The OpenCourseWare Metadiscourse (OCWMD) Corpus

Ghada Alharbi and Thomas Hain
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, United Kingdom
galharbi1@sheffiled.ac.uk, t.hain@sheffield.ac.uk

Abstract

This study describes a new corpus of over 60,000 hand-annotated metadiscourse acts from 106 OpenCourseWare lectures, from two different disciplines: Physics and Economics. Metadiscourse is a set of linguistic expressions that signal different functions in the discourse. This type of language is hypothesised to be helpful in finding a structure in unstructured text, such as lectures discourse. A brief summary is provided about the annotation scheme and labelling procedures, inter-annotator reliability statistics, overall distributional statistics, a description of auxiliary data that will be distributed with the corpus, and information relating to how to obtain the data. The results provide a deeper understanding of lecture structure and confirm the reliable coding of metadiscursive acts in academic lectures across different disciplines. The next stage of our research will be to build a classification model to automate the tagging process, instead of manual annotation, which take time and efforts. This is in addition to the use of these tags as indicators of the higher level structure of lecture discourse.

Keywords: metadiscourse, disciplines, OpenCourseWare lectures

1. Introduction

Academic lectures offer rich opportunities for studying a variety of complex discourse phenomena that help in building system to interpret the information in lectures discourse. Lectures contain regions where speakers introduce some concepts, emphasise others, interact with students, and engage in other interesting discourse phenomena. Thus, it is critical to locate these regions in order for a system to interpret lecture discourse. Moreover, lectures often involve strategies used to indicate these regions, which in turn reflect both topical and functional structure, as well as other high-level discourse dynamics. These strategies are known as *metadiscourse* (MD), which are linguistic expressions that are often referred to as discourse about discourse that has just occurred or is about to occur (Schiffrin, 1980; Crismore et al., 1993). Some examples of metadiscourse expressions include the *Introduction* ("Today I want to talk; Now moving on to"), *Conclusion* ("To conclude"), or *Previewing* ("We'll be coming to that").

Metadiscourse in spoken lectures poses interesting challenges to descriptive and theoretical models of discourse, as well as to downstream applications research, including the summarising of a meeting according to its activities (Niekrasz, 2012). Most recently, this has involved building presentation skills tools using Ted Talks (Correia et al., 2014).

Previous work by Alharbi et al. (2015) considered only five metadiscourse tags of the selected schema for about 47 university lectures. However, the present study describes a new corpus of hand-annotated metadiscourse, considering all the tags in the schema for roughly 106 lectures from different disciplines, namely Physics and Economics. The lectures were recorded at both Yale University and the Massachusetts Institute of Technology; this constituted as part of the OpenCourseWare initiative.

In this paper, a description of the OpenCourseWare Metadiscourse (OCWMD) is provided, in addition to other information on how the annotation procedure and post-annotation verification phases, such as inter-agreement measures, were conducted. The corpus is designed to be available online for research purposes, and in addition there is a plan for the future release of the automatic speech recognition transcriptions.

Future work will involve creating a classification model to automate the tagging process, which costs both time and efforts, and exploring different features for the model, such as lexical and prosodic information. In addition, we can check if we encounter any difference when we use the imperfect transcription that results from automatic speech recognition. Furthermore, we intend to use these MD tags to find higher level structure in lecture, as a form of table of contents suitable for web browsing.

2. Background

In the following we specifically list discourse analysis data and work that examines the function of the discourse for both written and spoken language.

In terms of written language, Marcu (2000) introduced the RST Discourse Treebank as a semantic-free theoretical framework of discourse relations; this was based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In RST, some relations are "intentional" whilst other are "subject matter" relations. Another related work is the contribution of Miltsakaki et al. (2008) to the Penn Discourse Treebank (PDTB) (Marcus et al., 1993); this was useful insofar as it classified discourse connectives according to their function. Teufel and Moens (2002) introduced a technique called *argumentative zoning* that assigns functions to sentences instead, with the aim of organising scientific articles into predefined zones, such as *Aim*, *Method* and *Background*.

