# Using the Crowd to Annotate Metadiscursive Acts

**Rui Correia[1,2], Nuno Mamede[2], Jorge Baptista[2,3], Maxine Eskenazi[1]**

[1] LTI - Carnegie Mellon University, Pittsburgh, USA
[2] INESC-ID, Lisbon, Portugal
[3] Universidade do Algarve, Faro, Portugal
rcorreia@cs.cmu.edu, Nuno.Mamede@inesc-id.pt, jbaptis@ualg.pt, max@cs.cmu.edu

**Abstract**

This paper addresses issues relating to the definition and non-expert understanding of metadiscursive acts. We present existing theory on spoken metadiscourse, focusing on one taxonomy that defines metadiscursive concepts in a functional manner, rather than formally. A crowdsourcing annotation task is set up with two main goals: (a) build a corpus of metadiscourse, and (b) assess the understanding of metadiscursive concepts by non-experts. This initial annotation effort focus on five categories of metadiscourse: INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING, and EMPHASIZING. The crowdsourcing task is described in detail, including instructions and quality insurance mechanisms. We report results in terms of time-on-task, self-reported confidence, requests for additional context, quantity of occurrences and inter-annotator agreement. Results show the crowd is capable of annotating metadiscourse and give insights on the complexity of the different concepts in the taxonomy.

**Keywords:** Metadiscourse, Crowdsourcing, Non-experts

## 1. Introduction

Metadiscourse is one of the basic functions of language. Commonly referred to as *discourse about discourse*, it is composed of rhetorical acts and patterns used to make the discourse structure explicit, acting as a way to guide the audience. Crismore et al. (1993) define metadiscourse 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 organize, interpret and evaluate the information given". Some examples of metadiscursive acts include introductions ("I'm going to talk about. . ."; "In this paper we present. . ."), conclusions ("In sum,. . ."), or emphasis ("The take home message. . ."; "Please note that. . .").

This study focuses on the function of metadiscourse in spoken communication. The functional analysis of such phenomena in discourse can contribute to tasks such as simplification or language understanding, and can be used for language learning purposes, such as presentation skill instruction. We describe the task of building a corpus of metadiscursive acts using crowdsourcing to annotate transcripts of presentations. By using non-experts, we expect not only to obtain the annotations of some metadiscursive acts, but also to get feedback on how those acts are perceived.

In this paper we start with background on metadiscursive theory, addressing how existing taxonomies and resources represent it (Section 2). Section 3 focuses on the choice of the material that the crowd will annotate with metadiscursive acts. Section 4 describes a preliminary annotation task aimed at testing the presence of some of the acts taken from our adopted metadiscourse taxonomy. Section 5 focuses on the setup of the crowdsourcing task, considerations regarding instructions, and quality control. The results obtained using the crowd and an ensuing discussion are presented in Sections 6 and 7. In Section 8, we conclude and present future directions.

## 2. Background

In the literature on discourse analysis we find studies that address function in discourse. For example, the contribution of Miltsakaki et al. (2008) to the Penn Discourse Treebank (PDTB) (Marcus et al., 1993) organized discourse connectives according to their function, considering categories such as giving examples (INSTANTIATION), making reformulations and clarifications (RESTATEMENT), comparing (CONTRAST), or showing cause (REASON). Another example is the RST Discourse Treebank (Marcu, 2000), a semantics-free theoretical framework of discourse relations based on Rhetorical Structure Theory (Mann and Thompson, 1988), which includes categories such as EXAMPLE, DEFINITION, or SUMMARY. Even though these projects explore function in discourse, they focus on written language and do not address the meta aspect of language.

The lack of work on the explicit nature of discourse motivated our decision to build a corpus targeting the function of metadiscourse in spoken communication. To accomplish that, we looked for definitions of metadiscourse.

