"Mama Always Had a Way of Explaining Things So I Could Understand": A Dialogue Corpus for Learning to Construct Explanations

Henning Wachsmuth *
Department of Computer Science
Paderborn University
henningw@upb.de

Milad Alshomary *
Department of Computer Science
Paderborn University
milad.alshomary@upb.de

Abstract

As AI is more and more pervasive in everyday life, humans have an increasing demand to understand its behavior and decisions. Most research on explainable AI builds on the premise that there is one ideal explanation to be found. In fact, however, everyday explanations are co-constructed in a dialogue between the person explaining (the explainer) and the specific person being explained to (the explainee). In this paper, we introduce a first corpus of dialogical explanations to enable NLP research on how humans explain as well as on how AI can learn to imitate this process. The corpus consists of 65 transcribed English dialogues from the Wired video series *5 Levels*, explaining 13 topics to five explainees of different proficiency. All 1550 dialogue turns have been manually labeled by five independent professionals for the topic discussed as well as for the dialogue act and the explanation move performed. We analyze linguistic patterns of explainers and explainees, and we explore differences across proficiency levels. BERT-based baseline results indicate that sequence information helps predicting topics, acts, and moves effectively.

1 Introduction

Explaining is one of the most pervasive communicative processes in everyday life, aiming for mutual understanding of the two sides involved. Parents explain to children, doctors to patients, teachers to students, seniors to juniors—or all the other way round. In explaining dialogues, one side takes the role of the *explainer*, the other the role of the *explainee*. Explainers seek to enable explainees to comprehend a given topic to a certain extent or to perform some action related to it (Rohlfing et al., 2021). This usually implies a series of dialogue turns where both sides request and provide different information about the topic. In line with the quote from the movie "Forrest Gump" in the title, how an explaining dialogue looks like is strongly affected by the specific explainer and explainee as well as by their interaction.

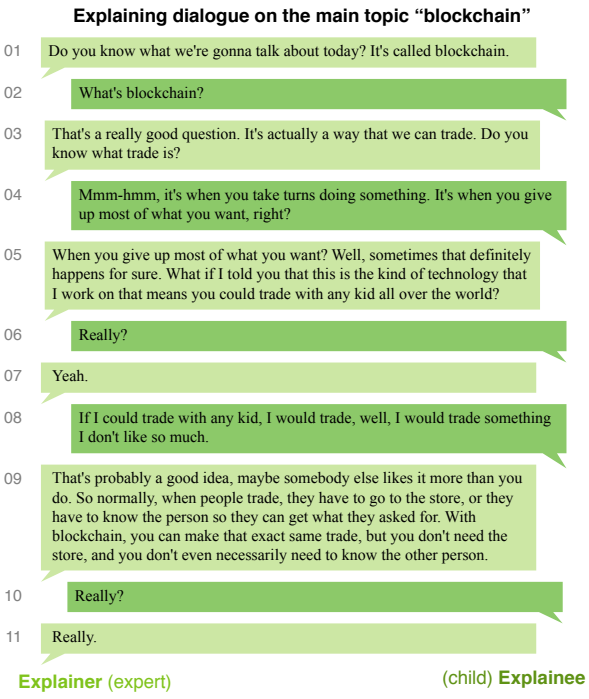

Figure 1: A short explaining dialogue from the video series *5 Levels*, included in the corpus presented in Section 3. Here, an expert explains blockchain to a child.

Consider the dialogue in Figure 1, where a technology expert explains the basic idea of blockchain to a 5-year old in a controlled setting. Beyond the explanations of the main topic (turns 05 and 09), the dialogue contains an explanation request (02), a test of prior knowledge (03), explanations from the explainee (04), and more. We observe that the explainer's explanations depend on the reaction of the explainee and that their level of depth is most likely adjusted to the explainee's proficiency.

The importance of studying how to explain has become apparent with the rise of research on explainable artificial intelligence, XAI (Barredo Arrieta et al., 2020). As AI finds its way into various

* Both authors contributed equally to this paper.

Proceedings of the 29th International Conference on Computational Linguistics, pages 344–354
October 12–17, 2022.

aspects of work and private life, humans interacting with respective systems, or being affected by them, have an increasing demand to understand their behavior and decisions. This demand has also been manifested in a *right to explanation* within the EU's General Data Protection Regulation (Goodman and Flaxman, 2017). Prior work on XAI largely starts from the premise that an ideal (monological) explanation exists for any behavior or decision, possibly dependent on the explainee at hand (Miller, 2019). According to Rohlfing et al. (2021), however, real explainability must account for the co-constructive nature of explaining emerging from interaction.

In natural language processing, early work modeled discourse structure of monological explanations (Bourse and Saint-Dizier, 2012), and a number of recent approaches generate respective explanations for XAI (Situ et al., 2021) and recommendation (Li et al., 2021). In contrast, the language of dialogical explanations is still understudied (details in Section 2). We argue that a better understanding of how humans explain in dialogues is needed, so that XAI can learn to interact with humans.

In this paper, we present a first corpus for computational research on how to explain in dialogues (Section 3). The corpus has been created as part of a big interdisplinary research project dealing with the construction of explainability.[1] It consists of 65 transcribed dialogical explanations from the American video series *5 Levels* freely published by the Wired magazine.[2] Five dialogues each refer to one of 13 science-related topics (e.g., "blockchain" or "machine learning"). They have the same explainer (an expert on the topic), but differ in the explainee's proficiency (from child to colleague).