A few studies focus on discourse function in speech. This motivated Correia et al. (2014) to design a corpus that could be used for exploiting the function of metadiscourse in Ted Talks. In order to accomplish this, the authors searched for definitions of metadiscourse in the related literature. For example, Luukka (1992) developed a schema that could

	Physics	Economics	Overall
# Lect	57	49	106
# Token	488k	422k	910k
# Words	9k	11k	20k
# Utterance.	32k	31k	63k

Table 1: Lecture Corpus Statistics

	MD Tag	Abbreviation
Metalinguistic	Repairing	REP
	Reformulating	REF
	Commenting on Linguistic Form/Meaning	CLF
	Clarifying	CLA
	Managing Terminology	MAT
Discourse Organisation	Introduction	INT
	Conclusion	CON
	Delimiting	DEL
	Contextualising	COT
	Enumerating	ENU
	Endophoric	PHO
	Reviewing	REV
	Previewing	PRE
Speech Acts	Emphasising	EMP
	Exemplifying	EXE
	Arguing	ARG
	Suggesting	SUG
Audience	Managing Comprehension	MAC
	Anticipating Audience's Response	AAR

Table 2: A list of the final MD tags used in developing the OCWMD corpus. This also shows the mapping between MD tags and their higher-level functions.

be applied in the context of both written and spoken academic discourse. This schema is based on three main categories: *Textual* (strategies related to the structuring of discourse), *Interpersonal* (related to the interaction with the different participants involved in the communication) and *Contextual* (covering references from audio-visual materials). Mauranen (2001) developed a schema for spoken language only. This author's taxonomy is characterised by three key categories: *Monologue*, *Dialogue* and *Interactive*. Both studies centre on the structuring of metadiscourse and the number of participants, but not its function. Ädel (2010) examined the functional approach of metadiscourse for both written and spoken language; the author introduced a schema consisting of 23 finer-level functional groups. These are further structured into four high-level tasks of *Metalinguistic Comments*, *Discourse organisation*, *Speech act labels* and *References to the audience*.

3. Data

The proposed annotation scheme was applied to source material drawn from two *OpenCourseWare* platforms, MIT[1] and YALE[2]; these are distributed under a Creative-Commons license. The lectures, presented by professional and highly skilled speakers, are available in the form of high quality video and audio data, transcripts, and subtitles. The objective of such posting follows the *OpenCourseWare* principle of wide accessibility. The decision to select these specific platforms was based on the high quality of the resources provided by these initiatives, as well as the availability of full lecture courses for a range of disciplines.

Another decision was necessary regarding the variety of disciplines to choose from, in order to formulate the final dataset. Lecture courses from two different disciplines, Physics and Economics, were chosen. This decision was primarily based on the availability of the lecture resources of similar introductory courses taught across the two different platforms, MIT OpenCourseWare and Open YALE Courses. For example, the Physics course from MIT OpenCourseWare is called Classical Mechanics, while the