Luuka (1992) developed a taxonomy for use in both written and spoken academic discourse. This taxonomy is composed of three main categories: TEXTUAL (strategies related to the structuring of discourse), INTERPERSONAL (related to the interaction with the different stakeholders involved in the communication) and CONTEXTUAL (covering references of audiovisual materials). Mauranen (2001), on the other hand, focused only on spoken discourse. This author's taxonomy is also composed of three categories with no further division: MONOLOGIC (similar to TEXTUAL in Lukka's taxonomy), DIALOGIC (similar to INTERPERSONAL in Lukka's taxonomy) and INTERACTIVE (related to question answering and other interactions with the speaker).

Luuka's and Mauranen's taxonomies organize metadiscourse in similar ways. However, both studies focus on

**METALINGUISTIC COMMENTS**
- Repairing
- Reformulating
- Commenting on Linguistic Form/Meaning
- Clarifying
- Manage Terminology

**DISCOURSE ORGANIZATION**

**Managing Topic**
- Introducing Topic
- Delimiting Topic
- Adding to Topic
- Concluding Topic
- Marking Asides
- Enumerating

**Managing Phorics**
- Endophoric Marking
- Previewing
- Reviewing
- Contextualizing

**SPEECH ACT LABELS**
- Arguing
- Exemplifying
- Other

**REFERENCES TO THE AUDIENCE**
- Managing Comprehension
- Managing Discipline
- Anticipating Response
- Managing the Message
- Imagining Scenarios

Figure 1: Ädel's taxonomy of metadiscourse.

the *form* of metadiscourse (i.e. number of stakeholders involved), not addressing its *function*.

A *functional* approach to metadiscourse can be found in the work of Ädel (2010) who unifies existing taxonomies under a framework that encompasses both spoken and written discourse. This framework was built using two academic-related corpora: MICUSP (Römer and Swales, 2009) – comprised of academic papers – and MICASE (Simpson et al., 2002) – a corpus of university lectures.

The categories and organization of Ädel's taxonomy of metadiscourse (Figure 1) reflect the author's concern about the unification of theories for both written and spoken discourse and the desire to describe metadiscourse in a functional manner. For these reasons, we have decided to adopt this taxonomy as a source of categories of metadiscourse. This taxonomy will be discussed further in Section 4.

## 3. Corpora

Having adopted a set of metadiscursive acts to annotate, we then needed to select a source of data where these strategies could be found. Two main sources of data were considered: classroom recordings and TED talks[1].

Analysis of the contents of these two sources led us to choose TED talks over classroom recordings. TED talks are consistently good quality presentations from good presenters. Each talk is carefully rehearsed beforehand, conveying one message in a short span of time (from 5 to 20 minutes). This contrasts with classroom recordings which are typically longer and where there is an order in which the classes should be listened too. Even if only self-contained classes are considered, they are targeted at a very specific audience and the topics are advanced and require a significant amount of previous knowledge. Secondly, TED talks are uniform in content. They contain high-quality audio and video material and are available in several languages. They are also updated daily and subtitled, providing a good source of transcribed material. Classroom recordings, on the other hand, are a more heterogeneous resource as far as source and recording conditions are concerned, making them harder to be automatically processed with the least amount of human intervention possible. Even though they are not further addressed in this paper, classroom recordings would be a good resource set to extend our TED findings at a later time.

At the time of the preparation of this annotation task there were 730 TED talks available in English with subtitles, synced at sentence level (a total of 180 hours, approximately).

## 4. Preliminary Annotation Task

A small preliminary annotation task was carried out to test the suitability of the combination of Ädel's taxonomy and the TED talks. The goal of this annotation was to find which metadiscursive categories are present in the TED talks. Ten TED talks were annotated with the tags from the chosen taxonomy (see Figure 1). The ten talks were randomly chosen, spanning a variety of topics and years. This annotation task was performed by the first author.

The following paragraphs, each named after the 4 main categories of the taxonomy, present the taxonomy itself and, at the same time, describe how each type of metadiscourse is distributed over the sample.

**Metalinguistic Comments** are composed of 5 metadiscursive acts: REPAIRING, REFORMULATING, COMMENTING ON LINGUISTIC FORM/MEANING, CLARIFYING and MANAGING TERMINOLOGY. Most of these categories are exclusive to spoken discourse. From this set, only CLARIFYING and MANAGING TERMINOLOGY (defining of concepts) were found consistently in the sample. We believe that the fact that the other tags were not found is due to the high degree of preparation of each talk (when compared to academic lectures).