To enable XAI to mimic human explainers, it has to learn what turn to make at any point in a dialogue. In discussion with humanities researchers, we model a turn for this purpose by the relation of its *topic* to the main topic (e.g., subtopic or related topic), its *dialogue act* (e.g., check question or informing statement), and its *explanation move* (e.g., testing prior knowledge or providing an explanation). We segmented the dialogues into a total of 1550 turns, and we let five independent professionals annotate each turn for these three dimensions.

In Section 4, we analyze linguistic patterns of explaining dialogues in the annotated corpus. We find clear signals for the explainer's alignment to the explainee's proficiency, such as the avoidance of deviating to related topics towards children. The roles of explainer and explainee are reflected in the varying use of dialogue acts and explanation moves, possibly stressed by the given setting.

To obtain baselines for the prediction of the three annotated dimensions, we evaluate three variants of BERT (Devlin et al., 2019) in 13-topic cross-validation on the corpus (Section 5). Our results reveal that modeling sequential dialogue interaction helps predicting a turn's topic, act, and move effectively. Improvements seem still possible, calling for more sophisticated approaches as well as for more explaining dialogue data in the future.[3]

In summary, the contributions of our paper are:

1. A manually annotated corpus for studying how humans explain in dialogical settings
2. Empirical insights into how experts explain to explainees of different proficiency levels
3. Baselines for predicting the topic, dialogue act, and explanation move of dialogue turns

2 Related Work

Explainable AI (XAI) largely focuses on the interpretability of learned models from the perspective of scientific completeness (Gilpin et al., 2018). Even though recent works tackle cognitive aspects, such as the trade-off between completeness and compactness (Confalonieri et al., 2019), Miller (2019) pointed out that this perspective is far away from the understanding of everyday explanations in the social sciences. Garfinkel (2009) argues that the key is to sort out what the explainer should actually explain, and Barredo Arrieta et al. (2020) stressed the importance of who is the explainee for XAI. Rohlfing et al. (2021) built on these works, but reasoned that explanations can only be successful in general, if they are co-constructed in interaction between explainer and explainee. The rationale is that explainees vary in their motives and needs, and they face different challenges (Finke et al., 2022). The corpus we present serves as a basis for studying the linguistic aspects of the explainer-explainee interaction computationally.

Natural language language processing (NLP) has notably dealt with the related genre of instructional texts, modeling their structure (Fontan and Saint-Dizier, 2008), extracting knowledge (Zhang et al.,

[1] Constructing Explainability, https://trr318.upb.de/en
[2] 5 Levels, https://www.wired.com/video/series/5-levels
[3] The corpus and the experiment code are freely available here: https://github.com/webis-de/COLING-22

2012), comprehending some meaning (Yagcioglu et al., 2018), or generating them (Fried et al., 2018). However, instructional text has a clear procedural style with distinctive surface features (Vander Linden, 1992), unlike explanations in general. For tutorial applications, Jordan et al. (2006) extracted concepts from explanation sentences, whereas Jansen et al. (2016) studied the knowledge needed for scientific explanations, and Son et al. (2018) identified causal explanations in social media. Towards a computational understanding of explaining, Bourse and Saint-Dizier (2012) modeled explanation structure with discourse relations (Mann and Thompson, 1988). In XAI and recommendation contexts, the generation of respective explanations is explored increasingly (Situ et al., 2021; Li et al., 2021).

However, our main goal is not to understand how to generate an explanation, but to model how people interact in an explanation process. For annotation, we thus rely on the widely accepted concept of dialogue acts (Stolcke et al., 2000; Bunt et al., 2010). Similar has been done for deliberative dialogues by Al Khatib et al. (2018). In addition, we model the *moves* that explainers and explainees make in their interaction, adapting the idea of rhetorical moves, in terms of communicative functions of text segments used to support the communicative objective of a full text (Swales, 1990). Wachsmuth and Stein (2017) proposed task-specific moves for monological arguments, but we are not aware of any work on moves for explanations, nor for dialogical settings.

Hence, we start by compiling data in this paper. Existing related corpora contain tutorial feedback for explanation questions (Dzikovska et al., 2012), answers to non-factoid questions (Dulceanu et al., 2018), and pairs of questions and responses from community question answering platforms (Nakov et al., 2017). Finally, the corpus of Fan et al. (2019) includes 270k threads from the Reddit forum *Explain like I'm Five* where participants explain a concept asked for in simple ways. While all these allow for in-depth analyses of linguistic aspects of explanations, none of them include explaining dialogues with multiple turns on each side. This is the gap we fill with the corpus that we introduce.

3 Data

This section introduces the corpus that we created to enable computational research on dialogical explanation processes of humans. We discuss our design choices with respect to the source and annotation, and we present detailed corpus statistics.

3.1 Explaining Dialogues on Five Levels

As source data, we decided to rely on explaining dialogues from a controlled setting in which two people explicitly meet to talk about a topic to be explained. While we thereby may miss some interaction behavior found in real-word explanation processes, we expect that such a setting best exhibits explaning dialogue features in their pure form.

In particular, we acquired the source dialogues in our corpus from *5 Levels*, an American online video series published by the Wired magazine. In each video of the series, one explainer explains a science-related or technology-related topic to five different explainees. The explainer is always an expert on the topic, whereas the explainees increase in terms of (assumed) proficiency on the topic:

1. a *child*,
2. a *teenager*,
3. an *undergrad* college sudent,
4. a *grad* student, and
5. a *colleague* in terms of another expert.

