DiaBiz.Kom – Towards a Polish Dialogue Act Corpus Based on ISO 24617-2 Standard

Marcin Oleksy, Jan Wieczorek, Dorota Drużyłowska, Julia Klyus, Aleksandra Domogała, Krzysztof Hwaszcz, Hanna Kędzierska, Daria Mikoś, Anita Wróż
Wrocław University of Science and Technology, Wrocław, Poland
{marcin.oleksy, jan.wieczorek, dorota.druzylowska}@pwr.edu.pl
{julia.klyus, aleksandra.domogala, krzysztof.hwaszcz}@pwr.edu.pl
{hanna.kedzierska, daria.dominiak, anita.wroz}@pwr.edu.pl

Abstract

This article presents the specification and evaluation of DiaBiz.Kom – the corpus of dialogue texts in Polish. The corpus contains transcriptions of telephone conversations conducted according to a prepared scenario. The transcripts of conversations have been manually annotated with a layer of information concerning communicative functions. DiaBiz.Kom is the first corpus of this type prepared for the Polish language and will be used to develop a system of dialogue analysis and modules for creating advanced chatbots.

1 Introduction

The rationale of the current research was predominantly connected with the lack of corpora including dialogue texts in Polish which could be used to train artificial intelligence for model creation. In order to bridge this gap, we decided to create a corpus that satisfies our expectations i.e. the one that contains dialogue samples from several different fields of business and is annotated for information concerning pragmatic functions. At the first stage of our work, we analysed the approaches adopted by other researchers, whose solutions are described in the "Related Works" section. Further, we describe the process of data collection and its manual annotation. A separate section is dedicated to "Key assumptions and limitations of the guidelines". The paper ends with "Corpus overview" and "Conclusions and future works".

2 Related work

While creating the Polish corpus of dialogue acts, we analyzed some pre-existing corpora. In theoretical perspective we refer to the ISO/DIS 24617-2, standardized annotations and the recommendations from the mentioned document. The ISO standard (Bunt et al., 2010, 2012) is based on particular innovations such as distinguishing between annotations and representations (according to ISO Linguistic Annotation Framework (LAF, ISO 24612:2009) and sets of dialogue participants, dimensions, communicative functions, functional segments and qualifiers (inventory of DiAML). Both manual and automatic annotation of dialogue segments are possible according to the ISO document and both have been tested in practice and described (Keizer et al., 2011; Petukhova et al., 2014; Bunt et al., 2016; Chowdhury et al., 2016; Ngo et al., 2017; Gilmartin et al., 2018). DialogBank is a language resource containing dialogues with gold standard annotations corresponding with the ISO 24617-2 standard (Bunt et al., 2016).

The development of annotation standards for particular corpora can be vividly exemplified by the case of the Switchboard Dialogue Act Corpus (the collection of telephone conversations). Telephone Speech Corpus (LDC97S62) was originally collected by Texas Instruments in 1990-1 and consists of approximately 260 hours of speech. The first release of the corpus was published in 1992-3[1]. Initially, the utterances in the corpus were annotated according to the DAMSL scheme for dialogue act analysis. Subsequently, NXT-format Switchboard Corpus was created with additional annotations according to an international standard ISO 64217-2:2012 (FANG et al., 2012). Conversion of one annotation system to another required matching of tags between them: DAMS consists of 59 combine tags while ISO – of 56 core tags. The re-annotation shows the significance of both standard scheme improvement and combining different standards on the same linguistic material.

Another work addressing the creation of corpora of dialogue acts concerns the DBOX corpus, created within the DBOX project and aimed at developing an interactive Question-Answering dialogue system (Petukhova et al., 2014). A more practical application of the project was to develop an

[1] https://catalog.ldc.upenn.edu/LDC97S62

Proceedings of the 29th International Conference on Computational Linguistics, pages 3631–3638
October 12–17, 2022.

interactive system used in computer games in three European languages. The authors collected 338 dialogues incorporating the continuous data collection method, i.e. they initially used the so-called Wizard-of-Oz paradigm with a human Wizard mirroring the system's behavior, and later replaced the Wizard with a complex dialogue system.

