Automatic Classification of Metadiscourse

Rui Pedro dos Santos Correia

CMU-LTI-18-011

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Prof. Maxine Eskenazi
Prof. Nuno João Neves Mamede
Prof. Diane J. Litman
Prof. Jorge Manuel Baptista
Prof. Jaime Carbonell
Prof. Robert E. Frederking
Prof. Isabel Maria Martins Trancoso

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies*

Resumo

Esta tese aborda questões relacionadas com a função metadiscursiva em discurso oral. Sendo uma das funções básicas da linguagem, o metadiscurso, normalmente referido como discurso sobre o discurso, é composto por atos retóricos e padrões que tornam a estrutura do discurso explícita, guiando o público durante o ato discursivo. O objectivo desta tese é o de detectar e classificar automaticamente o uso de metadiscurso em contextos de apresentação oral.

São discutidas teorias existentes sobre a vertente oral de metadiscurso, com especial foco numa taxonomia que define os conceitos metadiscursivos de uma forma totalmente funcional, ou seja, que atribui uma função discursiva a ocorrências de metadiscurso em vez de analisar exclusivamente a sua forma. Esta taxonomia é usada para anotar um conjunto de TED talks com funções metadiscursivas, usando *crowdsourcing*. Os resultados mostram que nem todas as categorias incluídas na anotação conseguem ser compreendidas da mesma forma por não-peritos.

Estas anotações são usadas para treinar classificadores de metadiscurso (um por categoria) que detectam e classificam funcionalmente ocorrências de metadiscurso em transcrições de apresentações. Esta tarefa de classificação é dividida estrategicamente em duas etapas. Primeiro, o treino de Máquinas de Vetores de Suporte, que geram uma lista de frases candidatas a conter metadiscurso. Em segundo lugar, aplicam-se Campos Aleatórios Condicionais aos candidatos para detectar os termos exatos usados pelo orador para o ato em questão. Em ambas as etapas são testados diferentes conjuntos de características (lexicais, sintáticas e semânticas) e analisadas quanto à sua adequação para a tarefa em questão.

A análise de desempenho desta classificação é discutida à luz de duas possíveis aplicações: como auxílio para tarefas de Processamento de Língua Natural (tais como sumarização e detecção de tópico), ou como parte de um currículo de técnicas de apresentação.

Abstract

This thesis addresses issues related to the function of metadiscourse in spoken language. Being one of the basic functions of language, metadiscourse, commonly referred to as discourse about discourse, is composed of rhetorical acts and patterns used to make the discourse structure explicit, acting as a way to guide the audience. The objective of the current thesis is to be able to automatically detect and classify the use of metadiscourse in presentational settings.

Existing theory on spoken metadiscourse is discussed, with special focus on a taxonomy that defines metadiscursive concepts in a fully functional manner, i.e. that assigns a discourse function to occurrences of metadiscourse rather than analyzing exclusively its form. This taxonomy is used to annotate a set of TED talks with metadiscursive functions, using crowdsourcing. Results show that not all categories included in the annotation can be annotated and understood in the same manner by non-experts.

The collected annotations are used to train metadiscourse classifiers (one per category) that detect and assign a function to occurrences of metadiscourse in presentation transcripts. This classification task was strategically divided in two steps. First, training of Support Vector Machines to generate a list of candidate sentences that can contain metadiscourse. And secondly, applying Conditional Random Fields to those candidates to detect the exact terms used by the speaker for the corresponding act. Different sets of features (lexical, syntactic and semantic) are used in both layers of the classifiers and discussed in light of their suitability to be used in the task at hand.

The performance analysis of this classification chain is discussed with respect to two possible applications: as an aid to common Natural Language Processing tasks (such as summarization and topic detection), and as part of a presentational skills curriculum.

Palavras chave
Keywords

Palavras Chave

Análise de Discurso

Metadiscurso

TED talks

Crowdsourcing

Aprendizagem Automática

Keywords

Discourse Analysis

Metadiscourse

TED talks

Crowdsourcing

Machine Learning

Acknowledgements

First and foremost, I would like to thank my advisors, Professor Nuno Mamede and Professor Maxine Eskenazi, for all their guidance, discussion, motivation, and for all the patience during the entire process.

I also want to thank all the members of L2F, for both providing me with essential tools for the development of this work and helping me with discussion. Special thanks go out to Professor Isabel Trancoso, for her guidance and always wise and practical advice, to Professor Jorge Baptista for his constant help, discussion and experience, to Wang Ling my colleague and room mate abroad, to Vera Cabarrão for all her availability and expertise, and to Tiago Luís for all his help and endless availability.

Many thanks to my family for their trust and help.

And finally, to my second family, my dear friends. To Luís Soares, who was always there, to Henrique Fernandes, for his unconditional friendship and portraits, to Bruno and Hugo, who always cared, to Mark Bechara, who was my rock, to Ângela Costa, for putting everything in perspective, and to Leandro Machado, who arrived late but became essential.

This work was made possible by Fundação para a Ciência e a Tecnologia, under grants with references UID/CEC/50021/2013 and SFRH/BD/51156/2010.

