

# First Steps Towards the Automatic Construction of Argument-Diagrams from Real Discussions

Daan Verbree<sup>a</sup>, Rutger Rienks<sup>a,1</sup>, Dirk Heylen<sup>a</sup>

<sup>a</sup> *Human Media Interaction (HMI)*

*University of Twente, Enschede, The Netherlands*

*{Verbree, Rienks, Heylen}@ewi.utwente.nl*

*Home page: <http://hmi.ewi.utwente.nl>*

**Abstract.** This paper presents our efforts to create argument structures from meeting transcripts automatically. We show that unit labels of argument diagrams can be learnt and predicted by a computer with an accuracy of 78,52% and 51,43% on an unbalanced and balanced set respectively. We used a corpus of over 250 argument diagrams that was manually created by applying the Twente Argument Schema. In this paper we also elaborate on this schema and we discuss applications and the role we foresee the diagrams to play.

## 1. INTRODUCTION

Argumentation has been proposed as constituting human kind's primary means of making progress [30]. It is pervasive in everyday life and plays an important role in human communication. Argumentation is situated in discussions, conversations and meetings, the arenas where one argues with another and one or more sides are attempting to win the approval of the opponent or of a designated audience.

Within organizations the outcomes of conversations or meetings are normally not much more than what participants are able to recall. When lucky some notes were taken, or more formal meeting minutes were made with a list of action items. Generally, a lot of energy and information that has been put into the actual outcome is never seen again.

We have tried to find an approach that is able to capture the lines of the deliberated arguments in meeting discussions. This approach, the TAS-schema, was introduced in [20] and promises to be a valuable technique for capturing organizational memory. The structure that the argument trees encapsulate reveals information about the trail or path that has been taken in a meeting. It shows the line of reasoning at specific moments in time. The method can aid querying and summarization systems and is being used in meeting browsers (See fig 1). The possibility of preserving the arguments and their coherence relations for future explorations make them potentially valuable documents that contain a tacit representation of otherwise volatile knowledge [2,16].

---

<sup>1</sup>Correspondence to: Rutger Rienks, University of Twente, Human Media Interaction, PO BOX 217, 7500 AE Enschede. Tel.: +31-53-4893740.

For end users of the representations, argument diagrams constitute a representation of the content of a conversation that leads to quicker comprehension, deeper understanding. They enhance the ability to detect weaknesses or flaws in the argumentation [22,10]. Furthermore it has been claimed that they aid the decision making process and that they can be used as an interface for communication to maintain focus, prevent redundant information and to save time [35,31].

In this paper we present our initial research efforts in this area. Before we elaborate in more detail on how we created a corpus of annotations in Section 3, Section 2 provides an introduction of the TAS-schema. Section 5 is devoted to the learnability of (a subset of) the schema and investigates if an automatic tagger can one day produce the actual schemes autonomously.

## 2. The Twente Argument Schema

The Twente Argument Schema (TAS) is a schema designed to define argument diagrams for meeting discussion transcripts. Following most of the existing diagramming techniques, application of the method results in a tree structure with labelled nodes and edges. The nodes of the tree contain complete speaker turns or parts of speaker turns whereas the edges represent the type of relation between the nodes. The complete label set is shown in Table 1.

Node labels	Relation labels
Statement	Positive
Weak statement	Negative
Open issue	Uncertain
A/B issue	Request
Yes/No issue	Specialization
	Elaboration
	Option
	Option exclusion
	Subject-to

**Table 1.** The labels of the Twente Argument Schema

The TAS trees are away to capture the most important conversational moves in dialogues in which participants discuss the pros and cons of certain solutions to a problem, marking the arguments in favor of or against the various solutions. TAS distinguishes acts in which issues are raised (questions put forward) from statements in favor of a particular position. The schema allows one to distinguish strong from weak statements. Three types of issues can be marked: open issues, issues for which a choice of solutions is presented, and yes/no issues. There are various kinds of relations that are marked. In many cases statements are not simply supporting or undercutting other statements but rather are (near) synonymous. So, besides a marking for positive/negative, also relations such as restatements, specializations or generalizations have been introduced. More details on the nodes and relation labels are provided below.

