Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Corpus with Speech Function Annotation: Challenges, Advantages, and Limitations

Lidiia Ostyakova[♡,◇] l.ostyakova@yandex.ru
Maria Molchanova[◇] molchanova.ma.63@gmail.com
Ksenia Petukhova[♡,◇] kpetyxova@mail.ru
Nika Smilga[♡,◇] smilgaveronika@gmail.com
Daniel Kornev[◇] danielko@deeppavlov.ai
Mikhail Burtsev[◇,‡] burtcev.ms@mipt.ru

[◇]Moscow Institute of Physics and Technology, Moscow
[♡]Higher School of Economics, Moscow
[‡]AIR Institute, Moscow

Abstract

Creating a corpus labeled with dialog acts is one of the most difficult tasks in corpus linguistics. Dialog acts can reflect various aspects of the utterances in the dialogues, but most often represent the pragmatic intentions of the speaker or the features of the dialog discourse. The pragmatic and discourse functions of one utterance can be interpreted in different ways depending on an annotator. We used speech function theory that are similar to DA theories to annotate casual conversations and analyzed labeled data. This article is devoted to the first steps and challenges of creating a corpus with speech function annotation.
Keywords: corpus linguistics, pragmatics, discourse, dialogue act theory.
DOI: 10.28995/2075-7182-2022-21-1129-1139

Корпус с аннотацией речевыми функциями: вызовы, преимущества и ограничения

Лидия Остякова[♡,◇] l.ostyakova@yandex.ru
Мария Молчанова[◇] molchanova.ma.63@gmail.com
Ксения Петухова[◇,‡] kpetyxova@mail.ru
Ника Смилга[♡,◇] smilgaveronika@gmail.com
Данила Корнев[◇] danielko@deeppavlov.ai
Михаил Бурцев[◇,‡] burtcev.ms@mipt.ru

[◇]Московский физико-технический институт, Долгопрудный
[♡]НИУ "Высшая школа экономики Москва
[‡]Институт искусственного интеллекта AIRI, Москва

Аннотация

Создание корпуса, размеченного диалоговыми актами, является одной из самых сложных задач в корпусной лингвистике. Диалоговые акты могут отражать различные аспекты высказываний в диалогах, но чаще всего отражают прагматические намерения говорящего или особенности диалогового дискурса. Прагматические и дискурсивные функции одного высказывания могут интерпретироваться по-разному в зависимости от особенностей восприятия каждого аннотатора. В данной статье мы использовали теорию речевых функций, которая имеет схожие черты с различными теориями диалоговых актов, для разметки диалогов на повседневные темы и проанализировали полученные данные. Эта статья посвящена первым шагам и вызовам на пути к созданию корпуса с аннотацией речевой функции.
Ключевые слова: корпусная лингвистика, прагматика, дискурс, теория диалоговых актов.

1 Introduction

Dialogue act (DA) tagging is an automatic analysis method that has many applications in computational linguistics: in modern human-machine dialogue systems (chat-bots), machine translation systems, and speech recognition. Dialogue act represents a communicative function of an utterance in the dialogue or

an abstract intent. As this analysis is commonly used in NLU tasks, there were many attempts to create tools for automatic DA tagging and corpora with relevant annotation based on different theories.

Due to a small amount of labeled data and the inconsistency of existing tagsets, there are still problems applying DA analysis in open-domain dialogue systems. To create a chat-bot that can support casual conversations with users, a tagset should provide multi-layer analysis considering different features of dialogues such as topic organization, discourse patterns, pragmatics, feedback, etc. As the most often used tagsets such as SWBD-DAMSL (Jurafsky et al., 1997), DiAML (Mezza et al., 2018), MIDAS (Yu and Yu, 2019) have different limitations that prevent them from being used in modern open-domain systems, we focused another approach to DA analysis based of the research of casual conversations. S.Eggins and D.Slade introduced speech function theory, an approach to DA analysis that provides a comprehensive, systematic discourse model of dialogues. (Eggins and Slade, 2004). Speech functions are similar to dialogue acts and represent abstract intentions in casual conversations.

(Mattar and Wachsmuth, 2012) was the first one who implemented speech function annotation in the dialogue system. Although a chat bot was task-oriented, this experiment demonstrated the potential of using speech function theory. In our early efforts to use speech functions as an aid in strategic dialogue management, we made an attempt to create the first open-source corpus to build a speech function classifier for DREAM socialbot (Baymurzina et al., 2021) during our participation in Amazon Alexa Prize 4. The classifier aimed at automatic analysis of the dialogues during human-machine interaction, while the speech function predictor provided a list of next logically and statistically meaningful moves. At the end of the challenge, we used enhanced versions of these components to build a recommendation system for a VS Code extension to aid scenario-driven dialogue designers in determining next steps in the dialogue (Kuznetsov et al., 2021). Nevertheless, the amount of labeled data wasn't enough to develop a state-of-the-art classifier.