Lisboa, September 8, 2018
Rui Correia

Contents

List of Figures

List of Tables

1 Introduction

Metadiscourse (also *metalanguage*, *signposting language*, or *text-referral*) is one of the primary 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."*

The quote above summarizes the three defining properties of metadiscourse:

1. it occurs in both written and spoken discourse;
2. it does not contribute to the content itself;
3. it is used by the speaker/writer to guide the audience through the communication event.

In other words, metadiscourse allows the speaker or writer to explicitly refer to events that exist in the realm of the discourse. Typical uses of metadiscourse in written or spoken language include the explicit highlight of important ideas (*"The take-home message is..."*), the announcement of the topic of the discourse (*"In this paper, we talk about..."*), or the illustration of ideas through examples (*"Consider, for instance..."*).

It is important here to clarify the distinction between metadiscourse and discourse function. Discourse functions (such as the ones referred to in the previous paragraph) can still exist in the discourse without being materialized through metadiscourse.

For example, a Ph.D. candidate writing their thesis may use bold-faced font to emphasize important concepts, instead of using a metadiscursive strategy such as *"It is important to note that. . . "*; they may surround examples with parenthesis instead of explicitly mention that they are exemplifying (*"For instance,. . . "*); or use bullet points to enumerate, not recurring to a strategy such as *"First, . . . ; secondly, . . . ; and finally, . . . "*

Similarly, in an oral presentation, these discourse functions can be performed with the use of intensity variations (to emphasize), specific pause patterns (to exemplify) and body/hand gestures (to enumerate). This distinction is critical, and this thesis is concerned only with the metadiscursive realization of such discourse functions.

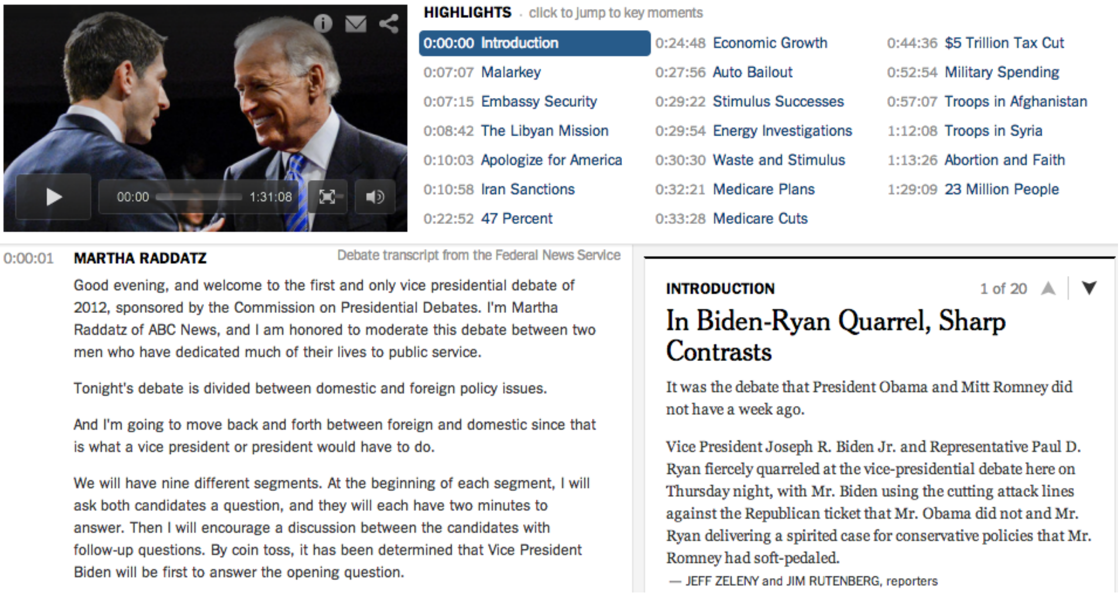
The goal of this work is two-fold:

- on the one hand, contribute to the understanding of metadiscourse as a phenomenon occurring in language;
- on the other hand, through a systematic approach, provide automatic mechanisms that detect and classify the phenomenon according to its function in the discourse where it occurs. At this level, the objective is to conclude on different Natural Language Processing (NLP) and Machine Learning strategies and features, and how they perform for the task at hand.

These two goals are not mutually exclusive and will interact throughout the document. The metadiscursive theoretical background and the understanding of its workings will support some of the decisions made regarding the automatic processing and, similarly, the systematic approach of classification with machine learning techniques will highlight some properties and characteristics of the phenomenon.

1.1 Motivation

The task of automatically classifying metadiscourse is situated in the field of discourse analysis. It parallels to some degree with the tasks of discourse structuring and segmentation, which deal with the separation of discourse into cohesive segments, going beyond the concept of sentences as units.



HIGHLIGHTS · click to jump to key moments

0:00:00	Introduction	0:24:48	Economic Growth	0:44:36	$5 Trillion Tax Cut
0:07:07	Malarkey	0:27:56	Auto Bailout	0:52:54	Military Spending
0:07:15	Embassy Security	0:29:22	Stimulus Successes	0:57:07	Troops in Afghanistan
0:08:42	The Libyan Mission	0:29:54	Energy Investigations	1:12:08	Troops in Syria
0:10:03	Apologize for America	0:30:30	Waste and Stimulus	1:13:26	Abortion and Faith
0:10:58	Iran Sanctions	0:32:21	Medicare Plans	1:29:09	23 Million People
0:22:52	47 Percent	0:33:28	Medicare Cuts		

0:00:01 **MARTHA RADDATZ** Debate transcript from the Federal News Service

Good evening, and welcome to the first and only vice presidential debate of 2012, sponsored by the Commission on Presidential Debates. I'm Martha Raddatz of ABC News, and I am honored to moderate this debate between two men who have dedicated much of their lives to public service.