TAS was constructed in a way that it preserves the conversational flow. By applying a left-to-right, depth-first walk through the resulting trees, the reader is able to read the nodes as they unfolded in time. This is realized by assuring that in principle every next contribution of a participant becomes a child of the previous contribution, unless the current contribution relates more to an ancestor. The resulting diagrams thus provide a comprehensive overview of the discussion relating the contributions from the individual participants. For a video about the TAS-schema and its applications see: [http://hmi.ewi.utwente.nl/video4ami/UT\\_argumentation.wmv](http://hmi.ewi.utwente.nl/video4ami/UT_argumentation.wmv). An example of a TAS argument diagram, embedded in a meeting browser application, is shown in Figure 1.

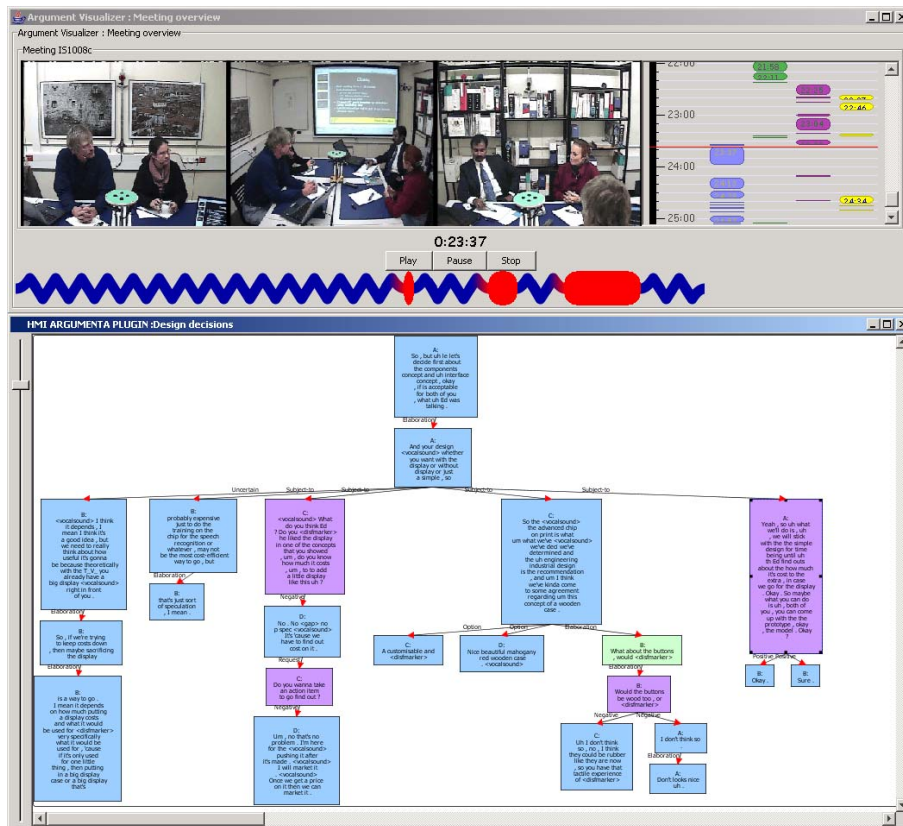


Figure 1. TAS Argument Diagrams in use as part of a meeting browser.

## 2.1. Related theories

There exist a number of different theories for labelling the contents of conversations. The TAS schema combines insights from various approaches among which are schemes for semantic and pragmatic relations between utterances such as Rhetorical Structure Theory (RST) [13], Dialog Act (DAs) annotation schemes [3], and theories or schemes that focus on the argumentative relations such as Toulmin's model [24], and the IBIS scheme [12]. For a more elaborate background about the motivations and ancestry of the

schema the reader is referred to [20] and [26]. In the following paragraphs we introduce TAS in more detail.

## 2.2. The Unit Labels

The content of the nodes are labeled with speech acts. If an utterance contains more than one act, they are split up into more than one node. In line with [7] backchannel utterances such as ‘uhhuh’ and ‘okay’ are filtered out and to be neglected, since they are generally used by listeners to indicate they are following along, and not necessarily indicating (dis)agreement. The nodes in our model are labeled either as issues or as statements.