Every video starts with a few introductory words by the expert, before one dialogue follows the other.[4] Transcriptions are already provided in the videos' captions. So far, the first season of the series is available with a total of 17 videos. Table 1 lists all explained topics (*main topics* henceforth) in these videos, along with explainer information.

At the time of starting the annotation process discussed below, only 14 of the 17 videos had been accessible, and one of these had partly corrupted subtitles. We thus restricted the annotated corpus to the remaining 13 videos, summing up to 65 dialogues that correspond to a video length of 5.35 hours. Later, we added all dialogues from the other four videos in unannotated form to the corpus.

Before annotation, we manually segmented each dialogue into its single turns, such that consecutive turns in a dialogue alternate between explainer and explainee. Overall, the 65 dialogues consist of 1550 turns (23.8 turns per dialogue on average), 790 from explainers and 760 from explainees. The turns span 51,344 words (33.1 words per turn). On

[4] It is noteworthy that the videos seem to have been cut a little, likely for the sake of a concise presentation. We assume that this mainly removed breaks between dialogue turns only. While it limits studying non-verbal interaction in explaining, the effect for textual analyses of the dialogues should be low.

#	Topic	Explainer	Expertise
1	Harmony	Jacob Collier	Musician
2	Blockchain	Bettina Warburg	Political scientist
3	Virtual reality	John Carmack	Oculus CTO
4	Connectome	Bobby Kasthuri	Neuroscientist
5	Black holes	Varoujan Gorjian	NASA astronomer
6	Lasers	Donna Strickland	Professor
7	Sleep	Aric A. Prather	Sleep scientist
8	Dimensions	Sean Carroll	Theoret. physicist
9	Gravity	Janna Levin	Astrophysicist
10	Computer hacking	Samy Kamkar	Security researcher
11	Nanotechnology	George Tulevski	Nanotec. researcher
12	Origami	Robert J. Lang	Physicist
13	Machine learning	Hilary Mason	Hidden Door CEO
14	CRISPR	Neville Sanjana	Biologist
15	Memory	Daphna Shohamy	Neuroscientist
16	Zero-knowl. proof	Amit Sahai	Computer scientist
17	Black holes	Janna Levin	Astrophysicist

Table 1: All 17 main topics explained in the *5 Levels* dialogues, along with the explainers and their expertise. The 65 dialogues of the 13 topics listed in black are annotated in our corpus; the rest is provided unannotated.

average, an explainer's turn is double as long as an explainee's turn (43.7 vs. 22.1 words). While the general data size is not huge, we provide evidence in Sections 4 and 5 that it suffices to find common patterns of explanation processes. Limitations emerging from the size are discussed in Section 6.[5]

3.2 Annotations of Explanatory Interactions

The corpus is meant to provide a starting point for XAI systems that mimic the explainer's role within dialogical explanation processes. Our annotation scheme supports this purpose and is the result of extensive discussions in our interdisciplinary project with a big team of computer scientists, linguists, psychologists, and cognitive scientists. Where possible, we followed the literature, but the lack of research on human interaction in explaining (see Section 2) made us extend the state of the art in different respects.

In particular, we focus on turn-level category labels that capture the basic behavior of explainers and explainees in explaining dialogues. Our scheme models the three dimensions of dialogue turns that we agreed on to be needed for a computational understanding of the behavior:

- the relation of a turn's *topic* to the main topic,
- the *dialogue act* performed in the turn, and
- the *explanation move* made through the turn.

[5]We also extracted the time code (start and end milliseconds) of each segment from the videos, for which one caption is shown. This may serve multimodal studies in the future.

We discuss the labels considered for each of the three annotation dimensions in the following. Since all labels apply to both explainer and explainee in principle, we refer to a speaker and a listener below.

Topic Even though the dialogues we target have one defined main topic to be explained, what is explained in specific turns may vary due to the dynamics of explaining interaction (Garfinkel, 2009). Since we seek to learn how to explain in general rather than any specificities of the concrete 13 main topics in the corpus, we abstract from the latter, modeling only the relation of the topic discussed in a turn to the dialogue's main topic. In particular, a turn's topic may be annotated as follows:

- t_1 *Main topic.* The main topic to be explained;
- t_2 *Subtopic.* A specific aspect of the main topic;
- t_3 *Related topic.* Another topic that is related to the main topic;
- t_4 *No/Other topic.* No topic, or another topic that is unrelated to the main topic.

Dialogue Act To model the communicative functions of turns in dialogues, we follow the literature (Bunt et al., 2010), starting from the latest version of the ISO standard taxonomy of dialogue acts.[6] In explaining, specific dialogue acts are in the focus, though. In collaboration with the interdisciplinary team, we selected a subset of 10 acts that capture communication on a level of detail that is specific enough to distinguish key differences, but abstract enough to allow finding recurring patterns:

- d_1 *Check question.* Asking a check question;
- d_2 *What/How question.* Asking a what question or a how question of any kind;
- d_3 *Other question.* Asking any other question;
- d_4 *Confirming answer.* Answering a question with confirmation;
- d_5 *Disconfirming answer.* Answering a question with disconfirmation;
- d_6 *Other answer.* Giving any other answer;
- d_7 *Agreeing statement.* Conveying agreement on the last utterance of the listener;
- d_8 *Disagreeing statement.* Conveying disagreement accordingly;
- d_9 *Informing statement.* Providing information with respect to the topic stated in the turn;
- d_{10} *Other.* Performing any other dialogue act.