Tonight's debate is divided between domestic and foreign policy issues.

And I'm going to move back and forth between foreign and domestic since that is what a vice president or president would have to do.

We will have nine different segments. At the beginning of each segment, I will ask both candidates a question, and they will each have two minutes to answer. Then I will encourage a discussion between the candidates with follow-up questions. By coin toss, it has been determined that Vice President Biden will be first to answer the opening question.

INTRODUCTION 1 of 20

In Biden-Ryan Quarrel, Sharp Contrasts

It was the debate that President Obama and Mitt Romney did not have a week ago.

Vice President Joseph R. Biden Jr. and Representative Paul D. Ryan fiercely quarreled at the vice-presidential debate here on Thursday night, with Mr. Biden using the cutting attack lines against the Republican ticket that Mr. Obama did not and Mr. Ryan delivering a spirited case for conservative policies that Mr. Romney had soft-pedaled.

— JEFF ZELENY and JIM RUTENBERG, reporters

Figure 1.1: New York Times interface of a debate.

Figure 1.1 shows an example of discourse segmentation, made by experts, which aims at enriching the online experience of watching a debate[1]. The interface displays segmentation information (top-right), transcript (bottom-left) and additional content (bottom-right).

By looking at the transcript, it is possible to find occurrences of metadiscourse that, when assigned to a function, can serve the purpose of dividing the discourse into meaningful segments. Additionally, as already mentioned, they add another layer of information since they signal the speaker's explicit intention on specific discourse functions, which can therefore aid the listener/reader to better understand the content at hand.

[1] Vice-presidential debate between Biden and Ryan in 2012
http://www.nytimes.com/interactive/2012/10/11/us/politics/20121011-vice-presidential-debate-biden-ryan.html
(visited in November 2017)

Below are three passages containing instances of metadiscourse extracted from the transcript:

- **Emphasis –** *"Here is the problem. Look at all the various issues out there and that's unraveling before our eyes. The vice president talks about sanctions on Iran."*;
- **Changing Topic –** *"Let's move to* Iran. *I'd actually like to move to* Iran *because there is really no bigger national security. . . "*;
- **Closing Discourse –** *"We now turn to the candidates for their closing statements. "*

Such content enriching capabilities constitute an example of a possible outcome of the current thesis. In more detail, as already mentioned, this work is comprised of two main goals. First, to introduce the concept of metadiscourse as an important and valuable part of discourse analysis in spoken language. And secondly, to test how its detection can be made in an automatic manner in order to, for instance, automate the generation of content similar to the one in Figure 1.1.

This particular combination of motivation and goals imposes two constraints on the scope of the thesis. First and foremost, this study will target the use of metadiscourse in the spoken settings only. This constraint aims at filling the gap between the extensive research specifically targeted at the written form, and seldom focus on the particularities and idiosyncrasies of spoken metadiscourse. In comparison to the written form, this setting introduces discourse elements that have an impact on how and for what purposes metadiscourse is used, such as the lack of time for planning/revision, or direct interaction with the audience.

The second constraint is that herein metadiscourse is to be approached functionally (as opposed to formally). In order to produce content that can be understood by the general audience (such as the one in Figure 1.1), it is important to package metadiscourse in a functional manner. In other words, the primary objective is to identify the rhetorical functions associated with each metadiscursive construction (whether an introduction, a conclusion, an example, *etc*), rather than to analyze metadiscourse according to its form or intrinsic properties (pronominal or non-pronominal, formal or informal).

Aside from content enriching capabilities, such an approach to metadiscourse can be the building block to other types of applications. One particular usage that is consistently referred to in the literature as important and valuable, is the possibility of using the phenomenon as a key concept in presentational skills curricula (Lyons, 1977; Auría, 2006; Ädel, 2010). Again, for this application to be feasible, the concepts at hand have to be grasped by non-experts.

The first research question of the current thesis follows from this constraint:

Can the general public understand the concept of metadiscourse?

On this front, this thesis will look into the existing theory of the phenomenon, justify the choice of a particular taxonomy over the remaining work, and investigate how non-experts react to the metadiscursive concepts they are exposed to.

The outcome of such exploration is a set of metadiscursive acts for which non-experts show consensus in what concerns the understanding of their functions in the discourse. Naturally, such discussion also presents concepts that are less consensual, as well as qualitative metrics that justify the decisions to follow-up the investigation of each of the metadiscursive tags.

The second research question, aligned with the second goal of the present study, has as precedence the conclusions drawn for the first question. It can be stated as:

To which extent can the identification and functional classification of metadiscourse be automated?

This question aims at exploring and comparing different techniques that can successfully classify metadiscursive phenomena. More than a simple analysis of performance, it is important to get insight on the nature of metadiscourse itself, *i.e.*, understand which features are representative of the phenomenon and how different approaches are capable of detecting and classifying its instances.

More objectively it encompasses identifying occurrences of metadiscourse as used in spoken language, and assigning a speaker intention to each occurrence of the phenomenon.

