Ranked multidimensional dialogue act annotation

Marcin Włodarczak

Bielefeld University

Abstract. In this paper we propose a dialogue act annotation system allowing ranking of communicative functions of utterances in terms of their subjective importance. It is argued that multidimensional dialogue act annotation schemes, while allowing more than one tag per utterance, implicitly treat all functions as equally important. Consequently, they fail to capture the fact that in a given context some of the functions of an utterance may have a higher priority than its other functions. The present approach tries to improve on this deficiency. Preliminary results of an annotation experiment suggest that ranking communicative functions accurately reflects the communicative competence of language users.

1 Introduction

Multifunctionality of utterances is often acknowledged in modern dialogue studies [1–3]. It is argued that participants simultaneously address several aspects of communication such as providing feedback, managing the turn-taking process and repairing faulty utterances. Various kinds of implicit functions are an additional source of multifunctionality [4]. The requirement for accounting for multifunctionality of utterances is, of course, also valid for dialogue act annotation schemes. There the notion of multifunctionality is usually introduced explicitly in the form of multidimensional annotation schemes, which allow an utterance to be labelled with more than one tag. However, in such schemes each utterance is represented as an unstructured set of tags. Consequently, they do not reflect the hierarchical organisation of utterance functions determined by speakers' communicative goals. The approach presented here tries to enrich the existing frameworks with a notion of ranking of communicative functions. Importantly, it allows more than one highest-ranking function and more than two different ranks.

The paper has the following structure. In the following section the notion of multidimensional tagsets is introduced. In Sec. 3 existing annotation frameworks are presented alongside the alternative approach proposed in the present paper. The design and the results of an annotation task conducted to validate this framework are presented in Sec. 4, and are followed by conclusions in Sec. 5.

2 Multidimensional Tagsets

Unlike in one-dimensional tagsets, which only allow one tag per utterance, in multidimensional tagsets each utterance can be labelled with multiple tags, each representing a different communicative function. We adopt here the formal definitions of both kinds of tagsets given in [2].

Definition 1. *A one-dimensional tagset is a set* $A = \{a_1, a_2, \ldots, a_N\}$*, each utterance being tagged with exactly one elementary tag* $a_n \in A$*.*

Definition 2. *A multi-dimensional tagset is a collection of dimensions (or classes, categories, etc.)* $\mathcal{T} = \{A, B, \ldots\}$ *where each dimension is in turn a list of tags, say* $A = \{a_1, a_2, \ldots, a_M\}, B = \{b_1, b_2, \ldots, b_N\}$*. When a multi-dimensional tagset is used, each utterance is tagged with a composite* label *or* tuple *of tags* $(a_i, b_j, \ldots)$*.*

Obviously, this is a highly idealised view since it requires that for each utterance a tag is specified in each dimension. If, as is most often the case (see [4]), this requirement is not met and a tag is specified only in some dimensions, the empty tag $\emptyset$ must be added to each of the dimensions In such cases, the empty label $(\emptyset, \emptyset, \ldots, \emptyset)$ must be ruled out. The set of possible labels is then $(A \times B \times C \times \ldots) - (\emptyset, \emptyset, \ldots, \emptyset)$.

Alternatively, rather than employ the notion of the empty tag, only those dimensions can be considered in which a non-empty tag is applicable. This is the approach adopted in [7]:

Definition 3. *A multidimensional dialogue act assignment system is a 4-tuple* $A = (D, f, C, T)$ *where* $D = D_1, D_2, \ldots, D_m$ *is a dialogue act taxonomy with 'dimensions'* $D_1, D_2, \ldots . D_m$*,* f *is a function assigning tags to utterances,* C *is a set of constraints on admissible combination of tags, which additionally allow a dialogue utterance to be assigned a tag in each of the dimensions, but never more than one tag per dimension, and* T *is a set of additional labels that* f *may assign to utterances—*T *contains such labels as* inaudible *or* abandoned[1]*.*

Notably, the set C should be kept relatively small to make *orthogonality* of dimensions as high as possible. This ensures that any combination of tags from different dimensions is admissible [8].

