Constructing the CODA Corpus: A Parallel Corpus of Monologues and Expository Dialogues

Svetlana Stoyanchev, Paul Piwek

Centre for Research in Computing, The Open University
Milton Keynes, UK
s.stoyanchev, p.piwek @open.ac.uk

Abstract

We describe the construction of the CODA corpus, a parallel corpus of monologues and expository dialogues. The dialogue part of the corpus consists of expository, i.e., information-delivering rather than dramatic, dialogues written by several acclaimed authors. The monologue part of the corpus is a paraphrase in monologue form of these dialogues by a human annotator. The corpus was constructed as a resource for extracting rules for automated generation of dialogue from monologue. Using authored dialogues allows us to analyse the techniques used by accomplished writers for presenting information in the form of dialogue. The dialogues are annotated with dialogue acts and the monologues with rhetorical structure. We developed annotation and translation guidelines together with a custom-developed tool for carrying out translation, alignment and annotation.

1. Introduction

Many, if not most, tasks in Natural Language Processing involve some kind of transformation. For example, Machine Translation (MT) and text simplification are both kinds of Text-to-Text (T2T) transformation. They take the information expressed in a text and present it in another text which best fits the readers' needs. In MT this amounts to changing the language of the text, whereas in text simplification – a type of paraphrasing – it consists of adjusting the text to the reading skills of the reader. The use of parallel corpora for creating translation or paraphrasing models is widespread in both MT and paraphrasing.

This paper introduces a new type of corpus for a recently developed type of transformation: automated Monologue to Dialogue (M2D) transformation (Piwek et al., 2007). M2D is motivated by the observation that most information is stored in monologue form (books, papers, leaflets, etc.), whereas there is ample empirical evidence that for various purposes, specifically education and persuasion, presentation of information in dialogue form is more effective than monologue (Craig et al., 2000; Lee et al., 1998; Suzuki and Yamada, 2004). For example, Craig et al. (2000) found that when information is presented to a student as a simulated dialogue between a student and tutor, as opposed to a monologue by a single tutor, students write more in a free recall test and ask twice as many deep-level reasoning questions. Additionally, generated dialogue can be presented by teams of animated agents for information presentation and entertainment (van Deemter et al., 2008), and in the context of serious games (Core et al., 2006).

Our aim is to put M2D transformations on an empirical footing. We are creating a parallel corpus of monologues and expository dialogues (dialogues presenting information to a reader) from which M2D transformations can be learned (semi-)automatically. Our corpus, the CODA corpus[1], consists of pairs $< M, D >$ where M is a monologue and D is an expository dialogue expressing the same information as the monologue. The corpus contains various annotations on M and D, and alignment information relating spans of M to spans of D.

The CODA corpus is a resource that promises to be useful beyond research on M2D. It contains many instances where declarative information is aligned with a question-answer pair. The community of researchers on Question Generation (QG) (Rus and Graesser, 2009) who aim at generating questions from declarative statements will benefit from the CODA corpus. So far, work on both M2D and QG has relied primarily on hand-crafted transformation rules. The corpus provides a valuable resource for automating the creation of such rules and grounding them in empirical data. In particular, we are interested in making sure that rules for M2D conversion yield dialogues similar to those created by professional dialogue authors.

In this paper, we describe the construction of the CODA corpus. We discuss the dialogue act annotation scheme for expository dialogues, annotation procedure, and a task-specific tool that we developed.

2. Sources

Our corpus construction starts from a collection of professionally authored dialogues. We wrote matching monologues for these dialogues. The decision to start with dialogues was based on the fact that it is much easier to find skilled monologue than skilled dialogue authors. Following on from a pilot study (Piwek, 2008), the selection of the dialogues was based on the following criteria:

- Authors should be professional writers; preferably their work should be widely acknowledged as world class.
- The core of the corpus will be made available to the research community as an Open Source, for this reason we drew mainly on text from the Gutenberg project which permits such reuse.

[1]CODA stands for 'COherent Dialogue Automatically generated from text', see computing.open.ac.uk/coda. The project is funded by the UK's Engineering and Physical Sciences Research Council under grant EP/G/020981/1

- Dialogues should be easy to paraphrase as monologue. This meant that we selected expository dialogues (which present a description or argument) and ruled out dramatic dialogue (e.g., plays and film scripts).