1.2 Structure of this Document

The remainder of this document is structured as follows:

- **Chapter 2** describes the existent theories of metadiscourse in spoken language, existent corpora addressing issues related to the structuring of discourse, and previous approaches to annotation and classification of metadiscourse-related phenomena;

- **Chapter 3** describes a crowdsourced annotation task aimed at building a corpus of metadiscursive acts – METATED. Considerations are made about the choice of material to annotate and the categories to use in the annotation task, along with the definition of instructions and training sessions targeted to non-expert annotators. The quantity and quality of the material collected are presented in detail. Finally, it describes an expert annotation task, where a set of experts validated the crowd's work, providing further insight on METATED;

- **Chapter 4** presents the classification task itself. METATED is used as training data to build metadiscursive classifiers. Here, the problem of classifying metadiscourse is divided into two main tasks: first, generating a list of candidate sentences that are thought to contain metadiscursive phenomena; and secondly, a finer grained classification, which takes those candidate sentences and highlights which of the words are performing the metadiscursive role. All decisions made throughout the process are discussed in the light of the results they achieve, including the choice of features and algorithms used for classification;

- **Chapter 5** contains a summary and discussion of the work accomplished in this thesis, highlighting its key contributions and proposing future work.

2 Background

To the current knowledge, human language is the only one with the property of being able to refer to itself (Lucy, 1993). This property is associated with the notion of *reflexivity*, introduced by Hockett (1963) as one of the *Design Features of Language* – a list of sixteen features that distinguished human communication from that of animals. It included traits such as *prevarication* (the ability to lie) or *displacement* (the ability to talk about what is not physically present). In another early study on metadiscourse, Silverstein (1976) distinguished between the notions of *metapragmatics* and *metasemantics*. Silverstein noticed that the language reflexive capabilities are primarily *metapragmatic* (used by the speaker to explicitly state the intentions and effects of his/her speech). In the field of *metasemantics* (the capability of language to comment on its own meaning or form), Lyons (1977) coined the terms *use* and *mention*, referring to the non-reflexive and reflexive use of language, respectively.

The topic of metadiscourse gained the attention of the research community during the 80s, mostly focusing on the presence of metadiscursive acts in written academic discourse (Kopple, 1985; Crismore, 1989). With respect to this written approach, to this date the most consensual theory is the one developed by Hyland and Tse (2004); Hyland (2004, 2005). Hyland's taxonomy is organized under two main categories (*Textual* and *Interpersonal Metadiscourse*) which then unfold into a total of 13 concepts, including *Logical Connectives* (such as *"in addition"*, *"but"*, *"thus"*), *Frame Markers* (*"to repeat"*, *"here we try to"*), and *Attitude Markers* (*"unfortunately"*, *"I agree"*).

It was only later, during the 90s and the 00s, that the spoken variety of metadiscourse started being explored and addressed in a systematic and data-driven manner.

This chapter's prime focus is precisely on work targeting spoken metadiscourse. It is organized into two main sections:

- **Section 2.1** presents the existent theories of metadiscourse in spoken discourse, describing and comparing the relevance of five different taxonomies, and discussing how each aligns to the goals of the current thesis;
- **Section 2.2** focus on previous NLP approaches to metadiscourse, presenting work on corpora building and on the classification and parsing of metadiscourse.

2.1 Metadiscursive Theory

Shannon and Weaver (1948) defined the most widely used communication theory. Shannon-Weaver's model defines seven elements of communication. The information that is being communicated from one end of the model to the other is the *message*. The message circulates in the model between the *information source* (who produces the message) and the *destination* (for whom the message is intended). The information source encodes the message via the *transmitter*, which generates a signal suitable to be transmitted over the *channel* (the medium used to send the signal). At the other end of the channel is the *receiver*, who performs the inverse operation of the transmitter: decoding the signal to be understood by the destination. The seventh element in the model is *noise*, *i.e.*, anything that can misconstrue the message (whether physical or semantic).

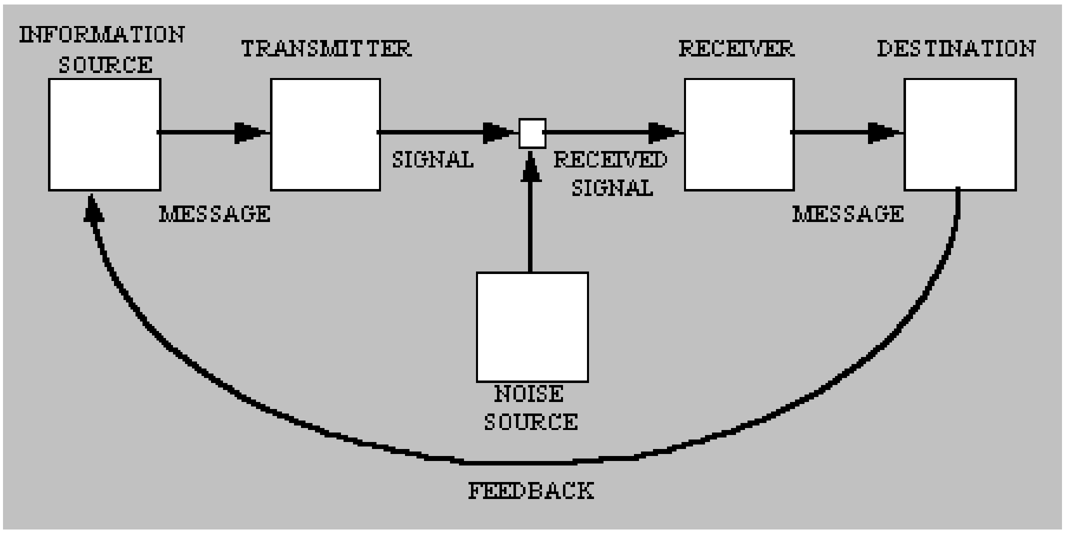


Figure 2.1: Wilbur Schramm's extension of Shannon-Weaver's model of communication.

