metaTED: a Corpus of Metadiscourse for Spoken Language

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 describes metaTED – a freely available corpus of metadiscursive acts in spoken language collected via crowdsourcing. Metadiscursive acts were annotated on a set of 180 randomly chosen TED talks in English, spanning over different speakers and topics. The taxonomy used for annotation is composed of 16 categories, adapted from Ädel (2010). This adaptation takes into account both the material to annotate and the setting in which the annotation task is performed. The crowdsourcing setup is described, including considerations regarding training and quality control. The collected data is evaluated in terms of quantity of occurrences, inter-annotator agreement, and annotation related measures (such as average time on task and self-reported confidence). Results show different levels of agreement among metadiscourse acts ($\alpha \in [0.15; 0.49]$). To further assess the collected material, a subset of the annotations was submitted to expert appreciation, who validated which of the marked occurrences truly correspond to instances of the metadiscursive act at hand. Similarly to what happened with the crowd, experts revealed different levels of agreement between categories ($\alpha \in [0.18; 0.72]$). The paper concludes with a discussion on the applicability of metaTED with respect to each of the 16 categories of metadiscourse.

Keywords: metadiscourse, spoken language, crowdsourcing

1. Introduction

Commonly referred to as *discourse about discourse*, metadiscourse 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...*).

This study focuses on the function of metadiscourse in spoken communication. As previously mentioned, metadiscourse reflects the explicit intention of the speaker and, therefore, its analysis uncovers a map of explicitly stated discourse functions. This idea contrasts with implicit information conveyed, for instance, by means of prosody.

For example, the speaker is able to enumerate simply by means of prosody, such as in the sentence *"they use several data modalities (localization, physiological, ...)"*, where the speaker will resort to pauses and intonation patterns commonly associated with the function of enumerating. On the other hand, if the items in an enumeration are complex and/or more important to the content, the speaker may explicitly signal them such as in *"They use several data modalities.* ***The first,*** *localization is ...* ***The second*** *is physiological ..."*.

Another example of the implicit *vs* explicit use of discourse functions is the way a speaker chooses to emphasize a point, which, on the one hand, can be done by simply increasing the intensity of speech (thus not using metadiscourse), or, on the other hand, can include explicit mentions to the importance of the idea, such as *"This is very important for you to understand"*.

The current paper describes the task of building a corpus of metadiscursive acts. It addresses the problem of annotating a sparse, multi-category phenomenon in a crowdsourcing framework. The collection and analysis of these explicit cues given by the speaker while presenting has a direct application for language learning purposes, more precisely in what concerns presentational skills, where metadiscourse can be used as a key concept during instruction. For the Natural Language Processing community, it contributes to the goal of natural language understanding, and, consequently, can be used to improve tasks such as simplification or translation.

This work is organized as follows:

- Section 2. presents background on the phenomenon of metadiscourse, with particular focus to previous efforts of annotation of metadiscourse-related concepts;
- Section 3., named after the resource, starts by providing support for the choice of metadiscursive theory and data sources. It contains considerations regarding the setup of the annotation task (interface, training, and quality control). It also presents the results of the annotation in both a quantitative and qualitative evaluation;
- Section 4. describes an expert validation task, where experts assessed a subset of the annotations provided by the crowd and decided if they corresponded, in fact, to the metadiscursive act at hand;
- Section 5. discusses the quality of the data obtained, relating both the crowd and the experts' answers, and infers on the applicability and usefulness of the data for each category considered;
- Section 6. concludes and presents future work directions.

2. Background

The first systematic approaches to metadiscourse were proposed by Williams (1981) and Meyer et al. (1980) and were further adapted and refined by Crismore (1983) and follow-up work (Crismore, 1984), in a taxonomy that is still broadly used today (Camiciottoli, 2003; van Aertselaer, 2008).