Based on these criteria we selected as core dialogues Mark Twain's "What is man?" and George Berkeley's "Three Dialogues between Hylas and Philonous" from the Gutenberg library[2], supplemented with a number of fragments from copyrighted dialogues, mainly by academic authors (e.g., David Lewis and Paul Feyerabend).

3. Dialogue Annotation and Transformation

To create the CODA corpus we segment and annotate dialogue turns, write monologue snippets, and map them to the corresponding dialogue segments. This section provides a general overview of the CODA annotation scheme and instructions. For further details, we refer to the CODA annotation manual.[3]

3.1. Segmentation

The corpus annotator first partitions the dialogue turns into segments. A segment can be an entire turn or a part of a turn that expresses a distinct dialogue act. For example, *Yes. It is diligently at work...* can be split into two segments: a positive answer *Yes* and an explanation *It is diligently at work....* We achieve 91% agreement between two annotators in the turn segmentation task.

3.2. Key and Decorative Segments

In a dialogue, the interlocutors exchange information with each other. Information that is directly relevant to the main purpose of the dialogue is classified in the CODA annotation scheme as *key* information. For instance, in a dialogue which consists of a discussion about some topic (say, whether holes exist as material objects), segments that present either side of the argument are labelled as key segments. Most segments in an authored dialogue are usually key segments. They are about the topic of the dialogue and their meaning needs to be preserved in the monologue. They can be copied verbatim to the monologue or paraphrased.

Apart from key segments, authored dialogue contains *decorative* segments. A decorative segment expresses 'dialogue control acts' in terms of Dynamic Interpretation Theory (DIT) (Bunt, 2000). From the point of view of the dialogue author they are often used to create a certain effect on the audience: creating a mood, attracting attention of the reader, or embellishing a dialogue. Examples of decorative segments are utterances for managing turn taking such as *Wait!* or *Just a moment*. Decorative segments also include exchanges which concern the *social context* of the dialogue. For example, decorative social dialogue may be found at the beginning or end of a fictional dialogue where characters establish acquaintance or say farewell. Decorative dialogue segments are not translated into monologue in CODA corpus because they do not carry content.

[2]www.gutenberg.org/

[3]computing.open.ac.uk/coda/AnnotationManual.pdf

3.3. Dialogue Act Annotation

For dialogue act annotation, we focus on key segments. These are the segments that will be translated into monologue. We have taken two existing dialogue annotation schemes, DAMSL (Core and Allen, 1997) and Dialogue Games (Carletta et al., 1997), and adapted them for our purposes. Both schemes were devised for modelling task-oriented spoken dialogue. In contrast, our dialogues are typically philosophical discussions and, most importantly, they have been authored and are not spontaneously spoken. The aim of the author is to use the dialogue between two fictional dialogue agents to present an idea to the reader.

The CODA dialogue act annotation tags are listed in Table 1. We have not included tags from the DAMSL and Dialogue Games scheme which are specific to task-oriented spoken dialogue (such as *instruct* or *commit to an action*). We have merged *Init-Explain* and *Resp-Explain* into a single *Explain* tag. An explanation move in authored dialogues is often both a response and initiation. Our initial evaluation showed a poor agreement for *Init-Explain* and *Resp-Explain* tags. We also realized that for the purpose of monologue-to-dialogue translation the distinction between *Init-Explain* and *Resp-Explain* is not important. In the monologue to dialogue translation task, when we generate an *Explain* dialogue move, we envision that syntactic and semantic rule for surface-level realization of *Resp-Explain* and *Init-Explain* to be the same. Hence, we combine *Init-Explain* and *Resp-Explain* into a single *Explain* tag. Additionally, we have created some new tags, which allow us to make more fine-grained distinctions between different types of requests for information (such as requests for factual information, i.e., yes/no questions, and requests for more complex explanations). This is motivated by the important role that questions play in discussions. As in DAMSL, we allow dialogue acts to be tagged simultaneously with both forward-looking (init) and backward-looking (response) acts. Annotators are required to assign a primary dialogue act tag (whichever act they deem most characteristic of the segment) and may add a secondary tag. To speed up annotation, options for secondary tags are automatically constrained by the choice of the primary tag. For example, for a segment with a primary tag *Init-Explain*, the possible secondary tags are *Resp-Agree* and *Resp-Contradict*.