3 Ranked Annotation System

As mentioned above, multifunctionality of utterances is a result of the fact that speakers simultaneously address several aspects of communication. Furthermore, it could be argued that depending on the context specific aspects might be more important than others, thereby forming a hierarchical ordering of functions, a

[1] It could be argued that a 5-tuple should be used instead. The additional element would define a domain of the function f—a set of utterances.

possibility hinted at already in [12]. However, it should be clear that multidimensional dialogue act schemes are not capable of capturing this notion. Instead, they implicitly treat all functions as equally important.

Surprisingly, the problem has received relatively little attention in literature. Bunt and Geertzen [13], discussing their modifications to the kappa statistic, remark that utterances may be argued to have a *primary function* and possibly several *secondary functions*, and note that disagreement about the former is usually more serious than about the latter.

Popescu-Belis observes that although multidimensional tagsets better reflect the multifunctionality of utterances, one-dimensional tagsets offer an advantage of having a much smaller search space, which leads to higher human and automatic annotation accuracy [5]. One of the ways of overcoming the trade-off between a rich pragmatic representation and a smaller search space is only considering the observed tag combinations. For example, the SWDB-DAMSL tagset [9] was developed by clustering 220 DAMSL [10] tag combinations which occurred in 205,000 utterances of the Switchboard corpus into 42 final mutually exclusive tags.

Instead, [5] proposes an alternative strategy. *Dominant Function Approximation* (DFA) assumes that a tagset specifies default values in every dimension based on linguistic and pragmatic grounds or on frequency counts, and states that at most one communicative function of an utterance is non-default (it is then called a *dominant function*). The author notes that while the DFA might be acceptable for current technological applications, it might not be sufficient for a detailed linguistic analysis.

Popescu-Belis tried to verify his hypothesis by checking the number of utterances with more than one non-default functions in existing annotations. Since the number was found to be relatively small (between 3 and 8%), the DFA seems to be correct. However, it could be argued that such findings might be a result of specific annotation guidelines, which often instruct annotators to only mark the most significant function. Indeed, it seems that the possibility of an utterance having several dominant functions cannot be ruled out *a priori*. Moreover, the binary distinction into dominant and default functions may well turn out to be too restrictive.

Alternative Approach. The present approach proposes to model the relative prominence of communicative functions by means of *greater or equal prominence* relation. The term *prominence* will be henceforth used to denote the significance of a communicative function relative to other functions of the same utterance. It is assumed that prominences of every two functions of the same utterance are comparable, i.e. it is possible to decide whether one of the functions is more prominent than the other or whether they are equally prominent. Consequently, the relation in question imposes a *non-strict linear order* on the set of functions of an utterance. Importantly, the ordering of functions is viewed here from the speaker's point of view, i.e. it is assumed that in a given context accomplishing some of the speaker's goals is of greater importance than accomplishing some other goals. The lower-ranking functions may either accomplish ancillary goals or be a means of accomplishing the higher-ranking goals. [17] suggest that entailment relations [4] between communicative functions might be a major factor influencing their relative prominence.

A set of functions of an utterance with equal prominences will be referred to as a *level of prominence*. It should be clear that each level of prominence is an equivalence class given an equivalence relation of *equal prominence*. Obviously, levels of prominence can be also ordered with respect to the prominence of their elements, i.e. one level of prominence precedes another level of prominence if the prominence of functions in the first is greater than the prominence of functions in the second (relation of *strict linear order*).

This approach might be thought of as a generalisation of the approaches outlined above by imposing fewer constraints on the number of levels of prominence. Specifically, multiple functions are allowed to have the same prominence, i.e. every level of prominence may have more than one element. One of the consequences of this is that many dominant (highest-ranking) functions are allowed. Therefore, the approach allows for more flexibility.

It should be also noted that, unlike in the DFA, the notion of default values is not employed here. Moreover, while the DFA was proposed to *simplify* the pragmatic representation of an utterance in order to improve the accuracy of automatic and manual tagging, the present approach aims at *enriching* the pragmatic representation for the needs of linguistic analysis.