Crismore's taxonomy is divided in two main categories: `Informational` and `Attitudinal` metadiscourse. The former deals with discourse organization, being divided in PRE-PLANS (preliminary statements about content and structure), POST-PLANS (global review statements), GOALS (both preliminary and review global goal statements), and TOPICALIZERS (local topic shifts). `Attitudinal` metadiscourse, as the name states, is used to show the speaker's attitude towards the discourse, and encompasses SALIENCY (importance), EMPHATICS (certainty degree), HEDGES (uncertainty degree), and EVALUATIVE (speaker attitude towards a fact).

This theory, while setting the first standard on the subject, does not cover the full spectrum of metadiscursive acts, particularly in relation to what concerns their function in spoken language (such as speaker-audience interaction). This motivated the examination of theoretical underpinnings dealing with spoken language.

Luukka (1992) developed a taxonomy for use with both written and spoken academic discourse, 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 process) and `Contextual` (covering references to audiovisual materials).

Mauranen (2001) focused only on spoken discourse. The resulting 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 of the audience with the speaker).

Auria (2006) focused on the use of spoken metadiscourse in academic settings, referring to it as a powerful linguistic resource in academic speech. The main concept behind this taxonomy is lecturer intention. It is divided in three categories: the I-PATTERN represents the speakers' overt presence when expressing their communicative intentions, while the WE-PATTERN and the POLITE DIRECTIVES are alternatives that seek to establish solidarity relationships between the speaker and the audience.

Even though the taxonomies proposed by Luukka (1992), Mauranen (2001), and Auria (2006) organize metadiscourse in similar ways (i.e. with respect to the number of stakeholders involved), their approaches focus solely on form, rather than function.

More recently, however, Ädel (2010) proposed a function-oriented take on metadiscourse, merging previous approaches under a framework that encompasses both spoken and written discourse. The resulting taxonomy 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 focus on function is emphasized by the author's concern for presentational skill education, stressing that a pedagogically packaged approach to metadiscourse can be beneficial, especially for non-native speakers of English. Ädel (2010) organizes a total of 23 concepts under four main categories `Metalinguistic Comments`, `Discourse Organization`, `Speech Act Labels`, and `References to the Audience`.

While the aforementioned studies discuss metadiscursive theory (its form and its function in language), they do not contribute to the goal of corpora building. Even in the cases where some kind of annotation was performed, they are not freely available and/or are comprised of a limited number of examples used only to support the category organization decisions.

Therefore, it is also important to look at approaches that represent extensive annotation efforts. From this standpoint, two distinct data-driven projects are broadly used and discussed.

One is the Penn Discourse TreeBank (PDTB) (Webber and Joshi, 1998), built directly on top of Penn TreeBank (Marcus et al., 1993), composed of extracts from the *Wall Street Journal*. PDTB enriched the Penn TreeBank with discourse connectives annotation (conjunctions and adverbials), and organized them according to meaning (Miltsakaki et al., 2008), considering categories such as giving examples (INSTANTIATION), making reformulations and clarifications (RESTATEMENT), comparing (CONTRAST), or showing cause (REASON).

The second project is the RST Discourse Treebank (RSTDT) (Marcu, 2000), a semantics-free theoretical framework of discourse relations, intended to be "general enough to be applicable to naturally occurring texts and concise enough to facilitate an algorithmic approach to discourse analysis". Similarly to PDTB, the RSTDT is a discourse-annotated corpus intended to be used by the NLP community, based on *Wall Street Journal* articles extracted from the Penn Treebank. The difference between PDTB and the RSTDT is the discourse framework: in the latter case this is the Rhetorical Structure Theory (Mann and Thompson, 1988), which includes categories such as EXAMPLE, DEFINITION, or SUMMARY.

Additionally, in the sequence of this work, Soricut and Marcu (2003) developed SPADE[1]. SPADE stands for Sentence-level PArsing for DiscoursE and, as the name states, processes one sentence at a time and outputs one discourse parse tree per sentence.

Even though PDTB and RSTDT make available two extensive corpora of different discourse functions, they have two drawbacks. Firstly, they do not address the metalinguistic aspects of language, i.e., do not make distinction between explicit and implicit use of the several discourse functions they analyze, as discussed in Section 1.; and, secondly, they are both built upon *Wall Street Journal* articles, meaning that they do not encompass any strategies or examples that may be characteristic of spoken presentational discourse.

