LEDA: a Large-Organization Email-Based Decision-Dialogue-Act Analysis Dataset

Mladen Karan[*], **Prashant Khare**[*], **Ravi Shekhar**[†], **Stephen McQuistin**[‡], **Colin Perkins**[‡], **Ignacio Castro**[*], **Gareth Tyson**[*§], **Patrick G.T. Healey**[*], **Matthew Purver**[*¶]
[*]Queen Mary University of London, [†]University of Essex, [‡]University of Glasgow
[§]Hong Kong University of Science & Technology, [¶]Jožef Stefan Institute
{p.khare, m.karan, i.castro, g.tyson, p.healey, m.purver}@qmul.ac.uk,
r.shekhar@essex.ac.uk, sm@smcquistin.uk, csp@csperkins.org

Abstract

Collaboration increasingly happens online. This is especially true for large groups working on global tasks, with collaborators all around the world. The size and distributed nature of such groups make decision-making challenging. This paper proposes a set of dialog acts for the study of decision-making mechanisms in such groups, and provides a new annotated dataset based on real-world data from the public mail-archives of one such organization – the Internet Engineering Task Force (IETF). We provide an initial data analysis showing that this dataset can be used to better understand decision-making in such organizations. Finally, we experiment with a preliminary transformer-based dialog act tagging model.

1 Introduction and Related Work

Motivation Online collaboration has been used for many years by large distributed organizations. The increasing availability of high-speed Internet connections and collaboration tools, along with the Covid-19 pandemic, are making it ever more prevalent. Large distributed organizations of this type often undertake important tasks. For example, the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C) are responsible for developing the technical standards that underpin the Internet. Consequently, understanding the decision-making processes in this type of organization is essential to increase transparency and accountability, to facilitate tracking of decisions and the reasoning behind them, and to understand alternatives that were considered (or not) and the voices that were (or were not) heard.

Goals Most studies of decision making in text (e.g. Hsueh and Moore, 2007; Fernández et al., 2008; Bui and Peters, 2010) rely on annotation and analysis of *Dialogue Acts* (DAs). We adopt this approach and label emails from public IETF mailing lists with DAs. Our aim is to answer the following research questions: **RQ1:** *What is an appropriate set of DAs to use for this annotation task?*; **RQ2:** *How do communication patterns change through the life-cycle of a decision discussion?*; and **RQ3:** *How do different types of participants differ in how they contribute to the process?* The overall goal of these questions is to better understand the mechanisms underlying the decision-making process in a large, distributed, collaborative organization.

Related Datasets The most notable email-based related dataset is the Enron Corpus (Klimt and Yang, 2004), covering over 200K messages of Enron employees in various positions within the organization. However, in-house emails of a single closed company are not representative of communication in larger, more diverse collaborations.

Datasets specifically relevant for studying decision making include AMI (McCowan et al., 2005) and ICSI/MRDA (Janin et al., 2003; Shriberg et al., 2004). However, the AMI dataset is not "real": it uses actors acting out small-group meetings on predefined topics. In contrast, the ICSI dataset is based on naturally occurring meetings at the International Computer Science Institute (ICSI). While both are annotated with general dialogue act labels, AMI also includes specific decision-oriented dialogue acts provided by Fernández et al. (2008). Despite this, they are not representative of interaction in large groups, or online collaborative settings. Consequently, we annotate a new dataset tailored to address our research questions. We denote it as **L**arge-organization **E**mail-based **D**ecision-dialogue-act **A**nalysis dataset – **LEDA**.

There are important differences between LEDA and AMI/ICSI. First, while AMI/ICSI are transcribed face-to-face, real-time, in-person, and small-group meetings. LEDA contains emails from mailing-lists, asynchronous, and from a large decentralized, globally spread group. Second, AMI/ICSI discuss mostly self-contained, focused topics (design, research-group progress); LEDA discusses the more long-term, complex task of designing Internet-standards. We further provide a more detailed comparison of LEDA with AMI in Appendix A.

Contributions First, we propose a taxonomy of DA labels for large-group email decision-making. Second, we provide a novel dataset labeled with DAs. Third, we provide data analyses exploring decision-making communication patterns within the IETF. Fourth, we provide a preliminary DA prediction model on this dataset, which can serve as a reference baseline model for future work.