[1] http://ocw.mit.edu/index.htm
[2] http://oyc.yale.edu

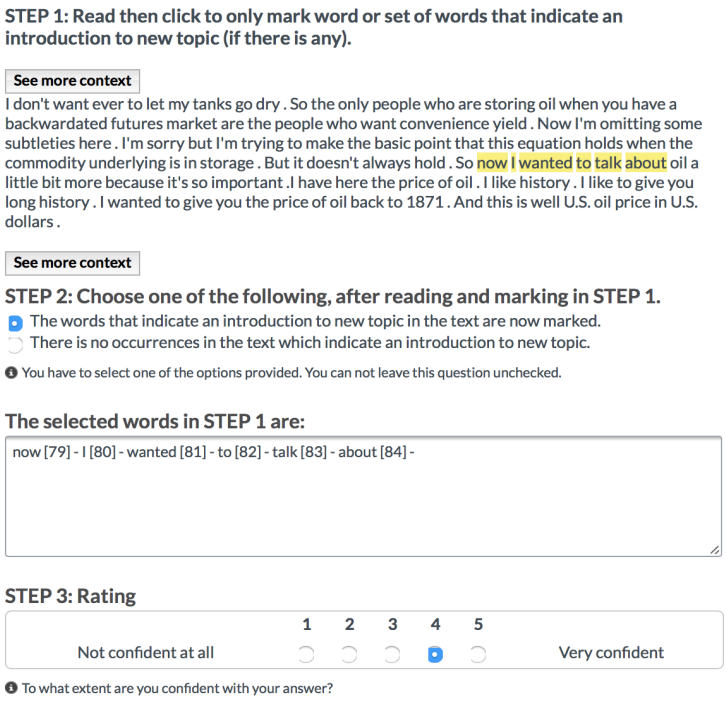

Figure 1: Example of the annotation interface used in annotating the category *Introduction*

course from Open YALE Courses is called Fundamentals of Physics. These courses cover approximately the same scientific material but different lecturers from different institutions teach them. Another reason for choosing these disciplines is to enable an investigation of whether there is a difference in detecting MD between Natural Science and Social Science lectures. Various corpus statistics for the new datasets are presented in Table 1.

4. Annotation

The proposed annotation schema was adopted from Ädel (2010), which provides a fairly suitable fit owing to the fact that this scheme signals functions in the discourse. Similar annotation procedures of previous work are followed here, too (Alharbi et al., 2015), where the focus was only on *Discourse Organisation* MD tags. In the following, a brief description of the participants, guidelines, annotation tool and preliminary study are presented.

4.1. Participants

Four expert participants were involved in this study, along with the first author of this paper. All annotators were students, two of whom are working towards a PhD in physics; the other two are working towards a PhD in economics. The participants were trained via different examples for each MD category before taking part in the final annotation study.

4.2. Guidelines

In order to facilitate the process for the annotators, categories are annotated one at a time, with only one segment with an average of 200 words (truncated to the closest end of utterance) per task, in order to manage annotator cognitive loads. Thus, for every category in the annotation schema, there are a total of 2,440 annotation tasks for the Physics lectures and 2,110 annotation tasks for the Economics lectures. As a result, different instructions sets have been designed for each of the 19 MD category in the schema. The instructions consist of three essential steps: first read the text segment and highlight MD tag occurrences; then confirm this finding in the second step; finally, the annotator is asked to rate their confidence score regarding this particular MD tag, on a scale out of 5. The use of the self-confidence score is to indicate how familiar the annotators are with the task at hand. These strategies were also followed in previous research, such as (Alharbi et al., 2015) and (Correia et al., 2014). It is worth noting that the final set of instructions was derived after several preliminary trials.

4.3. Tool

Annotation was conducted with the help of an annotation tool, which is also useful in outlining the annotation instructions as demonstrated in Figure 1. The tool was created and designed specifically for this task using HTML/XML languages and JavaScript functions. Moreover, specific mechanisms were provided in order to facilitate the work for the annotators, such as requesting them to highlight the target word or set of words that they consider are indications of the desired MD category. The same tool was also used in the annotation experiments conducted by Alharbi et al. (2015).

4.4. Trial

In order to assess how compatible Ädel's MD tags, the instructions, and the lecture data were, a preparatory anno-

