Lexical Level Distribution of Metadiscourse in Spoken Language

Rui Correia
L^{2}F, INESC-ID
Técnico Lisboa
Portugal
Rui.Correia@inesc-id.pt

Maxine Eskenazi
LTI
Carnegie Mellon University
USA
max@cs.cmu.edu

Nuno Mamede
L^{2}F, INESC-ID
Técnico Lisboa
Portugal
Nuno.Mamede@inesc-id.pt

Abstract

This paper targets an understanding of how metadiscourse functions in spoken language. Starting from a metadiscourse taxonomy, a set of TED talks is annotated via crowdsourcing and then a lexical grade level predictor is used to map the distribution of the distinct discourse functions of the taxonomy across levels. The paper concludes showing how speakers use these functions in presentational settings.

1 Introduction

Often referred to as discourse about discourse, metadiscourse is the linguistic material intended to help the listener organize and evaluate the information in a presentation (Crismore et al., 1993). Examples include introducing (*I'm going to talk about ...*), concluding (*In sum, ...*), or emphasizing (*The take home message is ...*).

This paper explores how this phenomenon is used in spoken language, in particular how it occurs across presentations with different vocabulary levels. Are these acts used independently of vocabulary complexity? Which ones are used more frequently in more lexically demanding talks?

Finding out the answer to these questions has not only direct applications in language learning, but can also give insight on features that can be used for automatically classifying metadiscourse.

Such classification establishes a link between discourse and lexical semantics, i.e., understanding the speaker's explicit intention can be of help in tasks such as word sense disambiguation. For instance, the word *means*, in most contexts used to signal a definition, can also be used to show entailment, such as in: *[...] these drugs [...] will reduce the number of complications, which* ***means*** *pneumonia and which* ***means*** *death.*

This paper is organized as follows. Section 2 presents related work on metadiscourse with focus to how it relates with grade level. Section 3 explains the choice of taxonomy of metadiscourse, describes the data and its annotation. Section 4 addresses the measure of vocabulary complexity used in this study, and the distribution of the data across different levels. Section 5 shows the results of mapping the metadiscourse functions according to vocabulary level. Finally, Section 6 has a discussion of the results and conclusions.

2 Related Work

The way discourse is used and organized in different grade levels started receiving attention in the early 80's. Crismore (1980) focused on the use of a set of logical connectives at different levels and disciplines (high school through university), showing difficulty of mastery. McClure and Steffensen (1985) examined how linguistic complexity, developmental, and ethnic differences conditioned the use of conjunctions in children (3^{rd} to 9^{th} grade), finding a correlation between correct use of conjunctions and reading comprehension.

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;1984) in a taxonomy that is still broadly used today. 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 at-

This work was supported through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and the PhD Fellowship by FCT with reference SFRH/BD/51156/2010.

Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 70–75,
Lisboa, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.

titude towards the discourse, and encompasses `saliency` (importance), `emphatics` (certainty degree), `hedges` (uncertainty degree), and `evaluative` (speaker attitude towards a fact).

Interestingly, it is in this early approach that we find the only attempt (to our knowledge) at understanding how metadiscourse occurs across grade levels. Crismore's decisions while building the taxonomy are supported with examples extracted from nine social studies textbooks (elementary through college). After an annotation process, Crismore discusses the statistics and occurrence patterns of the various categories of metadiscourse across grade levels and audience. `Goals` were used very rarely in all text books. `Pre-plans` increased as students got into middle school and junior high and then declined. `Post-plans` were used when `Pre-plans` were used, about half as often. There was no clear trend toward increased use of `Post-plans` in upper grade texts. `Topicalizers` were used only at college level. Finally, for `Attitudinal` metadiscourse the author shows that it occurred more in texts which also contained more `Informational` metadiscourse, and that there was a tendency for it to increase in higher grade levels.

Intaraprawat and Steffensen (1995) also touched on the topic of metadiscourse and its relations to level, analyzing how 12 English as second language students used organizational language in their essays. When dividing them in *good* and *poor*, the authors observed that good essays contained proportionally more metadiscourse.