[1] http://www.ted.com/

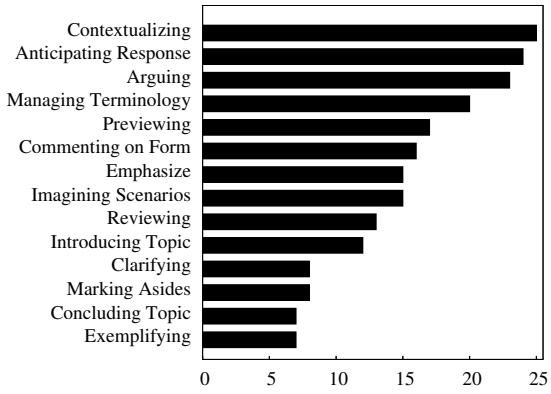


Figure 2: Occurrences of the most frequent tags.

**Discourse Organization** is divided in two other categories: *Manage Topic* and *Manage Phorics*. In *Manage Topic*, there are 5 metadiscursive acts: INTRODUCING TOPIC, DELIMITING TOPIC, ADDING TO TOPIC, CONCLUDING TOPIC and MARKING ASIDES. These structures were found consistently throughout the sample. The audience comes from a broad set of areas, and the speakers must wisely structure their discourse to convey their message. Additionally, the short time frame that is allotted for each talk demands an efficient use of language. The exceptions in this group were the tags DELIMITING TOPIC and ADDING TO TOPIC. The reason behind this may be the fact that TED talks have well-defined topics. The speakers tend to focus on what they want to talk about, going straight to the relevant points. *Manage Phorics*, the other subcategory under *Discourse Organization*, has four tags. PREVIEWING, REVIEWING and CONTEXTUALIZING are related to pointing to other locations in the current discourse. ENDOPHORIC MARKING contains references to physical elements (such as an image in the presentation), and was not considered in this preliminary task since it involved the integration of elements outside the discourse. The first three categories were well-represented in our sample.

**Speech Acts** contains 3 metadiscursive acts: ARGUING, EXEMPLIFYING, and OTHER (where the author included acts that were not frequent enough to generate a new tag). The first two tags were found frequently in the sample, and the category *other* was ignored since it did not represent a single concept.

**References to the Audience** is related to contact with the audience. Unlike in academic lectures, in TED talks the speaker typically does not interact with the audience. The message has to be conveyed without direct interaction, such as questions and checks for understanding. For these reasons, the tags MANAGE COMPREHENSION (check if the audience is in synch with the content of the presentation) and MANAGE DISCIPLINE (adjusting the channel asking for less noise, for example), were not found and therefore were not considered. The remaining 3 tags in this category ( ANTICIPATING RESPONSE, MANAGING THE MESSAGE and IMAGINING SCENARIOS), on the other hand, were found frequently throughout the sample.

Figure 2 shows the distribution of the most frequent tags found in the ten talk sample. From the resulting fourteen categories, a small subset was chosen for the initial annotation effort in which we tested the suitability of using non-experts to identify occurrences of metadiscourse. Three criteria dictated the set of tags used in this annotation task. We considered (a) the most frequent concepts in the literature on presentation skills, (b) the concepts that could be best explained to non-experts, and (c) the input from Carnegie Mellon's International Communications Center (entity that holds presentation skills workshops and is responsible for administering tests for non-native speakers applying for teaching assistant positions). The resulting set of five tags are: INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING and MANAGING THE MESSAGE. Additionally, under the category EXEMPLIFYING we decided to collapse both EXEMPLIFYING and IMAGINING SCENARIOS (since they both consist of illustrating an idea). For simplification, MANAGING THE MESSAGE (in Ädel's work, "typically used to emphasize the core message in what is being conveyed") will be referred to as EMPHASIZING.

