogue, and as a rule link dialogue units and signal speaker-identification (TURN TAKE) or speaker-continuation (TURN KEEP). For instance, consider the following excerpt[9]:

	utterance	naive	expert
A_1	to the left...	TAS:WH-A	TAS:WH-A TUM:KEEP
A_2	**and then** slightly around	TAS:WH-A	TAS:WH-A TUM:KEEP

The expert annotators fully exploited the phenomenon of multifunctionality in their annotations and assigned all

[5] The examples in this paper are all translated from Dutch.

[6] This excerpt originates from the human-machine part of the DIAMOND corpus.

[7] And the annotators were instructed accordingly.

[8] This excerpt originates from the OVIS corpus (H-M).

[9] This excerpt originates from the map task corpus (H-H).

functions they thought are applicable, whereas the naive annotators did not make use of this.

5. Effects of tagset complexity reduction

From the number of annotation pairs in Table 2 (column `#pairs`) it can be concluded that six dimensions were addressed much more often than others: Task, Auto-feedback, Allo-feedback, Turn Management, Time Management and Dialogue Structuring. Of these, both feedback dimensions and the Turn Management dimension have low agreement scores for the naive annotators, while Turn Management has a low agreement score for both groups of annotators. It was found that it is often difficult for annotators to determine the level of feedback (attention, perception, understanding, evaluation or execution), while for Turn Management the annotation guidelines were found to be unclear, as already mentioned (Note the low ap-ratios for this dimension for both groups).

These and other more detailed findings were used for designing a revised tagset as well as improving the annotation guidelines within the European project LIRICS[10] (see: Schiffrin and Bunt (2007)). Within this project, a test suite was developed of dialogues in several European languages which were annotated with the revised tagset. For English and Dutch the test suite dialogues were all annotated by two expert annotators. An analysis of the agreement between their annotations reveals that in all of the frequently addressed dimensions a very high agreement was reached (weighted kappa scores well above 0.9). By applying a mapping from the original DIT^{++} tagset to the revised LIRICS tagset the effects can be calculated that this revision should have on the agreements scores for both groups of annotators. The effect of the improvement of the annotation guidelines cannot be calculated in this way, but an estimation of that effect can be obtained by comparing the calculated improved agreement scores for the expert annotators with the scores that were found in the LIRICS project.

In DIT^{++} some of the dimensions contain one or multiple hierarchies of dialogue acts. The dialogue acts in such hierarchies are related in such a way that an act lower in a hierarchy is more specific than an act higher in the same hierarchy. For instance, in Figure 2 a `CHECK` is more specific than a `YN-QUESTION`, which is in turn more specific than a `INDIRECT-YN-QUESTION`.

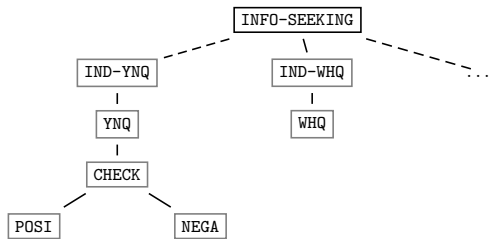

Figure 2: Two hierarchies of information-seeking general purpose functions.

Using the existing hierarchical structure, we could partially (or fully) 'collapse' a hierarchy and group acts together under a least specific parent act, flattening the hierarchy and making the tagset less complex. There are two major motivations for doing so. Firstly, by grouping dialogue acts together, disagreement that is the result of considering fine-grained distinctions is eliminated. Secondly, grouping dialogue acts can make inter-annotator agreement analysis less susceptible to very infrequently occurring, fine-grained dialogue acts which occur too infrequently to draw significant conclusions in evaluation. It should be remarked that collapsing a hierarchy to a general dialogue act is only justified when the general dialogue act is sufficiently fine-grained for the application of the tagset. There are various ways in which hierarchies can be collapsed to general dialogue acts. The dialogue acts proposed in the LIRICS project are based on acts in the DIT^{++} tagset but exhibit lower granularity, making it interesting to collapse DIT^{++} hierarchies to LIRICS dialogue acts in order to predict the performance of both annotator groups. Additionally, it would provide indicative inter-annotator agreement scores for dialogue acts in LIRICS. Because almost all hierarchies in the DIT^{++} tagset are either in the set of general-purpose communicative functions or in the feedback dimensions, we focus on these parts of the tagset. The grouping and mapping used for LIRICS are depicted in Figure 3.

As was to be predicted, the scores for both annotator groups improved after recalculating inter-annotator agreement and accuracy for the LIRICS dialogue acts. The differences in inter-annotator agreement are given in Table 5.

	naive annotators		expert annotators	
Dimension	DIT	LIRICS	DIT	LIRICS
task	0.56	0.65	0.82	0.86
auto feedback	0.36	0.71	0.82	0.88
allo feedback	0.33	0.46	0.81	0.85

Table 5: Agreement (in κ_{tw}) for LIRICS dialogue acts.

As can be seen from the table, the improvement for naive annotators is higher than that for expert annotators. When looking to the annotation it is not difficult to indicate why; for instance, in quite some cases of feedback — most notably those with feedback not being realised verbally — it is difficult to determine the feedback level, especially for naive annotators. By grouping all levels of feedback, this substantial source of disagreement got eliminated. The gain in accuracy turned out to be proportional to the relative gain in inter-annotator agreement, both for naive and expert annotators.

6. Discussion & conclusions

The statistics presented in Section 3. show that the scores for inter-annotator agreement are lower than those for annotation accuracy. This confirms that using inter-annotator agreement only when there is a possibility to use a gold standard would lead to underestimating the reliability of an annotation scheme.

We have seen in Table 2, inter-annotator agreement for naive coders is rather low where for expert annotators

[10] Linguistic Infrastructure for Interoperable Resources and Systems. See `http://lirics.loria.fr/`.