Currently, we do not require the annotators to assign dialogue acts to decorative segments. Decorative segments are mainly there to liven up the dialogue or emphasise specific information; they do not get translated into the monologue. We do, however, at a later stage plan to study the decorative segments in further detail, and possibly use them to formulate revision rules for dialogue along the lines described in Piwek and Van Deemter (2007).

3.4. Dialogue Annotation Inter-Annotator Agreement

We evaluate inter-annotator agreement between two annotators over *key* segments with matching boundaries (total number of segments is 72) for dialogue act annotation using the kappa coefficient (Cohen, 1960). First, we evaluate 2-way (k=2) agreement for individual tags occurring more than once. We use both primary and secondary dia-

Tag	Description
Explain	An explanation or a description of an idea
Initiating (forward-looking) Dialogue Acts	
Init-Factoid-InfoReq	A request for factoid information (who, when, where, which, what).
Init-YN-InfoReq	A question that syntactically requires yes/no answer
Init-Complex-InfoReq	A question that requires a complex answer or explanation (why, how, etc).
Init-Request-Clarify	A request for further explanation. *What do you mean ?* or *Explain*
Responding (backward-looking) Dialogue Acts	
Resp-Agree	Speaker shows an agreement (or partial agreement) with the previous statement.
Resp-Contradict	Speaker shows an disagreement with previous statement.
Resp-Acknowledge	Speaker acknowledges information in the previous statement without showing an opinion.
Resp-Answer-Yes	A positive answer to yes/no question.
Resp-Answer-No	A negative answer to a yes/no question. It is often followed by Init-Explain act that supports the negation: *No, ...*
Resp-Factoid	A short answer to a factoid question.
Other	A segment that does not fit into any of the above categories.

Table 1: Dialogue Act tags for expository dialogues.

Eval Type	Kappa
Individual Tags (N=72, k=2)	
Explain	.93
Init-YN-InfoReq	.95
Init-Complex-InfoReq	.92
Init-Request-Clarify	.88
Resp-Agree	.79
Resp-Contradict	.88
Resp-Answer-Yes	.88
Grouped Tags	
Overall (N=72, k=14)	.82
Init-Response-Explain (N=72, k=3)	.87
Inits(N=18, k=5)	1.0
Response (N=16, k=6)	.83

Table 2: Dialogue act inter-annotator agreement (N=number of cases, k=number of tags)

logue act tags in this evaluation. We achieve high agreement (kappa$>$0.8) for majority of the individual tags (see Table 2).

To compare our inter-annotator agreement with the agreement achieved in previous dialogue act annotation studies, we compute overall agreement and agreement of grouped tags. In this evaluation we only consider the primary dialogue act tag as it requires disjoint categories. We achieve a good overall agreement of kappa=.82, comparable with kappa=.83 in Carletta et al. (1997).

Next, we evaluate 3-way (k=3) tagging agreement by grouping all initiating and responding tags. The annotation agreement between the three groups (init, response, and explain) is kappa=.87, which is similar to Carletta's kappa=.89 between the 2-way tagging of grouped Init and Response tags. Finally, we evaluate agreement within initiating tags and within responding tags. The agreement within Initiating tags is kappa=1.0 and within responding tags is kappa=.83.

priority	RST relations
1	Explanation*(Evidence, Reason)*
2	Enablement
3	Cause
4	Evaluation *(Subjective, Inferred)*
5	Comment
6	Attribution
7	Condition-Hypothetical
8	Contrast
9	Comparison
10	Summary
11	Manner-means
12	Topic-Comment *(Problem-Solution, Statement-Response, Question-Answer, Rhetorical Question)*
13	Background
14	Temporal
15	Elaboration *(Additional, General-Specific, Example, Object-attribute, Definition)*
16	Same-unit
17	Joint

Table 4: Discourse relation tags in CODA corpus. Fine-grained relations are added *(in brackets and italicized)*.