[6]DIT++ Taxonomy of Dialogue Acts, https://dit.uvt.nl

Explanation Move Finally, we aim to understand the explanation-specific moves that explainers and explainees make to work together towards a successful explanation process. Due to the lack of models of explaining dialogues (see Section 2, we started from recent theory of explaining (Rohlfing et al., 2021). Based on a first inspection of a corpus sample, we established a set of 10 explanation moves that a speaker may make in the process, at a granularity similar to the dialogue acts:[7]

- e_1 *Test understanding.* Checking whether the listener understood what was being explained;
- e_2 *Test prior knowledge.* Checking the listener's prior knowledge of the turn's topic;
- e_3 *Provide explanation.* Explaining any concept or a topic to the listener;
- e_4 *Request explanation.* Requesting any explanation from the listener;
- e_5 *Signal understanding.* Informing the listener that their last utterance was understood;
- e_6 *Signal non-understanding.* Informing the listener that the utterance was not understood;
- e_7 *Providing feedback.* Responding qualitatively to an utterance by correcting errors or similar;
- e_8 *Providing assessment.* Assessing the listener by rephrasing their utterance or giving a hint;
- e_9 *Providing extra info.* Giving additional information to foster a complete understanding;
- e_{10} *Other.* Making any other explanation move.

We note the hierarchical nature of the scheme with respect to dialogue acts and explanations; for example, d_1–d_3 could be merged as well as e_1–e_2. While some acts and moves are much more likely to be made by an explainer or an explainee, we did not restrict this to avoid biasing the annotators.[8]

3.3 Crowd-based Annotation Process

The restriction of the annotations to a manageable number of turn-level labels was also made to make the annotation process simple enough to carry it out with independent people. In particular, we hired five freelancers, working as content editors and annotators on the professional crowdworking platform *Upwork*. All were native speakers of English with a 90%+ job success rate on the platform. We clarified the task individually with each of them.

We provided guidelines based on the definitions above, along with general explanations and some examples. Using Label Studio,[9] we developed a task-specific user interface where each dialogue was shown as a sequence of turns and one label of each dimension could be assigned to a turn (if multiple labels seemed appropriate, the best fitting one). Each annotator labeled all 1550 turns. We paid $ 1115 for an overall load of 85 hours, that is, $ 13.12 per hour on average (with minor differences for annotators due to bonuses and varying durations).

Agreement In terms of the conservative measure Fleiss' κ, the inter-annotator agreement among all five was 0.35 for the topic, 0.49 for dialogue acts, and 0.43 for explanation moves. While these values indicate moderate agreement only, they are in line with related subjective labeling tasks of short texts such as news sentences (Al Khatib et al., 2016) and social media arguments (Habernal et al., 2018). Moreover, we exploited the multiple labels we have per turn to consolidate reliable annotations, as described in the following.

Output Annotations For consolidation, we rely on MACE (Hovy et al., 2013), a widely used technique for grading the reliability of crowdworkers based on their agreement with others. The MACE competence scores of the annotators suggest that all did a reasonable job in general, lying in the ranges 0.30–0.76 (topic), 0.58–0.82 (dialogue acts), and 0.45–0.85 (explanation moves) respectively. We applied MACE' functionality to derive one aggregate output label for each dimension from the five annotations weighted by competence scores.

3.4 The Wired Explaining Dialogue Corpus

Table 2 presents detailed general statistics of the three annotation dimensions. More insights into the distribution of annotations across proficiency levels follow in Section 4.

With respect to topic (t_1–t_4), about half of all turns explicitly discuss the *main topic* (27.7%), a *subtopic* (5.7%), or a *related topic* (16.8%). Explainees much more often mention none of these (62.8% vs. 37.3%), underlining the leading role of the explainer in dialogue setting.

[7] We decided to leave a distinction of different explaining types (such as causal or analogy-based explanations) to future work, as it does not match the level of detail in our scheme.

[8] For dialogue acts d_3, d_6, and d_{10} as well as explanation move e_{10}, the annotators had to name the label in free text. We provide these as part of the corpus, we give individual examples of other moves and acts in Section 4.

[9] Label Studio, https://labelstud.io

Label	Explainer #	Explainer %	Explainee #	Explainee %	Total #	Total %
t_1 Main topic	**301**	**38.1**	129	17.0	430	27.7
t_2 Subtopic	52	6.6	36	4.7	88	5.7
t_3 Related topic	142	18.0	118	15.5	260	16.8
t_4 Other/No topic	295	37.3	**477**	**62.8**	**772**	**49.8**
d_1 Check question	183	23.2	62	8.2	245	15.8
d_2 What/How question	77	9.7	38	5.0	115	7.4
d_3 Other question	3	0.4	10	1.3	13	0.8
d_4 Confirming answer	14	1.8	40	5.3	54	3.5
d_5 Disconfirm. answer	3	0.4	21	2.8	24	1.5
d_6 Other answer	2	0.3	23	3.0	25	1.6
d_7 Agreeing statement	75	9.5	190	25.0	265	17.1
d_8 Disagree. statement	2	0.3	10	1.3	12	0.8
d_9 Informing statement	**391**	**49.5**	**305**	**40.1**	**696**	**44.9**
d_{10} Other	40	5.1	61	8.0	101	6.5
e_1 Test understanding	56	7.1	0	0.0	56	3.6
e_2 Test prior knowledge	111	14.1	1	0.1	112	7.2
e_3 Provide explanation	**409**	**51.8**	**270**	**35.5**	**679**	**43.8**
e_4 Request explanation	47	5.9	95	12.5	142	9.2
e_5 Signal understanding	37	4.7	104	13.7	141	9.1
e_6 Signal non-underst.	1	0.1	16	2.1	17	1.1
e_7 Provide feedback	61	7.7	224	29.5	285	18.4
e_8 Provide assessment	10	1.3	1	0.1	11	0.7
e_9 Provide extra info	26	3.3	22	2.9	48	3.1
e_{10} Other	32	4.1	27	3.6	59	3.8
Σ	790	100.0	760	100.0	1550	100.0