Lastly, the concept of the ordering of communicative functions can be easily incorporated into the definition of Multidimensional Dialogue Act Assignment System (Def. 3) to capture the notion of the *Multidimensional Ranked Dialogue Act Assignment System*:

Definition 4. *A Multidimensional Ranked Dialogue Act Assignment System is a 5-tuple $A = (D, f, R, C, T)$ where D, f, C and T are as before, and R is a relation of* greater of equal prominence *holding between functions represented as tags which f assigns to an utterance.*

4 Experiment

The following experiment was conducted to investigate how many dominant functions and how many levels of prominence are identified by annotators. It is based on an analogous experiment proposed by Popescu-Belis [5], namely participants were asked to order functions assigned to segments with respect to their relative prominence. However, unlike in the original design, minimal constraints were imposed on the ordering of functions of utterances. Since approaches like the DFA impose much stricter constraints on an annotation scheme, they would be supported if under these conditions the proportion of utterances with more than one dominant function and more than two levels of prominence was relatively low. Otherwise, the alternative approach outlined above would be more appropriate.

4.1 Experimental settings

HCRC Map Task Corpus [14] was used. Map task dialogues are task related dialogues in which participants cooperate to reproduce a route drawn in one participant's map on the other participant's map. Differences between the maps are introduced to make the task more difficult. The total duration of the data selected for the experiments equalled 4 minutes and 43 seconds.

The tagset chosen for the experiment was the DIT^{++} dialogue act taxonomy [11]. It consists of ten dimensions related to managing the task domain (*Task/Activity*), feedback (*Allo-* and *Auto-feedback*), time requirements (*Time Structuring*), problems connected with production of utterances (*Own* and *Partner Communication Management*), attention (*Contact Management*), discourse structure (*Discourse Structuring*) and social conventions (*Social Obligations Management*).

The data were segmented into functional segments [2] in accordance with [16], and annotated by two experts. 136 functional segments were identified. Full agreement had been reached with regard to segmentation and annotation. Importantly, entailed feedback functions [4] were included in the annotations.

Four naive annotators took part in the experiment. The annotators were undergraduate students at Tilburg University. They had been introduced to the annotation scheme and the underlying theory while participating in a course on pragmatics. The course comprised approximately three hours of lectures and a few small annotation exercises on data other than map task dialogues.

All annotators accomplished both tasks individually, having received the materials (transcriptions and sound files) in electronic form. Time for the task was not limited. To encourage high quality of annotations the students were motivated by an award of 10% of the total grade for the pragmatic course.

The participants' task was to order utterance functions to order the functions assigned to utterances with respect to their relative importance. The ordering was done by assigning each function a numerical value from the set of *consecutive* natural numbers, starting from "1" as the most prominent function. The lowest possible rank was, therefore, equal to the number of utterance functions. However, more than one function could be assigned the same numerical value.

4.2 Results and Discussion

Since it was observed that participants failed to rank functions of some segments, the total number of analysed rankings was equal to 293 (243 and 55 for segments with two and three functions respectively). Cohen's kappa [18] was calculated for 54 segments (44 and 10 with two and three functions respectively) ranked properly by all four participants.

Inter-rater agreement values for functions assigned specific ranks are given in Tab. 1 and 2. As can be observed, mean kappa values indicate fair to moderate

[2] [15] defines a *functional segment* as a "minimal stretch of communicative behaviour that has one or more communicative functions."

agreement. It should be borne in mind, however, that while the participants had some experience using the DIT^{++} tagset, they were completely naive with regard to ranked annotation. It could be, therefore, hoped that more experienced annotators could achieve much higher agreement. Moreover, kappa values do not seem to decrease substantially across ranks. Indeed, while the number of segments with three functions was rather low, in Tab. 2 agreement for the third rank was higher than for the second rank. These results contrast sharply with the assumptions of the DFA, which would predict that agreement values should drop across ranks.

Table 1. Kappa coefficient values for functions assigned specific ranks in two-functional segments.

Annotators	Rank 1	Rank 2
1 & 2	0.46	0.35
1 & 3	0.64	0.67
1 & 4	0.34	0.32
2 & 3	0.27	0.21
2 & 4	0.41	0.37
3 & 4	0.47	0.49
Mean	0.43	0.40

Table 2. Kappa coefficient values for functions assigned specific ranks in three-functional segments.