3.5. Monologue Authoring

Once the dialogue has been segmented and annotated, the annotator composes monologue snippets which express the information of the key dialogue segments. The annotator is instructed to keep lexical and syntactic content of monologue snippets as close as possible to the corresponding dialogue segments. Groups of one or more segments (e.g., a question followed by an answer) are translated into snippets (declarative sentences).

Table 3 shows an example of an annotated dialogue aligned with a parallel monologue translation. In this example, the monologue contains five snippets. Each snippet maps to

	Utterance	Dialogue Act	Monologue snippet
YM	Do you really believe that mere public opinion could force a timid and peaceful man to –	Init-YN-InfoReq	Mere public opinion could force a timid and peaceful man to go to war.
OM	Go to war ?	Init-YN-InfoReq	
OM	Yes.	Resp-Answer-Yes	
OM	Public opinion can force some men to do ANYTHING	Explain	Public opinion can force some men to do ANYTHING.
YM	Anything?	Init-YN-InfoReq/Init-Request-Clarify	
OM	Yes – anything	Resp-Answer-Yes	
YM	I do not believe that	Resp-Contradict	
YM	Can it force a right-principled man to do a wrong thing ?	Init-YN-InfoReq	It can force a right-principled man to do a wrong thing.
OM	Yes.	Resp-Answer-Yes	
YM	Can it force a kind man to do a cruel thing ?	Init-YN-InfoReq	It can force a kind man to do a cruel thing.
OM	Yes.	Resp-Answer-Yes	
YM	Give an instance	Init-Request-Clarify	For instance , Alexander Hamilton...
OM	Alexander Hamilton ...	Explain	

Table 3: Example of a dialogue by Mark Twain's 'What is man?' segmented, annotated with dialogue act, and translated to monologue.

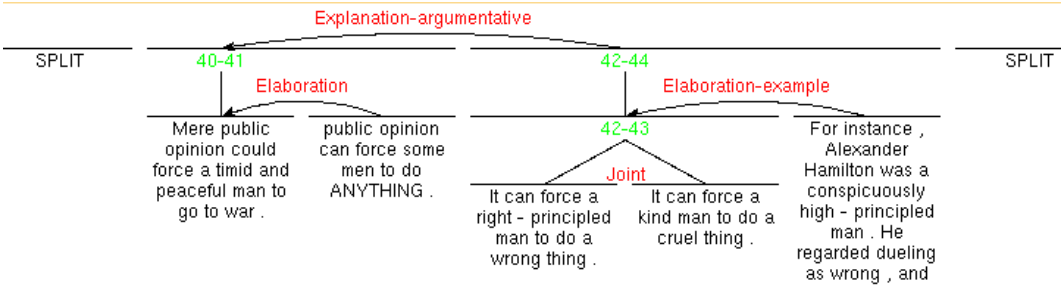

Figure 1: A section of translated monologue (from Table 3) annotated with discourse relations.

a sequence of dialogue segments. The mapping from segments to snippets is a many-to-one relation.

3.6. Tool Description

We built the CODA *D2MTool*, a graphical user interface for segmenting, tagging, and translating dialogues into monologues (see Figure 2). The main window of the D2MTool displays dialogue turns (on the left), tagged dialogue turns (middle), and monologue (bottom right). By clicking on a turn, a user opens the turn annotation window that allows segmentation and dialogue act annotation of the turn. Once the turn is segmented, the D2MTool automatically assigns a unique id to each segment. To translate dialogue segments, the user selects a set of segments and enters their ids in the mapping index text box (top right). The text of these segments then appears in the text box labelled 'Enter monologue snippet' and can be added. When the monologue snippet has been completed it can be added and appears in the bottom right snippets display.

4. Monologue Annotation

All annotations are done with our ultimate goal in mind: to create a collection of transformation rules which take as input patterns in monologue (syntactic and discourse structure) and transform the underlying monologue into sequences of dialogue acts. Such rules can then be used for transforming monologue automatically into dialogue. We also plan to use off-the-shelf syntactic parser and co-reference resolution tools to annotate syntactic structure and co-reference in the monologue.