Considering all our previous experience on this topic, we decided to make a second attempt to create a corpus with speech function annotation for the extensive analysis of casual conversations. So, this article is devoted to the first steps and challenges of creating a collection of gold standard dialogues with such an annotation.

2 Related Work

The term **dialogue act** has got many interpretations due to the variety of DA theories and their different applications. The most common definition of a dialogue act is a communicative action that an utterance performs in the dialogue. In computational linguistics, dialogue acts have an additional interpretation and serve as abstract intentions that are important for building systems that can analyze and predict the communicative behavior of users. According to A.Popescu-Belis, all the DA annotation schemes can be differed by such key parameters as (Popescu-Belis, 2005):

- **A number of dimensions**

  A number of dimensions indicate the complexity of the DA annotation scheme. Existing methods of analyzing dialogue interaction include various dimensions describing the pragmatic intentions of interlocutors; discourse organization of the dialogue; social obligations between speakers; auto-correction; etc. Mostly, modern approaches to DA tagging (DiAML (Mezza et al., 2018), MIDAS (Yu and Yu, 2019)) are multi-layer because more information for an extensive analysis of dialogues is required. Nonetheless, a large number of dimensions leads to uneven annotation and representation issues, which should be taken into account when using schemes in NLU tasks.

- **Dialogue domain**

  Depending on the task that the system should complete, tagsets can include special labels that are not appropriate for other tasks. For example, the HCRC Maptask Coding Scheme was designed for analyzing communication between people guessing the place on the map during dialogue games. The scheme includes a tag READY, described as a move that occurs after the close of a dialogue game and prepares the conversation for a new game to be initiated (Anderson et al., 1991). The tag READY can't be applied in other cases, as dialogue games

are not typical for all conversations. With the development of dialogue systems, there were several attempts to create annotation schemes for analyzing casual conversation (e.g., MIDAS (Yu and Yu, 2019)). Besides that, the ISO standard for DA annotation was designed as a domain-free scheme (Mezza et al., 2018).

- **Level of segmentation**

  The segmentation type in the DA annotation schemes is dependent on the purpose for which it was created. Utterance segmentation is used for such tasks as automatic speech recognition, taking into account dialogue context and prosodic features, while in other cases, only sentence segmentation is possible. Sometimes, DA annotation schemes represent several segmentation levels for researching the complex structure of casual conversations. For instance, some dimensions in the ISO scheme describe particular segments: Turn Management is used for turn-level annotation; Time Management denotes pauses or stalling in the dialogue; etc. (Mezza et al., 2018)

Due to all these differences, existing corpora are not consistent with each other. Moreover, there is a small amount of labeled casual conversations for training automatic DA taggers to make an extensive analysis. Working on speech function annotation, we orient on the most often used DA theories for tagging conversations within open-domain dialogues systems (see Table 1).

Annotation Scheme	Dimensions	Classes	Dialogue Domain	Segmentation Level
MIDAS	2	26	casual conversation	utterance level
DiAML	7	49	domain free	several levels
SWBD-DAMSL	7	42	casual conversation	utterance level
Speech Functions	5	33	casual conversation	several levels

Table 1: Comparison of DA Tagsets

SWBD-DAMSL This annotation scheme is extensive and includes 42 classes. All tags are divided into 7 dimensions depending on their functionality (e.g., Task Management, Self and Other-talk). This annotation scheme is not hierarchical and doesn't reflect differences between tag dimensions. The utterances can be annotated with only one tag. SWBD-DAMSL was originally created for analyzing casual conversations for the improvement of automatic speech recognition, but it is also used in task-oriented dialogue systems. Switchboard Dialogue Act Corpus that contains more than 1000 recordings of telephone calls is labeled according to the SWBD-DAMSL annotation scheme (Jurafsky et al., 1997).

DiAML Dialogue Annotation Mark-up Language is an ISO standard that was created as a domain free DA annotation scheme. There are several dimensions that include overall 49 dialogue acts: Time Management, Discourse Structure, Auto-Feedback, Partner Communication Management, Social Obligations, Own Communication Management, Turn Management, Task (a communicative act). Despite the fact that this scheme provides detailed analysis of the dialogues, tags are not so suitable for representation in NLU challenges due to their multi-functionality. There is not so much data labeled with DiAML tags but there were attempts to compare this annotation scheme to the others (Mezza et al., 2018).

