Annotating Meta-discourse in Academic Lectures from Different Disciplines

Ghada Alharbi, Raymond W. M. Ng, Thomas Hain
Department of Computer Science, University of Sheffield, United Kingdom
galharbi1@sheffield.ac.uk, wm.ng@sheffield.ac.uk, t.hain@sheffield.ac.uk

Abstract
The use of discourse structure was shown to be effective in various applications. Meta-discourse is often used as an expression to signal discourse structure. Previous work focused on using the meta-discourse structure in written texts, or spoken material in very clean conditions. This paper presents a meta-discourse annotated corpus in a more challenging educational context. The corpus comprises of academic lectures from two different disciplines: physics and economics. The schema used focuses on five categories: Introduction, Conclusion, Previewing, Reviewing and Enumerating. The annotation task is described in detail, including instructions and strategies used by expert annotators. Annotation results are reported in terms of inter-annotator agreement, self-reported confidence and number of occurrences. Results show that meta-discourse is frequently used in academic lectures and this is observed in the two selected disciplines. Further analysis of the corpus is conducted showing that some of these categories, namely Introduction and Previewing, are correlated with labelled topic boundaries, which is also consistent in both disciplines. This finding shows the potential for using meta-discourse information in topic segmentation task.

Index Terms: Meta-Discourse, Annotation, Academic Lectures, Disciplines

1. Introduction
Education nowadays is no longer confined to learning in the classroom. With the popularity of the internet and mobile devices, ubiquitous learning has been made possible. Education materials are easily accessible online and learners can learn in their place of choice, according to their own schedule. Under the Massive Open Online Course (MOOC) initiative, many educational resources such as MIT OpenCourseWare\(^1\), Open Yale Courses\(^2\) and Stanford Online\(^3\) have become freely available. Despite the abundance of education materials, however, questions on how to organise this vast quantity of material in a systematic way have not yet been addressed sufficiently. By organisation, we mean the extraction of some high-level information from the materials, allowing learners to acquire further information, such as a summary, the hierarchical structures of a talk, the linking of relevant content in a talk or among multiple talks and the establishment of milestones where important concepts are presented, etc. This kind of organisation is vital to educational resources in terms of facilitating learners.

Most of the existing research work focuses on the use of lexical information in learning materials. In our research we investigate the usefulness of discourse information, in particular meta-discourse. Meta-discourses are linguistic expressions that are often referred to as discourse about discourse that has just occurred or is about to occur [1]. [2] has formally defined this language as “linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organise, interpret and evaluate the information given”. This kind of language has a privileged place in discourse analysis because it reflects the discourse structure. Some examples of meta-discourse expressions include Introduction (“Today I want to talk; Now moving on to”), Conclusion (“To conclude”), or Previewing (“We’ll be coming to that”).

Meta-discourse was shown to be effective in a range of applications; for example: summarising a meeting according to its activities [3], modelling argumentative zoning in scientific research articles [4], and most recently, building presentation skills tools using Ted Talks [5]. However, coping with lecture recordings is considered a challenging task because of the heterogeneity of speaker styles, audio channels and quality [5]. Furthermore, to the best of our knowledge there is no available resource on the meta-discourse of lectures. Thus, this study investigates the feasibility of detecting meta-discourse phenomena in academic lectures from different disciplines.

In this work we present a methodology to incorporate expert annotations and generate meta-discourse information to add to the open source lecture database. The paper concentrates on lecture materials in two subjects – physics and economics. This work follows standard procedures in related work used in other domains [5], in particular the annotation procedures and post-annotation verification such as inter-agreement measures. Details of these will be outlined further in the following sections. The paper further includes a preliminary analysis study of annotation results, which demonstrates the potential usefulness of meta-discourse information in the topic segmentation. The methodology of annotation, the annotated data and analysis is planned to be made available to the public.

The rest of the paper is organised as follows: A general overview of meta-discourse is provided in Section 2. The annotation experiments conducted to create the corpora of meta-discourse in academic lectures are presented in Section 3. The annotation results are outlined in Section 4. The correlations with topic boundaries are discussed in Section 5. Finally, conclusions are drawn and recommendations for further research are made in Section 6.

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.