Annotators	Rank 1	Rank 2	Rank 3
1 & 2	0.75	0.38	0.41
1 & 3	0.29	0.37	0.55
1 & 4	0.49	0.21	0.54
2 & 3	0.31	0.23	0.14
2 & 4	0.51	0.08	0.44
3 & 4	0.51	0.56	0.44
Mean	0.48	0.30	0.42

By comparison, [19] reports results of a ranking experiment using a simplified versions of the DIT^{++} tagset and ten completely naive raters. Perhaps not surprisingly, some the the observed kappa values were lower than those in the present study. More interestingly, however, the value for the first rank was substantially higher than for the remaining ranks. Specifically, it was found that for two-functional segments inter-rater agreement was equal to 0.39 for the first rank and 0.1 for the second rank. In the category of utterances with three functions kappa values equalled 0.18, 0.04 and 0.04 for the ranks of one, two and three. Notably, the values for the ranks of two and three are identical, which might indicate that functions with these ranks did not differ much with regard to their relative prominence. While these results are in accordance with the DFA, it should be noted that participants had no experience not only with ranking but with the tagset itself. This suggests that the DFA could prove more useful when completely naive annotators are used.

Proportions of utterances with different numbers of identified levels of prominence are presented in Fig. 1. Overall, in 97% of segments the number of identified levels of prominence was equal to the number of segment functions. Only in three out of 243 two-functional segments, and five out of 55 three-functional segments was it otherwise. Since minimally two levels of prominence were identified in three-functional segments, at most two functions were assigned the same rank. However, all these cases came from the same annotator, and might, therefore, be highly idiosyncratic.

The DFA predicts that the proportion of utterances with more than two levels of prominence should be small. Obviously, since utterances with two functions can be assigned the maximum of two distinct ranks, only three-functional segments are of interest in this respect. Although there were relatively few such segments, as much 91% of them would not be represented correctly if more restrictive annotation guidelines, such as the DFA, were adopted.

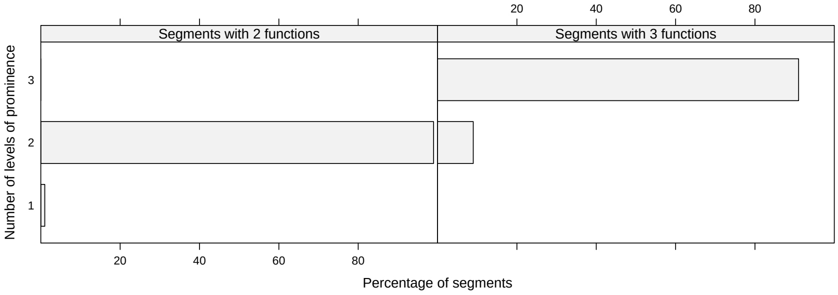

Fig. 1. Proportions of segments with different numbers of levels of prominence

Fig. 2 presents proportions of utterances with different numbers of identified dominant functions (i.e. functions assigned the rank of one). Here the overwhelming tendency is for a segment to have exactly one such function. This was the case for 99% of two-functional segments and 91% of three-functional segments. The remaining cases again came from the same annotator.

Considering the results regarding the numbers of dominant functions and levels of prominence together, it should be said that there is a very strong tendency for each function to be assigned a different rank. The relation of *greater or equal prominence* is, therefore, in most cases a relation of *greater prominence*, i.e. it is a relation of *strict* linear order. Consequently, the DFA is only partially correct. It is right in predicting one dominant function per segment but does not differentiate between the prominences of non-dominant functions. However, it is interesting to note that whenever the same rank was assigned to two functions, it was in fact the first rank in all but one case.