Issues can also be found in the IBIS model. There, they are represented as questions [12] as they can be seen as utterances with a direct request for a response. Kestler distinguishes two fundamental types of question with respect to conversational moves. These are *yes-no questions* and *why questions* [11]. A yes-no question admits only two kinds of answer, be it either supportive, or negative but rules out the uncertainty *option* ‘I don’t know’. The *why questions* are a subclass of a more general type of *open question*. The number of positions participants can take on such an issue depends on the set of possible options enabled by the type of question or issue.

In our Schema we have defined three different labels for our nodes to represent the issues: The ‘**Open issue**’, the ‘**A/B issue**’ and the ‘**Yes-No issue**’. The open issue allows any number of possible replies possibly revealing positions or options that were not considered beforehand. This in contrast with the A/B issue, that allows participants to take a position for a number of positions which should be known from the context (c.f. ‘Would you say ants, cats or cows?’). The yes-no issue, in line with the yes-no question directly requests whether the participants positions agree or disagree with the issue. A *why question* in TAS is modelled as an open question with a clarification relation (see below).

The positions that participants take are generally conveyed through the assertion of a **statement**. The content of a statement always contains a proposition which can be a description of facts or events, a prediction, a judgement, or an advice ([27]). Statements can vary in their degree of force and scope. Meeting participants may indicate that they are not sure if what they say is actually true. In [24] *qualifiers* provide an indication of the force of *claims*. As [28] points out, the force of an argument can also be derived from lexical cues such as the words ‘likely’ and ‘probably’. Such statements, in which the speaker does not commit himself fully to the opinion are labeled as ‘**weak statements**’ in TAS.

## 2.3. The Relation Labels

In normal texts and conversations, the statements forms a coherent whole, partly, because they are connected through semantic and pragmatic relations which in Rhetorical Structural Theory are called *rhetorical relations* [13]. The TAS schema concentrates on typical argumentative relations in conversations.

When engaged in a discussion or debate, the elimination of misunderstandings is one of the prerequisites to understand each other and hence to proceed [15]. According to Neass, participants in a discussion eliminate misunderstandings by clarifying, or spec-

ifying their statements. These moves can e.g. be observed in the criteria definition phase, of the decision making process.

The **'Generalization/Specialization'** label can be applied when a particular issue generalizes or specializes another issue. The contribution 'Which animal is the most intelligent?' can be specialized with the following proceeding contribution 'Is an ant or a cow the most intelligent animal?' which again can be specialized if one for instance asks 'Are ants the most intelligent animal?'. It is also possible that a person is not satisfied with the information or the argument explained. He can then explicitly invite the previous speaker to elaborate on his earlier statements. For these situations we define the relations **'Request'**. The **'Elaboration'** label is used if a person continues his previous line of thought and adds more information to it.

Whenever an issue is raised, an exchange of ideas about the possible solutions occurs in the decision making process. As questions call for answers, issues call for opinions expressed through statements. Whenever a statement is made as a response to an open-issue or an A/B-issue it might reveal something about the opinion of the participant on the solution space. In general a participant provides an **'Option'** to settle the issue at hand. For example when a speaker asks 'Which animal is the most intelligent?' and the response from someone else is 'I think it's an ant' the option relation is to be applied. The opposite of the option relation is the **'Option-exclusion'** relation, and it is to be used whenever a contribution excludes a single option from the solution space.

With respect to a yes/no-issue the contributions that can be made are not intended to enlarge or to reduce the solution space, but to reveal one's opinion to the particular solution or option at hand. Contributions from participants are either supporting, objecting to the issue, or express uncertainty. For this purpose the labels **'Positive'**, **'Negative'** and **'Uncertain'** are introduced. The positive relation can exist for example between a yes/no-issue and a statement that is a positive response to the issue or between two statements agreeing with each other. When one speaker states that cows can be eliminated as being the most intelligent animals and the response from another participant is that cows don't look very intelligent, then the relation between these statements is positive. The negative relation is to be applied in situations where speakers disagree with each other or when they provide a conflicting statement as a response to a previous statement or a negative response to a Yes/No-issue. In a case where it is not clear whether a contribution is positive or negative, but that there exists some doubt on the truth value of what the first speaker said, the uncertain relation is used.