Schramm (1954) expanded Shannon-Weaver's model incorporating human behavior in the communication process. Figure 2.1 shows a representation of the Schramm's model, displaying a circular communication model between the source and destination, made possible with the inclusion of the element *feedback*: information that comes from the destination to the source.

Varying the properties of these elements allows one to define different communication settings. For instance, altering the *channel* allows distinguishing spoken language from its written form. Changing the media in which the message circulates affects other elements in the model. Both *noise* and *feedback* are characteristics of spoken communications (and presentations in particular). In spoken language, the immediacy of production affects the *noise* element, since planning and corrections occur in real-time. The fact that the audience can contribute to the message in real-time (by asking questions, applauding or laughing) on the other hand, affects the *feedback*. Chafe and Danielewicz (1987) highlight this two-fold distinction, noting that situational settings affect processing considerations (restrictions of real-time production vs. opportunity for editing) and the degree of involvement between the speaker/writer and the audience.

These different settings give origin to actual differences in style and expression between speech and writing. According to Biber (1986), who summarizes the literature on the differences of spoken vs. written communication, writing can be seen as more detached and contextualized (*e.g.* has more nominalizations and passives), more elaborated and expanded (with more occurrences of relative clauses and infinitives), and as having a more explicit level of expression (differences in word length and type/token ratio). On the other hand, speech is typically more informal (with more contractions, deletions of relative pronouns, use of informal emphatics), more interactive and involved (more occurrences of first and second pronouns), and also more situated in a physical/temporal context.

These situational differences also affect the way metadiscourse is used in spoken language (as opposed to writing). For example, speakers may use metadiscourse to manage comprehension (*"Can you hear me back there?"*) or to correct a point (*"Sorry, what I meant was. . . "*).

For those reasons, this chapter focuses for the most part on research that looks at metadiscourse as used in spoken language. More specifically, this section presents the theories of metadiscourse that consider only spoken discourse (sections 2.1.3 and 2.1.4) or that adopt a unified approach, discussing both written and spoken varieties (2.1.1, 2.1.2 and 2.1.5).

2.1.1 Luukka (1992)

Focusing on academic discourse, Luukka developed a taxonomy of metadiscourse that dealt with both written and spoken varieties. In this work, Luukka used a small corpus of five papers delivered at a conference and considered two versions of each document: the written text submitted for the proceedings, and the transcript of the oral presentation.

By analyzing both strategies of presentation of the same content, Luukka created a taxonomy of metadiscourse that unifies both varieties. The guiding principle of Luukka's categorization is the distinction between strategies used for discourse organization and for interaction with the audience. The proposed taxonomy is comprised of three categories:

- ***Textual –*** strategies related to the structuring of discourse;
- ***Interpersonal –*** related to the interaction with the different stakeholders involved in the communication;
- ***Contextual –*** covering references to audiovisual materials.

2.1.2 Mauranen (2001)

Mauranen is one of the most active authors in the area of metadiscourse, with a large body of work on the topic (Lindemann and Mauranen, 2001; Mauranen, 2002, 2003, 2010), including its relation to presentations (Mauranen, 2013a), and oral proficiency in L2 (Mauranen, 2013b).

In Mauranen (2001), the author develops a taxonomy for both written and spoken language adopting a splitting approach, with different taxonomies for each variety. Mauranen uses the Michigan Corpus of Academic Spoken English (MICASE), developed at the University of Michigan's English Language Institute (Simpson and Swales, 2001), composed of 200 hours of lecture courses, seminars, and student presentations, with speakers ranging from senior lecturers to undergraduate students. For those reasons, MICASE contains both monologic (one-way communication from sender to receiver) and dialogic types of spoken discourse (includes feedback from the receiver), contrasting with Luukka's five conference paper corpus.

The author's taxonomy is composed of three categories, with no further subdivision:

- ***Monologic –*** structuring of the speaker's own discourse (similar to *textual* in Luukka's taxonomy);
- ***Dialogic –*** referring to audiences interventions or answering questions (identical to *interpersonal* in Luukka's taxonomy);
- ***Interactive –*** eliciting participation from the audience and manipulating the roles of the stakeholders (also related to *interpersonal* in Luukka's taxonomy).

The identification of the stakeholder who took the discourse initiative is the guiding principle under the division proposed by Mauranen. It is also interesting to notice the similarities between Luukka and Mauranen's approaches. As the first taxonomies that tried to categorize the use of metadiscourse in spoken language, they are both guided by one of the principles that distinguish spoken from written communications, *i.e.*, the fact that the audience (or receiver) can contribute to the message in real-time (the already mentioned immediacy of feedback).