2 Dataset

Our data consists of emails from the IETF mailing list archive.[1] The IETF is a typical example of decision making in a large, distributed, online collaborative community; it has rich metadata available via the IETF DataTracker;[2] and the data is publicly available with appropriate consent.[3]

IETF background The IETF is a large, open, voluntary organization tasked with developing Internet standards (Flanagan, 2019; McQuistin et al., 2021; Khare et al., 2022). It is comprised of *working groups* (WGs), each focusing on a relatively narrow field: e.g., RMCAT[4] WG focuses on specific Real-time Media Congestion Avoidance Techniques. Each WG has one or more participants as *chairs*. During its development, an Internet standard is called a *draft*. Drafts are discussed in the mailing lists (the archive has >2M emails, predominantly in English, between 56k participants over 20 years) and in several live meetings yearly. After sufficient revision and review, a draft becomes an Internet standard.

Data preparation The email archive consists of threads (sets of emails connected with reply-to relations, forming a tree-like structure). Given a particular draft, we extract all threads with at least one message that mentions the draft in either the subject or body. We do this for four drafts, chosen by an IETF expert to span a range of technical areas. We opted for entire threads over a smaller number of drafts (rather than more drafts but with partial threads) to ensure a full view of the draft discussion and agreement process over its life-cycle.

We then preprocess all messages, splitting them into *Quote*, *Signature*, or *Normal* segments using custom heuristics developed for this data. A *Normal* segment contains text written by the author of the message. A *Quote* segment contains text written by someone else, which is being quoted. A *Signature* segment contains signatures (name, company name, website). *Normal* segments are useful for analysis, while the rest introduce noise. We also keep track of quoting relations between segments.

Label set calibration As our starting point, we take the DA labels defined in the ISO 24617-2 standard (Bunt et al., 2012). Cross-referencing with labels in datasets from related work and manual inspection of the IETF data suggested that much of the complexity in the standard is not needed for our goals. This was confirmed in several initial rounds of annotations where we observed considerable confusion between the very fine grained ISO 24617-2 DAs on our data. After each iteration, we simplified the label set by removing irrelevant labels for email communication (e.g., rhetorical devices such as pauses) and aggregating hard to distinguish labels (e.g., accepting a request and agreeing to an opinion). Table 1 presents our two-level taxonomy with three coarse grained labels divided into eleven fine-grained ones, which was obtained after four rounds of calibration.

Annotation Annotation of each segment with DA labels was carried out by seven student annotators, all with a background in linguistics. A segment can be assigned several DAs simultaneously (a multi-label setting). During the calibration rounds, annotators provided feedback which helped modify the taxonomy and instructions. For the final annotation, they were provided a detailed set of instructions and an annotation tool specifically developed in-house.

Table 1 reports data statistics and inter-annotator agreement (IAA). Each thread is annotated by at least two annotators. To measure IAA, we considered both Fleiss' Kappa and Krippendorff's Alpha, but neither supports multi-label annotation. Instead, we consider one annotator's labels as "gold labels," and another's as "classifier predictions." We calculate the F1 score for all annotator pairs and average them. This calculation is performed on a subset of 15 threads labeled by all annotators. For some

[1] http://mailarchive.ietf.org/arch/
[2] http://datatracker.ietf.org/
[3] www.ietf.org/privacy-statement/
[4] http://datatracker.ietf.org/wg/rmcat/

labels, the annotation is inherently difficult, as reflected in the IAA. Manual inspection reveals that many of these disagreements may be impossible to completely resolve as the task is subjective (Uma et al., 2021). For example, *ClarificationElicitation* is more often implicit (*"I don't see why ..."*) than explicit (*"Can you explain why ..."*), introducing disagreement. However, recent work (Pavlick and Kwiatkowski, 2019; Basile et al., 2021; Leonardelli et al., 2021) shows it is viable to design models and evaluation measures that account for this inherent ambiguity instead of trying to resolve it. Accordingly, we release all individual annotators' labels with the text data and code. [5] While covering only four drafts, LEDA is of substantial size (8230 segments, 2196 messages, 363 authors), with the drafts hand-picked by an IETF expert to ensure they are representative. We focus on trends that are very prominent and supported by statistical significance tests. Finally, an inspection of plots for individual drafts revealed that the main trends outlined in the remaining sections were consistent across all four drafts.