A similar approach was adopted by the authors of The ADELE Corpus of Dyadic Social Text Conversations (Gilmartin et al., 2018) who created a corpus consisting of 193 dialogues resumed with the purpose of initiating interactions with other people. Correspondingly to the DBOX corpus, the ADELE corpus was predominantly constructed with the view of training a spoken dialogue system that could easily engage in a conversation during a role-playing computer game. Both in the case of DBOX and ADELE, the obtained dialogues were manually annotated with dialogue act information in accordance with the ISO 24617-2 dialogue act annotation scheme, which was supplemented with additional dimension (for DBOX) as well as several additional dimension-specific functions and general-purpose functions (for both corpora).

Other related works which are worth mentioning include the Italian Luna Human-Human Corpus, which is a collection of 572 dialogues in the hardware/software helpdesk domain. The dialogues are conversations of the users engaged in problem solving tasks; a subset of 50 dialogues was annotated with the use of dialogues acts.

Furthermore, the DiaBiz.Kom corpus correlates with the DialogBank corpus, which is mentioned as the current golden annotation standard. Most dialogues from the DialogBank corpus were taken from other corpora and re-segmented and re-annotated. All annotations were double-checked for inconsistencies, errors and omissions. The data include samples which may be considered illustrative examples for annotations (Bunt et al., 2016). What is noteworthy here is the fact that suggestions and remarks with regard to limitations and extensions of the ISO standard put forth by the authors of the DialogBank are often subsequently implemented in the updated versions of ISO (Bunt et al., 2018)).

Another point of reference was the corpus of Vietnamese data using sources from IARPA Babel Vietnamese Language Pack (Ngo et al., 2018). The corpus includes 28 selected conversations whose transcripts were manually segmented in turns and then annotated. The agreement scores is 0.84 Fleiss'kappa measure.

In comparison to the previously collected corpora DiaBiz.Kom is much more extended in terms of the number of dialogues, and it covers different fields of communication. All the data were deliberately created to adhere to the research purposes and practical applications. As a consequence, DiaBiz.Kom could be considered the only corpus which is to be used in all main business communication fields. Also, especially in comparison to Switchboard Dialogue Act Corpus, the DiaBiz.Kom corpus uses much more up-to-date language materials. Over the last 30 years the languages have been vastly influenced by overwhelming technological development especially by social networks that have severely modified communication strategies and behaviours. The innovation of our approach is based mainly on the detailed consideration of the mutual influence of dialogue dimensions and communicative functions, as well as on the designation of the new functions not included in the previously used standards. Finally, DiaBiz.Kom was not only fully manually annotated, but also verified in the agreement procedure, which enhances the credibility of the corpus.

3 Data

DiaBiz.Kom corpus development is an annotation effort performed simultaneously with DiaBiz corpus creation (Pęzik et al., 2022). DiaBiz is a large, multi-modal corpus of Polish telephone conversations conducted in varied business settings, comprising 3,766 call center interactions from eight different domains, i.e. banking, energy services, telecommunication, insurance, medical care, debt collection, tourism and car rental. The phone-call interactions were based on 110 distinct customer service call scripts. They were then transcribed and enriched with punctuation. The selected dialogues from DiaBiz corpus are the basis for DiaBiz.Kom annotation.

4 Annotation Procedure

In the first place, the annotations included communicative functions and dimensions. The annotation process was divided into two main stages: (1) initial phase and (2) the final annotation of DiaBiz.Kom corpus. Both stages were performed by a team of qualified linguists with the use of the Inforex system (Marcińczuk et al., 2017).

Initially, the first version of the annotation guidelines was developed with an aim of achieving an appropriate level of inter-annotator agreement. In order to ensure high data quality, we have performed several iterations of manual annotation prior to the annotations performed on the final corpus. Three main sources were successively used as a dialogue base for manual annotation: LUNA corpus, samples of real-life data, and test sample from DiaBiz corpus. Moreover, the team of linguists was systematically expanded, so that we received feedback from annotators not involved in the early stages of guideline development. This was done to avoid a situation in which many of the rules of conduct were not verbalized, but rather were based on the annotator's practical experience. All these efforts aimed at making the guidelines as complete as possible. We calculated inter-annotator agreement by applying *Positive Specific Agreement* measure (Hripcsak and Rothschild, 2005). The first stage was continued until achieving the satisfactory level of the inter-annotator agreement, which involved 8 iterations of manual annotation.