2.1.3 Thompson (2003)

The premise in Thompson (2003) is the comprehension of lectures in the real world, which motivates the analysis of how students and teachers use metadiscourse in classrooms. Throughout this study, the author focuses on showing the misalignment between the curricula for English for Academic Purposes (EAP) courses and the real practices of discourse organization and intonation in real-world communications.

Therefore, Thompson uses a corpus of six authentic undergraduate university lectures and five EAP published listening skills materials, comparing them and highlighting the mismatch between what is being taught as good practices of presentations and what is, in fact, used in lectures in real academic settings.

As a result of this comparison process, the author formulated a taxonomy of metadiscourse for academic lectures, which categorizes metadiscourse into three main groups:

- ***Content Markers –*** used to give information about the lecture to come;
  - *"In what is left of this hour, which is actually half an hour, I hope to give you a sort of brief idea [. . .]"*
- ***Structuring markers –*** used to outline the structure and sequence of the lecture;
  - *"I'll start with water [. . .] And then I'll move on to farms [. . .]"*
- ***Metastatements –*** used to organize the communication event itself (not its content).
  - *"Right. So, with that let me start the lecture."*

Additionally, Thompson further divides each category into three levels: *global*, *topical* and *sub-topical*. Each level indicates at what granularity the metadiscourse marker is operating. This distinction reflects the natural granularity and diversity of topics existent in every communication event, allowing the modeling of the interaction between the different sections that compose each lecture.

2.1.4 Auría (2006)

Auría also focused on the use of spoken metadiscourse in academic settings, comparing it to conversational language and with the written register. In Auría (2006), the author found that metadiscourse is more prominent in events where knowledge is being transmitted, since lecturers, seeking maximum comprehension, explicitly signal their communicative intentions. Another interesting conclusion from this work is the higher density of metadiscursive acts found in longer lectures. Longer and larger classes often imply larger audiences – not only in size but also in the variability of previous knowledge and cognitive capabilities. As a result, teachers tend to show higher concern in drawing attention towards discourse organization for more effective comprehension.

The central concept behind Auría's taxonomy is lecturer intention. While analyzing the MICASE corpus (described in Section 2.1.2), the author proposes the following division:

- ***I-pattern –*** expressions that use the first person singular nominative pronoun, such as *I'm gonna* or *I wanna*;
- ***We-pattern –*** expressions that use the first person plural nominative pronoun, such as *We'll* or *We're gonna*;
- ***Polite Directives –*** other expressions, such as *Let's* or *Let me*.

According to the author, 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.

2.1.5 Ädel (2010)

Ädel is another author with an extensive body of work on metadiscourse in both written and spoken form, including metadiscourse used in argumentative writing, (Ädel, 2003, 2005), a reference book on metadiscourse in L1 and L2 (Ädel, 2006), a position paper (Ädel and Mauranen, 2010), and a study on metadiscourse used in feedback (Ädel, 2017).

In Ädel (2010), the author packages her research in a taxonomy that unifies existing theories of metadiscourse. Ädel's framework encompasses both spoken and written discourse and is built using two academic-related corpora: Michigan Corpus of Upper-level Student Papers (MICUSP) (Römer and Swales, 2010) – comprised of academic papers – and the already mentioned MICASE (Simpson and Swales, 2001) – a corpus of university lectures.

Ädel stresses the importance of a pedagogical approach to metadiscourse, stating that *"[a]nyone using spoken and written academic English needs to be intimately familiar with the rhetorical acts and recurrent linguistic patterns involved in metadiscourse, both for comprehension and for production."*

This focus on comprehension resulted in a taxonomy that, in contrast with other theories, commits to represent metadiscourse concerning its function rather than its form.

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
- **Managing Phorics**
  - Enumerating
  - 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 2.2: Ädel's taxonomy of metadiscourse.

Figure 2.2 summarizes Ädel's taxonomy of metadiscourse. It is composed of four main categories (*Metalinguistic Comments*, *Discourse Organization*, *Speech Act Labels*, and *References to the Audience*), further divided according to their discourse function. The remaining sections are going to describe in detail Ädel's taxonomy, illustrating the different categories with examples extracted from the original paper.

Metalinguistic Comments

In this top-level category, the author distinguishes between five discourse functions: REPAIRING, REFORMULATING, COMMENTING ON LINGUISTIC FORM/MEANING, CLARIFYING and MANAGING TERMINOLOGY.

- REPAIRING refers to the need to correct prior statements that the speaker thinks s/he conveyed in an imprecise or wrong manner. As expected, examples of this function could only be found in the spoken corpus MICASE, and include *"I'm sorry"*, or *"maybe I've should have said"*;

- REFORMULATING is associated with the speaker's desire to provide an alternative term to a previously exposed idea, not because it was wrong but because it adds value to the content. Although more frequent in spoken discourse, this function was also found in written language. An example is *"let me rephrase a little"*:

- COMMENTING ON LINGUISTIC FORM/MEANING relates to comments on word choice or meaning, and can be found in both discourse varieties (*"we can therefore say that "statue" is a word that..."*). This discourse function is related to the *mention* notion in *metasemantics* (as opposed to *use*) introduced by Lyons (1977), and referred to at the beginning of this chapter;

- CLARIFYING is found in both written and spoken language, and is used to avoid misinterpretations (*e.g. "I'm not claiming that ..."* or *"I should note for the sake of clarity ..."*);

- Finally, the last function in this category, MANAGING TERMINOLOGY, is also related to the *mention* concept and occurs in both varieties of discourse. As the name of the function states, it is used to give definitions (*e.g. "which we might as well define now"* or *"we will be using the following definition..."*).