Regarding annotation of metadiscourse, and discourse in general, two distinct data-driven projects are broadly referred to and used. 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). Given its goal to reach out to the NLP community and serve as training data, the resulting senses taxonomy is composed of low-level and fine-grained concepts.

In another approach, Marcu (2000) developed the RST Discourse Treebank, 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 RST Discourse Treebank 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 RST Discourse Treebank is the discourse organization framework, which in the case of the RST Discourse Treebank is the Rhetorical Structure Theory (Mann and Thompson, 1988).

All these approaches however, focus exclusively on written language. This was the motivation behind building our own corpora of metadiscourse in spoken language (see Section 3).

3 Metadiscourse Annotation

For this experiment we look at how metadiscourse is used in spoken English. We chose TED[1], a source of self-contained presentations widely known for its speakers' quality, and for targeting a general audience. A random sample of 180 talks was used, spanning several years and topics.

Our examination of theoretical underpinnings dealing with spoken language revealed that most approaches focus on the number of stakeholders involved, and never discuss function (Luukka, 1992; Mauranen, 2001; Auria, 2006). However, Ädel (2010) merges previous approaches in a taxonomy built upon MICUSP and MICASE (Römer and Swales, 2009; Simpson et al., 2002), corpora of academic papers and lectures, respectively.

Consequently, Ädel's taxonomy was adapted according to the categories that appeared in the TED talks. More precisely, we consider 16 acts:

- `COM` – *Commenting on Linguistic Form/Meaning*
- `CLAR` – *Clarifying*
- `DEF` – *Definitions* (originally *Manage Terminology*)
- `INTRO` – *Introducing Topic*
- `DELIM` – *Delimiting Topic*
- `CONC` – *Concluding*
- `ENUM` – *Enumerating*
- `POST` – *Postponing Topic* (originally *Previewing*)
- `ARG` – *Arguing*
- `ANT` – *Anticipating Response*
- `EMPH` – *Emphasizing* (originally *Managing Message*)
- `R&R` – collapse of *Repairing* with *Reformulating*
- `ADD` – collapse of *Adding to Topic* with *Asides*
- `EXMPL` – collapse of *Exemplifying* with *Imagining Scenarios*
- `RECAP` – *Recapitulating* (subdivision of the original *Reviewing*)
- `REFER` – *Refer to Previous Idea* (subdivision of the original *Reviewing*)

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

Category	occur	conf	α
ADD	93	3.88	0.15
ANT	312	3.61	0.24
ARG	283	3.51	0.32
CLAR	265	3.82	0.15
COM	203	3.10	0.33
CONC	45	4.36	0.44
DEF	169	4.04	0.29
DELIM	26	4.21	0.31
EMPH	330	3.31	0.18
ENUM	343	3.74	0.49
EXMPL	179	3.62	0.38
INTRO	220	3.40	0.40
POST	20	4.17	0.32
RECAP	29	3.33	0.18
REFER	76	3.93	0.32
R&R	224	3.57	0.16

Table 1: Annotation results in terms of occurrence (**occur**), confidence (**conf**) and agreement (α).

Crowdsourcing was used to annotate metadiscourse (Amazon Mechanical Turk[2]). There was one task per metadiscursive category. This decreased the workers' cognitive load per task. Each of the 180 talks was divided into segments of 500 words (truncated to the closest end of sentence). This configuration generated 742 Human Intelligent Tasks (HITs) for each category (not counting gold standard HITs). To annotate a given category, workers had to first pass a training session. Upon successful completion, they were asked to read each segment and select the words that signal the existence of the metadiscursive function in question. For agreement calculation and quality control purposes, each segment was annotated by 3 different workers.

Table 1 presents the annotation results, in terms of number of occurrences found (by majority vote), the average self-reported confidence on a 5-point Likert scale (1 equals to *not confident at all* and 5 equals to *completely confident*)[3], and inter-annotator agreement. Herein, two workers are in agreement when the intersection of the words they select is not empty. We used Krippendorff's alpha since it adjusts itself better to small sample sizes than Cohen's Kappa, for example (Krippendorff, 2007). A value of zero indicates complete disagreement, and $\alpha = 1$ shows perfect agreement.