MD Tag	Transcript
REV	**Let me remind you that everything I did so far** in class came from analyzing the Lorentz transformations.
EMP	**And you should be really** on top of those two marvelous equations because all the stuff we are doing is a consequence of that.
DEL|REV	**I won't go over how** we derive them because **I've done it** more than once.
REV	**But I remind you that** if you've got an event that occurs at x_t for one person and to a person moving to the right at velocity the same event will have coordinates $\dot{x} = x - u_t$.
NONE	This is it .
EMP	**This is the key** .
NONE	From this by taking differences of two events you can get similar equations for coordinate differences.
REF	**In other words** if two events are separated in space by x for one person and $\dot{x}$ for another person and likewise in time then you get similar formula for differences.
NONE	So differences are related the same way that coordinates themselves are.
NONE	But I will write it anyway because I will use it sometimes one way and sometimes the other way.
NONE	Even this one you can think of as a formula for a difference except one of the coordinates or the origin.
NONE	So I put this to work.
NONE	I got a lot of consequences from that.
REV	**You remember that**
EXE|REV	**For example** **I said before** take a clock that you are carrying with you.
NONE	Or let's take a clock that I'm carrying with me and let's see how it looks to you.
EXE	**And let's say** it goes tick and it goes tick one more time.

Figure 2: Example from physics lecture yale-phy0020014. 'MD': metadiscourse label (multiple tags are separated by '|'). For the purposes of illustration, each tag in the table is indicated by a unique font colour and the corresponding phrases have been highlighted with the same colour. The first author of the paper annotated these expressions.

tation study was conducted initially. This was intended to determine frequency estimates for tags appearing in the lecture data for Physics and Economics. In light of this, decisions relating to the matter of which MD tags were to be included were based on the frequency of the tags in this trial study. The tags 'Marking Aside', 'Adding to topic' and 'Managing Audience Discipline' were excluded from the final set of MD tags, due to their low occurrences in the trial study. Further modifications were also made to Ädel's schema. For example, the tag 'Exemplifying' was originally named 'Managing the message', and 'Imagining Scenario' was integrated with the 'Exemplifying' category. Thus, 19 MD tags were included in the final set of MD tags, and this excluded the non-label case. Table 2 presents this list, in combination with abbreviations for each tag that will be used herein. Ädel (2010) provides a precise definition for each MD tag in the selected list.

5. Annotated Example

An example from one of the physics lectures is shown in Figure 2; it will serve as an illustration of some of the types of MD that one is able to observe in the corpus. The example also provides a taste of the frequency and complexity of MD information. For instance, two of the utterances in the example display more than one tag, as was the case with the third utterance, which has two tags: 'Delimiting' and 'Reviewing'. In addition, of the 17 utterances within the excerpt, 8 have no labels. This is typical of many regions in the corpus, which show no MD tag at all. This case seems typical in related work, for example in identifying MD in a presentation style corpus (Correia et al., 2014) and in detecting speech acts in messages (Qadir and Riloff, 2011).

6. Reliability

In order to measure the degree of agreement between annotators, we used one of the most commonly applied metrics in NLP research (Carletta, 1996): *Kappa*. In particular, Fleiss's kappa coefficient κ (Fleiss, 1971), since there are three annotators for each annotation task as described above. A complete agreement corresponds to $\kappa = 1$, and no agreement corresponds to $\kappa \leq 0$. Similarly to the trail study, the decision of whether or not an utterance includes the discourse function of a certain MD tag is made based on the overall agreement between annotators; this is because the primary objective of this study is to detect if an utterance contains an instance of metadiscursive acts. Thus, agreement between annotators is considered to exist if the intersection (in terms of number of words) between their annotations is not void. A stricter approach for computing the agreement at word-level is reported in Madnani et al. (2012) for an extraction task of such metadiscursive acts expressions.

Table 3 presents the inter-annotator agreement and self-reported confidence scores for both Physics and Economics lectures. Agreement results show that experts have no trouble in labelling some of the MD tags. This could be at-