The final relation of our set is applied when the content of a particular contribution is required in order to figure out whether another contribution can be true or not. We termed this the **Subject to** relation. It is related to the concession relation in Toulmin's model. It is applied for example in the situation where someone states 'If you leave something in the kitchen, you're less likely to find a cow' and the response is 'That depends if the cow is very hungry'.

### 3. Creating a corpus of Meeting Discussions

TAS was initially devised to create argumentation diagrams for the meetings recorded in the Augmented Multiparty Interaction (AMI) Project. The AMI project is focused on the use of advanced signal processing, machine learning models and social

interaction dynamics to improve human-to-human communications. In particular the development of tools and models that provide insight into the decision making process are of primary concern. Over one hundred hours of meetings was captured for the AMI meeting corpus. All the meetings followed a script that described the global theme and the global structure of the meeting. There were no constraints on the way participants gave content to their contribution.

The recordings consist of four people meetings constituting a design team from a small company, RealReactions. In these meetings, the participants, take four different roles: a project manager (PM), user interface specialist (UI), marketing expert (ME), and industrial designer (ID). The teams design a new kind of remote control from start to finish over a series of four meetings. Transcriptions were created for all the meetings in the AMI corpus, following strict annotation guidelines [14]. For more information about the AMI corpus, see [4].

To perform the manual TAS annotations, the annotation tool *ArgumentA* was designed by using a number of components described in [18]. *ArgumentA* allows annotators to select text on a transcription-view pane and label them. The label is assigned by selecting the unit text with the mouse from the transcription pane and then pressing a button that makes a label selection window pop-up from which the unit label can be picked. The labelled units appear on a canvas where they can be attached to the graph via an intuitive drag and drop interface. Once attached, a popup window appears from which the relation-label can be chosen. The resulting trees can be saved in different XML formats.

Three annotators were trained in several iterations. Apart from collectively developing the schema, elaborate discussions were held after a number of training sessions about when and why to pick a particular label in that particular case. The corpus, as it stands, comprises a total of 256 annotated discussions (diagrams) including over 5000 unit labels and 5000 relation labels.

#### 4. Reliability of the TAS Schema

With respect to the issue of reliability one should first note that it is very well possible to end up with several diagrams from one discussion as there are likely to be more than one possible interpretation. [33] for instance showed that various different argument diagrams can be instantiated by one single text. Moreover, in Rhetorical Structure Theory (RST) [13], which addresses similar issues as the TAS scheme, the suggestion is made that the analyst should make *plausibility judgements* rather than absolute analytical decisions, implying that more than one reasonable analysis may exist.

To measure the reliability of the scheme we therefore compared the unit labels on pre-segmented discussions for four meetings (12 discussions) between two annotators. The reliability issue for the relation part of the scheme is still under investigation. It turned out that, especially in first trials the value of Cohen's kappa ( $\kappa$ ) [6] were rather low (0.50) as a lot of confusion existed amongst the labels 'other' and 'statement'. This was resolved by a consensus definition, after which  $\kappa$  rose to a more acceptable value (0.87).

We also experimented with other ways to obtain reliability score based on more data. We applied techniques comparable to those introduced in [23], by setting out the results

of a classifier trained on (unit label) annotations of one annotator against the values provided by another annotator. (See Section 5).

## 5. Tagging the TAS-unit labels

Eventually we aim to build a system that can automatically detect discussion segments, tag individual contributions with TAS-unit-labels, depict and label the relations between the units using the TAS-relation-labels and generate a visualization of the argument diagram. In this section we report on our first experiments related to the automatic classification of the TAS unit labels.

### 5.1. Features

Except for the *lastlabel* feature, we only used lexical features.

*? and OR* A good indicator for an issue is a question mark. The *?-feature* gives a binary value whether a question mark is present or not. If a question mark is available, the number of times the word *or* appears is counted and used as a feature. (If the classification is based on transcripts derived from automatic ASR, a substitute for the question mark feature is needed.)

*Length* The length (number of words) of each segment is a feature. This feature helps to make a distinction between the *statement* and *other* labels.

*Last Label* Since discussions have the property of having some coherence we might expect that given the label of a segment the conditional chance of the label of the next segment might differ from the unconditional chance. Therefore the *lastlabel* feature, which is a bigram of the previous two labels, is used.