## 5. Crowdsourcing

It has been shown that the quality of the crowdsourcing results can approach that of an expert labeler, while requiring less monetary- and time-related resources (Nowak and Rüger, 2010; Zaidan and Callison-Burch, 2011; Eskenazi et al., 2013). However, this advantage comes at a cost. Unlike experts, using the crowd requires setting up training and quality assurance mechanisms to eliminate noise in the answers. Additionally, it is necessary to approach problems in a different way, such as dividing complex jobs in subtasks to reduce cognitive load (Le et al., 2010; Eskenazi et al., 2013).

In our case, the reasons behind using crowdsourcing go beyond time and money. It allows the assessment of the crowd understanding. By designing a task requiring the annotation of metadiscourse, we are building a corpus of the phenomenon and understanding how non-experts comprehend metadiscursive concepts.

In the remainder of this section, we will describe the setup of a crowdsourcing annotation task (run on Amazon Mechanical Turk[2]).

The first decision concerned the amount of text that workers would annotate in each HIT (Human Intelligence Task – the smallest unit of work someone has to complete in order to be paid). Each HIT had to be simple and to allow workers to do it in the fastest way possible. However, metadiscursive phenomena are not local, requiring understanding the context. With this in mind, we decided to use segments of approximately 300 words. This limit was influenced by the design of the interface of the annotation task, taking into consideration that all the text should be visible on the screen without having to scroll down (scrolling increases time-on-task, influencing the answer rate). To make it monetarily worthwhile for a worker to chose our task, we included four segments per HIT, shown in a 2 by 2 matrix. Figure 3 shows

[2]https://www.mturk.com/mturk/welcome

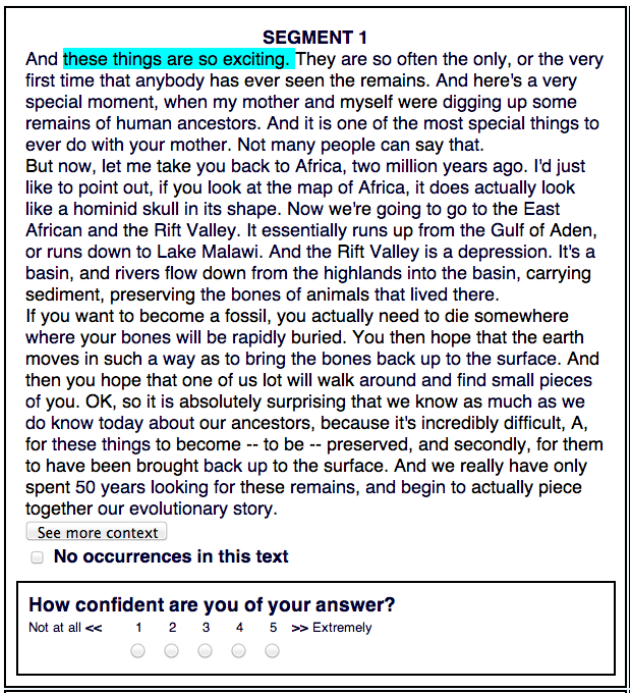


Figure 3: One of the four segments in a HIT

the interface for one of the segments in a HIT. This configuration generated 2,461 HITs (or 9,844 segments) per category. It is also important to notice the presence of the button *See more context* in Figure 3. This feature allowed the workers to see the surrounding text of the segment in the talk (before and after), in case they needed additional context to support their decision.

The second consideration concerns the design of the instructions. Knowing that a metadiscursive act is a complex notion which workers may have never heard of, we decided that each HIT would target only one category, instead of requiring the identification of all five categories in each segment in one single passage. This decision lessens the cognitive load for the workers at each point. The instructions, for the *emphasis* task read as follows:

> *When making a presentation, to guide the audience, we often use strategies that make the structure of our talk explicit. Some strategies are used to announce the topic of the talk ("I'm going to talk about…"; "The topic today will be…"), to conclude a topic or the talk ("In sum,…"; "To conclude,…"), to emphasize ("The take home message is …"; "Please note that…"), etc. We believe that by explaining and explicitly teaching each of these strategies we can help students improve their presentation skills.*
>
> *In this task, we ask you to focus on the strategies that the speaker uses to EMPHASIZE A POINT. Your job is to identify the words that the speaker uses to give special importance to a given point, to make it stand out, such as "more important", "especially", or "I want to stress that…". The passages you mark will be used on a presentation skills virtual tutor, showing students how professional speakers EMPHASIZE a point.*