MIDAS scheme It was created specifically for the analysis of human-machine interaction. Tags are inherited from ISO standard and SWBD-DAMSL schemes, but their number has been reduced to 26 compared to previous tagsets. MIDAS is a hierarchical structure and includes two dimensions reflecting functional (e.g., greeting, thanks) and semantic queries (e.g., yes/no questions) that can be divided into minor classes of tags. However, MIDAS hierarchy is not reflected in the tagsets. Although this method of dialog analysis supports multiple labeling, some utterances may have only one tag (Yu and Yu, 2019). There is available data containing MIDAS labels, and an open-source algorithm for predicting MIDAS labels can be used.

Utterance	DiAML	SWBD-DAMSL	MIDAS
A: Okay.	DS:opening	o (other)	other
B: All right.	AutoF:autoPositive	o (other)	other
B: Uh...	TiM:stalling, TuM:turnKeep	qy (yes/no question)	yes/no question
B: Do you have any friends that have children?	Ta:propositionalQuestion	qy (yes/no question)	yes/no question
A: I do have friends that have children.	Ta:answer	na (affirmative non-yes answer)	positive answer
A: Yes.	Ta:answer	ny (yes answer)	positive answer

Table 2: Comparison of Approaches to DA Annotation

When the three approaches to DA annotation are compared, it is clear that the schemes differ in terms of segmentation level (see Table 2). While the second and third utterances are considered to be separate segmentation units in DiAML annotation, they are treated as a single utterance in MIDAS and SWBD-DAMSL. The MIDAS tagset is the easiest to comprehend because there are no abbreviations in comparison to other schemes. MIDAS annotation reflects several dimensions in the same manner as SWDA-DAMSL does, but DiAML highlights specific dimensions and provides a thorough analysis of the dialogue that makes it the most representative one and suitable for annotation of open-domain dialogues.

3 Speech Functions

Speech functions, the same as dialogue acts, are presented as a tool for annotation of the communicative intentions of interlocutors in casual conversations. S.Eggins and D.Slade in (Eggins and Slade, 2004) introduced an approach connecting dialogue turns and cross-dialogue discourse structure patterns that are specific for casual conversation as the higher-level abstraction. At the level of turns, S.Eggins and D.Slade extended M.K. Halliday's concept of speech functions that express pragmatic goals of speakers and can be used as an enhanced alternative to dialogue acts (Halliday et al., 2014). They introduced a taxonomy including 45 speech functions that was developed on the basis of the casual conversation research. As the taxonomy was created specially for describing the structure of daily dialogues, it's appropriate to use the scheme for open-domain chat-bots. Moreover, speech functions taxonomy is a multi-layer hierarchical annotation scheme in which all the tags are mutually exclusive. Each speech function consists of several layers representing different dimensions: Turn Management, Discourse Structure, Topic Organisation, Feedback, Communicative Act, or Pragmatic Purpose (see Figure 1).

Figure 1: Distribution of Dimensions in Speech Function

There are three high-level types of discourse moves in the taxonomy:

1. **Opening moves**

   Opening moves are used for introducing new topics into dialogues or starting a conversation. This group includes six speech functions: Open.Attend, Open.Command, and four distinct Open.Initiate functions (see Table 3). In the Open.Initiate group, speech functions define a specific topic that is realized in utterances by demanding or providing factual and evaluative information. Opening moves are significant for defining a level of Discourse Structure in the dialogue. According to S.Eggins and D.Slade, every Opening move indicates not only another topic or the beginning of interaction between interlocutors within a conversation but also a

discourse pattern (see Figure 2). Depending on the number of Opening moves, a dialogue can include one or more discourse patterns (Eggins and Slade, 2004).

2. **Sustaining moves**
   Sustaining moves, which include four speech functions, are used to supplement a current topic with details and clarifications provided by the same speaker. They don't contribute to the development of the topic, but rather enrich the information discussed within it while the speaker's role is not delegated to another interlocutor.
3. **Moves of Reaction**
   Moves of Reaction are dialogue turns in which a speaker changes or a response to the interlocutor's previous utterance occurs. These moves are arranged in a more complex way than the others since they include many layers. They are divided into two groups to represent different approaches to topic development. React.Respond group of speech functions leads a conversation to its completion as they do not contribute to the appearance of new challenges (e.g., questions changing conversational flow) throughout the dialogue, whereas React.Rejoinder group, in contrast to the first one, aids in the discussion development. The following layer of Reaction moves is assigned to the Feedback level and indicates whether one interlocutor supports or opposes the other (Support and Confront dimensions in the speech function tags).