Results show that non-experts have trouble identifying some metadiscourse acts. 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's (2012) work on metalanguage, also show low agreement for sparser acts (0.09; 0.39), even when annotated by experts and considering only four categories.

Confidence results show all categories scoring above the middle of the scale (3). Workers showed less confidence for *Commenting on Linguistic Form/Meaning* (COM), which corresponds to the speaker commenting on their choice of words (confidence score of 3.1). On the other hand, workers showed the highest confidence for *Concluding Topic* (CONC), *Delimiting Topic* (DELIM) and *Postponing Topic* (POST), interestingly three categories that mark the change of topic in a talk.

4 Lexical Complexity

Evaluating linguistic complexity involves many aspects of language, such as lexis, syntax, semantics (Pilán et al., 2014; Dascalu, 2014). This paper, however, is concerned with the lexical complexity component only. Comparing the occurrences of metadiscourse across different vocabulary levels allows one to analyze its use independently of the syntactic structures that the speaker uses.

Although there is no commonly accepted measure of lexical complexity (Thériault, 2015), strategies typically rely on word unigrams to assure that only lexical clues are captured, since already capture grammatical properties (Vermeer, 2000; Heilman et al., 2007; Yasseri et al., 2011; Vajjala and Meurers, 2012). A drawback of such solutions is their inability of representing multiword expressions, like fixed phrases or idioms.

This study uses the predictor described in Collins-Thompson and Callan (2004), which is available online[4]. This approach is a specialized Naive Bayes classifier with lexical unigram features only (for the previously mentioned reasons), which creates a model of the lexicon for each grade level – between 1^{st} and 12^{th}.

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

[3] Confidence measured on a sample of 100 segments

[4] http://reap.cs.cmu.edu/demo/readability2012/

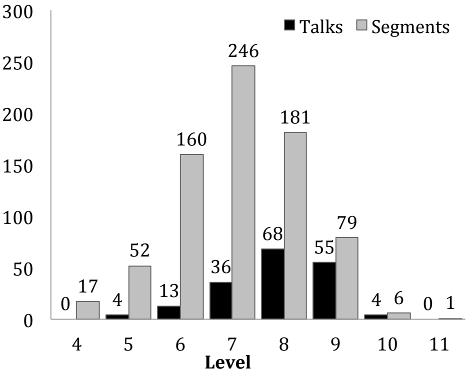

Figure 1: Level distribution of the TED talks.

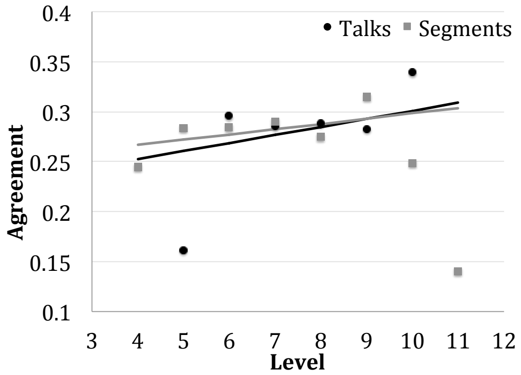

Figure 2: Agreement distribution and correlation.

The training data is composed of 550 English documents evenly distributed across the 12 American grade levels, containing a total of 448,715 tokens and 17,928 types. The documents were drawn from a wide variety of subject areas such as fiction, non-fiction, history, science, etc.

All documents, comprised of both readings and student work, were collected online. Their level classification was directly extracted from the information contained in the web page that hosted them (for instance, a document extracted from a specific classroom page).

The system developed by Collins-Thompson and Callan (2004) first performs morphological stemming and stopword removal. Then, for a given passage P, the classifier computes the likelihood that the words of P were generated from the representative language models of each level. The level where the likelihood is higher is the level that is attributed to P. The classifier performed at a correlation of 0.79 between the real and predicted levels (in a 10-fold cross validation setting).

It is important to note that this level prediction is used herein to distinguish between easier and more complex talks, more than to assign a specific grade level. In other words, one focuses at finding out which metadiscursive functions are used more often in talks with less demanding vocabulary with comparison to more complex ones (or vice-versa), never discussing occurrence at a specific level.