3 Analysis of gold-standard labels

3.1 Draft life-cycle

. To address RQ1, we divide the period between the cansubmission and publication of a draft into five equal time intervals (T1 - T5), each representing 20% of the period. We visualize the distribution of DAs falling into each of the periods. in Figure 1.[6]

Answer and *Question* are more common in the early phases, likely due to more new issues being raised and unresolved issues discussed.

ContextSetting and *Extension* are very frequent, increasingly so towards the end phases; we conjecture this is because those phases cover more complex issues requiring more background description.

The frequency of *ProposeAction* is stable throughout the cycle and noticeably higher than *StateDecision*. This may imply that participants prefer to discuss actionable options rather than explicitly deciding on a single one.

3.2 Different groups

To explore RQ2, we categorize the participants as: (1) authors of the draft being discussed, or not; (2) influential — following (Khare et al., 2022), having top-10% centrality in the email interaction graph — or not; (3) chairs of any IETF WG, or not; (4) everyone (all participants). Figure 2 gives a visualization of DA distributions for each group.

Authors vs. non-Authors Authors are more social, give more answers, and ask fewer questions (including clarification questions). Also, they use fewer *NeutralResponse*, *Extension*, and *ContextSetting*, indicating shorter, more focused messages. These trends imply they take a more reactive role in the discussion. Finally, they make the most decisions in the discussion, as would be expected.

Influential vs. non-Influential Influential people use *Answer*, *Agreement*, and *NeutralResponse* more, making them generally more responsive. They use less *Extension*, *ContextSetting* and Thanking, implying a concise, focused communication style. As expected, since they are in charge of the writing process, they make more decisions and propose slightly more actions.

Chairs vs. non-Chairs Similar to influential praticipants, chairs use *NeutralResponse* more than non-Chairs. However, they use more *ContextSetting* and *Extension*, and do more *Thanking*. We find this is because chairs send a lot of emails initiating and managing discussions and review assignments. Such emails are often composed of many small segments and contain a lot of these labels.

Feedback to questions We further explored how likely the different groups are to have their questions answered. From the labeled data we obtain percentages for authors (22%), chairs (51%), influential (34%), and everyone (37%). Authors have the lowest ratio, possibly because their questions are, on average, more complex. The chairs, while they tend not to ask many questions, are the most likely to to get an answer. This is expected, as it is difficult to ignore a question from someone in that position. Surprisingly, the difference between ratios of influential participants and everyone are not statistically significant.[7] Another surprising finding is that, on average, around two thirds of all questions appear to remain unanswered.

3.3 Other observations

ClarificationElicitation is almost nonexistent, implying either very little misunderstandings or un-

[5] `https://github.com/sodestream/acl2023-email-da-dataset.git`

[6] In both figures InformationProviding is omitted because it dominates the plot and obscures other trends.

[7] We used a z-test with significance level 0.05.

Label	Description	Example	Count	IAA
InformationProviding	Any type of providing information	-	7643	.86
Agreement	Agreeing with opinion or accepting a task	That's a good idea.	651	.74
Answer	Answering a question	It is 42 bytes.	655	.73
ContextSetting	Providing context before other DAs	Imagine the case when ...	2212	.25
Disagreement	Disagreeing with opinion on rejecting a task	I don't think so.	365	.68
Extension	Natural continuation of the previous one.	Moreover, it's faster.	3007	.65
NeutralResponse	Response without clear (dis)agreement	Your idea seems interesting.	2066	.71
ProposeAction	Propose an actionable activity	We should update the text.	2225	.65
StateDecision	Explicitly express a decision	We will incorporate this.	359	.63
InformationSeeking	Any type of seeking information	-	1146	.84
ClarificationElicitation	Expresses need for further elaboration.	Could you explain again ...	326	.29
Question	Any type of question.	How big is the header?	865	.86
Social	Social acts (thanking, apologizing etc.)	-	1040	.67
Thanking	Conveying thanks.	Thanks for the comment.	249	.98

Table 1: Labels at the higher (bold) and lower levels of the taxonomy with corresponding counts and inter-annotator agreement.