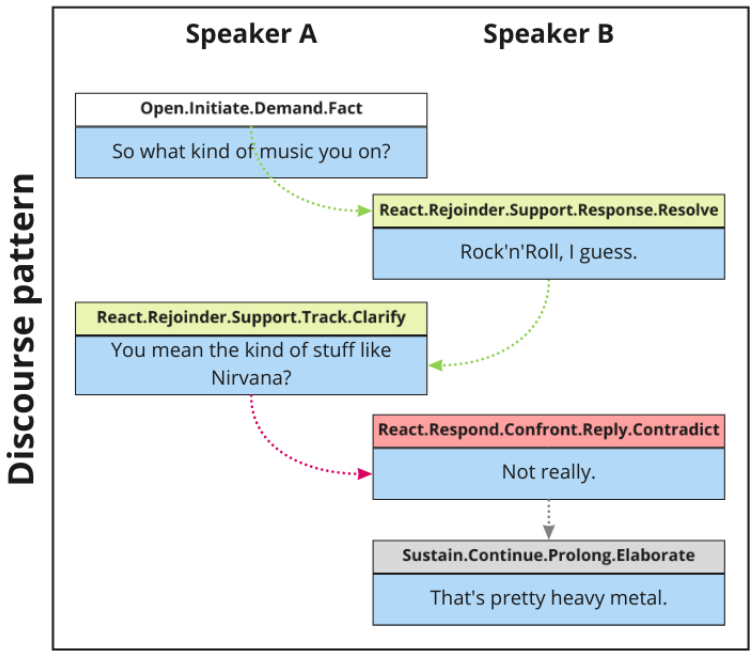

Figure 2: Discourse Pattern. Green pointers are used for moves that add to the topic's development, whereas red pointers are used for those that do not. Sustain.Continue moves are denoted by grey pointers. The colours of the frames represent different interlocutors' feedback: red - disagreement, green - support, grey - neutral.

The advantage of speech functions is that they provide a suitable representation with only one tag for a particular utterance, expressing several annotation layers. This method of DA analysis provides the ability to constrain a comprehensive, systematic discourse model of dialogues. Despite the fact that the distribution of dimensions in speech function tags is uneven, the taxonomy quite fully reflects the structure of dialogues at the level of topics and discourse and the abstract intentions of speakers in particular utterances. Compared to other DA taxonomies (see Table 1), the speech function annotation scheme has grammatical criteria for tag identification but doesn't feature them in the tags.

4 Creating Gold Standard for Speech Function Annotation

DA annotation is considered to be one of the most challenging tasks in corpus linguistics because one utterance in the dialogue can be defined differently by several annotators due to their own experience and understanding of the dialogue context. Therefore, our primary goal at this point was to check whether it is possible to reach high inter-annotator agreement when labeling dialogues with speech functions. As the speech functions taxonomy includes pragmatic and discourse features that appear to be too abstract for recognition, there was a need to develop comprehensive guidelines for annotators specifying semantic and grammatical features of utterances and providing proper examples for each tag. The created guidelines were then verified, first in terms of design usability and then in terms of the possibility of achieving high inter-annotator agreement. We provided an extensive analysis of the most common mistakes made in the annotation, denoted most problematic groups of speech functions to improve our future experiments with the corpus.

4.1 Preparatory stages

There were several stages prior to working on guidelines design, such as 1) revising the taxonomy of speech functions; 2) choosing a corpus for annotation and data preprocessing.

4.1.1 Revising Taxonomy of Speech Functions

First, the entire taxonomy had to be revised because S.Eggins and D.Slade had created a tagset of 45 speech functions for analyzing human casual conversations. Their study was based on transcriptions of students' daily conversations in the cafe (Eggins and Slade, 2004). The tagset was reduced to 32 labels considering the specificity of human-machine interaction (see Table 3). For example, to avoid redundancy, we excluded a group of three Sustain.Continue.Append tags that have functionality similar to Sustain.Continue.Prolong ones but are used only for the description of noun groups. S.Eggins and D.Slade also supplied information on the different types of questions in the Open.Initiate.Demand group of labels: Closed (yes/no questions) and Open (wh-questions) (Eggins and Slade, 2004). We opted not to include this annotation level in our taxonomy since it represents grammatical aspects of sentences and does not contribute to pragmatics of the utterance.