	MD Tag	Physics		Economics	
		κ	Confidence	κ	Confidence
Metalinguistic	REP	0.76	3.80	0.73	3.60
	REF	0.83	3.91	0.75	3.68
	CLF	0.70	3.74	0.73	3.80
	CLA	0.69	3.56	0.66	3.35
	MAT	0.77	3.86	0.79	3.82
	Total	0.75	3.75	0.73	3.65
Discourse Organisation	INT	0.85	3.98	0.80	4.00
	CON	0.79	4.00	0.77	3.80
	DEL	0.80	3.93	0.74	3.76
	COT	0.70	3.86	0.71	3.63
	ENU	0.78	3.76	0.81	3.89
	PHO	0.75	3.89	0.78	3.78
	REV	0.80	3.97	0.81	3.98
	PRE	0.77	3.81	0.76	3.85
	Total	0.78	3.90	0.77	3.84
Speech Acts	EMP	0.81	4.13	0.84	4.20
	EXE	0.82	3.97	0.85	4.00
	ARG	0.75	3.89	0.67	3.55
	SUG	0.74	3.83	0.72	3.69
	Total	0.78	3.95	0.77	3.86
Audience	MAC	0.71	3.78	0.69	3.71
	AAR	0.80	3.99	0.74	3.90
	Total	0.76	3.89	0.72	3.80
	Overall	0.77	3.87	0.75	3.79

Table 3: Results in terms of inter-annotator agreement (Fleiss's kappa κ) and the self-reported confidence scores.

tributed to the fact that each tag is dealt with individually. This is in contrast to previous work in Alharbi et al. (2015) on a subset of the dataset presented here and using only discourse organisation tags, such as *Introduction*, which reports lower inter-annotator agreement. This is because the knowledge of the annotators increased and the guidelines for annotators were improved. However, in the present study there are some tags where even expert annotators have struggled to detect them compared to others, such as the CLARIFYING (CLA) tag for both disciplines and, the ARGUING (ARG) tag for the economics discipline. Previous annotation work relating to similar phenomena reported similar situations (Correia et al., 2014). Self-reported confidence scores show all tags scoring above the middle of the scale (5). This generally indicates that annotators are familiar with the annotation task. In addition, annotators showed less confidence in marking *Metalinguistics* tags than other tags in the schema. Another important finding is that there is a direct association between having high agreement between annotators and high self-confidence scores.

7. Distributional Statistics

Basic statistics of MD tags for 106 lectures from the two different disciplines (Physics and Economics) are provided. Excluding the non-MD tag NONE, we find 11,466 total tags. This includes the utterances that have more than one tag. Table 4 shows the distribution of the tags in more detail. MD tags of the '*Discourse Organisation*' type seem to be the most frequent compared to the other tags. This finding is also consistent across the two disciplines, with nearly 2500 occurrences each. Another important finding is that the '*Interaction with the Audience*' MD tags are less frequent and, again, this is consistent for both disciplines. These observed similarities could indicate that MD tags are general expressions and, furthermore, that the discipline knowledge may have no discernible effect on its occurrences.

8. Auxiliary Information

Further useful information is also included with the corpus. Word-level time information will be available, based on alignments from an automatic speech recogniser. Automatic transcriptions will also be provided. We recommend various ways in which one can group the large set of labels into a smaller set of classes, depending on the research focus. For example, a convention was provided in the gold standard set in order to map each MD tag to its higher-level functions. This would formulate four general tags in total. Finally, we aim to attach other information that that may be useful when attempting to develop the automatic modeling of prosody.

9. Conclusion

For the interpretation of a lecture, it is critical to have a primitive concept of the purpose of each utterance. The

	MD Tag	Physics		Economics	
		#	%	#	%
Metalinguistic	REP	83	1.36	94	1.72
	REF	181	2.97	71	1.30
	CLF	18	3.00	37	0.68
	CLA	285	4.68	300	5.50
	MAT	545	8.95	319	5.85
	Total	1112	20.96	821	15.05
Discourse Organisation	INT	208	3.42	346	6.35
	CON	104	1.71	123	2.26
	DEL	87	1.43	82	1.50
	COT	22	0.36	31	0.57
	ENU	571	9.38	583	10.70
	PHO	123	2.02	195	3.58
	REV	834	13.70	685	12.57
	PRE	536	8.81	396	7.27
	Total	2485	40.83	2441	44.80
Speech Acts	EMP	1234	20.27	1070	19.63
	EXE	842	13.83	885	16.24
	ARG	43	0.71	14	0.26
	SUG	11	0.20	22	0.40
	Total	2130	35.01	1991	36.53
Audience	MAC	218	3.58	110	2.02
	AAR	113	1.86	45	0.83
	Total	331	5.44	155	2.85
	Overall	6058	18.93	5408	17.45