4.1. Discourse Structure Annotations

The monologue text is manually annotated with discourse structure following Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Carlson et al., 2001). There are, however, also three differences between the CODA discourse annotations and RST. The first difference is in the tag set: we use coarse-grained tags for majority of RST relations in order to balance between tag diversity and burden on the annotators. For the relations occurring more frequently in our corpus (Explanation, Evaluation, Topic-Comment, and Elaboration), annotators had an option of using fine-grained tags. However, when an annotator is not sure which fine-grained tag to assign, s/he may back-off to a coarse-grained tags. This decision was inspired by the annotation scheme of Penn-Treebank (Prasad and others, 2008) where the annotators choose one out of

	Utterance	Dialogue Act	Monologue snippet
Example 1. Do not split the monologue snippet into EDUs			
OM:	As a rule it will listen to neither a dull speaker nor a bright one. It refuses all persuasion.	Init-Explain	As a rule it will listen to neither a dull speaker nor a bright one. It refuses all persuasion
Example 2. Split the monologue snippet into two EDUs			
OM YM	He felt well? One can not doubt it	Init-YN-Request Resp-Answer-Yes/ Resp-Explain	[One can not doubt that] [he felt well]

Table 5: Two examples of translated dialogue from Mark Twain's 'What is man?'.

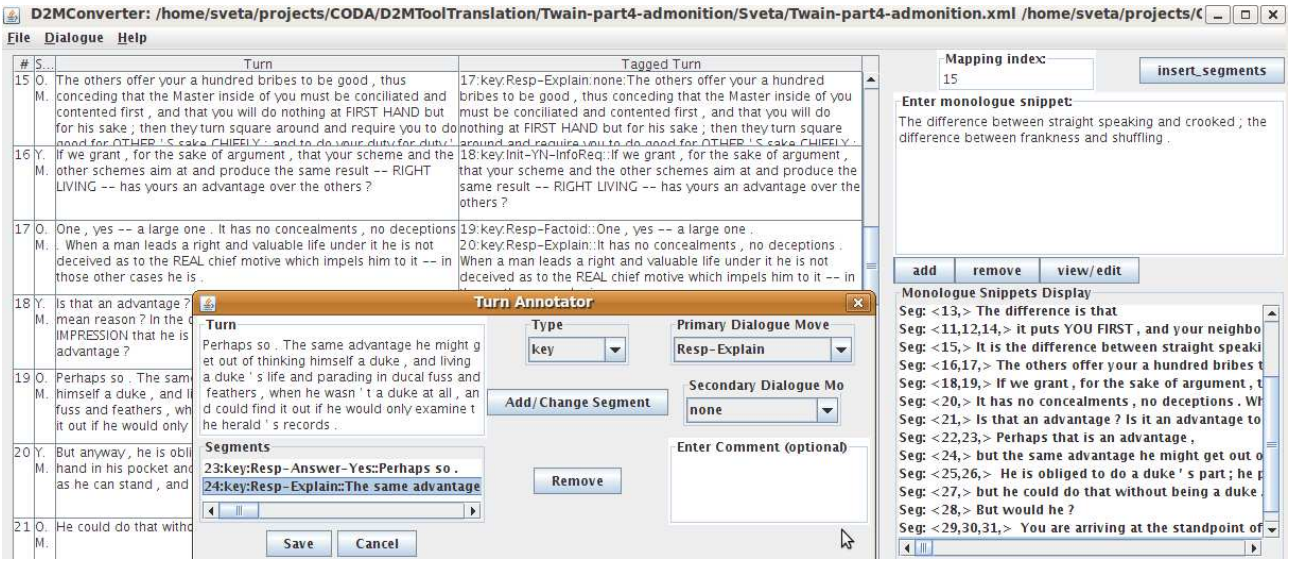

Figure 2: CODA D2MTool: graphical user interface for annotating dialogue and translating it into monologue.

three levels of granularity for each case. Discourse annotation tags used for annotating the CODA corpus are listed in Table 4. The second difference with RST was inspired by Wolf and Gibson (2005) who observe that the discourse structure underlying coherent text is not always a proper tree. A node in monologue structure may be a parent for multiple other nodes. The third difference with RST is that we do require a single tree to cover the entire monologue text. Monologue in our corpus is a direct translation from dialogue. Dialogue turns are grouped in order to be translated into a coherent paragraphs of monologue. The dialogue translator/annotator identifies these indicating *SPLIT*s between paragraphs. A discourse annotator labels discourse relations only within paragraphs, not between paragraphs. Consequently, the discourse structure of CODA monologues is a sequence of RST trees.[4]

We use RST annotation tool (O'Donnell, 2000) for manual annotation of discourse structure. Figure 1 shows the discourse structure of the monologue segments in Table 3.