Table 2: Corpus distribution of annotated topics (t_1–t_4), dialogue acts (d_1–d_{10}), and explanation moves (e_1–e_{10}) separately for explainer and explainee turns and in total. Per type, the highest value in a column is marked bold.

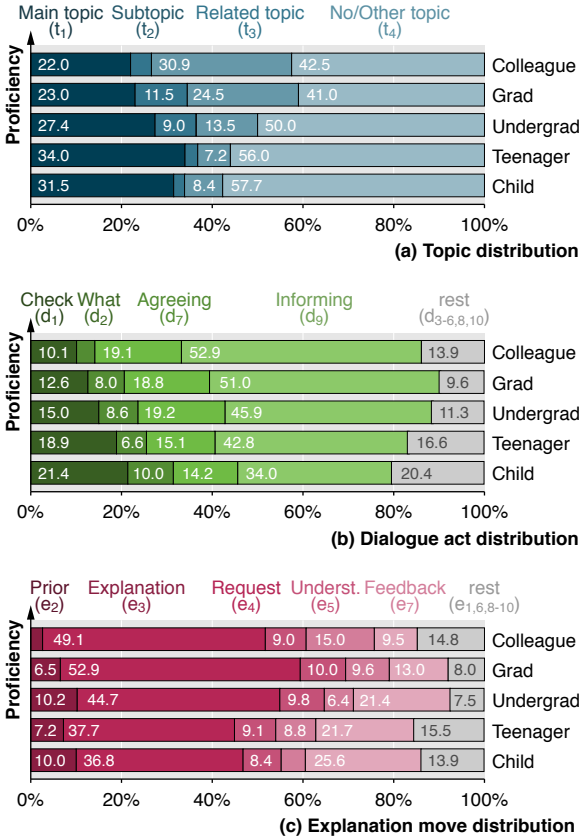

Figure 2: Distribution of topic, discourse act, and explanation act annotations in the corpus, depending on the proficiency of the explainee (from *Child* to *Colleague*).

For dialogue acts (d_1–d_{10}), we see that, quite intuitively, *informing statements* (44.9%) are dominant in explaining dialogues on both sides (explainer 49.5%, explainee 40.1%). However, also *agreeing statements* (17.1%) as well as *check questions* (15.8%) play an important role. The low frequency of *other questions* (0.8%) and *other* (6.5%) suggests that the selected set of dialogue acts cover well what happens in the given kind of dialogues, even though our annotators identifid sum acts, such as *disagreeing statements* (0.8%), rarely only.[10]

Similar holds for the explanation moves (e_1–e_{10}): only 3.8% of all 1550 turns belong to *other*.[11] As expected, the core of explaining is to *provide explanations* (43.8%), also explainees do so in 270 turns (35.5%). Besides, they often *provide feedback* (29.5%). Explainers rather *test prior knowledge* (14.1%) and *test understanding* often (7.1%), but also provide feedback sometimes (7.7%).

[10]Notable examples of other dialogue acts the annotators observed include *greetings* (e.g., "Hi, are you Bella?"), *casual chat* ("What do you do?"), and *gratitude* ("Thank you.").

[11]Here, other cases include *inquiry* ("Hi, are you Bella") and *introduction* ("Bella, I'm George, nice to meet you.").

4 Analysis

One main goal of the presented corpus is to learn how humans explain in dialogical settings. This section analyzes commonalities and differences regarding meta-information available in the corpus.

4.1 Explaining across Proficiency Levels

First, we explore to what extent explaining differs depending on the proficiency of the explainee. Figure 2 shows the distributions of the three annotated dimensions separately for the five given explainee levels. For dialogue acts and explanation moves, we distinguish only the most frequent labels and merge all others into a class *rest*.

With respect to topic, we see that particularly the discussion of *related topics* grows notably with the explainee's proficiency, from 8.4% of all annotations for children to 30.9% for colleagues. Conversely, the *main topic* is mentioned less in dialogues with more proficient explainees; the same holds for *no/other topic*. *Subtopics* are considered mainly with grads (11.5%) and undergrads (9.0%), possibly related to the way they learn.

Topic Sequences	Explainer	Explainee	Total
(Main, Rel, Main)	**24.6%**	7.7%	**15.4%**
(Main, Rel, Main, Rel, Main)	–	–	7.7%
(Main)	12.3%	**18.5%**	6.2%
(Rel, Main, Rel, Main, Rel, Main)	–	–	4.6%
(Main, Rel)	3.1%	10.8%	4.6%
(Rel, Main, Rel, Main)	3.1%	–	3.1%
(Main, Sub, Main)	–	–	3.1%
(Main, Sub, Main, Rel, Main)	4.6%	3.1%	3.1%

Table 3: Relative frequencies of all recurring sequences of *main*, *sub*, and *rel*ated topic in the corpus' dialogues and in the explainers and explainees' parts alone.