[1] http://www.isi.edu/licensed-sw/spade/

3. metaTED

The corpus presented in this work – metaTED – was built having in mind the aforementioned limitations of existing research:

- it targets the **metalinguistic aspects of language**, being a representation of the explicit cues that reveal speaker intention;
- it aims at illustrating the phenomenon of metadiscourse as used in **spoken language**;
- it adopts a purely **functional approach**, with metadiscursive concepts being associated with their role in discourse.

The source material chosen for annotation, as the name of the resource indicates, was the set of TED talks[2]. These presentations were chosen for being self-contained, widely known for their speakers' quality, and for targeting a general audience. These aspects contrast with classroom recordings, for example, which are typically longer, and targeted at a very specific audience, requiring a significant amount of previous knowledge.

With respect to the theory that serves as basis for the annotation, the proposal form Ädel (2010) was chosen, which follows from the analysis in Section 2. The final set of concepts that compose metaTED resulted from an analysis of which acts from Ädel (2010) could be found in the TED talks. Some of the categories of the original taxonomy were discarded given their low representation or non-existence (such as strategies related to managing the communication channel). Other categories, that did not have enough representation by themselves, were collapsed when it was possible to define them under a broader concept. Finally, there was also the case of one category subdivision (when concepts were better explained individually).

As a result, a final set of 16 discourse functions described in metaTED was achieved, and it is composed as follows:

- `ADD` – collapsed from *Adding to Topic* and *Marking Asides*
- `ANT` – *Anticipating Response*
- `ARG` – *Arguing*
- `CLAR` – *Clarifying*
- `COM` – *Commenting on Linguistic Form/Meaning*
- `CONC` – *Concluding*
- `DEF` – *Definitions* (originally, *Manage Terminology*)
- `DELIM` – *Delimiting Topic*
- `EMPH` – *Emphasizing* (originally *Managing Message*)
- `ENUM` – *Enumerating*
- `EXPL` – collapsed from *Exemplifying* and *Imagining Scenarios*
- `INTRO` – *Introducing Topic*
- `POST` – *Postponing Topic* (originally, *Previewing*)
- `RCAP` – *Recapitulating* (subdivision of *Reviewing*)
- `REF` – *Refer to Previous Idea* (subdivision of *Reviewing*)
- `R&R` – collapsed from *Repairing* and *Reformulating*

[2]https://www.ted.com/talks

3.1. Annotation Setup

The annotation of metaTED was done through crowdsourcing, on Amazon Mechanical Turk (AMT)[3]. 16 different tasks were uploaded, one per category, so as to lessen the workers' cognitive load at any given point. A set of 180 talks was submitted for annotation, totaling 23,348 sentences and 418,368 tokens.

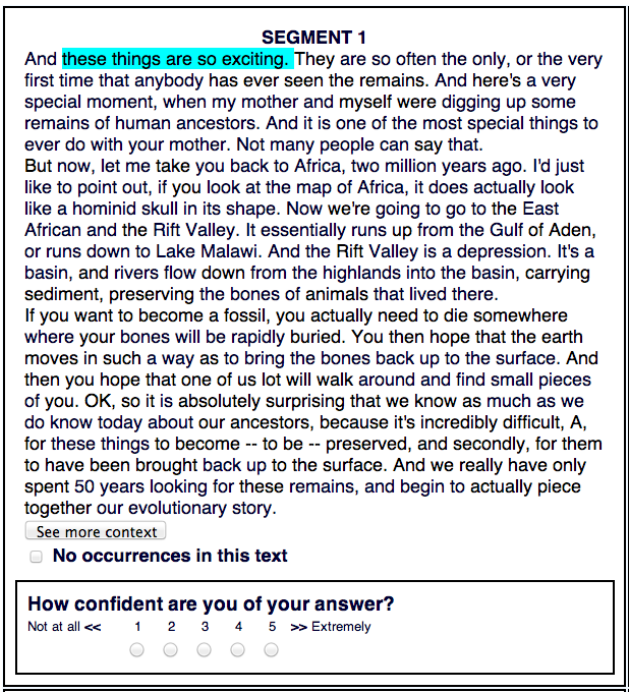

Figure 1: Interface and example annotation (in blue) for the category `EMPH`.