*N-gram points* The n-gram-point feature is used to reduce the number of features. At first, all bi-, tri- and quadri-grams are computed for all segments. Then, for each label a predictivity score is computed and the X most predictive n-grams are selected. The predictivity score is equal to the product of the times the ngram occurs in nodes labeled X and the part of this 'ngram-space' occupied by nodes of type X. For example, the score for the ngram 'what do you' (see table 2) for type *statement* is  $\frac{3}{3+0+100+97+2+0} \times 3 = 0.045$ .

Using the ngrams selected points, an utterance is assigned ngram points by computing all ngrams in an utterance and enumerating all the occurrences of all ngrams per order and label. If for example the trigrams listed in Table 2 are found in an utterance and the occurrences of the ngrams in the training set are as shown in the table, than this utterance will get 69 points for the *statement - trigram* feature, 31 for the *weak statement - trigram* feature and so on.

*POS-ngram points* The POS n-gram-point features are quite similar to the n-gram point features. But instead of attributing points to words, points are attributed to n-grams of Part-of-Speech tags.

trigram	statement	weak statement	open issue	a/b issue	y/n issue	unknown
what do you	3	0	100	97	2	0
do you think	3	1	97	92	100	0
we have to	63	30	50	1	93	4

**Table 2.** Examples of a trigrams found in an utterance and available in the training set

Perl scripts were used to extract the features *? and OR*, *Length*, and *Last Label* from our XML-format. The construction of n-grams was done using the N-gram Statistic Package (NSP) [1]. Using the Stanford Part-of-Speech tagger all segments were tagged to make POS-n-gramming possible [25].

## 5.2. Baseline

The corpus as it stands is unbalanced, consisting of 4245 *statements*, 199 *weak statements*, 244 *open issues*, 72 *a/b issues*, 460 *yes/no issues* and 3061 *others*. As a baseline we have used the implementation of a one-rule classifier resulting in a correct score of 69.1%. To see how our features would perform on a balanced corpus we also constructed a balanced corpus, having an equal number of nodes for each unit type. The baseline was again computed using a one-rule classifier, which resulted in an accuracy of 28.33%.

## 5.3. Results

We tried out different Machine learning techniques to produce our results, but looked into most detail at Weka’s **J48** implementation of the C4.5 decision tree algorithm [17], since this classifier gave the best results as a baseline classifier compared to seven other classifiers available in Weka. Furthermore Weka’s **DecisionTable** and **MultilayerPerceptron** were used on our most promising results. All our results were obtained after a 10 fold cross-validation. Here we only present our best results, a more extensive presentation of experiments and results can be found in [32].

Our best result on the unbalanced corpus is 78.52% which shows an improvement of 9.4% on the best baseline. The combined confusion matrix produced by the J48, (Table 3) shows that improvement could be obtained by features that distinguish between utterances with the label *statement* or *unknown*. The table also shows that a label such as *ab\_issue* is often incorrectly classified as it has only few occurrences.

a	b	c	d	e	f	< -- classified as
19	15	22	1	0	15	a = ab_issue
7	116	47	9	0	65	b = open_issue
8	31	3722	388	36	60	c = statement
1	9	668	2365	2	16	d = unknown
0	2	162	21	11	3	e = weak_statement
15	45	121	9	1	269	f = yn_issue
header						

**Table 3.** Confusion matrix of unbalanced J48-classifier on our best result

On the balanced corpora our best result was 51.43% which shows an improvement of 23.1% on the best baseline.

#### 5.4. Elaborating on The Reliability Issue

In section 4  $\kappa$ -measures were computed for the TAS annotation of the HUB corpus. Two problems met there were the small amount of discussions that could be compared and the absence of utterances of type *A/B issue* in each annotation. To get more insight in the reliability of our corpus we performed experiments where the J48 classifier was trained using parts of the corpus annotated by one annotator (row) and was tested on a part of the corpus annotated by another annotator (column). This resulted in the performances shown in table 4. When both training and test sets were picked from the same annotator, we used 10-fold cross-validation.

Trained / Tested on	Annotator 1	Annotator 2	Annotator 3
Annotator 1	84.4%	75.7%	70.3%
Annotator 2	75.6%	79.5%	66.2%
Annotator 3	67.0%	66.2%	82.2%

**Table 4.** Performance amongst annotators

Such a table presents an alternative view on the reliability scores.