Table 4: A statistical summary of all the tags in the gold standard dataset for each discipline, showing the number of occurrences (#) and the frequency of each tag relative to all tags (%).

general strategies of a system trying to infer a lecture discourse or understand it will be tied to these primitives. For example, if the purpose of an utterance is to emphasise some concept, the system then will easily determine what information is important to highlight for students, by fetching those utterances labelled as important. In another example, a system may use utterances based on their functions to find higher level forms, such as table of content and discourse structures analogous to paragraphs and chapters.

Lectures often involve strategies used to indicate the functions of utterances. These strategies are known as *metadiscourses* (MD), which are a set of linguistic expressions that signal different functions in the discourse. The proposed MD scheme consists of a set of categories that allow the process of labelling utterances with suitable functional categories in an academic context. The experimental results show the reliability of the annotation scheme and confirm that MD as a linguistic phenomenon occurs frequently in academic lectures from different disciplines. The annotated corpus with MD labels would provide a valuable resource in the study of discourse as well as a source of training and testing for a discourse analysis system using MD labels in its utterance representation. Future work will involve creating a classification model to automate the tagging process, which costs both time and efforts, and exploring different features for the model, such as lexical and prosodic information. In addition, we can check if we encounter any difference when we use the imperfect transcription that results from automatic speech recognition. Furthermore, we intend to use these MD tags to find higher level structure in lecture, as a form of table of contents suitable for web browsing.

10. Acknowledgements

The authors would like to acknowledge the valuable comments and suggestions of the reviewers, which have improved the quality of this paper. We also thank the annotators for taking part in this study. Many special thanks and appreciation go to the Royal Embassy of Saudi Arabia Cultural Bureau in London for funding this research.

11. Bibliographical References

Ädel, A. (2010). Just give kind of map of where we are going: A taxonomy of metadiscourse in spoken and written academic english. *Nordic Journal of English Studies*, 9(2):69–97.

Alharbi, G., Ng, R. W. M., and Hain, T. (2015). Annotating Meta-discourse in Academic Lectures from Different Disciplines. pages 161–166.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2).

Correia, R., Mamede, N., Baptista, J., and Eskenazi, M.

(2014). Using the crowd to annotate metadiscursive acts. *In Proceedings 10th Joint ISO-ACL SIGSEM*, page 102.

Crismore, A., Markkanen, R., and Steffensen., M. S. (1993). Metadiscourse in persuasive writing a study of texts written by American and Finnish university students. *Written communication*, 10(2):39–71.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Luukka, M.-R. (1992). Metadiscourse in academic texts. In *Conference on Discourse and the Professions*, volume 28, Uppsala, Sweden.

Madnani, N., Heilman, M., Tetreault, J., and Chodorow, M. (2012). Identifying high-level organizational elements in argumentative discourse. In *Proceedings of NAACL'12: HLT*, pages 20–28.

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

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Mauranen, A. (2001). Reflexive academic talk: Observations from micase. In *Corpus linguistics in North America Selections from the 1999 symposium*, pages 165–178.

Miltsakaki, E., Robaldo, L., Lee, A., and Joshi, A. (2008). Sense annotation in the penn discourse treebank. In *Proceedings of the LREC'08*.

Niekrasz, J. (2012). *Toward Summarization of Communicative Activities in Spoken Conversation*. Ph.D. thesis, University of Edinburgh.

Qadir, A. and Riloff, E. (2011). Classifying sentences as speech acts in message board posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 748–758. Association for Computational Linguistics.

Schiffrin, D. (1980). Meta-talk: Organizational and evaluative brackets in discourse. *Socioogical Inquiry*, 50(3?4):199–236.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:2002.