For written language, [7] introduced the RST Discourse Treebank as a semantic-free theoretical framework of discourse relations based on Rhetorical Structure Theory (RST) [8]. In

\(^1\)http://ocw.mit.edu/index.htm
\(^2\)http://oyc.yale.edu
\(^3\)http://online.stanford.edu/courses

10.21437/SLaTE.2015-28
### 3. Annotation Experiments

#### 3.1. Data

For conducting our work in this paper data from MOOC resources was chosen, including MIT OpenCourseWare and Open Yale Courses. Here sets of courses (a coherent series of lectures) is available for download. The selected materials are accessible free-of-charge via the website for Open Yale Courses, and 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 MOOC principle of wide accessibility. After careful consideration of the available materials, courses from the disciplines of physics and economics are chosen to form the corpus under investigation. When preparing the annotation experiments, the overall number of lectures were forty-seven of which twenty-four were Physics lectures and twenty-three were Economics lectures. Table 1 shows statistics describing the new dataset.

#### 3.2. Schema

In keeping with [5], the schema for meta-discourse annotation from [6] was adapted, as it explicitly addresses the function of meta-discourse. As outlined in Section 2, the schema proposed by [6] included 4 high-level functions and 23 categories. The three high-level functions of Metalinguistic comments, Speech act labels and References to the audience, relate primarily to lecture content or interaction with students. This study employs solely the high-level function of discourse organisation

---

Table 1: Lecture Corpus Statistics

<table>
<thead>
<tr>
<th></th>
<th>#Lect</th>
<th>Avg. # Segments Per Lect</th>
<th>#Total Segments</th>
<th>#Total Words</th>
<th>#Uniq. Words</th>
<th>#Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>24</td>
<td>6</td>
<td>144</td>
<td>284k</td>
<td>6k</td>
<td>18k</td>
</tr>
<tr>
<td>Economics</td>
<td>23</td>
<td>7.1</td>
<td>172</td>
<td>231k</td>
<td>9k</td>
<td>15k</td>
</tr>
<tr>
<td>Overall</td>
<td>47</td>
<td>6.5</td>
<td>316</td>
<td>515k</td>
<td>15k</td>
<td>33k</td>
</tr>
</tbody>
</table>

Table 2: Meta-discourse categories for lecture discourse organisation provided by [6]. Examples of each category are provided from both physics and economics lectures used in this study.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Examples</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>used to open the subtopic</td>
<td>So I want to now talk about</td>
<td></td>
</tr>
<tr>
<td>CON</td>
<td>used to close the subtopic</td>
<td>I wanted to conclude with</td>
<td></td>
</tr>
<tr>
<td>ADD</td>
<td>used to explicitly comment on the addition of a subtopic</td>
<td>also let me add</td>
<td></td>
</tr>
<tr>
<td>DEL</td>
<td>used to explicitly state how the topic is constrained</td>
<td>but I'm not going to get into</td>
<td></td>
</tr>
<tr>
<td>MAS</td>
<td>used to open or close a &quot;topic sidetrack or digression&quot;</td>
<td>This is an aside</td>
<td></td>
</tr>
<tr>
<td>COT</td>
<td>used to comment on the situation of speaking</td>
<td>If I have some time</td>
<td></td>
</tr>
<tr>
<td>ENU</td>
<td>used to show specific parts of the discourse are ordered in relation to each other</td>
<td>The first thing I want to talk about is</td>
<td></td>
</tr>
<tr>
<td>PHO</td>
<td>used to point to a specific location in the discourse</td>
<td>let's look at this plot</td>
<td></td>
</tr>
<tr>
<td>PRE</td>
<td>used to point forward in the discourse</td>
<td>We'll come back later to</td>
<td></td>
</tr>
<tr>
<td>REV</td>
<td>used to point backward in the discourse</td>
<td>Last time I talked about</td>
<td></td>
</tr>
</tbody>
</table>

this, categories such as Evaluation, Elaboration, and Background are included. Another related work is the contribution of [9] to the Penn Discourse Treebank (PDTB) [10] which classified discourse connectives according to their function. The work considered functional categories such as giving examples (Instantiation), making reformulations and clarifications (Restatement), comparing (Contrast), or showing cause (Reason). [4] introduce a technique called argumentative zoning that assigns functions to sentences instead with the aim to organise scientific articles into predefined zones, such as Aim, Method and Background. Following this initial work, many studies have introduced a range of new schema for scientific articles in terms of different domains and tasks, as can be observed in [11, 12, 13]. The primary objective of these experiments was discourse. However, in examining the function of the discourse, they have only applied the method to written language. Few experiments look at discourse function in spoken discourse. This motivated [5] to design a corpus that is useful for exploiting the function of meta-discourse in Ted Talks. To accomplish this, the authors looked for definitions of metadiscourse in the literature. For example, [14] developed a schema for use in 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). [15] developed a schema for spoken language only. This author’s taxonomy also proposes three key categories: Monologue (similar to Textual in the scheme proposed by [14]), Dialogue (similar to Interpersonal in the scheme proposed by [14]) and Interactive (related to the interactions with the speaker).