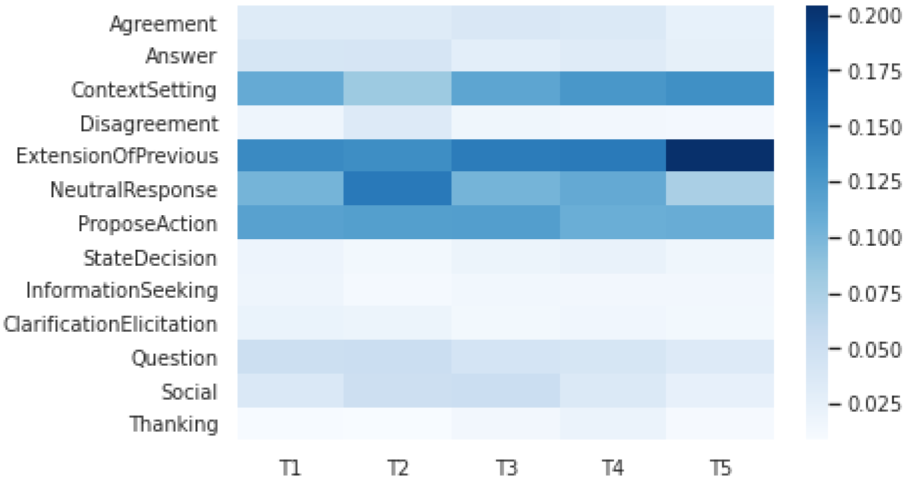

Figure 1: DA distribution across time. Each column is a DA distribution in a particular time period of the draft life-cycle. Colors convey the probability mass assigned to a DA in emails from that period.

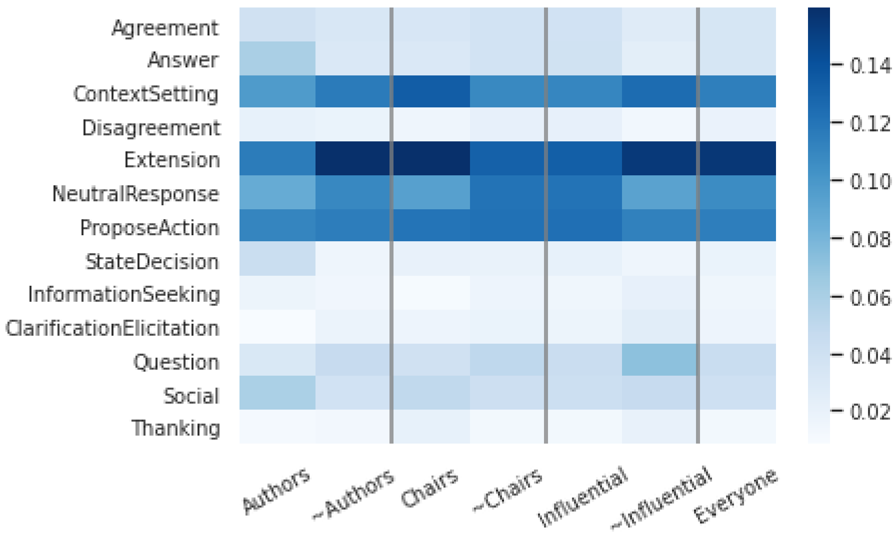

Figure 2: DA distribution for different groups. Each column is a DA distribution of a particular group. Colors convey the probability mass of a DA for that group.

willingness to explicitly voice it. Research on misunderstandings in dialog (Aberdeen and Ferro, 2003) implies it is likely the latter.

Most participants tend to use *NeutralResponse*, as opposed to *Agreement* or *Disagreement*, and between the latter two they prefer *Agreement*. This tendency is confirmed by related research on agreement (Stromer-Galley and Muhlberger, 2009).

ContextSetting, *Extension*, and *NeutralResponse* are, expectedly, very frequent. This implies there are a lot of boilerplate explanations around segments with more relevant DAs.