4.1.2 Data for Speech Function Annotation

As we build a corpus for further automatic speech function classification within open-domain dialogue systems, it is critical to consider the type of interaction (human or human-machine) represented in the data as well as the domain of the dialogues, i.e., whether conversations are casual or task-oriented. Thus, human-machine dialogues with daily topics discussed would be preferable for the annotation with speech functions. However, available datasets are comprised of relatively short and simple human-machine dialogues that might limit the benefits of the future corpus in terms of the discourse structure research and the diversity of speech functions.

The first attempt to label casual conversations with speech functions was made on the basis of Santa Barbara Corpus of Spoken American English, which consists of 60 transcriptions of the naturally-occurring spoken dialogues (Kuznetsov et al., 2021). Three face-to-face dialogues were preprocessed and then labeled with speech functions, resulting in a small dataset with approximately 1700 manually annotated utterances [1]. Two annotators reached an inter-annotator agreement of kappa = 0.71 on 1200 utterances which is considered to be a good result. Nevertheless, that is not enough for building a sufficient automatic speech function classifier. Despite promising inter-annotated agreement and a wide range of speech functions represented in the dialogues, Santa Barbara corpus is not suitable for enlarging because each dialog contains between 300 and 900 utterances. This interferes with the annotator's comprehension of the dialogue's context and is also incomparable to dialogues within human-machine interaction.

Taking into account all of the mistakes made during our first attempt to create a dataset, DailyDialog corpus was chosen for further expanding speech function labeling (Li et al., 2017). It contains over 13 000 human dialogues of varying lengths and has already been annotated with the MIDAS scheme.

[1] https://github.com/lostyakova/speech_function/blob/main/labeled_data.json

Speech Function	**Communicative Role**
Open.Attend	attention seeking
Open.Initiate.Demand.Fact	demand factual information
Open.Initiate.Demand.Opinion	demand evaluative information
Open.Initiate.Give.Fact	give factual information
Open.Initiate.Give.Opinion	give evaluative information
Open.Command	make a request, an invitation or command
Sustain.Continue.Prolong.Extend	offer additional or contrasting information
Sustain.Continue.Prolong.Elaborate	clarify and restate
Sustain.Continue.Prolong.Enhance	qualify previous move by giving details
Sustain.Continue.Monitor	check that audience is still engaged
React.Rejoinder.Confront.Challenge.Counter	dismiss addressee's right to his/her position
React.Rejoinder.Confront.Response.Re-challenge	question relevance of a prior move
React.Rejoinder.Support.Challenge.Rebound	dismiss addressee's right to his/her position
React.Rejoinder.Support.Response.Resolve	provide clarification
React.Rejoinder.Support.Track.Check	elicit repetition of a misheard element
React.Rejoinder.Support.Track.Clarify	verify information heard
React.Rejoinder.Support.Track.Confirm	confirm information heard
React.Rejoinder.Support.Track.Probe	volunteer further details
React.Respond.Confront.Disengage	show unwillingness to interact
React.Respond.Confront.Reply.Contradict	negate prior information
React.Respond.Confront.Reply.Disagree	provide negative respond to question
React.Respond.Confront.Reply.Disawow	deny acknowledgement of information
React.Respond.Support.Develop.Elaborate	clarify and restate a prior move
React.Respond.Support.Develop.Enhance	qualify previous move by giving details
React.Respond.Support.Develop.Extend	offer additional or contrasting information
React.Respond.Support.Engage	show willingness to interact
React.Respond.Support.Register	display attention to the speaker
React.Respond.Support.Reply.Acknowledge	indicate knowledge of information given
React.Respond.Support.Reply.Affirm	provide positive response to the question
React.Respond.Support.Reply.Accept	accept the offered goods or services
React.Respond.Support.Develop.Enhance	deny acknowledgement of information
React.Respond.Support.Reply.Agree	indicate support of information given

Table 3: Speech functions and their communicative roles in the dialogue

Mostly, conversations in the DialyDialog corpus are casual, but sometimes task-oriented dialogues can be found. The corpus needed preprocessing before the annotation stage as there were misprints and duplicates in the data. We annotated 30 dialogues from the DailyDialog corpus with an average length of 13 utterances to produce a collection of gold-standard data and assess inter-annotator agreement while improving guidelines.