Monologue snippets may contain multiple clauses (see Table 5). These clauses should be split into separate elementary discourse units (EDUs) according to RST. In CODA, we split a monologue snippet into EDUs only if the snippet maps to multiple dialogue segments. For the CODA corpus we are interested in dialogue-to-monologue mappings where the dialogue side involves changes of speaker. In the first example in Table 5 the monologue snippet contains multiple clauses. It is not segmented because it maps to a single *Explain* dialogue segment. In the second example, the snippet is segmented into EDUs because it maps to two dialogue segments. This example creates a mapping between two dialogue segments and a discourse relation in the monologue.

RST relations	kappa
Contrast	.87
Elaboration	.42
Explanation	.28
Explanation+Elaboration	.60
Evaluation	.61
Attribution	1.0
Condition	.62
Topic-Comment	.73
Overall	.62
Overall (merged Exlanation and Evaluation)	.68

Table 6: Inter-annotator agreement based on N=52 tags by two annotators

[4]Initially we tried to annotate relations for the whole monologue text; we achieved, however, extremely low agreement on higher levels of discourse structure.

4.2. Monologue Inter-Annotator Agreement

Two discourse annotators labelled monologue translation of 85 turns from Twain's dialogue. Prior to discourse annotation the dialogue was translated and segmented by one of the annotators. In total 52 labels were assigned by both annotators relating matched spans of the monologue. Table 6 shows inter-annotator agreement kappa values for the coarse-level tags that occurred more than once in the corpus and overall agreement.[5] The overall agreement between two annotators reaches a moderate kappa=0.62. We observe the highest disagreement between *Elaboration* and *Explanation* tags which in isolation reach a very low kappas of 0.42 and 0.28 respectively. When the two annotators discussed the disagreements, they realized that the relations in the cases of disagreement between *Explanation* and *Elaboration* are ambiguous. Hence, we merge *Explanation* and *Elaboration* tags. The overall agreement reaches kappa=0.68.

4.3. Current Status

We have annotated and translated to monologue 800 turns from the CODA dialogue corpus. We have manually parsed with discourse structure monologue translations of 259 turns. Figure 3 shows distribution of dialogue act tags in the dialogue annotations. Figure 4 shows distribution of RST relations in monologue-to-dialogue mapping.
We aim to translate and annotate a total of 1000 turns by May 2010.

5. Conclusion

We described the CODA corpus, a parallel corpus of dialogues and expository monologues. Collection of the CODA corpus is a first step towards data-driven automated generation of dialogues from text. The corpus will also be useful for the Question Generation task. To construct the corpus, we designed a dialogue act annotation scheme specifically for expository dialogues adapting existing dialogue annotation schemes. We also developed the D2MTool for writing aligned monologue for expository dialogue. We achieved good inter-annotator agreement for segmentation and dialogue act tagging tasks and reasonable agreement for (RST) discourse annotation of monologue. We describe detailed evaluation of our dialogue and monologue annotation schemes and show examples of analysed dialogues and translated monologue.

6. References

H. Bunt. 2000. Dialogue pragmatics and context specification. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*, volume 1 of *Natural Language Processing*, pages 81–150. John Benjamins.

J. Carletta, A. Isard, and J. C. Kowtko. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31.

L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.

J. Cohen. 1960. A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20:37–46.

M. Core and J. Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme. In *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machine*.

M. Core, D. Traum, H. C. Lane, W. Swartout, S Marsella, J. Gratch, and M van Lent. 2006. Teaching negotiation skills through practice and reflection with virtual humans. *Simulation: Transactions of the Society for Modeling and Simulation*, 82:685–701.

S. Craig, B. Gholson, M. Ventura, A. Graesser, and the Tutoring Research Group. 2000. Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education*, 11:242–253.