## 6. Discussion and Future work

### 6.1. Relation with DA-Tagging

The classification task described in this paper is very similar to dialog-act tagging. Research in this field mostly concentrates on cues that are either manually [8] or automatically [19] selected. The biggest difference for our approach in comparison to earlier dialogue act classifying approaches is the use of an ngram selection method. This method selects the most predictive ngrams from the total set of ngrams acquired. We have also experimented with *compressed* feature sets. The compression decreases the size of our feature vector and therefore also decreases our computing time. This of course, by itself not an advantage, unless we maintain accuracy. In addition to the compression, we also made use of n-grams of POS-tags as has previously been done in research on the generation of backchannels in a spoken dialogue system [5]. Using the same ngrams an accuracy of 78.52% was obtained without making use of compression and a result 77.20% when using compression. These results are based on the use of the J48 classifier.

### 6.2. Research on other ngram-selecting methods

Our work has mostly concentrated on ngrams of words and POS-tags. Results of the experiments show that for each classifier the ngram-selecting method strongly influences the performance. More research on scoring algorithms might result in better ngram selection methods and therefore a better performance on the classification task. It is not

just the selection of the right ngrams that influences the performance of our classifiers based on ngrams, however. Also the points attributed to a feature when a ngram is present are important. In our study we have used the number of occurrences of an ngram as a feature value. It might be worth the effort to research other possible values one could assign to an ngram.

### 6.3. Researching the punctuation features

The use of the presence or absence of a question mark as a feature could be regarded as a form of ‘cheating’, since in automatic speech recognition it is very hard to recognize whether an utterance is a question or not and thus deciding on placing a question mark in the output or not (See e.g. [9]). Since we like to have a classification of a discussion using TAS to be applicable to discussion transcribed using automatic speech recognition we are considering the omission of this particular feature. Ongoing work investigates the influence of the *? and or* feature on the performance.

### 6.4. Applications

A plug-in has been developed for the JFerret meeting browser [34]. Users are able to access the discussions depicted on a meeting time line. For each discussion the resulting argument diagram appears allows a quick grasp of the content of the on-going discussion. Clicking on the nodes in the diagram shifts the browser directly towards the corresponding moment in the meeting.

Eventually the possible applications for meetings annotated with the TAS schema are endless. They can be used for automatic summarization purposes, or aid processes aiming to find out who adhered to a specific opinion at any given moment. They can be used to see who proposed the accepted solution, or who objected to most of the discussed points. Managers can use the diagrams to investigate what went well or wrong in the discussion and which arguments were made in favor or against a specific proposal. For more information about the sorts of applications we foresee to emerge refer to [21].

### 6.5. Future Work

There are currently three lines of research that we are engaged in with respect to the Argumentation Schema.

Up till now we have focused on node classification only. We are currently working on relation classification as well. Our first approach to the classification of relations are discussed in [32].

In the end, the system we would like to have the system work in real time. We are therefore considering to run tests directly on the ASR output.

Finally, investigations have started to measure the actual benefit of the use of argument diagrams in a meeting browser. Does presenting a Argument Diagram really improve the system? (i.e. are user queries answered quicker with a higher satisfaction rate?) This is certainly an important topic [29].

## 7. Conclusions

This paper showed some of the first steps we have taken to derive at the automatic generation of argument diagrams. A corpus containing over 250 argument diagrams deriving from real-meeting discussions has been created. Machine learning experiments on automatic tagging the unit-labels resulted in a performance of 78.52% on our unbalanced and an average of 51.43% on our balanced test set using a J48 classifier.

## 8. ACKNOWLEDGEMENTS

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-188).