4 Automated Dialogue Act Tagging

We provide a preliminary DA tagging model to investigate the predictability of our DA tags, and to serve as a baseline for future work. We use a hierarchical sequence model, inspired by work in DA tagging for spoken dialogue (e.g. Li et al., 2019): the input is a sequence of segments (each one a sequence of words), and the output is a sequence of predictions, one 14-dimensional vector for each input segment, representing DA probabilities.

Each input segment is encoded into a vector; we use the [CLS] token of BERT (Devlin et al., 2019). The sequence of segment vectors is then passed to a Bidirectional-LSTM (Hochreiter and Schmidhuber, 1997); each BiLSTM hidden state vector is passed through a linear layer (shared for all time steps) to produce the output prediction vector sequence. The loss function is binary cross-entropy averaged across all labels and all elements of the sequence.

The model is implemented using PyTorch (Paszke et al., 2019) and scikit-learn (Pedregosa et al., 2011). We used a learning rate of 2^{-5}, batch-size of 32, and LSTM hidden-layer size of 256. All other hyper-parameters are left at default values. We experiment with two variants of BERT: bert-base and bert-base-ietf (fine-tuned using language modeling loss on the entire IETF mail archive).

We split the data into train (60%), validation (20%), and test threads (20%). We report results on test threads by the model best on the validation threads. The input sequences for the model are the possible root-to-leaf paths in the input threads, following (e.g. Zubiaga et al., 2016).[8]

Results are given in Table 2. Predicting higher-level labels is easier, as expected. For lower-level labels, performance is worst for labels that are conceptually more subjective (as reflected by IAA) or have very few examples.

[8]This will cause segments that are part of several paths to be processed multiple times and assigned multiple label hypotheses; we take the most common label in this case.

	bert-base			bert-base-ietf		
Label	P	R	F_1	P	R	F_1
InfProviding	.89	.96	.93	.88	.97	.93
Agreement	.67	.72	.69	.47	.67	.55
Answer	.44	.40	.41	.35	.49	.41
ContextSetting	.38	.67	.49	.36	.67	.47
Disagreement	.14	.24	.17	.10	.29	.15
Extension	.64	.72	.67	.66	.62	.64
NeutralResponse	.45	.52	.48	.43	.52	.47
ProposeAction	.47	.72	.57	.44	.67	.53
StateDecision	.39	.28	.47	.19	.30	.23
InfSeeking	.85	.87	.86	.78	.84	.81
ClarificationEl.	.25	.46	.33	.21	.51	.30
Question	.78	.98	.87	.84	.88	.86
Social	.33	.67	.44	.45	.52	.48
Thanking	.75	.99	.86	.33	.92	.48
Macro-average	.53	. 66	.59	.46	63	.52

Table 2: Precision, Recall, and F1 on the test set.

Curiously, bert-base-ietf performs comparably to or worse than bert-base. We hypothesize the reason for this may be the specific language of the IETF (technical discussions). It may cause the additional language model training step to make the bert-base-ietf model forget information generally useful for DA tagging. On the other hand, this information is retained in bert-base. If this is the case, it would hurt the performance of bert-base-ietf after further fine-tuning on the DA tagging task. However, we leave investigation of this and other hypotheses for this unexpected result to future work.

5 Conclusion

We have presented a taxonomy of dialogue acts (DAs) and a labeled dataset of emails. Moreover, we provided a data analysis and a preliminary DA prediction model. We hope this dataset will be useful to facilitate further research on the interaction behavior of participants in online collaboration settings. Future work could include a more detailed investigation into the underlying reasons for the observed trends. Another possibility is looking into the interaction of DAs and the participant interaction graph as described by (Khare et al., 2022). Finally, to get further insights, it would be interesting to annotate segments of with a particular DA with additional labels, e.g., explicit/implicit for *Agreement* or different sub-types of *Question*.

6 Limitations

One of the main limitations is that we focus solely on the IETF. Consequently, we can never be completely sure how well our findings generalize to other similar organizations without further annotation.

We are also limited by not conducting a hyperparameter search on our models. We omit this step as the main goal is not maximizing performance, but rather data annotation and analysis. In a similar vein, it is likely possible to increase performance by using a more advanced model that is either trained on dialogue-like data or is specifically designed to exploit phenomena specific to dialogue (e.g., having speaker embeddings).