Figure 1 shows the interface for one of the tasks, in this case for the category `EMPH`, where workers clicked on the words that represented the function at hand. In line with common practices of crowdsourcing (micro-tasks), each of the 180 transcripts was divided into segments of 500 words, generating a total of 742 tasks per category. The choice of providing a larger segment, instead of only one sentence, as an example, has to do with the fact that metadiscourse is not a local phenomenon, requiring the surrounding context to be detected. The button *"See more context"* in Figure 1 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.

For quality control purposes (a) only native English speakers with rate of previously accepted work $\geq$95% were considered, (b) training sessions with category explanation, examples, counter-examples, and targeted feedback were set up for each category, (c) answers were compared to golden standards, (d) and workers were asked for a self-confidence report on a 5-point Likert scale for each segment.

Additionally, for reliability and to be able to report agreement, three workers annotated each pair category-segment. A more detailed version of the instructions and interface can be found in Correia et al. (2014b).

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

Category	Workers in Agreement				Expansion	Self-reported	Avg.	Observ.	α2+3	α
	1	2	3	2+3	(%)	confidence	Time	agr (%)		
ADD	923	102	33	135	3.51	3.88 (1.10)	01:55	97.14	0.65	0.16
ANT	1,426	356	100	456	3.80	3.61 (1.02)	01:56	95.07	0.65	0.24
ARG	1,538	322	223	545	4.56	3.51 (1.18)	02:02	94.77	0.72	0.31
CLAR	1,975	283	58	341	3.46	3.82 (0.90)	02:27	93.57	0.62	0.15
COM	738	271	85	356	1.39	3.10 (0.76)	02:08	97.19	0.66	0.34
CONC	153	52	34	86	18.67	4.36 (0.78)	01:12	99.42	0.75	0.43
DEF	836	189	68	257	5.62	4.04 (0.85)	02:27	97.13	0.67	0.28
DELIM	132	28	12	40	1.85	4.21 (0.79)	01:53	99.58	0.70	0.30
EMPH	2,023	336	80	446	4.16	3.31 (0.98)	02:17	93.41	0.52	0.18
ENUM	1,067	368	346	714	2.50	3.74 (0.70)	02:01	95.95	0.79	0.49
EXPL	771	195	140	335	2.50	3.62 (0.72)	01:58	97.31	0.77	0.39
INTRO	732	239	131	370	5.08	3.40 (1.17)	01:33	97.31	0.73	0.39
POST	184	23	24	47	4.20	4.17 (0.69)	01:55	99.45	0.80	0.32
RCAP	202	32	4	36	11.78	3.33 (0.76)	02:15	99.35	0.58	0.16
REF	411	83	42	125	3.16	3.93 (0.54)	01:50	98.63	0.72	0.29
R&R	1,493	233	46	279	3.35	3.57 (0.96)	02:23	95.03	0.61	0.16

Table 1: Annotation results in terms of collected data quantity and quality.

3.2. Annotation Results

Table 1 shows the results of the crowdsourcing task in terms of number of instances, annotation statistics, and inter-annotator agreement rates.

The first four columns represent the number of sentences where metadiscourse was identified. This information is organized by how many workers agreed on each instance. For example, for the category ADD, there were only 33 occurrences that were selected by all three workers, 102 occurrences selected by two of the workers, and 923 occurrences marked by only one worker. The column *2+3* represents the majority vote, i.e., the number of sentences that were signaled by 2 or more workers.