The second stage (the final annotation of DiaBiz.Kom corpus) is currently underway. The inter-annotator agreement is constantly monitored and remains high. The figure presents the improvement of the average level of inter-annotator agreement. It is currently at the level of 0.78 (for annotation borders and categories) and 0.86 (for annotation borders). Every dialogue included in DiaBiz.Kom corpus is annotated by 3 specialists: 2 independently working annotators and a super-annotator who resolves all annotation inconsistencies (for current number of annotations see Appendix A, Table 3).

During the two stages of annotation, we used essentially the same annotation categories (i.e., those specified in the ISO 24617-2 standard). The main difference between the two stages was that during the first annotation stage we annotated a greater variety of texts, coming from diverse sources but greatly resembling the target texts which were later annotated at the second stage. Dividing the processes into stages allowed us to test the model in a variety of domains. Thanks to this solution, we did not adjust the guidelines to data acquired or produced in one specific way. Furthermore, during the first annotation stage the agreement level between annotators was not particularly high. In order to improve the inter-annotator agreement, we decided to work on texts coming from other sources than the target corpus. As a result, the DiaBiz.Kom annotation was quite consistent from the very beginning, and the need for corrections for the first iterations was significantly reduced (the second phase consisted of five iterations). Once the annotation was established (in joint discussions of professional linguists), the super-annotators returned to the previously annotated documents. The correctness of the texts was additionally verified at the dimension marking stage, and in the future – it will also be double-checked at the relation marking stage. Consequently, the material will be verified several times with a small chance of guidelines misinterpretations.

5 Key assumptions and limitations of the guidelines

Even though the annotation guidelines were constantly developed throughout the project, we decided to follow a set of certain unchanging assumptions. The increasing annotator agreement was the result of new specifications that were successively added to the guidelines. Importantly, we were persistently mindful of the versatility of the guidelines, which was primarily aimed at facilitating various possible applications of the corpus in the future. This approach, however, also imposed certain limitations on our work. Below we present the main assumptions as well as some selected issues which we encountered.

One of the main assumptions involved the choice of the communication function for a given utterance as primarily influenced by its goal, effect and the context in which it is set. The form of the annotated statement is considered less important – it may lead to the proper function, but it cannot fully determine its choice. The above mentioned situations may be illustrated with the following examples.

a) *Czy w czymś jeszcze mogę pani pomóc? ('Is there anything else I can help you with?')* [Interaction Structuring]

b) Agent: *Zna Pani swój numer klienta? ('Do you know what your client number is?')* [Propositional Question, dimension: Task]

Client: *Tak 'Yes.'* [Answer, dimension: Task]

Formally, the utterance in (a) points to be interpreted as Questions, but due to its conventionalized form and structuring role in the dialogue, it is marked as Interaction Structuring. Further, when there is a discrepancy between the intention and the effect (reaction), as illustrated in (b), we assign the

specific function on the basis of the direct reaction.

Expressions that can naturally perform different functions depending on the context (e.g. lexemes, such as dobrze 'well', tak 'yes') have been approached more thoroughly in our guidelines, which presently include specific contexts alongside with the plethora of examples illustrating their use in a given function. The goal that the sender wants to achieve is a key criterion here. If the interlocutor's utterance is aimed at obtaining or transmitting some information, it is assigned an appropriate function from the Information-transfer group, even if the form of this statement may initially indicate a function belonging to the Action-discussion group (c) and (d).

c) Proszę powiedzieć, na kogo zarejestrowany jest ten numer. ('Please tell me who this number is registered to.')

d) Proszę w pierwszej kolejności o imię i nazwisko. ('First of all, please give me your name and surname.')

The agent wants to obtain some information from the client, and the usage of the word proszę ('please') is only meant to make the question more polite.