We also acknowledge that many emails are longer than 512 tokens which is the limit of our BERT model and thus might have been cut short. However, most of the emails do fit into this limit.

7 Ethical Considerations

The IETF conditions participation by agreements and policies that explicitly state mailing list discussions and Datatracker metadata will be made publicly available.[9] In our analysis we use only this publicly available data. We have discussed our work with the IETF leadership and confirmed it is conforming to all their policies.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was supported by the UK EPSRC under grants EP/S033564/1 and EP/S036075/1 (Sodestream: Streamlining Social Decision Making for Enhanced Internet Standards). Purver was also supported by the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103).

References

John Aberdeen and Lisa Ferro. 2003. Dialogue patterns and misunderstandings. Technical report, MITRE Corp. McLean VA.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Trung Bui and Stanley Peters. 2010. Decision detection using hierarchical graphical models. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 307–312.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex C Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. Technical report, University of Southern California Los Angeles.

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.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 156–163.

H Flanagan. 2019. RFC 8700: Fifty Years of RFCs.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Pei-Yun Hsueh and Johanna D Moore. 2007. What decisions have you made?: Automatic decision detection in meeting conversations. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 25–32.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Prashant Khare, Mladen Karan, Stephen McQuistin, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, and Ignacio Castro. 2022. The web we weave: Untangling the social graph of the IETF. In *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM)*, pages 500–511, Palo Alto, CA, USA. AAAI Press.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

[9] For details see https://www.ietf.org/about/note-well/ and the IETF privacy policy available at https://www.ietf.org/privacy-statement/.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China. Association for Computational Linguistics.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100. Citeseer.

Stephen McQuistin, Mladen Karan, Prashant Khare, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, Waleed Iqbal, Junaid Qadir, and Ignacio Castro. 2021. Characterising the IETF through the lens of RFC deployment. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 137–149.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. Technical report, International Computer Science Institute, Berkeley CA.

Jennifer Stromer-Galley and Peter Muhlberger. 2009. Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy. *Political communication*, 26(2):173–192.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2438–2448, Osaka, Japan. The COLING 2016 Organizing Committee.

A Appendix A: Comparison with the AMI dataset

In this section, we compare our dataset with the AMI dataset (McCowan et al., 2005). The counts are given in Table 3, after removing those AMI DA categories that make sense in AMI's spoken, face-to-face setting but do not exist in the email given the text modality and non-synchronous nature (e.g. Stall, Backchannel). The distributions are roughly similar. The main difference is a lot more *ClarificationElicitation* and *Answer* in AMI. The former may reflect the explicitly decision-oriented setting of AMI (actors were tasked with making design decisions on how to build a remote control, and therefore decisions and clarity were the primary focus), and/or its synchronous speech, which participants must clarify immediately (while email can be studied over more time before replying). The latter may reflect the fact that AMI is built on live face-to-face conversations, thus leaving an articulated question ignored and unanswered would be considered rude, while in email communication, this is less problematic.

B Appendix B: Computing resources

The prediction model experiments (two of them – bert-base and bert-base-ietf) were run on a single Nvidia QUADRO RTX 6000 GPU for 100 epochs each. For both experiments, one epoch took approximately 4 minutes. In preliminary experiments, we found the models with our hyperparameters need 14GB of video memory. They can, however, run with less memory with reduced batch size. Alternatively, larger batches could be emulated using several smaller batches and gradient accumulation (this is not implemented in our code).

AMI		This work	
label	count	label	count
Inform	33484	InformationProviding	7643
Assess	21391	Answer	655
Suggest / Offer	10921	ProposeAction	2225
Elicit-Inform / Elicit-Offer-Or-Suggestion / Elicit-Assessment	7191	Question	865
Comment-About-Understanding / Elicit-Comment-Understanding	2560	ClarificationElicitation	326
Be-Positive	2210	Agreement	651
Be-Negative	98	Disagreement	365

Table 3: Comparison of label distributions between AMI and the dataset proposed in this work. We consider only labels that have a rough equivalent in both datasets.

C Appendix C: Annotation details

The annotators come from diverse backgrounds but were primarily chosen as skilled linguists from the population of graduate and Ph.D. level linguistics students. They all lived in the UK and were paid an hourly wage that was slightly above average for similar tasks in the UK.