Regarding the percentage of times workers asked for additional context (column *Expansion (%)*), the categories CONC and RCAP show significant differences from all other categories. Workers asked for more context approx. 19% and 12% of the time, respectively. On the other hand, the categories that seem to be more local, not needing so much more supplementary context to be identified (besides the 500 words given), are COM and DELIM, where additional context was asked for less than 2% of the time.

The next column shows the average self-reported confidence on a 5-point Likert scale and corresponding standard deviation (reported on a subset of 100 segments). All categories scored above the middle of the scale (3), with workers showing less confidence for COM, which corresponds to the speaker commenting on their choice of words or on the definition of terms. Contrarily, workers show the highest confidence for CONC, DELIM and POST, interestingly three categories that signal the change of topic in a talk.

Regarding time on task, no significant variations were observed, most categories requiring about 2 minutes per segment. The only exception seems to be CONC, taking only about one minute per segment. It is interesting to notice that this was the category where workers most expanded context and achieved the second best inter-annotator agreement.

The last three columns on Table 1 report different measures of agreement: observable agreement (percentage of items agreed upon), Krippendorf's α ignoring the occurrences marked by one worker only, and α considering all data. Krippendorff's alpha was used since it adjusts itself better to small sample sizes than Cohen's Kappa (Krippendorff, 2007). As with Cohen's κ, perfect agreement corresponds to $\alpha = 1$, while $\alpha = 0$ corresponds to the agreement that can be expected by chance.

Herein, two workers are in agreement when the intersection of the words they select 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"*.

The last column shows that non-experts have the most trouble while identifying instances of CLAR, ADD, RCAP, R&R, and EMPH, all with $\alpha < 0.20$. The categories CONC and ENUM, on the other hand, show the highest level of agreement.

Metadiscourse is a sparse phenomenon, even more so when dealt with one category at a time. It follows that the probability of two workers selecting the same passage by chance is very low. This quantity is taken into account when calculating agreement, and consequently, the case where one worker selects a word and others do not is severely penalized. Previous annotation attempts on similar phenomena, such as Wilson (2012) work on metalanguage, show agreement values in the order of $[0.09; 0.39]$ for sparser acts, even when annotated by experts and considering only four categories.

The impact of the occurrences marked by one worker only (out of three) can be observed by comparing the two last columns on Table 1. When filtering out the answers selected exactly by one worker only, which corresponds to assume majority vote, annotator agreement goes up drastically ($\alpha \in [0.58; 0.80]$), with only one category ranking bellow the 0.6 threshold, commonly referred to as substantial agreement.

Given this difference in agreement between the last two columns on Table 1, it is important to understand which were the main sources of disagreement. Below are grouped five situations that contributed to this disparity:

Variance in interpretation
In categories such as `Emphasizing` and `Arguing` it was possible to observe that workers approached the annotation from different standpoints. For instance, regarding `EMPH`, some workers signaled occurrences where the emphasis was put in a very subtle form (such as *"An important result..."*), while others marked only cues that were much more explicit (such as *"What I really want you to take home"* or *"The real important issue here is..."*).

Span of occurrences
Another source of disagreement was the fact that some instances are spread out along different sentences, such as in the case of the categories `Clarifying` or `Enumerating`. This type of problem was more severe for `CLAR` since one commonly used structure is of the form *"I'm not saying that... What I really mean is..."*. While these two statements are part of the same instance of a clarification, they can be spread out in the discourse (including being separated in two different 500-word segments). Looking at the data, it is possible to see that some workers selected only the first or second parts of the occurrence. Not knowing *a priori* if these cases are actually part of the same occurrence or consist of two separate instances, it is impossible for the inter-annotator agreement metric here used to capture this phenomenon.

Cognitive load
When designing the annotation task, as pointed out in the beginning of this section, it was decided to merge some of the categories with lower representation together under a common concept (when such a concept existed), such as the case of the categories `ADD`, `EXMPL`, and `R&R`. This however may have added to the cognitive load of workers and hindered the annotation.