Discourse Organization

Discourse Organization is further divided into two subcategories: *Manage Topic* and *Manage Phorics*. The functions that compose the subcategory *Manage Topic* are similar to the ones described by Thompson (2003) (see Section 2.1.3). They are: INTRODUCING TOPIC, DELIMITING TOPIC, ADDING TO TOPIC, CONCLUDING TOPIC and MARKING ASIDES.

- INTRODUCING TOPIC and CONCLUDING TOPIC are used by the speaker to open or close the current topic and can naturally be found in both written and spoken discourse;
- DELIMITING TOPIC refers to strategies used to impose constraints on the topic of the talk such as in *"I have restricted my discussion to. . . "* (in written form) or *"We won't go into that."* (in spoken discourse);
- ADDING TO TOPIC covers the explicit additions to the content that can occur in both varieties of communication (*e.g.*, *"we might add that. . . "*);
- Finally, MARKING ASIDES is the only function of this subcategory that can only be found in the spoken corpus of English lectures. It is used as a digression, to add content on a slightly different topic (*e.g.*. *"I want to do a little aside here."*).

Manage Phorics, the other subcategory under *Discourse Organization*, is comprised of five functions: ENUMERATING, PREVIEWING, REVIEWING, CONTEXTUALIZING and ENDOPHORIC MARKING.

- ENUMERATING is used to make the organization of the discourse explicit (similar to *structuring markers* in Thompson (2003) – see Section 2.1.3), being found in written and spoken form (*e.g.*. *"We're gonna talk about mutations first."* or *"I have two objections against this."*);
- PREVIEWING and REVIEWING are used to point forward and backward in the discourse, as in *"As I discuss below. . . "* and *"We have seen two different arguments. . . "*;

- ENDOPHORIC MARKING is similar to the *Contextual* category in Luukka's work (see Section 2.1.1) and is used to point to tables, images, and other audiovisual materials such as in *"If you look at question number one..."*;

- Finally, CONTEXTUALIZING is used to comment on the situation of writing or speaking such as *"There's still time for another question."* or *"I have said little about..."*.

Speech Acts Labels

This category contains three functions: ARGUING, EXEMPLIFYING, and OTHER (where the author included acts that were not frequent enough to generate a new label).

- ARGUING is used in speech or writing to support an idea explicitly (like in *"I argue that..."*);

- EXEMPLIFYING, as the name states, is used to explicitly introduce an example (*"I will use the example..."*).

Audience References

In the last category of Ädel's taxonomy there are five discourse functions, all related to the interaction between speaker/writer and audience: MANAGE COMPREHENSION, MANAGE DISCIPLINE, ANTICIPATING RESPONSE, MANAGING THE MESSAGE, and IMAGINING SCENARIOS.

- MANAGE COMPREHENSION is used by the speaker to check for understanding and to test the communication conditions, such as in *"You know what I mean?"* and *"Can you guys hear?"*;

- MANAGE DISCIPLINE refers to events where the speaker instructs the audience to do something (usually intended to improve the communication channel, as in *"Can we have a little bit of quiet?"*).

- ANTICIPATING RESPONSE is similar to the function CLARIFYING (in *Metalinguistic Comments*) but here involves a reference to the audience (as in *"You guys probably end up thinking. . . "* and *"The reader might wonder why. . . "*);

- MANAGING THE MESSAGE is used to emphasize the main message, such as in *"What I want you to remember is. . . "*;

- Finally, IMAGINING SCENARIOS is a more engaging version of the function EXEMPLIFYING (in *Speech Act Labels*) where the speaker/writer invites the audience to share a given perspective (*e.g.*, *"Suppose you are a researcher."* or *"Imagine the following situation."*).

2.2 NLP Approaches to Metadiscourse

As mentioned previously, this section describes work resultant from Natural Language Processing (NLP) research that addressed either discourse functions in general, or metadiscourse in spoken language in particular. There are three types of works being discussed:

- Section 2.2.1 deals with existent corpora that target discourse and its function;
- Section 2.2.2 presents work addressing annotation and classification of similar phenomena (more precisely the *use-mention* paradigm and the detection of *shell language* in argumentative discourse);
- Section 2.2.3 enumerates existing working tools that deal with metadiscourse.

2.2.1 Corpora

This section presents two corpora that are related to discourse, even though not targeting metadiscourse in particular.

The first is the Penn Discourse Treebank (PDTB), built directly on top of Penn TreeBank (Marcus et al., 1993) – a corpus widely used in the NLP community for training data-driven parsing algorithms, composed of extracts from the *Wall Street Journal*. PDTB was built by enriching the Penn TreeBank with discourse connectives and respective arguments (Webber and Joshi, 1998), organizing them into four categories:

- **Subordinating conjunctions –** *when, because, as soon as, now that*;
- **Coordinating conjunctions –** *and, but, or, nor*;
- **Subordinators –** *provided (that), in order that, except (that)*;
- **Discourse adverbials –** *instead, therefore, on the other hand, as a result*.