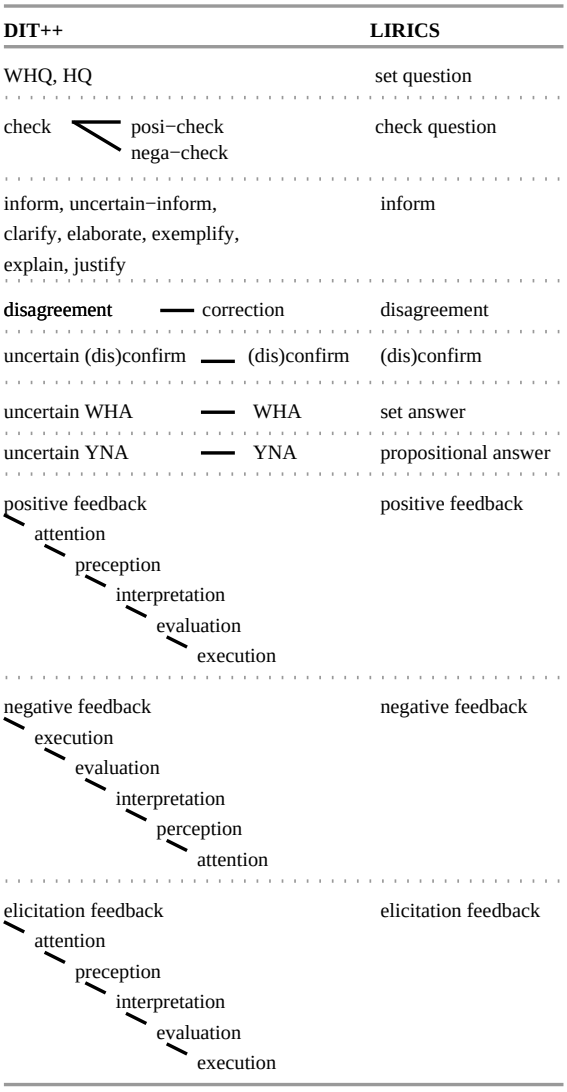

Figure 3: Grouping and mapping of dialogue acts, where lines indicate hierarchical relations.

agreement is high (mostly > 0.8). When looking at annotation accuracy it was found that calculating reliability based on inter-annotator agreement only results in an indication of reliability that is too low. We can conclude that both inter-annotator agreement and annotation accuracy statistics are informative in determining how reliably a scheme can be used for annotation. Calculation of the latter indicator presupposes that on expert level a ground truth can be established, meaning that the concepts in the scheme should not be too subjective and should be sufficiently well-defined. The expectations that inter-annotator agreement and accuracy scores are both higher for expert annotators are confirmed.

Remarkably, it occurred that naive annotators showed higher inter-annotator agreement for the dimension *Social-obligations Management* and higher tagging accuracy for the dimension *Contact Management*. For both cases this difference is explained by the interaction of the score with the ap*-ratio*. Naive annotators disagree more (with each other and with the gold standard) whether or not to annotate in a specific dimension, but the cases in which there is agreement are mostly the easy ones to annotate. Conversely, expert annotators show more agreement on when to annotate in a specific dimension, but as a result are also addressing more difficult cases.

When reducing the granularity of the DIT^{++} tagset by collapsing its hierarchies to obtain the LIRICS dialogue acts, evaluation scores for naive annotators improved substantially more than those for expert annotators but the latter group has better scores. This confirms the intuition that on less complex tagsets the difference between naive and expert annotators becomes smaller.

Some objections to using a weighted metrics, such as κ_{tw}, are discussed in (Artstein and Poesio, *to appear*). In their thorough overview of inter-coder agreement used in computational linguistics, it is concluded that weighted metrics are not easy to interpret. However, while it is true that the absolute value of the weighted kappa is not easy to interpret, for the analyses presented in this paper only the differences between κ_{tw}-values for different annotators are essential. Moreover, we would like to stress once more that quantitative indicative figures such as agreement scores should be complemented with qualitative analyses including co-occurrence matrices[11].

In conclusion, we can summarise by stating that differences in both inter-annotator agreement and tagging accuracy between naive and expert annotators against the gold standard are considerable, and that the annotations of both groups provide complementary insights in reliability to each other concerning clarity and accessibility of the tagset, and fundamental conceptual issues. In comparing both annotator groups, it turned out that for multidimensional dialogue act taxonomies it is essential to distinguish agreement on whether or not to annotate in a dimension from agreement on the dialogue act or communicative function within a dimension.

Acknowledgements

This research was partly supported by the Netherlands Organisation for Scientific Research (NWO), grant 017.003.090.

7. References

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

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

Ron Artstein and Massimo Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics, To appear*. See: `http://cswww.essex.ac.uk/Research/nle/arrau/icagr.pdf`

[11] To provide such an in-depth analysis is beyond the scope (and aim) of this paper; see also (Geertzen, 2006).

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

Harry Bunt. 2006. Dimensions in dialogue annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1444–1449.

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.

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

Jeroen Geertzen and Harry Bunt. 2006. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 126–133.

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

Jeroen Geertzen. 2007. DitAT: a flexible tool to support web-based dialogue annotation. In Jeroen Geertzen, Elias Thijsse, Harry Bunt, and Amanda Schiffrin, editors, *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS)*, pages 320–323.

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

J. Richard Landis and Gary G. Koch. 1977. A one-way components of variance model for categorical data. *Biometrics*, 33:671–679.

Amanda Schiffrin and Harry Bunt. 2007. Defining a preliminary set of interoperable semantic descriptors. LIRICS Project Deliverable D4.2, Tilburg University.

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.