Both studies centre on the structuring of meta-discourse and the amount of participants, but not its function. Adel [6] examined the functional approach of meta-discourse for both written and spoken language, and 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. Two similar academic corpora from different disciplines were used in that study: MICUSP, consisting of academic papers [16], and MICAES, consisting of academic lectures [17]. This study investigates whether it is feasible to adopt Adel’s schema of meta-discourse [6], in a similar way to the study conducted by [5] on Ted Talks, but applied to academic lectures. By doing so, our work identifies and highlights the challenges and strategies needed to study such phenomena in academic lectures. The functions and descriptions of Adel’s schema of metadiscourse will be discussed further in Section 3.
and its ten categories. Clearly these categories by definition show some relevance to the task of topic segmentation which is investigated in Section 5. These categories are Introduction (INT), Conclusion (CON), Adding to Topic (ADD), Delimiting Topic (DEL), Marking Asides (MAS), Contextualizing (COT), Enumerating (ENU), Endophoric Marking (PHO), Previewing (PRE) and Reviewing (REV). The following paragraphs further explain these categories in detail along with meta-discourse examples as briefly demonstrated in Table 2.

Lecturers often use the Introduction (INT) category to open new subtopics. For instance in a physics lecture, Newton’s First Law, Newton’s Second Law and Newton’s Third Law would be the subtopics of a lecture on the topic of Newton’s Laws. Examples of this category are “So I want to now talk about” and “Let’s now move on to”, which clearly indicate a shift in the discourse. In contrast, the Conclusion (CON) category is normally used to conclude or summarise subtopics of the lecture, such as “to conclude” or “to summarise”. The Adding to Topic (ADD) category on the other hand is used to add to the current subtopics (e.g. “I should add too that”). Delimiting Topic (DEL) expressions are used to establish a constraint in presenting the subtopic. For instance, the lecturer may use expressions to demonstrate that: “We’re not gonna deal with all eight here” and “We won’t go into that, that’s a little too much for us to consider”. Marking Asides (MAS) is used to open or close aside comments that are not related to either topic or subtopics of the lecture such as the expression “I want to do a little aside here”.

Enumerating (ENU) is used to show how specific parts of the discourse are ordered in relation to each other. An example of this category would be an expression like “we’re going to talk about consumer first”. Endophoric Marking (PHO) is used to point to a specific location in the discourse; it refers to cases that occur before or after the current point (unlike Previewing (PRE) and Reviewing (REV), as for example when the student is instructed to look at a table, or turn to a specific point in a book [6]. Finally, the category Contextualizing (COT) is used to comment on the situation of speaking, and thus contains traces of the production of the discourse. In this category, there is spelled-out justification for choices made in planning or organising the discourse. An example of this is “We’re doing pretty well on time so let’s”.

### 3.3 Participants and Agreement Measures

Five experts participants were involved in this study of which four are the annotators and one is the first author of the paper who annotated the two datasets.

All annotators are students, two of which are working towards a PhD in physics and the other two are working towards a PhD in economics. During the pilot study of the experiment, the annotators familiarised themselves with the annotation scheme, which included various examples of every category. The agreement measure selected is the one most commonly applied in NLP research [18], namely, Fleiss’ Kappa coefficient κ [19]. Complete agreement corresponds to κ = 1, and no agreement (other than chance) corresponds to κ ≤ 0.

### 3.4 Pilot Study

To assess how well Ådel’s taxonomy, the instructions, and lecture data are compatible, a preparatory annotation study was conducted first. This was intended to determine frequency estimates with which categories appear in the lecture data for physics and economics. For this initial study, five lectures are selected at random from each discipline. The complete set of schema categories was used for annotation (see Table 3). All five participants took part in this initial study. Decisions on occurrence of an event were made based on a majority vote. Table 3 shows results that highlight the rare occurrence of three categories in the sample in both physics and economics. One of these categories, Marking Aside (MAS) appears only once in the economics lectures. Low frequency for this category was also observed in Ted Talks [5]. Furthermore, the number of times that Adding to Topic (ADD), Contextualizing (COT), and Endophoric Marking (PHO) appear in the entire sample for physics and economics is only 7, 5 and 8, respectively. The Delimiting Topic (DEL) category contains insufficient samples in physics for further analysis. Based on the frequency of these categories in our sample, only five of the ten categories of the selected schema are used further in the complete annotation of the corpus.