What also needs to be emphasized is that due to the nature of the annotated dialogues, some of the functions described in the ISO/DIS 24617-2 standard were not used (e.g. functions from the Turn Management group), although they were included in the guidelines. As a result, it will be possible to apply them also to other types of dialogues in the future. The nature of the dialogues is related to the difficulty of the texts and this is also expressed by the degree of agreement between annotators (see Appendix A, Table 2). The annotation process showed that particular functions are performed in different ways depending on the type or theme of the dialogue. In the process of working on a given group of dialogues, a situation regularly occurred when certain detailed solutions were developed, which seemed to be completely inappropriate and inapplicable for the next set of texts. Over time, it has been noticed that this repeated situation is dictated by objective reasons. Below we will discuss two illustrative examples of such limitations. First, there are different schemes used for banking dialogues, different – for debt collection, sales, medicine, etc. A characteristic example may present the construction of a banking dialogue, in which the employee is obliged to verify the customer's identity at the beginning of the conversation by asking them a series of questions, the so-called TestQuestions (name and surname, PESEL number, customer number, mother's maiden name, etc.). This element does not occur, for example, in the debt collection dialogues: the client's identity is not strictly verified, as the employee knows who they are calling: most often they just ensure the data available to them are valid (e.g. in the form of PropositionalQuestion: Czy dodzwoniłem się do pani Anny Nowak? 'Have I reached Anna Nowak?' or CheckQuestion: Rozmawiam z panią Anną Nowak, tak? 'I am talking to Ms Anna Nowak, right?'). Second, the choice of a dialogue function is often determined by the relationship between the interlocutors: whether it is based on reciprocity ("equal with equals"), or rather hierarchical, and if hierarchical, who is superior and who is somewhat subordinate to the interlocutor? The following utterances can pose a very clear example: Bardzo proszę o rozłożenie mojej zaległości na raty. 'I would very much like to request that my arrears be spread out in installments' (the debtor is the sender) and Bardzo proszę o natychmiastowe uregulowanie zaległości na numer podany w mailu. 'I strongly request that you pay the arrears immediately to the number provided in the email.' (the debt collector is the sender). Despite the fact that both statements are built on the same syntactic structure, in the former case we are dealing with a Request, while in the latter – with an Instruct (understood as a command).

6 Corpus overview

The aim is to develop a well balanced corpus of annotated dialogues. Thus, we decided to annotate 10 dialogues for each script. As a result DiaBiz.Kom corpus will consist of 1100 annotated dialogues: 260 for banking domain, 150 for energy services, 180 for telecommunication, 110 for insurance, 140 for medical care, 100 for debt collection, 100 for tourism and 60 for car rental. The annotation process continues. All the dialogues (for current statistics see the Table 1) are annotated with communicative functions. The Inforex system enables to export the data using various formats (xml, json, conll or txt). The corpus sample is available under CC BY-NC-ND 4.0 license at: http://hdl.handle.net/11321/886.

There are 138.968 annotated functional segments within DiaBiz.Kom at this stage (see Appendix A, Table 3). The annotations distribution results from the nature of the dialogues. Some

Domain	Dialogues	Tokens
Banking	264	327.731
Debt collection	100	109.189
Energy services	150	131.698
Insurance	110	116.151
Medical care	140	145.765
Car rental	60	71.265
Telecom-munications	180	157.701
Tourism	100	218.465
All	1.104	1.277.965

Table 1: Current size of DiaBiz.Kom annotation in 2+1 system. The numbers refer to the dialogues with final annotation.

communicative functions appear less frequently, e.g. Turn Management functions. We actually recorded few such cases where the annotation of the functions within this group was obligatory. The limited number of such situations may have resulted from the fact that we annotated only those segments whose primary function was to manage dialogue turns. Such an approach was determined by the implicit nature of Turn Management functions (e.g., according to ISO 24617-2: "every time someone starts speaking, this can be interpreted as the performance of a turn-taking act; every time someone stops speaking, this can be interpreted as a turn-release act"). Implied functions were not annotated manually. That is, we did not annotate Turn Management functions in the situations where the speaker, for instance, communicated that they were ready to continue the dialogue (the function we used in such situations was Contact Indication as its definition was more extensive), e.g.