J. Lee, F. Dinneen, and J. McKendree. 1998. Supporting student discussions: it isn't just talk. *Education and Information Technologies*, 3:217–229.

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

M. O'Donnell. 2000. RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory. In *Procs of International Natural Language Generation Conference (INLG 2000)*, Mitzpe Ramon, Israel.

P. Piwek and K. van Deemter. 2007. Generating under Global Constraints: the Case of Scripted Dialogue. *Journal of Research on Language and Computation*, 5(2):237–263.

P. Piwek, H. Hernault, H. Prendinger, and M. Ishizuka. 2007. T2D: Generating Dialogues between Virtual Agents Automatically from Text. In *Intelligent Virtual Agents: Proceedings of IVA07*, LNAI 4722, pages 161–174. Springer Verlag.

P. Piwek. 2008. Presenting Arguments as Fictive Dialogue. In *Proceedings of 8th Workshop on Computational Models of Natural Argument (CMNA08)*, Patras, Greece, July.

R. Prasad et al. 2008. Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

V. Rus and A. Graesser, editors. 2009. *The Question Generation Shared Task and Evaluation Challenge*. The University of Memphis. Available at: http://www.questiongeneration.org/.

S. V. Suzuki and S. Yamada. 2004. Persuasion through overheard communication by life-like agents. In *Procs of the 2004 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 225–231, Beijing, China, September.

K. van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann. 2008. Fully generated

[5] We did not have enough data to evaluate fine-grained agreement.

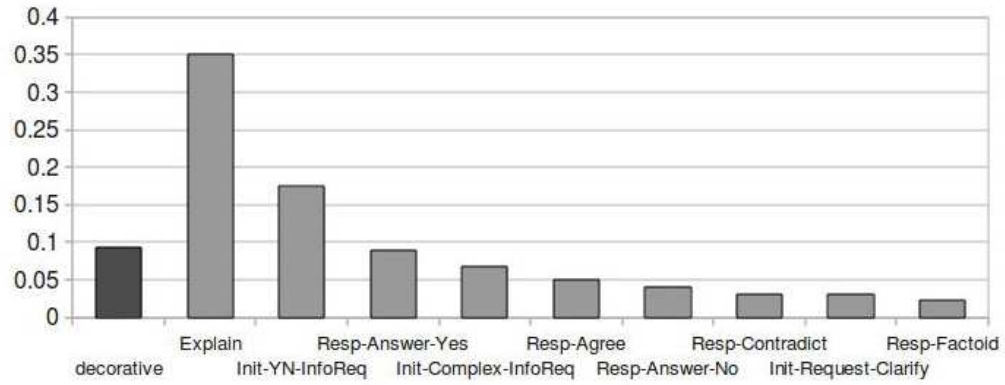

Figure 3: Distribution of Dialogue Acts in CODA corpus

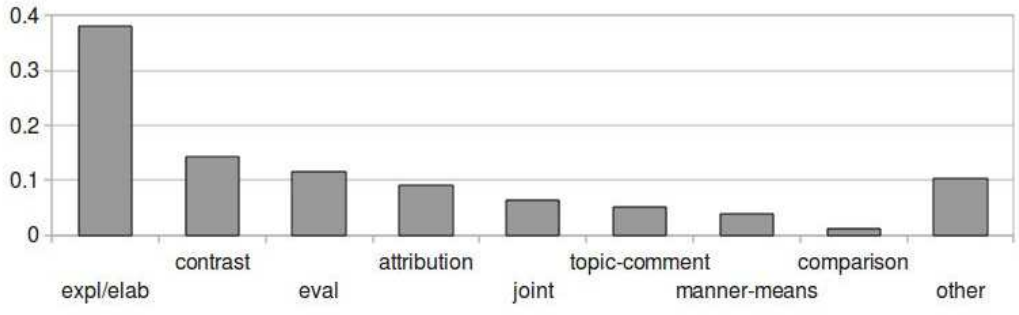

Figure 4: Distribution of RST relations in CODA corpus

scripted dialogue for embodied agents. *Artificial Intelligence Journal*, 172(10):1219–1244.

F. Wolf and E. Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 2(31):249–288.