4.2 Design of Guidelines

Any DA annotation is quite complicated due to the complex structure of taxonomies with multiple dimensions, the use of abbreviations, and label inconsistency. To be able to annotate dialogues with DA tags, an annotator must be involved in the research of a certain taxonomy and its tags that's a time-consuming process. So, our objective was to develop guidelines in such a way that an annotator could easily navigate through 32 speech functions, each of which should have an extensive description and relevant examples from casual conversations.

The quality of DA-annotated corpora is typically measured using Fleiss' kappa, which is calculated

over a group of multiple annotators (Hoek and Scholman, 2017). Fleiss' kappa only indicates differences in labeling between annotators, not their errors. When the inter-annotation agreement is close to one, it is considered to be almost perfect. Due to the subjective nature of DA annotation, the average inter-annotator agreement across different DA corpora is 0.75. As guidelines for annotation with speech function should help to avoid or minimize disagreements among annotators, we used Fleiss' kappa to evaluate the applicability of different ways to present needed information about tags to annotators. Inter-annotation agreement of full (e.g., Sustain.Continue.Prolong.Extend) and cut labels (e.g., Sustain.Continue.Prolong) labels was compared during four stages of guidelines development to analyze a level of dimensions in speech functions where annotators disagreed with each other (see Figure 3).

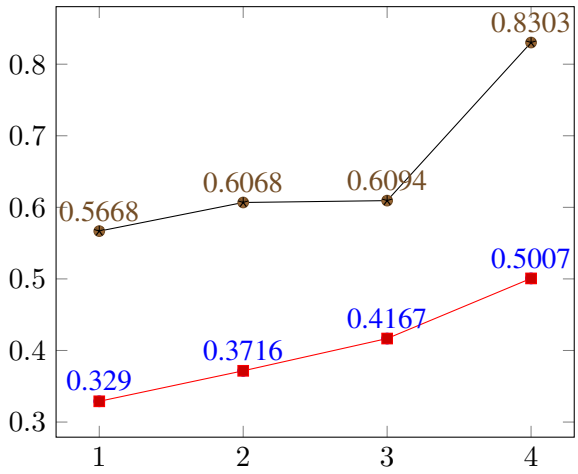

Figure 3: Fleiss' kappa for cut and full Speech Function labels during developing guidelines

There were four stages of improvement for the speech function annotation guidelines. Five annotators were involved in the experiment, three of whom were professional linguists. **In the first step**, annotators used "Analyzing casual conversations" as an annotation instruction (Eggins and Slade, 2004). A detailed description of speech functions is presented there in several tables with examples from casual conversations, considering the type of discourse move. Although there were some modifications and clarifications for certain speech functions made, the source of instructions remained the same **at the second stage**. The inter-annotator agreement on the final two iterations demonstrates the outcome of two different methods for guidelines design. **The third version** of the guidelines was designed in the form of cards. This method could be used by annotators to select tags based on grammatical aspects of utterances. There were, for example, cards describing only the speech functions that are more often performed by questions. Such guidelines sped up the annotation process but did not result in higher inter-annotator agreement.

At the fourth stage, the final version of the instruction was created as a graph, which considerably increased inter-annotator agreement. Nodes in the graph contain simple questions about utterances in dialogues that assist an annotator in selecting the most appropriate speech function (see Figure 4). Such a representative method makes it easier for the annotator to navigate through all speech functions and increases the pace of the labeling process. Then, for each utterance-related question, examples from the DailyDialog dialogues were provided, allowing non-professional annotators to participate in the tagging process on the crowdsourcing platform. Using this version of the guidelines, 25 casual conversations from the DailyDialog corpus were annotated as gold standard dialogues.

4.3 Analysis of labeled Data

As inter-annotator agreement doesn't point out mistakes of annotators that need to be defined for the further work on guidelines, we analysed the most problematic groups of speech functions that are hard to recognize. Inter-annotator agreement for cut labels is 0.83 that is considered to be a good result while Fleiss' kappa for full labels is still not so high(see Figure 3). This means that pragmatic purposes

in speech functions are difficult to distinct from each other. Taking into account distribution of labels across labeled conversations (see Figure 5) for the future gold standard, the most frequent and crucial for discourse analysis tag groups were analyzed in detail. It is worth noting that the distribution of label groups is comparable to the first corpus annotated with speech functions.