Labels	Act/Move Pair	Explainer	Explainee	Total
d_9/e_3	Informing/Explanation	**45.9%**	**31.3%**	**38.8%**
d_7/e_7	Agreeing/Feedback	3.9%	14.2%	9.0%
d_7/e_5	Agreeing/Understanding	3.5%	9.1%	6.3%
d_1/e_2	Check/Prior	10.5%	–	5.4%
d_1/e_4	Check/Request	2.7%	6.8%	4.7%
d_2/e_4	What/Request	3.0%	4.5%	3.7%
d_{10}/e_{10}	Other/Other	2.8%	2.6%	2.7%
d_1/e_1	Check/Understanding	5.1%	–	2.6%
d_4/e_7	Confirming/Feedback	1.4%	3.7%	2.5%
d_9/e_7	Informing/Feedback	0.5%	4.2%	2.3%

Table 4: Relative frequencies of the ten most frequent pairs of dialogue act and explanation move in the corpus and the differences for explainers and explainees.

Explainer			Explainee		
Word	**Frequency**	**Ratio**	**Word**	**Frequency**	**Ratio**
here	0.16%	4.20	yes	0.21%	5.12
around	0.12%	4.03	mean	0.14%	4.20
space	0.24%	3.32	stuff	0.11%	3.11
light	0.18%	2.96	oh	0.16%	2.75
earth	0.10%	2.65	yeah	0.65%	2.70
us	0.15%	2.39	many	0.12%	2.39
want	0.14%	2.28	interesting	0.12%	2.11
going	0.22%	2.19	well	0.21%	1.94
point	0.11%	2.11	like	1.10%	1.85
thing	0.18%	1.93	no	0.18%	1.83

Table 5: The top-10 words used specifically by explainers and explainees, respectively, along with the relative frequency (minimum 0.1%) and specificity ratio (e.g., explainees say "yes" 5.12 times as often as explainers).

For dialogue acts, the key difference lies between the proportion of *informing statements* and the number of questions asked (d_1 and d_2). Whereas the former monotonously goes up from 34.0% (child) to 52.9% (colleague), particularly the use of *check questions* is correlated inversely with proficiency, used mainly to test prior knowledge and to check understanding. A similar behavior can be observed for explanation moves. There, *providing feedback* shrinks from 25.6% to 9.5%, while *providing explanations* mostly grows, with peak at grads (52.9%). In contrast, how often people *request explanations* remains stable across proficiency levels.

4.2 Interactions of Topics, Moves, and Acts

Interactions of the annotated dimensions happen between the turns and within a turn. We analyze one example of each here, and, due the limited data size, we look at topics separately from dialogue act and explanation moves.

Inspired by the flow model of Wachsmuth and Stein (2017), Table 3 shows all eight sequences of topics that occur more than once among the 65 dialogues. Each sequence shows the ordering of topics being discussed, irrespective of how often each topic is mentioned in a row. Most dialogues start and end with the main topic, often in alternation with related topics, such as *(Main, Rel, Main)* in 15.4% of all cases (sometimes also with subtopics). The ordering of what *explainers* talk about is similar, whereas *explainees* often focus on the main topic only (18.5%).

Table 4 lists the top-10 pairs of acts and moves. *Informing statements* that *provide explanations* are most common across both explainers (45.9%) and explainees (31.3%). *Agreeing statements* (d_7) and *check questions* (d_1) cooccur with multiple moves, and especially *providing feedback* happens via different dialogue acts. As expected in the given setting, explainees never check for prior knowledge or understanding (d_1/e_2, d_1/e_1). Instead, they agree by providing feedback or signaling understanding (d_7/e_7, d_7/e_5) much more often than explainers.

4.3 Language of Explainers and Explainees

Finally, we investigate basic differences in the language of the two sides: We determine the words that are often used by explainers (at least 0.1% of all words) and rarely by explainees, or vice versa.

Table 5 presents the 10 most specific words on each side. Aside from some topic-specific words (e.g., "light"), the explainer's list includes typical words used in meta-language, as in this explanation to a teenager: "I *want* to know if you agree, sleep is the coolest *thing* you've ever heard of." On the explainee's side, we find multiple reactive words, such as "oh" and "interesting", but also indicators of vagueness, as in this colleague's response to an explanation of hacking: "So all kind of older logic and *stuff like* that. So, I *mean*, it's sort of based on, *like*, you're presented the little MUX chip."

5 Experiments

The second goal of the corpus is to serve the creation of XAI systems that mimic human explainers. As an initial endeavor, this section reports on baseline experiments on the computational prediction of topics, dialogue acts, and explanation moves.

5.1 Experimental Setup

We evaluate three models based on BERT (Devlin et al., 2019), along with a simple majority baseline, for predicting each dialogue turn dimension in 13-fold cross-topic validation: For each main topic, we trained one model on the other 12 topics and tested it against the labels of the respective dimension. We average the resulting F_1-scores over all 13 folds.[12] Figure 3 illustrates the three BERT variants.

BERT-basic The first model simply adds a classification head to BERT. It takes as input the dialogue's main topic and the turn's text, x_i (separated by `[SEP]`), as well as the label y_i to predict (topic t_i, dialogue act d_i, or explanation move e_i). We trained the model for five epochs, optimizing its F_1-score on the turns of two main topics. We balanced the training set using oversampling to prevent the model from only predicting the majority label.