Since the idea is to do one pass over all the segments for each one of the tags, we designed different sets of instructions for each one of the five metadiscourse categories. It is important to notice that the first paragraph of the instructions above was only included after some preliminary trials. Its inclusion was intended to reveal a concrete example of the applicability of our work, as a way of motivating workers. The inclusion of this paragraph increased the response rate. After the instructions there is a section with examples and counterexamples derived from the preliminary annotation task. Finally, at the bottom of the page, before the presentation segments, there is a succinct set of steps that explain the interface and how to use it to annotate the passages:

> **STEP 1:** For each of the extracts below, click on EVERY word that the speaker uses to EMPHASIZE A POINT. There may be zero, one ore more instances in each extract.
>
> **STEP 2:** The words you click on will display a light blue background. If you change your mind, you can click on the word again to deselect it.
>
> **STEP 3:** If you need more information to support your decision, you can click "*See more context*" below the segment to see the its surrounding context in the talk.
>
> **STEP 4:** If the speaker does not emphasize any point in the extract, select the "*No occurrences in this text*" checkbox below the text.
>
> **STEP 5:** Click the SUBMIT button once you are finished.

The last set of considerations had to do with quality control. We took advantage of the AMT prerequisites feature to filter out workers who were not native-speakers of English and find those who had a high reliability rate ($\geq$95%). Workers who satisfied the prerequisites and accepted the HIT were then guided through a four-segment training session. The training tested if the worker read the instructions and examples carefully, and was capable of performing this task. Only upon successful completion of the four training segments were the workers allowed to access real HITs in the category they were just certified on.

This training strategy is effective in filtering out *bots*, however it does not prevent malicious workers from giving random answers to the real HITs. For that reason, and in line with what is done in much of the crowdsourcing community, we defined a gold standard for each of the five metadiscursive tags. In every four HITs, at least one segment was compared to an expert annotation. The gold standard segments were very similar to the examples provided, and failing one of them raised a flag for the worker. This information was then checked before accepting or rejecting that annotator's work. Workers also noted their confidence level for each segment on a 5-point Likert scale (see Figure 3).

| Category | time (m) | Confidence | Context Requests (%) |
|---|---|---|---|
| ASD | 10 | 3.60 | 5.52 |
| INT | 3.7 | 3.95 | 1.32 |
| CONC | 3.5 | 4.00 | 37.09 |
| EXMPL | 6.2 | 3.94 | 4.81 |
| EMPH | 6.3 | 3.99 | 1.14 |

Table 1: Results in terms of time-on-task, self-reported confidence score and percentage of context expansion requests for MARKING ASIDES (ASD) INTRODUCING TOPIC (INT), CONCLUDING TOPIC (CONC), EXEMPLIFYING (EXMPL) and EMPHASIZING (EMPH).

A final mechanism to assure quality consisted of submitting the same HIT to 3 different workers, using a majority vote scheme.

Prior to publishing all the HITs in each category, we uploaded a small sample of 100 HITs to test the suitability of the instructions and interface. This trial phase allowed us to modify the instructions and examples for each category if necessary, and to test if the workers were able to understand and identify the metadiscourse act.

## 6. Results

This section presents the results of the annotation for each of the five metadiscursive acts. In Table 1 we report the results in terms of average time-on-task in minutes; self-reported confidence score on a 5-point Likert scale; and percentage of segments in which workers expanded context (by clicking on the *See more context* button). Table 2 indicates the number of occurrences of the metadiscourse tag; and inter-annotator agreement ($\kappa$). We used the Fleiss' kappa (Fleiss, 1971) as a measure of annotator agreement. Complete agreement corresponds to $\kappa = 1$, and no agreement (other than chance) corresponds to $\kappa \leq 0$. Herein, annotators agree if the intersection of the words selected by each of them is not empty. For example, two workers agree when one selects "Today, I would like to say that" and the other misses some of the words, selecting "I would like to say".