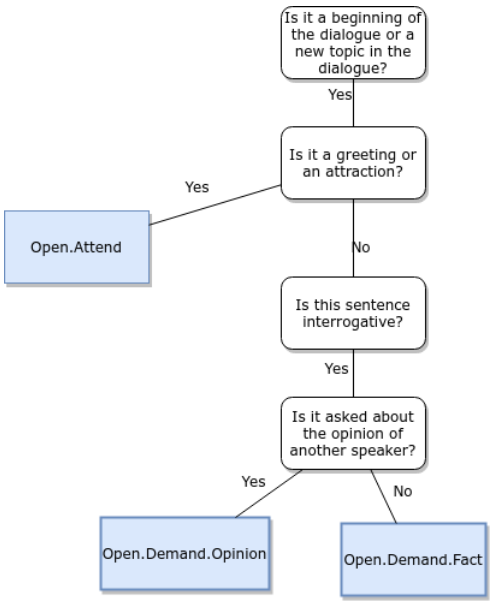

Figure 4: Part of the Instruction for Opening Moves at the 4th stage

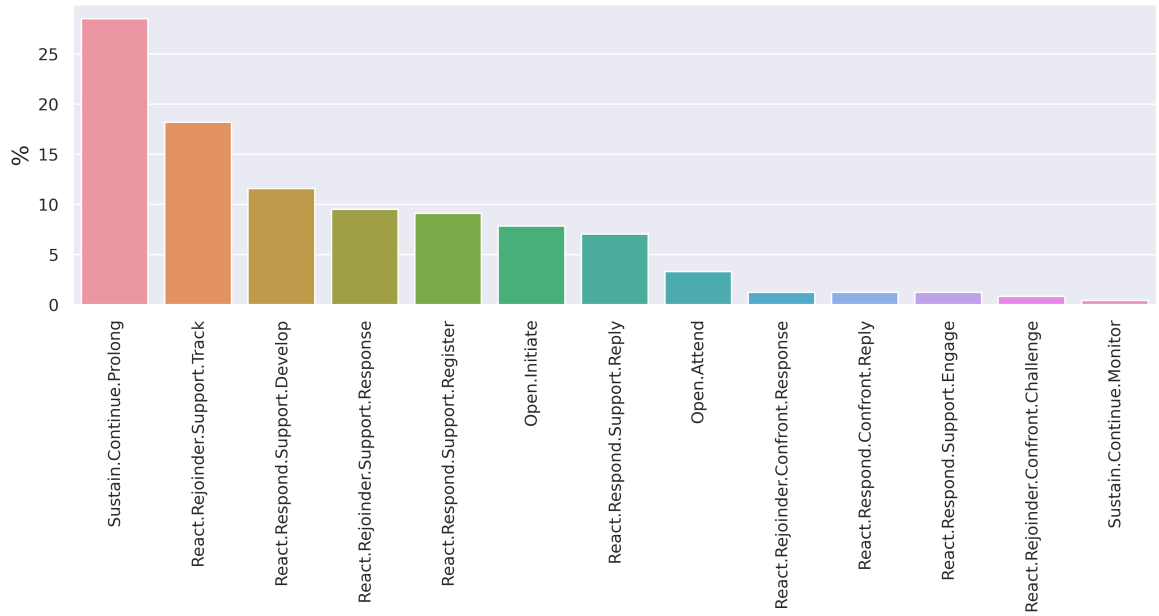

Figure 5: Distribution of cut labels across annotated data

Opening moves Although the number of Opening moves does not prevail in the dialogues, this group of speech functions is essential for proper conversation analysis. The main role of opening moves is to determine the borders of discourse patterns, which are considered high-level components of the dialogue following speech function theory. Even expert annotators may struggle to recognize a topic change in

casual conversations due to the unpredictability of the discourse flow and different understanding of topics in the dialogue.

According to the labeled data, 64 % of Opening tags are used by annotators incorrectly. In the example (1), *"I just happen to have a question for you guys."* is an utterance where a topic change happens but annotators choose other labels. In addition to this, annotators often can't define whether the utterance can be referred to factual or evaluative information. Two annotators labeled this utterance *"You seem to be in a hurry."* as Open.Initiate.Give.Fact although the word *seem* is evaluative and indicates that this is an opinion. In our experiment, human accuracy was ≈ 0.7 while opinion and fact classification models can perform this task better with the state-of-the-art accuracy of more than 0.95 %.

(1)

Speaker1: Oh , don't let that worry you.
Speaker1: If that were true , China wouldn't have such a large population.
Speaker2: *I just happen to have a question for you guys.*
Speaker2: Why do you cook the vegetables ?

(2)

Speaker1: I had a good time.
Speaker1: *You seem to be in a hurry.*
Speaker1: Don't let me hold you up .