Category confusability
When looking at the intersection of annotations between categories, three pairs of categories stood out. Workers had a hard time distinguishing between (a) `Clarifying` and `Repairing & Reformulating`; (b) `Defining` and `Commenting on Linguistic Form/Meaning`, and (c) `Recapitulating` and `Referring to Previous Idea`. The definition and differences between these categories can, in fact, be subtle, which may justify the low-level agreement.

Lack of attention
While workers' answers were compared to a golden standard set (every four tasks), and workers who constantly missed them were removed, there were still some clear occurrences of metadiscourse that have not been signaled. For example, the pattern *"by the way"*, a mark of making an aside that was included in the examples during training, has not always been spotted by all workers.

Cat.	α	1		2		3	
		#	TP	#	TP	#	TP
ADD	0.40	256	0.14	34	0.35	10	1.00
ANT	0.48	236	0.36	45	0.76	19	0.95
ARG	0.62	213	0.18	55	0.64	31	0.87
CLAR	0.28	258	0.09	35	0.37	7	0.71
COM	0.46	218	0.21	65	0.48	17	0.71
CONC	0.72	153	0.32	52	0.75	34	0.88
DEF	0.64	216	0.14	61	0.36	23	0.35
DELIM	0.46	132	0.51	28	0.71	12	0.92
EMPH	0.61	243	0.20	44	0.59	13	0.69
ENUM	0.63	189	0.09	59	0.41	55	0.84
EXPL	0.49	190	0.34	56	0.88	54	1.00
INTRO	0.57	202	0.32	69	0.72	29	0.97
POST	0.67	174	0.13	23	0.39	24	0.88
RCAP	0.18	202	0.09	32	0.28	4	0.25
REF	0.59	217	0.27	56	0.84	27	0.89
R&R	0.59	249	0.12	46	0.39	5	0.80

Table 2: Results of expert revision in terms of agreement (α), occurrence number (#) and true positive rate (TP).

4. Expert Validation

The variation between instances marked by one, two or all three workers (see Table 1) served as motivation to validate the data with experts, and thus gain further insight on the annotations: How many of the cases selected by only one worker are really false positives? What is the rate of true positives for the occurrences selected by all three workers?

Four experts were asked to assess the crowd's annotations: they were given a highlighted occurrence previously marked by the crowd, and decided whether it corresponded to the category at hand or not. Experts validated a sample of 300 occurrences of each category (with the exceptions of `CONC`, `DELIM`, `POST` and `RCAP`, where the total number of occurrences does not meet the 300 threshold). For occurrences marked by more than one worker (columns *2* and *3* in Table 1), experts were presented with the union of the workers' answers. They were also asked to focus on the existence or non-existence of the function at hand, being permissive about the boundaries of the selection. Two experts revised each occurrence. In case of disagreement, a third opinion was requested.

Table 2 shows, for each category, the total inter-annotator agreement achieved by the experts, the number of instances evaluated and corresponding true positive rate. It is important to highlight that agreement here is not directly comparable with the values in Table 1, given the difference in the tasks (identification vs. correction).

For most categories, experts achieved an inter-annotator agreement above 0.40. The exceptions were `Clarifying` and `Recapitulating` with significantly lower agreements (0.28 and 0.18 respectively). These results mimic what happened previously, with these categories being those where the crowd performed the worst. Also in line with the workers' performance, experts agreed the most for the category `Concluding Topic`, with an inter-annotator agreement of 0.72.

The remaining columns on Table 2 show how experts evaluated the crowd's decisions. As previously, results are separated in terms of number of workers involved in the selection of a particular occurrence. Ideally, if following a majority vote rule, the True Positive (TP) rate under the column *1* should be 0 (experts reject all occurrences marked by one worker only), while the TP rate under the columns *2* and *3* should be 1 (experts validate occurrences marked by at least two workers).

As expected, for most categories, there is a growing trend of TP rate with respect to the number of workers in agreement, i.e., the more workers that agreed on a given occurrence, the more likely it is for experts to accept it. Exceptions are DEF and RCAP, with experts even rejecting the majority of the instances selected by all 3 workers. For all other categories, experts accept more than 70% of the occurrences selected by all three workers, reaching perfect agreement ($TP = 1$) for the categories ADD and EXPL.