For the remainder of this study, the analysis takes place on two levels: *whole talk* and *segment*. The level predictor will be used on the 180 talks as a whole and on the 742 segments that compose them. The second strategy is a finer-grained local decision, since not all parts in a talk identified as high level are necessarily also complex.

Figure 1 shows level distribution: in black, the predictions when submitting the full talks to the classifier, and in light-gray, the segments prediction. In both cases we observe a normal distribution. It is interesting to notice the difference in the mode of the two cases. Most talks were assigned to a level corresponding to 8^{th} grade when submitted as a whole. However, when partitioned in segments, the most frequent level is the 7^{th}.

To exclude the hypothesis that annotators' performance was impacted by the complexity of the vocabulary, we examined how the vocabulary level of the talks relates with agreement.

Figure 2 shows how inter-annotator agreement is distributed. The correlation of the two variables is $\rho = 0.39$ for the talks and $\rho = 0.30$ for the segments, showing that vocabulary complexity does not negatively affect the capacity of two workers to agree on the annotation. In fact, the opposite trend was observed: workers agree more on segments with higher level vocabulary. This may be due to a higher degree of attention when facing more challenging content.

These results confirm that metadiscourse is independent of the content itself, and its structures can be detected independently of the propositional content in which they are inserted and for which they are used.

5 Results

With a set of 180 talks and 742 segments, annotated with 16 categories of metadisocurse, and automatically assigned to a level according to the lexical predictor described previously, one can now map the occurrences of the different acts across levels and conclude on how its use varies with lexical level of the content.

Category	Occur avg. (%)	Correlation	
		by talk	by segment
ADD	0.60	**0.95**	**0.50**
ANT	1.20	(0.48)	**(0.85)**
ARG	1.13	**0.63**	**0.68**
CLAR	1.47	**0.58**	(0.16)
COM	1.54	**0.78**	**0.70**
CONC	0.37	(0.07)	**(0.73)**
DEF	1.13	**0.63**	**0.85**
DELIM	0.18	**0.54**	0.12
EMPH	1.90	0.47	(0.27)
ENUM	3.15	0.09	0.23
EXMPL	1.47	0.43	**0.50**
INTRO	1.61	0.37	0.22
POST	0.21	(0.21)	(0.01)
RECAP	0.16	0.15	**(0.50)**
REFER	0.54	**0.94**	(0.33)
R&R	1.23	**0.85**	**0.68**

Table 2: Average occurrence and level correlation.

Table 2 shows the probability of a sentence containing a given metadiscursive act (*Occur avg. (%)*) and how each category correlates with level at both the talk and segment levels. Correlations are weighted for the amount of sentences in each level to decrease the impact of outliers in levels with few cases. Negative correlations are shown between brackets, significant correlations in bold, and high correlations are bold and underlined.

Adding Information (ADD) correlated at both talk and segment level, registering the highest correlation of all at talk level (0.95). Higher frequencies of ADD seem to be associated to talks with higher level vocabulary. This same pattern was also observed for R&R, which tends to occur in talks/segments assigned to higher grade levels.

Commenting on Linguistic Form/Meaning (COM), and *Definitions* (DEF) also showed significant correlation at both levels. However, these categories have strong correlations at segment level, i.e., they do not only occur more frequently in higher level talks, but also in segments that contain words typically found in higher levels.

Anticipating Response (ANT) registered the strongest negative correlation both at sentence and talk levels (-0.85;-0.48). As talks are assigned to higher lexical levels, less instances addressing the audience's previous knowledge are found. As one would expect, the more complex the vocabulary and topic of a talk is, the less assumptions are made about what the audience knows.

Arguing (ARG) shows moderate correlation at both levels. The more complex the vocabulary of a talk/segment is, the more the speaker feels the need to defend a point or prove his position.