It is important to notice the absence of the tag MARKING ASIDES in Table 2. All the categories with the exception of MARKING ASIDES produced satisfying results in the trial sample of 100 HITs uploaded prior to submitting the entire set of talks. This fact lead us to discard the asides-related category. This will be discussed in detail in Section 6.1.

### 6.1. Marking Asides

As mentioned, the annotation of MARKING ASIDES was discontinued due to the inconclusive results obtained during the AMT trial phase. The first indicator of unsuccessful annotations was the slow response rate. The 100 HITs were up for one week during which less than 50% were completed. In the remaining categories, the sample was fully completed in less than two days. This slow response rate could be due to the small amount of HITs that were uploaded (workers tend to focus on tasks that have a significant amount of HITs online, in order to minimize training time and maximize payment). However, the four other categories were also first presented with 100 HITs and completed much faster.

| Category | # occurrences | $\kappa$ |
|---|---|---|
| INT | 1,159 | 0.64 |
| CONC | 628 | 0.60 |
| EXMPL | 1,327 | 0.72 |
| EMPH | 2,580 | 0.58 |

Table 2: Number of occurrences and inter-annotator agreement (Fleiss' kappa) for the completed categories: INTRODUCING TOPIC (INT), CONCLUDING TOPIC (CONC), EXEMPLIFYING (EXMPL) and (EMPH).

We looked for other indicators and decided that the crowd could give us some insight on the understanding of the concept MARKING ASIDES. Workers were spending 10 minutes on average for each HIT, contrasting with the 4 to 6 minutes the other tasks took. Self-reported confidence scores were also the lowest of the five categories: 3.60, as opposed to 4.00 for the category CONCLUDING TOPIC. Finally, the workers wrote comments that clearly showed the task was hard, justifying the slow response rate and lack of confidence. Workers wrote: *"I am nervous that I am not doing these correctly *at all*"*; *"I hope that this is what you are looking for"*; and *"a little difficult"*.

### 6.2. Introducing a Topic

The task of annotating introductions resulted in an inter-annotator agreement of 0.64. Workers took on average 3.7 minutes to complete each HIT and identified over 1,000 instances of INTRODUCING TOPIC in our set of talks. It is important to note that speakers sometimes introduce several topics throughout a single talk, and therefore there can be more occurrences of INTRODUCING TOPIC than the total number of talks in the set (in this case 730). A final interesting point was the low number of times that workers asked for more context: only in 1.32% of the segments.

### 6.3. Concluding a Topic

The annotation of conclusions provided results that resembled the previous category: a slightly lower inter-annotator agreement ($\kappa = 0.60$), and similar average time-on-task and self-reported confidence. An important difference comes from the percentage of segments for which annotators asked to see the surrounding context: 37% of the segments. This might be an indication that conclusions are less local, needing a wider context to be identified. It is also important to notice that the number of occurrences of conclusions (628) is lower than the number of talks. This aligns with what we encountered in the preliminary annotation task (7 conclusions over 10 talks) and is related to the fact that the speakers do not always explicitly conclude (particularly true for shorter talks).

### 6.4. Exemplifying

In this category, workers spent on average two more minutes per HIT than while annotating instances of INTRODUCING TOPIC and CONCLUDING TOPIC. This results from the greater quantity of occurrences detected (1,327).

The more occurrences a category has, the more time workers will spend clicking on them. As previously described, this category collapses two metadiscursive acts as defined in Ädel's taxonomy: EXEMPLIFYING and IMAGINING SCENARIOS. Despite the collapse of tags, in this category annotators reached the highest agreement ($\kappa = 0.72$), which corroborates our decision to combine the two tags.

### 6.5. Emphasizing

While annotating occurrences of EMPHASIZING, the relationship between average time-on-task and number of instances was similar to the one found for the previous category. Workers spent on average 6.3 minutes per HIT and identified over 2,500 occurrences. While identifying emphasis, workers asked for the lowest amount of additional context amongst the five categories (1.14). EMPHASIZING was also the category where workers achieved the lowest inter-annotator agreement (0.58). This result may be due to the fact that this category is the only one in which there is a scale of intensity related to the concept, i.e., different workers might have different thresholds for considering that the speaker is emphasizing.