BERT-sequence Turns made in explaining dialogues depend on previous turns, for example, a conclusion on the *main topic* may be preceded by a *related topic* (see Table 3). In the second model, we exploit such dependencies with turn-level sequence labeling: Given the sequence $(x_1, \ldots, x_n)$ of all turns in a dialogue, the input to predicting a label y_i of x_i is the turn's history $(x_1, \ldots, x_{i-1})$ along with all previously predicted labels $(y_1, \ldots, y_{i-1})$ of the same dimension. For each turn, we encode the history in a `CLS` embedding with BERT. Then, we pass all labels and `CLS` embeddings through a CRF layer to model the label's dependencies.

BERT-multitask Finally, the interaction of topic t_i, act d_i, and move e_i in a turn may be relevant. For example, an *informing statement* likely *provides an explanation* (see Table 4). Our third model thus learns to classify all three dimensions jointly in a multitask fashion, based on multitask-NLP.[13] We trained one multitask model each with one of the three dimensions as main task and the others as auxiliary tasks, oversampling with respect to the main task. To this end, we employ a shared BERT encoder and three classification heads, one for each task. The final loss is the weighted average of the three classification losses, with weight 0.5 for the main task and 0.25 for both others. We trained the models for 10 epochs allowing them to converge.

[12] All models start from the `bert-based-uncased`, and are trained with a learning rate of $2e^{-5}$ and a batch size of 4.

[13] Multitask NLP, https://multi-task-nlp.readthedocs.io

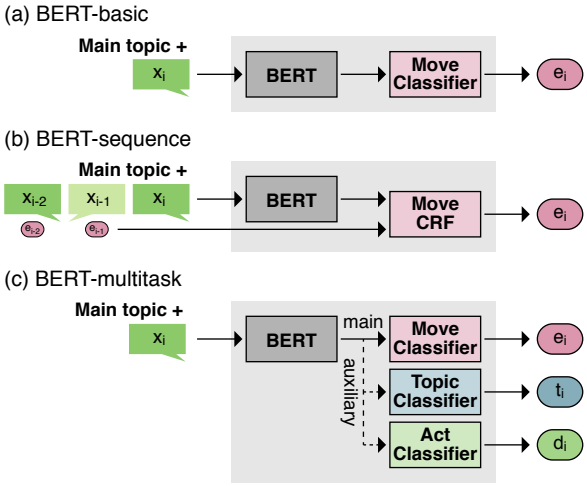

Figure 3: Sketch of the three evaluated models, here for predicting a turn's explanation move, e_i: (a) *BERT-basic* labels a turn in isolation. (b) *BERT-sequence* takes the labels of previous turns into account. (c) *BERT-multitask* classifies all three turn dimensions simultaneously.

Approach	Main T. (t_1)	Sub- T. (t_2)	Related T. (t_3)	No/Oth. T. (t_4)	Macro F_1-Score
BERT-basic	0.58	0.11	**0.44**	**0.89**	0.51
BERT-sequence	**0.61**	**0.13**	**0.44**	**0.89**	**0.52**
BERT-multitask	0.43	0.04	0.36	0.81	0.41
Majority baseline	0.00	0.00	0.00	0.66	0.17

Table 6: Topic prediction results: The F_1-scores of the evaluated BERT models for each considered relation to the main topic, t_1–t_4, as well as the macro-averaged F_1-score. The best value in each column is marked bold.

5.2 Results

Tables 6–8 show the individual and the macro F_1-scores for all three dimensions.

BERT-sequence performs best across all three labeling tasks, highlighting the impact of modeling the sequential interaction in dialogues. It achieves a macro F_1-score of 0.52 for topics, 0.47 for dialogue acts, and 0.43 for explanation moves. However, likely due to data sparsity, some labels remain hard to predict, such as *Subtopic* (t_2), *disagreement statements* (d_8), and *provide assessment* (e_8).

BERT-basic beats BERT-sequence on a few labels, such as *signal non-understanding* (e_8), but

Approach	Check Q. (d_1)	What/H. Q. (d_2)	Other Q. (d_3)	Confirm. A. (d_4)	Disconf. A. (d_5)	Other A. (d_6)	Agree. St. (d_7)	Disagr. St. (d_8)	Inform. St. (d_9)	Other (d_{10})	Macro F_1-Score
BERT-basic	**0.76**	**0.73**	0.00	0.33	**0.67**	0.00	0.51	0.00	**0.87**	0.57	0.44
BERT-sequence	**0.76**	0.72	0.00	**0.35**	**0.67**	0.00	****0.69**	0.00	**0.87**	**0.61**	**0.47**
BERT-multitask	0.54	0.49	0.00	0.29	0.59	0.00	0.53	**0.09**	0.84	0.44	0.38
Majority baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.00	0.06

Table 7: Dialogue act prediction results: The F_1-scores of the evaluated BERT models for each considered dialogue act, d_1–d_{10}, as well as the macro-averaged F_1-score. The best value in each column is marked bold.