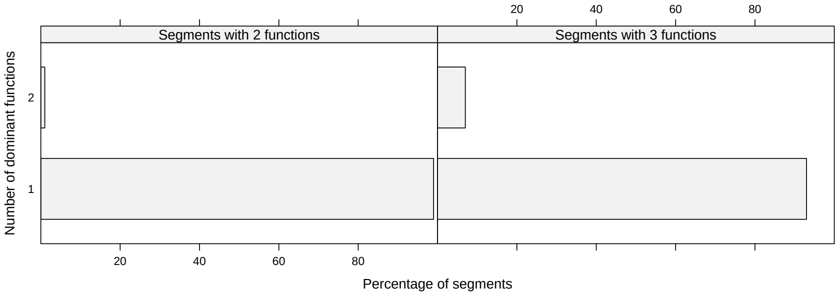

Fig. 2. Proportions of segments with different numbers of dominant functions

Figure 3 presents distributions of functions of two-functional segments belonging to specific dimensions across ranks. While functions from most dimensions are assigned the ranks of one and two with comparable frequencies, there is a noticeable difference between frequencies of *Turn Management* and *Feedback* functions. Specifically, *Feedback* functions are the most frequent of functions assigned the rank of one (38%), and *Turn Management* functions are the second most frequent (29%). By contrast, among functions assigned the rank of two it is the other way round with *Turn Management* functions comprising 43%, and *Feedback* functions comprising 30%. Additionally, *Task Management* functions have a higher frequency among the functions ranked second (18%) than among those ranked first (11%). Two-tailed Fisher's exact test was conducted to test whether the proportions between functions depend on the rank. The result was statistically significant with a p-value of 0.01.

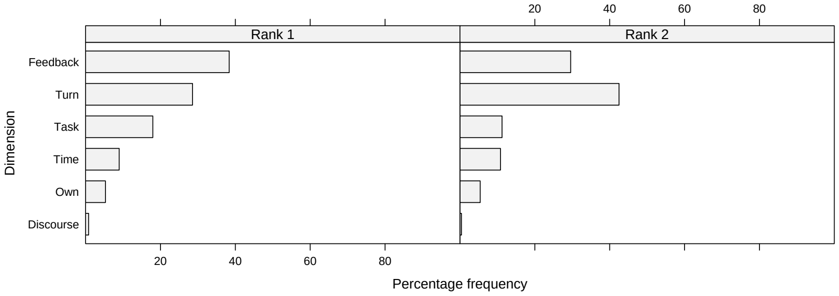

Fig. 3. Frequency distribution of ranks assigned to functions from specific dimensions in two-functional segments. The dimension names were abbreviated as follows: *Feedback–Auto-* and *Allo-feedback* clustered together, *Turn–Turn Management*, *Task–Task Management*, *Time–Time Management*, *Own–Own Communication Management*, *Discourse–Discourse Management*.

Analogous result for utterances with three functions are presented in Fig. 3. Here, except for minor differences among low frequency *Own Communication Management*, *Time Management*, and *Discourse Management* functions, the greatest differences concern functions from the *Feedback*, *Turn Management* and *Task* dimensions. While *Feedback* functions have the highest frequency across all three ranks but their dominance over the other two dimensions varies greatly depending on the rank. Among functions assigned the rank of one *Feedback* functions make up 41%, *Turn Management* functions make up 25%, and *Task* functions make up 14%. This difference is even larger among functions ranked second with respective frequencies of 57%, 14% and 10%, but is almost nonexistent in the category of functions ranked third, where their frequencies equal 28%, 26% and 26%. Two-tailed Fisher's exact test was again conducted to test whether the proportions between functions depend on rank. The result was statistically insignificant with a p-value of 0.06.

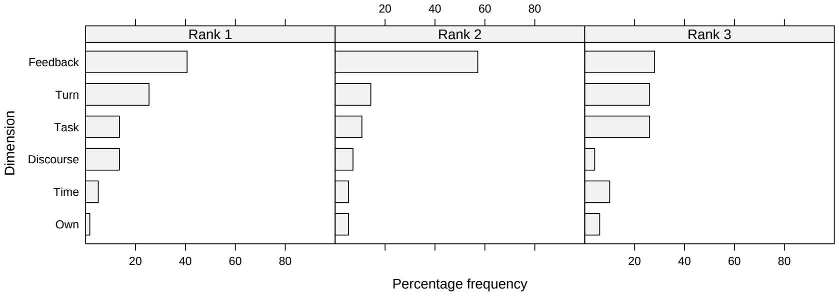

Fig. 4. Frequency distribution of ranks assigned to functions from specific dimensions in three-functional segments. For the explanation of the dimensions names abbreviations see Fig. 3.

5 Conclusions

The results reported above show clearly that in a great majority of cases the number of identified levels of prominence tends to be equal to the number of segment functions. In other words, each function is usually assigned a different rank. Therefore, the relation proposed in Sec. 3 was in most cases a relation of *strict* linear order. Apart from that, frequencies of functions from respective dimensions were found to depend on rank in case of two-functional segments, and to be independent of it in case of three-functional segments. However, since the analysed dataset (and, in particular, the number of segments with three functions) was relatively small, these results should be treated as preliminary.