Clarifications (CLAR) correlate moderately with the level of the talk but show a negative correlation trend with the segment. This shows that while talks with more demanding vocabulary have more clarifications, they are not necessarily located in lexically complex segments. This pattern is also observed for *Conclusions* (CONC), *Recapitulations* (RECAP), and *References to Previous Ideas* (REFER), all with negative segment correlations. Interestingly, the four categories are related to paraphrasing (whether summarization or simplification). The high correlation for CONC in particular (0.73) shows that a segment that contains a conclusion tends to have simpler vocabulary.

Results for *Delimiting Topic* (DELIM) and *Exemplify* (EXMPL) are at the frontier of low and moderate agreement. The remaining categories (*Emphasizing*, *Enumerating*, *Introducing* and *Postponing Topic*) did not correlate with level ($\rho < 0.5$) and seem to occur independently of the level of the vocabulary of the talk or segment.

6 Conclusions

This study used an empirical approach to understand how metadiscourse is used across different levels in spoken language. It employs a set of TED talks and a functional theory of metadiscourse. Crowdsourcing was used to annotate 16 metadiscourse functions. Comparing annotations with a vocabulary classifier showed that some but not all categories correlate with vocabulary level.

Strategies of topic management (delimiting, introducing, postponing) and broadly used functions (examples, emphasis, enumerations) occur at the same rate in all levels, not correlating with level.

Results also show that functions related to paraphrasing are more frequent in higher level talks, but not necessarily in segments containing the highest level vocabulary. In fact, the occurrence of a strategy that aims at language simplification contributes itself for lower level classification. This shift in correlation polarity from talk to segment level suggests that these strategies do not occur in close context with the ideas they are simplifying.

Contrastingly, functions that manage vocabulary (commentaries and definitions) seem to appear in the context of the vocabulary they address.

Future work includes using the annotation to build metadiscourse classifiers. As observed, the vocabulary level of the talk/segment can be a valuable feature for classification.

References

Annelie Ädel. 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.

Carmen PL Auria. 2006. Signaling speaker's intentions: towards a phraseology of textual metadiscourse in academic lecturing. *English as a GloCalization Phenomenon. Observations from a Linguistic Microcosm*, 3:59.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200.

Avon Crismore, Raija Markkanen, and Margaret S Steffensen. 1993. Metadiscourse in persuasive writing a study of texts written by american and finnish university students. *Written communication*, 10(1):39–71.

Avon Crismore. 1980. Student use of selected formal logical connectors across school level and class type.

Mihai Dascalu. 2014. Analyzing discourse and text complexity for learning and collaborating. *Studies in computational intelligence*, 534.

Michael J Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, pages 460–467.

Puangpen Intaraprawat and Margaret S Steffensen. 1995. The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing*, 4(3):253–272.

Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental Papers (ASC)*, page 43.

Minna-Riitta Luukka. 1992. Metadiscourse in academic texts. In *Text and Talk in Professional Contexts. International Conference on Discourse and the Professions, Uppsala, 26-29 August*, pages 77–88.

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

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT press.

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

Anna Mauranen. 2001. *Reflexive academic talk: Observations from MICASE.*

Erica F McClure and Margaret S Steffensen. 1985. A study of the use of conjunctions across grades and ethnic groups. *Research in the Teaching of English*, pages 217–236.

Bonnie JF Meyer, David M Brandt, and George J Bluth. 1980. Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading research quarterly*, pages 72–103.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the penn discourse treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286. Springer.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 174–184.

Ute Römer and John M. Swales. 2009. The michigan corpus of upper-level student papers (MICUSP). *Journal of English for Academic Purposes*, April.

Rita C Simpson, Sarah L Briggs, Janine Ovens, and John M Swales. 2002. The michigan corpus of academic spoken english. *Ann Arbor, MI: The Regents of the University of Michigan.*

Mélissa Thériault. 2015. The development of lexical complexity in sixth-grade intensive english students.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. Association for Computational Linguistics.

Anne Vermeer. 2000. Coming to grips with lexical richness in spontaneous speech data. *Language testing*, 17(1):65–83.

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

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

T Yasseri, A Kornai, and J Kertész. 2011. A practical approach to language complexity: a Wikipedia case study. *PloS one*, 7(11):e48386–e48386.