## References

- [1] S. Banerjee and T. Pedersen. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February 2003.
- [2] S. Buckingham Shum. Negotiating the construction and reconstruction of organisational memories. *Journal of Universal Computer Science*, 3(8):899–928, 1997.
- [3] H.C. Bunt. Conversational principles in question-answer dialogues. *The theory of questions*, 1979.
- [4] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*, 2005. AMI-108.
- [5] N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 51–58, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
- [6] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (20):37–46, 1960.
- [7] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 669–676, 2004.
- [8] J. Hirschberg and D. Litman. Empirical studies on the disambiguation of cue phrases. *Comput. Linguist.*, 19(3):501–530, 1993.
- [9] J. Huang and G. Zweig. Maximum entropy model for punctuation annotation from speech. *Proc. Eurospeech*, 2002.
- [10] G. Kanselaar, G. Erkens, J. Andriessen, M. Prangma, A. Veerman, and J. Jaspers. Designing argumentation tools for collaborative learning. In *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, pages 51–73. Springer Verlag, London, UK., 2003.
- [11] J.L. Kestler. *Questioning Techniques and Tactics*. McGraw-Hill, 1982.
- [12] W. Kunz and H.W.J. Rittel. Issues as elements of information systems. Working Paper WP-131, Univ. Stuttgart, Inst. Fuer Grundlagen der Planung, 1970.

- [13] W.C. Mann and S.A. Thompson. Rhetorical structure theory: toward a functional theory of text organization. *Text*, 8:243–281, 1988.
- [14] J. Moore, M. Kronenthal, and S. Ashby. Guidelines for AMI speech transcriptions. Technical report, IDIAP, Univ. of Edinburgh, February 2005.
- [15] A. Neass. *Communication and argument. Elements of applied semantics*. George Allen & Unwin Press, 1966.
- [16] V. Pallotta, J. Niekrasz, and M. Purver. Collaborative and argumentative models of meeting discussions. In *Proceeding of CMNA-05 international workshop on Computational Models of Natural Arguments (part of IJCAI 2005)*, July 2005.
- [17] J. Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann, San Mateo, CA, USA, 1993.
- [18] D. Reidsma, D.H.W. Hofs, and N. Jovanovic. A presentation of a set of new annotation tools based on the nxt api. Poster at Measuring Behaviour 2005, 2005. AMI-105.
- [19] N. Reithinger and M. Klesen. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pages 2235–2238, 1997.
- [20] R.J. Rienks and D. Heylen. Argument diagramming of meeting conversations. In A. Vinciarelli and J.-M. Odobez, editors, *Multimodal Multiparty Meeting Processing, Workshop at the 7th International Conference on Multimodal Interfaces (ICMI)*, pages 85–92, Trento, Italy, October 2005.
- [21] R.J. Rienks, A. Nijholt, and P. Barthelmess. Pro-active meeting assistants : Attention please! In *Social Intelligence Design*, Osaka, Japan, March 2006.
- [22] D. Schum and A. Martin. Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 17(1):105–152, 1982.
- [23] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. "of all things the measure is man" automatic classification of emotion and intra labeler consistency. In *ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing*, 2005.
- [24] S. Toulmin. *The uses of argument*. Cambridge University Press, 1958.
- [25] K. Toutanova, D. Klein, and C. Manning. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
- [26] E. Van der Weijden. Structuring argumentation in meetings : Visualizing the argument structure. Master's thesis, University of Twente, November 2005.
- [27] F. Van Eemeren, R. Grootendorst, and F. Snoeck Henkemans. *Argumentation*. Lawrence Erlbaum Associates, 2002.
- [28] F.H. Van Eemeren. A glance behind the scenes: The state of the art in the study of argumentation. *Studies in Communication Sciences*, 3(1):1–23, 2003.
- [29] T.J. Van Gelder. How to improve critical thinking using educational technology. In *Proceedings of the 18th annual conference of the Australasian Society for Computers in Learning in Tertiary education*, pages 539–548, 2001.
- [30] T.J. van Gelder. Argument mapping with reason!able. The American Philosophical Association Newsletter on Philosophy and Computers, 2002.
- [31] A. Veerman. *Computer-supported collaborative learning through argumentation*. PhD thesis, University of Utrecht, 2000.
- [32] A. T. Verbree. On the structuring of discussion transcripts based on utterances automatically classified. Master's thesis, University of Twente, June 2006.
- [33] D.N. Walton. *Argument Structure, A pragmatic Theory*. University of Toronto Press, 1996.
- [34] P. Wellner, M. Flynn, and M. Guillemot. Browsing recorded meetings with ferret. In *In Proceedings of MLMI'04*. Springer-Verlag, 2004.
- [35] J. Yoshimi. Mapping the structure of debate. *Informal Logic*, 24(1), 2004.