### 3.5 Tool and Instructions

Annotation is conducted with the help of an online annotation tool, which is also useful in outlining the annotation instructions. The online tool was created and designed specifically for this task using HTML/XML languages and JavaScript functions.

There is one segment with an average of 200 words per task and, in order to facilitate the process for the annotators, categories are annotated one at a time. Thus, for every category in the annotation schema, there are a total of 1,420 annotation tasks for the physics lectures and 1,150 annotation tasks for the economics lectures. Moreover, the annotators are requested to highlight words they consider to be related to the desired category. The annotation interface for the category Introduction is
demonstrated in Figure 1.

3.6. Gold Standards

Similar to the pilot study, the results of the annotation activity are used to determine all categories in the design for every sentence and the decision as to whether or not a sentence includes function structure of a certain category is made based on majority vote. Since this study is aimed to detect if an utterance contains an instance of meta-discursive acts. Thus, the 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 [20] for an extraction task of such meta-discursive which is out of the scope of this study.

4. Annotation Results

Table 4 shows the results for inter-annotator agreement, self-reported confidence scores, number of event occurrences and agreement. Overall, the expert annotators mostly agreed on the occurrence of meta-discourse categories. However, the annotators found difficulty in annotating the category Enumerating as shown in the inter-annotator agreement $\kappa$ values in Table 4 in both disciplines. In the following, the number of occurrences of each meta-discourse category is examined in detail.

**Introduction:** For the physics lectures the result of inter-annotator agreement is 0.68; for the economics lectures the score is 0.71. These scores indicate a relatively high level of agreement between annotators in accomplishing this task. The number of occurrences of this category (112 and 195 for physics and economics, respectively) is higher than the number of lectures (47, in both disciplines). This indicates that lecturers use this category to introduce multiple subtopics within a single lecture. [5] undertook an experiment along the same lines; for the same category in Ted Talks there were agreements of 0.64, with 1,159 occurrences. This was done by employing the Amazon Mechanical Turks (AMT) crowdsourcing platform, without the need for high domain coder knowledge as seen in this study.

**Conclusion:** The findings on the annotation of topic conclusion were consistent in both disciplines. This consistency can also be seen in the confidence rate of the annotators, with a score of 3.95 for physics and 3.80 for economics. However, even though there are similarities, the score for agreement between annotators was significantly less, with 0.65 for physics and 0.63 for economics. It is also important to note that the number of occurrences of conclusions (53) and (49) is higher than the number of lectures in physics and economics, respectively. This is due to the fact that the lecturers might conclude on topic segments and not just at the end of lectures. This is also aligned with our findings regarding the correlation with labelled topic boundaries, which is explained in more detail in Section 5.

**Enumerating:** Here, annotators had relatively poor performance in both disciplines; this led to fewer appearances of this category in both the economics and physics lectures. This is due to the fact that the lecturer in these corpora might infrequently provide outlines of their lectures and instead go straight to the points of their discussion. This is also observed in the two disciplines. This category, as well as the following two categories, were not explored in [5].

**Reviewing:** Aside from yielding higher agreement rates among annotators, compared to the previous category, this category has a significant frequency of occurrence in both disciplines. The self-reported scores also indicate that this task is understandable for the annotators in both disciplines. The large number of occurrences of this category in both disciplines is due to the fact that these corpora are complete courses containing multiple lectures related to each other. Thus, the lecturer needs to link their information in order to provide the student with the big picture.

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Confidence</th>
<th># Occurrences</th>
<th>Confidence</th>
<th># Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.68</td>
<td>3.99</td>
<td>0.71</td>
<td>4.00</td>
</tr>
<tr>
<td>CON</td>
<td>0.65</td>
<td>3.95</td>
<td>0.63</td>
<td>3.80</td>
</tr>
<tr>
<td>ENU</td>
<td>0.59</td>
<td>3.75</td>
<td>0.61</td>
<td>3.85</td>
</tr>
<tr>
<td>PRE</td>
<td>0.70</td>
<td>3.85</td>
<td>0.68</td>
<td>3.99</td>
</tr>
<tr>
<td>REV</td>
<td>0.73</td>
<td>4.00</td>
<td>0.69</td>
<td>3.95</td>
</tr>
<tr>
<td>Overall</td>
<td>0.67</td>
<td>3.91</td>
<td>0.68</td>
<td>3.92</td>
</tr>
</tbody>
</table>

Table 4: Results in terms of inter-annotator agreement using Fleiss’ kappa \( \kappa \), self-reported confidence rating and occurrences of every discipline.