Agent: *Dzień dobry. ('Good morning.')* [initGreeting, dimension: Social Obligations Management]

Client: *Tak, słucham.'Yes, I'm listening.'* [contactIndication, dimension: Contact Management]

All these decisions were preceded by a number of joint discussions of professional linguists over specifically extracted samples from the target corpus (i.e., the examples that had the potential to fall into the category of Turn Management functions). Also, it is significant to mention that Turn Management functions are more natural to polylogues and the annotated corpus consisted solely of dialogues.

7 Conclusions and future work

In this paper we have outlined DiaBiz.Kom – the first corpus, which contains dialogues of various domains with gold standard dialogue act annotations in the Polish language to satisfy the criteria set by machine learning applications. A crucial feature of this resource is the manual layer annotation of information about communication functions (based on ISO standard). The achieved inter-annotator agreement provides a way to use the corpus for the purpose of machine learning. Further development work on DiaBiz.Kom will aim at adding annotation layers – especially those that specify the communicative intent of the speaker (using frame semantics) – and, subsequently, those that determine parameters congruent with the ISO standard (communicative dimensions and relations between annotations). The next step consists in expanding the existing corpus with supplementary dialogues using active learning techniques.

Acknowledgements

The DiaBiz.Kom corpus was developed in the project entitled "CLARIN - Common Language Resources and Technology Infrastructure", which is financed under the 2014-2020 Smart Growth Operational Programme, POIR.04.02.00-00C002/19. We would also like to thank the VoiceLab[2] company for their consultations.

References

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Harry Bunt, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Fang. 2016. The DialogBank. In *Proceedings of the Tenth International Conference*

[2] https://voicelab.ai

on Language Resources and Evaluation (LREC'16), pages 3151–3158, Portorož, Slovenia. European Language Resources Association (ELRA).

Harry Bunt, James Pustejovsky, and Kiyong Lee. 2018. Towards an ISO standard for the annotation of quantification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shammur Absar Chowdhury, Evgeny Stepanov, and Giuseppe Riccardi. 2016. Transfer of corpus-specific dialogue act annotation to ISO standard: Is it worth it? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 132–135, Portorož, Slovenia. European Language Resources Association (ELRA).

Chengyu FANG, Jing CAO, Harry BUNT, and Xiaoyue LIU. 2012. The annotation of the switchboard corpus with the new iso standard for dialogue act analysis. In *Proceedings of the Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation ; Conference date: 03-10-2012 Through 05-10-2012.

Emer Gilmartin, Christian Saam, Brendan Spillane, Maria O'Reilly, Ketong Su, Arturo Calvo, Loredana Cerrato, Killian Levacher, Nick Campbell, and Vincent Wade. 2018. The ADELE corpus of dyadic social text conversations:dialog act annotation with ISO 24617-2. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Simon Keizer, Harry Bunt, and Volha Petukhova. 2011. *Multidimensional Dialogue Management*, pages 57–86. Springer Berlin Heidelberg, Berlin, Heidelberg.

Michał Marcińczuk, Marcin Oleksy, and Jan Kocoń. 2017. Inforex — a collaborative system for text corpora annotation and analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 473–482, Varna, Bulgaria. INCOMA Ltd.

Thi-Lan Ngo, Pham Khac Linh, and Hideaki Takeda. 2018. A Vietnamese dialog act corpus based on ISO 24617-2 standard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Thi-Lan Ngo, Son-Bao Pham, Khac-Linh Pham, Xuan-Hieu Phan, and Minh-Son Cao. 2017. Dialogue act segmentation for vietnamese human-human conversational texts. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 203–208.

Volha Petukhova, Martin Gropp, Dietrich Klakow, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz, Steffen Liersch, and Anna Schmidt. 2014. The DBOX corpus collection of spoken human-human and human-machine dialogues. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 252–258, Reykjavik, Iceland. European Language Resources Association (ELRA).

Piotr Pęzik, Gosia Krawentek, Sylwia Karasińska, Paweł Wilk, Paulina Rybińska, Anna Cichosz, Angelika Peljak-Łapińska, Mikołaj Deckert, and Michał Adamczyk. 2022. DiaBiz – an annotated corpus of polish call center dialogs. In *Proceedings of the Language Resources and Evaluation Conference*, pages 723–726, Marseille, France. European Language Resources Association.