Extend, Elaborate, Enhance One of the challenges for annotators was to distinguish between speech functions belonging to Sustain.Continue.Prolong and React.Respond.Support.Develop groups that are among the top three most frequently encountered in conversations.. These two groups of moves are similar in that they represent speech functions used to continue the narration. The only difference between them is a speaker change that occurs in the case of using React.Respond.Support.Develop moves while it is irrelevant for Sustain.Continue.Prolong speech functions.

There are three types of such moves: **Extend** (offer additional or contrasting information), **Enhance** (qualifies previous move by giving details about time, a place, etc.), **Elaborate** (clarifies or restates previous moves). In the example (3), it was not obvious what label of these three to choose so there were 3 different variants for the utterance *"I find it very relaxing."*. Following the description of speech functions provided by S.Eggins and D.Slade, a right answer should be Sustain.Continue.Prolong.Elaborate as a speaker clarifies what was meant in the previous utterance.

The next fragment of the dialogue (4) was also difficult in terms of defining a right pragmatic purpose expressed in the last utterance. As it gives information about location that leads an annotator to the wrong conclusion that the last utterance can be labeled as Sustain.Continue.Prolong.Enhance. Furthermore, this sentence doesn't restate a previous move of the speaker that means the only appropriate tag will be Sustain.Continue.Prolong.Extend. Overall, human accuracy for this particular task is not satisfying and amounts ≈ 0.55.

(3)

Speaker1: You prefer classical music , don' t you ?
Speaker2: Yes , I do.
Speaker2: *I find it very relaxing.*

(4)

Speaker1: Oh, no!
Speaker1: What should I do now?
Speaker2: Don't worry.
Speaker2: *You can get off at the next stop and walk across the street and take the Bus 151 to the opposite direction.*

5 Conclusion

The analysis of data labeled with speech functions reveals that it is difficult to annotate casual conversations since annotators with enough expertise disagree on many aspects, particularly those referring to the

last layers in speech functions reflecting pragmatic purposes. Despite the fact that speech function theory is based on abstract categories, it is nevertheless appropriate for DA analysis because inter-annotator agreement for cut labels is rather high (0.83).

Concerning the last levels of speech functions, we plan to give additional real examples of the dialogues and an analysis of complex cases in the guidelines. Moreover, in order to identify excellent annotators who make mistakes while choosing the wrong last layers, we'll develop a list of acceptable error variants for each tag that do not impact inter-annotator agreement on cut labels.

Some dimensions of speech functions, such as opinions and facts, can be categorized more accurately using state-of-the-art classifiers. As there are no known ways to label some group speech functions (for example, Sustain.Continue.Prolong group) using modern models, one of our future goals will be to perform experiments to define features that can be extracted from utterances to distinguish pragmatic purposes automatically.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Dilyara Baymurzina, Denis Kuznetsov, Dmitry Evseev, Dmitry Karpov, Alsu Sagirova, Anton Peganov, Fedor Ignatov, Elena Ermakova, Daniil Cherniavskii, Sergey Kumeyko, Oleg Serikov, Yury Kuratov, Lidiya Ostyakova, Daniel Kornev, and Mikhail Burtsev. 2021. Dream technical report for the alexa prize 4. *4th Proceedings of Alexa Prize*.

Suzanne Eggins and Diana Slade. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.

Michael Alexander Kirkwood Halliday, Christian MIM Matthiessen, Michael Halliday, and Christian Matthiessen. 2014. *An introduction to functional grammar*. Routledge.

Jet Hoek and Merel Scholman. 2017. Evaluating discourse annotation: Some recent insights and new approaches. *// Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (isa-13)*.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. *// 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, P 88–95. IEEE.

Denis Kuznetsov, Dmitry Evseev, Lidia Ostyakova, Oleg Serikov, Daniel Kornev, and Mikhail Burtsev. 2021. Discourse-driven integrated dialogue development environment for open-domain dialogue systems. *// Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, P 29–51.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Nikita Mattar and Ipke Wachsmuth. 2012. Small talk is more than chit-chat. *// Annual Conference on Artificial Intelligence*, P 119–130. Springer.

Stefano Mezza, Alessandra Cervone, Giuliano Tortoreto, Evgeny A Stepanov, and Giuseppe Riccardi. 2018. Iso-standard domain-independent dialogue act tagging for conversational agents. *arXiv preprint arXiv:1806.04327*.

Andrei Popescu-Belis. 2005. Dialogue acts: One or more dimensions. *ISSCO WorkingPaper*, 62.

Dian Yu and Zhou Yu. 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*.