For the cases that were selected by exactly two workers (column *2*), experts validate more than half the occurrences for 9 of the categories, with EXPL and REF reaching TP rates of above 80%. Below the 50% threshold are the categories ADD, CLAR, COM, DEF, POST, RCAP and R&R, with experts showing to be more strict on what to consider metadiscourse.

Finally, occurrences that were selected by only one worker are consistently rejected. For DELIM however, experts accepted more than half (51%) of the instances.

5. Discussion

The metaTED corpus fills a gap in current research, providing a reference for the explicit cues used by speakers to organize their discourse in spoken language. It is composed of a set of 16 categories that were labeled in a crowdsourcing framework, and therefore, is a representation of non-expert awareness on metadiscursive strategies.

The annotation effort that took place during this work shows that not all acts in the same taxonomy can be understood in the same manner by annotators. Metadiscourse proved to be a hard concept to annotate given the characteristics and similarities of some of the categories adopted.

Table 3 provides a high-level judgment of the quality of the corpus assembled, in terms of quantity of data and agreement. Agreement is represented in the scale suggested in Landis and Koch (1977) (< 0 *no agreement* – $[0;\ 0.20]$ *slight* – $[0.21;\ 0.40]$ *fair* – $[0.41;\ 0.60]$ *moderate* – $[0.61;\ 0.80]$ *substantial* – $[0.81;\ 1]$ *almost perfect*).

The first column in Table 3 shows the categories for which there are at least 200 occurrences where there was consensus between non-experts. Ten metadiscursive acts fulfill this criterion, which serves as an indicator, for example, of the possibility of using the data in NLP-related tasks.

The last two columns in Table 3 provide a representation of the reliability of the data in metaTED by category. The categories ADD, CLAR, and RCAP have serious problems of consensus, for both the crowd and for the experts. On the other end of the spectrum are the categories CONC and ENUM, where agreement was the highest for both non-experts and experts.

Category	>200 occurr.	worker α	expert α
ADD		slight	fair
ANT	✓	fair	moderate
ARG	✓	fair	substantial
CLAR	✓	slight	fair
COM	✓	fair	moderate
CONC		moderate	substantial
DEF	✓	fair	subtantial
DELIM		fair	moderate
EMPH	✓	slight	substantial
ENUM	✓	moderate	substantial
EXPL	✓	fair	moderate
INTRO	✓	fair	moderate
POST		fair	substantial
RCAP		slight	slight
REF		fair	moderate
R&R	✓	slight	moderate

Table 3: metaTED high-level judgement by category, regarding quantity of annotation and annotator agreement.

6. Conclusions

This paper described the building of the metaTED corpus, a collection of functionally oriented metadiscourse acts annotated in a crowdsourcing setting for spoken language data. This corpus represents non-experts' awareness of what metadiscourse is and what function it has in language. Adopting a theory of metadiscourse (Ädel, 2010) and a set of TED talks' transcripts, a resulting set of 16 categories were submitted for annotation on Amazon Mechanical Turk. Several quality control mechanisms were set in place to filter out unwanted work. The crowd showed different levels of understanding regarding the different categories that were presented for annotation, with inter-annotator agreement varying between $[0.15; 0.49]$.

When validating the crowd's work, experts behaved similarly, regarding the categories with the best/worst performance, also showing different levels of agreement ($\alpha \in [0.18; 0.72]$). They also confirmed that the amount of workers agreeing on a given instance is a good indicator of correctness.

Given the aforementioned idiosyncrasies of the corpus, and in order to allow other researchers to make better use of this data, this resource is made available through the LRE Map[4] with all the metadata associated with the annotation (annotator ID, time-on-task, expansion information, self-reported confidence).

Additional work with this corpus can be found in Correia et al. (2014a), a small experiment on automatic classification of metadiscourse with an earlier version of the corpus; and in Correia et al. (2015), where it was exploited for understanding the use of metadiscourse in different levels of English proficiency.