This can be observed among physics lecturers predominantly, who are re-examining content of earlier lectures, prior to starting the new lecture topic. This can once again be attributed to the fact that lecturers often tend to show students how the information in the set of lectures is linked and related.

In summary, the experiment results point towards metadiscourse phenomena occurring frequently in academic lectures and this finding is consistent in the two selected disciplines. This study is part of ongoing research; thus we plan to include other categories from Ådel’s schema and to include other disciplines, namely, computer science and biology.

### 5. Correlation with Topic Boundaries

A preliminary further study investigated whether metadiscourse information can be used in recovering of higher level information. The following concentrates on topic boundary information. The labelled topic boundaries were taken from the transcription of these lectures; they were provided by the lecturers themselves and are available online on the Open Yale Courses website. An example of topic segmentation in a lecture covering “Newton’s Laws”, then “Newton’s First Law”, “Newton’s Second Law”, and “Newton’s Third Law” would be the chosen sub-topics. We analysed the correlation between metadiscourse categories in the lecture corpus and labelled topic boundaries, and automatically extracted metadiscourse categories that are statistically correlated with labelled topic boundaries. For every annotated category in the lecture corpus, the number of its occurrences near any topic boundary (with a window size of 5 seconds on either side of the target boundary, inclusive) are counted, and set against those further away. These counts are obtained for all different metadiscourse categories available. The chi-square test allows the calculation of the significance of the near-against distinct-statistics by comparing with the overall statistics, where the null hypothesis is assumed. In Table 5 the counts in the Introduction meta-discourse category and the overall counts in economics talks are listed. The computed \( \chi^2 \) value is 11.51. The introduction category is indicative of the presence of a topic boundary.

Table 5: \( \chi^2 \) test for each category in each discipline. Boldface indicates that the \( \chi^2 \) value is significant at the level of \( p < 0.01 \).

<table>
<thead>
<tr>
<th>Category</th>
<th>Physics</th>
<th>Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>66.62</td>
<td>11.51</td>
</tr>
<tr>
<td>CON</td>
<td>0.90</td>
<td>7.51</td>
</tr>
<tr>
<td>ENU</td>
<td>1.48</td>
<td>0.20</td>
</tr>
<tr>
<td>PRE</td>
<td>4.39</td>
<td>8.94</td>
</tr>
<tr>
<td>REV</td>
<td>22.45</td>
<td>9.62</td>
</tr>
</tbody>
</table>

Table 6: Results of \( \chi^2 \) test for each category in each discipline. Boldface indicates that the \( \chi^2 \) value is significant at the level of \( p < 0.01 \).

The Introduction category occurred 112 and 195 times in physics and economics, respectively, whereas Reviewing occurred 208 and 153 times in physics and economics, respectively. The other categories, Conclusion and Previewing, do not correlate with the topic boundaries labelled by physics lecturers and this can be attributed to the lecturer style in physics lectures. However, this is not the case in economics lectures as these meta-discourse categories correlated with labelled topic boundaries. The Enumerating category does not correlate with labelled topic boundaries in both disciplines as this meta-discourse category could be used by the lecturer to order parts of the discourse; this ordering often occurred after introducing the topic.

5. Conclusions

In this work we presented the methodology of generating metadiscourse information on academic lecture data using expert annotations. A publicly available meta-discourse database was created. The materials are based on academic lectures in two academic subjects: physics and economics. It was found that re-
liable, consistent annotations of five meta-discourse categories (Introduction, Conclusion, Previewing, Reviewing and Enumerating) could be obtained. Results show that meta-discourse phenomena occur frequently in lecture resources across the two disciplines. Meta-discourse features can be utilised to derive higher-level knowledge in lectures, which in turn can facilitate the organisation and accessing of such educational materials that are often found in abundance online for students. This was illustrated through preliminary statistical analysis correlating the features with topic segmentation. The resource is valuable to future work in the research community on content analysis and discourse organisation of educational materials.

7. Acknowledgments
We are grateful to Oscar Saz Torralba for his helpful advice. We thank the annotators for taking part in this study. This work was partially funded by the Royal Embassy of Saudi Arabia Cultural Bureau in London.

8. References