In the light of these findings it must be said that the DFA is right in predicting that most segments have just one highest-ranking function but it fails to account for distinctions among lower-prominence functions. It is, of course, a question of specific research goals whether the resulting underspecification is considered acceptable. Regarding the notion of default values assumed in the DFA, the fact that each function was assigned a different rank in most of the three-functional segments seems to suggest that the usefulness of this notion is limited. Additionally, contrary to the assumptions of the DFA, inter-annotator agreement values were found to be similar across all ranks.

Obviously, the results obtained here should ideally be confirmed in a larger scale annotation experiment. In addition, a number of issues not discussed here could also be investigated. For example, rather that analyse frequencies of functions across ranks globally, relative prominences of specific *combinations* of functions could be analysed. This, in turn, should shed more light on the problem of default functions assumed in the DFA.

6 Acknowledgements

I would like to thank Harry Bunt and Volha Petukhova (Tilburg University) for granting me access to their data.

References

1. Bunt, H. (2000) *Dialogue pragmatics and content specification* In H. Bunt. and W. Black (eds.), Abduction, Belief and Context in Dialogue. Amsterdam: Benjamins
2. Popescu-Belis, A. (2005). *Dialogue Acts: One or More Dimensions?* ISSCO Working Paper, 62, University of Geneva
3. Allwood, J. (2000). *An activity-based approach to pragmatics* In H. Bunt. and W. Black (eds.), Abduction, Belief and Context in Dialgoue. Amsterdam: Benjamins
4. Bunt, H. (2009). Multifunctionality and multidimensional dialogue semantics. *Proceedings of DiaHolmia 2009*, Stockholm
5. Popescu-Belis, A. (2008). Dimensionality of dialogue act tagsets: an empirical analysis of large corpora. *Language Resource and Evaluation 42 (1)*
6. Bunt, H. (2006). Dimensions in Dialogue Act Annotation *Proceedings of LREC 2006*, Paris
7. Bunt, H. and Girard, Y. (2005). *Designing an Open, Multidimensional Dialogue Act Taxonomy.* In C. Gardent and B. Gaiffe (eds.) DIALOR'05, Proceedings of the Ninth International Workshop on the Semantics and Pragmatics of Dialogue, Nancy
8. Petukhova, V. and Bunt, H. (2009). The independence of dimensions in multidimensional dialogue act annotation. *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado
9. Jurafsky, D., Shriberg, E. and Biasca, D. (1997). *Switchboard SWBD-DAMSL Labelling Project Coder's Manual. Draft 13.* Technical Report 97-02. University of Colorado, Institute of Congnitive Science.
10. James Allen and Mark Core. (1997). *DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1).* Technical Report, Multiparty Discourse Group, Discourse Resource Initiative

11. Bunt, H. (2009). The DIT^{++} taxonomy for functional dialogue markup. *Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009)*, Budapest
12. Jakobson, R. (1960). *Linguistics and Poetics.* In Sebeok (ed.) Style in Language, 350–377, MIT Press: Cambridge, MA
13. Geertzen, J. and Bunt, H. (2006). Measuring annotator agreement in a complex hierarchical dialogue act scheme. *Proceedings of SIGDIAL 2006*, Sydney
14. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34
15. ISO (2009). *Semantic annotation framework (SemAF), Part 2: Dialogue acts.* ISO CD 24617-2, October 2009. ISO, Geneva.
16. Geertzen, J., Petukhova, V. and Bunt, H.. (2007). A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification. In: *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, pp. 140–149.
17. Włodarczak, M., Bunt, H. and Petukhova, V. (2010). Entailed feedback: evidence from a ranking experiment. In P. Łupkowski and M. Purver (eds.) *Aspects of Semantics and Pragmatics of Dialogue. Semdial 2010, 14th Workshop of the Semantics and Pragmatics of Dialogue*, Poznań
18. Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol.20, No.1, pp. 37–46.
19. Włodarczak, M. (2009). *Ranked multidimensional dialogue act annotation.* MA thesis, Adam Mickiewicz University, Poznan.