A Appendix. Tables

	Debt collection	Insurance	Medical Care	Car rental	Telcom
Information-seeking					
setQuestion	0.73 (68)	0.70 (99)	0.82 (115)	0.64 (47)	0.70 (55)
checkQuestion	0.61 (32)	0.35 (33)	0.74 (47)	0.71 (28)	0.35 (12)
choiceQuestion	0.64 (12)	0.67 (15)	0.87 (52)	0.50 (5)	0.67 (16)
propositionalQuestion	0.73 (76)	0.66 (57)	0.70 (77)	0.83 (101)	0.66 (120)
Information-providing					
inform	0.58 (477)	0.57 (489)	0.61 (477)	0.61 (394)	0.57 (367)
answer	0.73 (155)	0.67 (197)	0.75 (239)	0.71 (167)	0.67 (155)
confirm	0.72 (30)	0.44 (27)	0.71 (44)	0.69 (18)	0.44 (11)
Directives					
request	0.28 (17)	0.67 (17)	0.65 (49)	0.57 (18)	0.67 (26)
suggest	0.31 (29)	0.58 (18)	0.17 (16)	0.36 (11)	0.58 (25)
acceptOffer	0.67 (1)	1.00 (5)	0.25 (7)	1.00 (1)	1.00 (4)
Commisives					
offer	0.12 (8)	0.38 (12)	0.10 (13)	0.50 (1)	0.38 (23)
acceptRequest	0.00 (7)	0.53 (7)	0.59 (16)	0.32 (12)	0.53 (11)
Discourse Structuring					
interactionStructuring	0.68 (161)	0.69 (247)	0.62 (253)	0.56 (163)	0.69 (150)
Feedback					
autoPositive	0.72 (184)	0.78 (261)	0.71 (305)	0.66 (169)	0.78 (17)
alloPositive	0.57 (9)	0.50 (19)	0.60 (13)	0.86 (3)	0.50 (7)
all categories	0.74 (2827)	0.73 (2909)	0.75 (2999)	0.74 (2480)	0.73 (2107)

Table 2: Inter-annotator agreement (*PSA*) for selected communicative functions regarding 5 domains. The agreement is based on annotations of the same two annotators performed on 20 dialogues within each domain. The number in the brackets corresponds to the number of final annotations submitted by the independent super-annotator.

Communicative function	Number of annotations	Communicative function	Number of annotations
Information-seeking functions		**Contact Management Functions**	
checkQuestion	2.158	contactCheck	123
choiceQuestion	959	contactIndication	3.284
propositionalQuestion	4.063	**Discourse-structuring functions**	
setQuestion	4.223	interactionStructuring	14.195
testQuestion	1.201	opening	1.093
Information-providing functions		**Feedback functions**	
agreement	886	alloNegative	11
answer	9.741	alloPositive	517
confirm	1.680	autoNegative	73
correction	14	autoPositive	12.892
disagreement	102	feedbackElicitation	462
disconfirm	71	**Own- and Partner-Management Functions**	
inform	24.090	completion	27
Directives		correctMisspeaking	12
acceptOffer	309	retraction	189
addressOffer	9	selfCorrection	4.504
declineOffer	76	selfError	422
instruct	1.082	**Social Obligations Management Functions**	
request	1.558	acceptApology	35
suggest	1.138	acceptThanking	35
Commissives		apology	350
acceptRequest	619	compliment	7
acceptSuggest	211	congratulation	2
addressRequest	13	initGoodbye	1.151
addressSuggest	6	initGreeting	1.161
declineRequest	24	initSelfIntroduction	1.148
declineSuggest	35	returnGoodbye	1.261
offer	668	returnGreeting	955
promise	653	returnSelfIntroduction	364
		sympathyExpression	186
		thanking	3.185
		Time-management functions	
		pausing	410
		stalling	35.317
		Turn-management functions	
		turnAssign	2
		turnGrab	6
		All	**138.968**

Table 3: Statistics for the final annotations within the DiaBiz.Kom corpus (current state).