Future work includes using metaTED to build classifiers of metadiscourse that will identify and assign a function to the explicit cues given by speakers in a presentation transcript.

[4] http://www.resourcebook.eu

7. Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (UID/CEC/50021/2013), Instituto Superior Técnico (BL184/2015), and Carnegie Mellon|Portugal program (SFRH/BD/51156/2010).

8. Bibliographical References

Ädel, A. (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.

Auria, C. P. (2006). Signaling speaker's intentions: towards a phraseology of textual metadiscourse in academic lecturing. In Carmen Pérez-Llantada et al., editors, *English as a GloCalization Phenomenon. Observations from a Linguistic Microcosm*, volume 3, pages 59–86. Universitat de València.

Camiciottoli, B. C. (2003). Metadiscourse and ESP reading comprehension: An exploratory study. In *Reading in a foreign language*, volume 15(1), pages 28–44. University of Hawaii, National Foreign Language Resource Center.

Correia, R., Mamede, N., Baptista, J., and Eskenazi, M. (2014a). Toward Automatic Classification of Metadiscourse. In *Proceedings Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL, Warsaw, Poland*, pages 262–269. Springer.

Correia, R., Mamede, N., Baptista, J., and Eskenazi, M. (2014b). Using the Crowd to Annotate Metadiscursive Acts. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, Reykjavik, Iceland*, pages 102–108.

Correia, R., Eskenazi, M., and Mamede, N. (2015). Lexical Level Distribution of Metadiscourse in Spoken Language. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem), Lisbon, Portugal*, pages 70–75.

Crismore, A., Markkanen, R., and Steffensen, M. S. (1993). Metadiscourse in persuasive writing a study of texts written by american and finnish university students. In *Written Communication*, volume 10(1), pages 39–71. Sage Publications.

Crismore, A. (1983). Metadiscourse: What it is and how it is used in school and non-school social science texts. In *Center for the Study of Reading Technical Report; no. 273*. Champaign, Ill.: University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Crismore, A. (1984). The rhetoric of textbooks: Metadiscourse. In *J. Curriculum Studies*, volume 16(3), pages 279–296. Taylor & Francis.

Krippendorff, K. (2007). Computing krippendorff's alpha-reliability. *Departmental papers (ASC)*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. In *Biometrics*, volume 33(1), pages 159–174. JSTOR.

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

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, volume 19(2), pages 313–330. MIT Press.

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

Meyer, B. J., Brandt, D. M., and Bluth, G. J. (1980). Use of top-level structure in text: Key for reading comprehension of ninth-grade students. In *Reading research quarterly*, volume 16(1), pages 72–103. JSTOR.

Miltsakaki, E., Robaldo, L., Lee, A., and Joshi, A. (2008). Sense annotation in the Penn Discourse Treebank. In *Computational Linguistics and Intelligent Text Processing: 9th International Conference, CICLing, Haifa, Israel, February 17-23*, pages 275–286. Springer.

Römer, U. and Swales, J. M. (2009). The Michigan Corpus of Upper-level Student Papers (MICUSP). In *Journal of English for Academic Purposes*. The Regents of the University of Michigan, Ann Arbor, MI, April.

Simpson, R. C., Briggs, S. L., Ovens, J., and Swales, J. M. (2002). The Michigan corpus of academic spoken English. The Regents of the University of Michigan, Ann Arbor, MI.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 149–156. Association for Computational Linguistics.

van Aertselaer, J. N. (2008). Contrasting English-Spanish interpersonal discourse phrases. A corpus study. In John Benjamins, editor, *Phraseology in foreign language learning and teaching*, pages 85–100. Amsterdam.

Webber, B. and Joshi, A. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. In *Coling/ACL workshop on discourse relations and discourse markers*, pages 86–92.

Williams, J. M. (1981). Ten lessons in clarity and grace. *University of Chicago Press, Chicago*.

Wilson, S. (2012). The creation of a corpus of English metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, pages 638–646. Association for Computational Linguistics.