Approach	Test U. (e_1)	Test P.K. (e_2)	Provide Ex. (e_3)	Request Ex. (e_4)	Signal U. (e_5)	Signal N.U. (e_6)	Provide Fe. (e_7)	Provide As. (e_8)	Provide E.I. (e_9)	Other (e_{10})	Macro F_1-Score
BERT-basic	**0.27**	**0.64**	**0.84**	0.60	0.29	**0.34**	0.51	0.00	**0.11**	0.50	0.41
BERT-sequence	**0.27**	**0.64**	**0.84**	**0.64**	**0.33**	0.21	****0.60**	**0.15**	0.08	**0.56**	**0.43**
BERT-multitask	0.21	0.54	0.80	0.40	0.16	0.32	0.53	0.00	0.08	0.35	0.34
Majority baseline	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06

Table 8: Explanation move prediction results: The F_1-scores of the evaluated BERT models for each considered explanation move, e_1–e_{10}, as well as the macro-averaged F_1-score. The best value in each column is marked bold.

cannot compete overall. *BERT-multitask* performs worst among the three models. We attribute this to the data imbalance: While oversampling helps with respect to the main task, it does not benefit the label distribution of the auxiliary tasks. Also, optimizing the loss weights of the three tasks may further aid multitask learning, but such an engineering of prediction models is not the focus of this work.

6 Conclusion

How humans explain in dialogical settings is still understudied. This paper has presented a first corpus for computational research on controlled explaining dialogues, manually annotated for topics, dialogue acts, and explanation moves. Our analysis has revealed intuitive differences in the language of explainers and explainees and their dependence on the explainee's proficiency. Moreover, baseline experiments suggest that a prediction of the annotated dimensions is feasible and benefits from modeling interactions. With these results, we lay the ground towards more human-centered XAI. We expect that respective systems need to learn to how to explain depending on the explainee's reactions, and how to proactively lead an explaining dialogue to achieve understanding on the explainee's side.

A limitation of the corpus lies in the restricted corpus size caused by the availability of source data, preventing deeper statistical analyses and likely rendering a direct training of dialogue systems on the corpus hard. Also, it remains to be explored what findings generalize beyond the controlled setting of the given dialogues. Future work should thus target both the scale and the heterogeneity of explaining data, in order to provide the pervasive communicative process of explaining the attention it deserves.

Acknowledgments

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), partially under project number TRR 318/1 2021 – 438445824 and partially under SFB 901/3 – 160364472. We thank Meisam Booshehri, Hendrik Buschmeier, Philipp Cimiano, Josephine Fisher, Angela Grimminger, and Erick Ronoh for their input and feedback to the annotation scheme. We also thank Akshit Bhatia for his help with the corpus preparation as well as the anonymous freelancers on Upwork for their annotations.

7 Ethical Statement

We do not see any immediate ethical concerns with respect to the research in this paper. The data included in the corpus is freely available. All participants involved in the given dialogues gave their consent to be recorded and received expense allowances, as far as perceivable from the Wired web resources. As discussed in the paper, the three freelancers in our annotation study were paid about $13 per hour, which exceeds the minimum wage in most US states and is also conform to the standards in the regions of our host institution. In our view, the provided prediction models target dimensions of dialogue turns that are not prone to be misused for ethically doubtful applications.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Sarah Bourse and Patrick Saint-Dizier. 2012. A repository of rules and lexical resources for discourse structure analysis: the case of explanation structures. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2778–2785, Istanbul, Turkey. European Languages Resources Association (ELRA).

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).

Roberto Confalonieri, Tarek R. Besold, Tillman Weyde, Kathleen Creel, Tania Lombrozo, Shane T. Mueller, and Patrick Shafto. 2019. What makes a good explanation? Cognitive dimensions of explaining intelligent machines. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pages 25–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu, and Seokhwan Kim. 2018. PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210, Montréal, Canada. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Josefine Finke, Ilona Horwath, Tobias Matzner, and Christian Schulz. 2022. (de)coding social practice in the field of xai: Towards a co-constructive framework of explanations and understanding between lay users and algorithmic systems. In *Artificial Intelligence in HCI*, pages 149–160, Cham. Springer International Publishing.

Lionel Fontan and Patrick Saint-Dizier. 2008. Analyzing the explanation structure of procedural texts: Dealing with advice and warnings. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 115–127. College Publications.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Alan Garfinkel. 2009. *Forms of Explanation: Rethinking the Questions in Social Theory*, revised edition. Yale University Press, New Haven & London, New Haven; London.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. ArXiv: 1806.00069.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1930–1940. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia. Association for Computational Linguistics.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

Pamela W. Jordan, Maxim Makatchev, and Umarani Pappuswamy. 2006. Understanding complex natural language explanations in tutorial applications. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 17–24, New York City, New York. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada. Association for Computational Linguistics.

Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728.

Xuelin Situ, Ingrid Zukerman, Cecile Paris, Sameen Maruf, and Gholamreza Haffari. 2021. Learning to explain: Generating stable explanations fast. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5340–5355, Online. Association for Computational Linguistics.

Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. 2018. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3350–3359, Brussels, Belgium. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

John M. Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Keith Vander Linden. 1992. The expression of local rhetorical relations in instructional text. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 318–320.

Henning Wachsmuth and Benno Stein. 2017. A universal model for discourse-level argumentation analysis. *Special Section of the ACM Transactions on Internet Technology: Argumentation in Social Media*, 17(3):28:1–28:24.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Ziqi Zhang, Philip Webster, Victoria Uren, Andrea Varga, and Fabio Ciravegna. 2012. Automatically extracting procedural knowledge from instructional texts using natural language processing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 520–527, Istanbul, Turkey. European Language Resources Association (ELRA).