## 7. Discussion

The results obtained in this annotation task show that, once trained, non-experts can understand concepts of metadiscourse and identify them on TED presentations. However, this is not true for all of the categories we proposed to annotate. The category MARKING ASIDES was discarded during the trial phase on AMT since workers manifested signs of not understanding the task.

After the experiment took place, we looked into the instructions for this category to understand why workers were not able to annotate it. One of the counterexamples stressed the difference between MARKING ASIDES (where the speaker digresses to a topic sidetrack, such as in "Just a little side note here. . .") and ADDING TO TOPIC (where the speaker explicitly adds to the current topic, such as in "Let me add that. . ."). This distinction may have added to the worker's cognitive load. They were not only asked to be aware of another category in the taxonomy, but also required to focus on a subtle difference. The solution to this problem may be the division of the category in two. This can be done with a first pass collapsing both concepts under a more general notion, such as *adding information*, and a second pass where workers now only see instances that were detected in the first pass and decide if the addition of information is on or off-topic.

Another interesting result from this experiment is the need for additional context in different metadiscursive acts. Workers were able to identify occurrences of INTRODUCING TOPIC and EMPHASIZING in a window of 300 words without requesting for additional context. On the other hand, identifying conclusions was the task where more context was needed. The fact that workers expanded context in 37% of the segments might result from the necessity to first understand which topic is being presented, before deciding on the occurrence of its conclusion.

## 8. Conclusion and Future Work

In this paper, we have described an annotation task that took place on Amazon Mechanical Turk, where workers focused on a predefined set of metadiscourse categories to annotate text extracted from TED talks. We started from a set of 730 presentations and a taxonomy of metadiscourse and described the considerations for setting up a crowdsourcing annotation task aimed at finding metadiscursive concepts in the talks. The task was successful for four of the five categories that were submitted.

In future work, we plan to continue this annotation effort, extending it to the remaining categories of Ädel's taxonomy, and refining unsuccessful attempts (i.e. MARKING ASIDES) to meet the workers' cognitive load. We plan to extend this analysis to other languages, more precisely to European Portuguese, comparing the use of metadiscourse between the two languages. Finally, we aim at using the resulting annotation as training data for an automatic metadiscourse classifier.

## Acknowledgments

This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and by FCT project CMU-PT/HuMach/0053/2008.

## 9. References

Ädel, Annelie. (2010). Just to give you kind of a 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.

Crismore, Avon, Markkanen, Raija, and Steffensen, Margaret S. (1993). Metadiscourse in persuasive writing. *Written communication*, 10(1):39.

Eskenazi, Maxine, Levow, Gina-Anne, Meng, Helen, and Parent, Gabriel. (2013). *Crowdsourcing for Speech Processing*. John Wiley & Sons.

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

Le, John, Edmonds, Andy, Hester, Vaughn, and Biewald, Lukas. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 21–26.

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

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

Marcu, Daniel. (2000). *The theory and practice of discourse parsing and summarization*. The MIT press.

Marcus, Mitchell P, Marcinkiewicz, Mary Ann, and Santorini, Beatrice. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mauranen, Anna. (2001). Reflexive academic talk: Observations from MICASE. In *Corpus linguistics in North*

*America: Selections from the 1999 symposium*, pages 165–178.

Miltsakaki, Eleni, Robaldo, Livio, Lee, Alan, and Joshi, Aravind. (2008). Sense annotation in the Penn Discourse Treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286. Springer.

Nowak, Stefanie and Rüger, Stefan. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 557–566. ACM.

Römer, Ute and Swales, John M. (2009). The Michigan Corpus of Upper-level Student Papers (MICUSP). *Journal of English for Academic Purposes*, April.

Simpson, Rita C., Briggs, Sarah L., Ovens, Janine, and Swales, John M. (2002). The Michigan Corpus of Academic Spoken English.

Zaidan, Omar F. and Callison-Burch, Chris. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1220–1229.