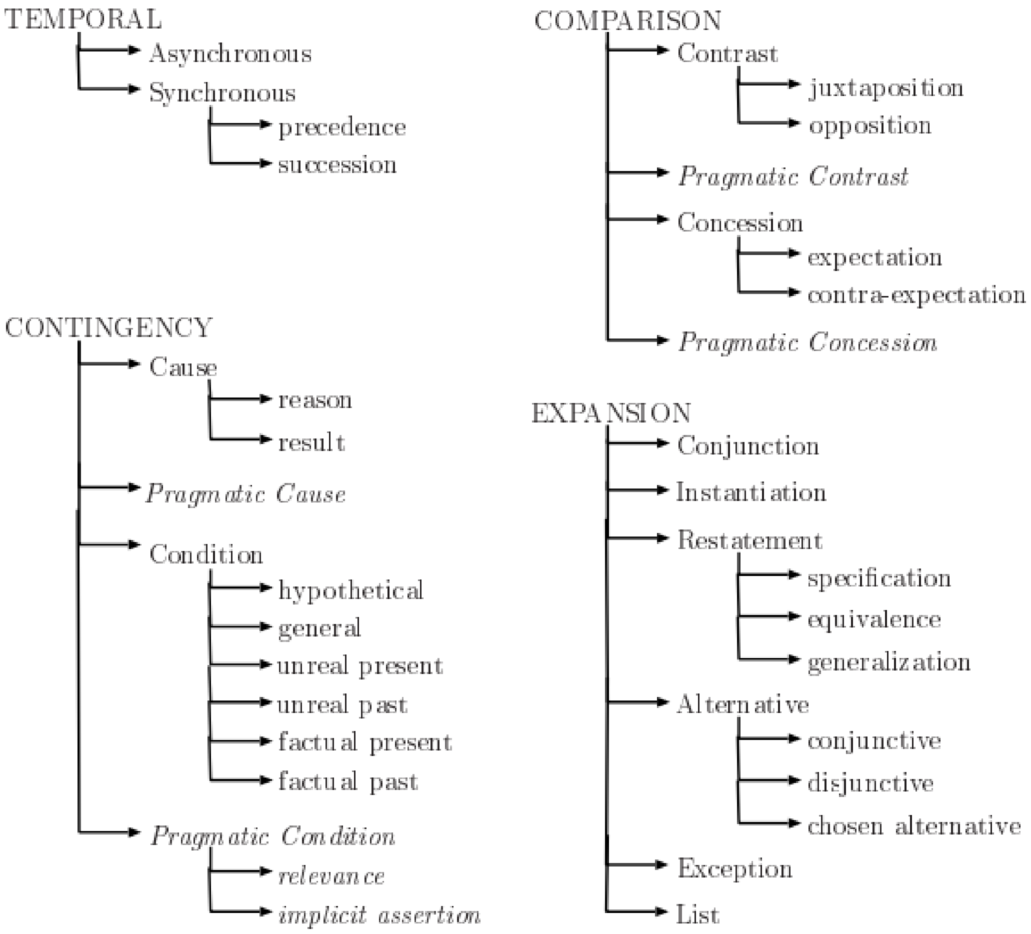


Figure 2.3: Hierarchy of senses in Penn Discourse Treebank (PDTB)

After this first approach, Miltsakaki et al. (2008) reorganized those categories according to their meaning. The resulting taxonomy of senses can be found in Figure 2.3. As in Penn Treebank, PDTB is intended to reach out to the NLP community and serve as training corpora in supervised learning approaches to discourse. The proposed sense categorization reflects this intention of automaticity, classifying discourse connectives with low-level and fine-grained discourse concepts.

Despite PDTB's lower level categorization of discourse, it is still possible to find some common ground with Ädel's functional taxonomy described in Section 2.1.5. For instance, the category EXPANSION::INSTANTIATION from PDTB somehow relates to EXEMPLIFYING in Ädel's taxonomy, and the category EXPANSION::RESTATEMENT links to REFORMULATING.

The second discourse related effort on corpora building is the Rhetorical Structure Theory (RST) Discourse Treebank (Marcu, 2000). 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 used. In the latter, such organization is the Rhetorical Structure Theory – a semantics-free theoretical framework of discourse relations developed by Mann and Thompson (1988). Marcu claims RST to be *"general enough to be applicable to naturally occurring texts and concise enough to facilitate an algorithmic approach to discourse analysis."*

ATTRIBUTION
- Attribution
- Attribution-Negative

BACKGROUND
- Background
- Circumstance

CAUSE
- Cause
- Result
- Consequence

COMPARISON
- Comparison
- Preference
- Analogy
- Proportion

CONDITION
- Condition
- Hypothetical
- Contingency
- Otherwise

CONTRAST
- Contrast
- Concession
- Antithesis

ELABORATION
- Elaboration
- Example
- Definition

ENABLEMENT
- Purpose
- Enablement

EVALUATION
- Evaluation
- Interpretation
- Conclusion
- Comment

EXPLANATION
- Evidence
- Argumentative
- Reason

JOINT
- List
- Disjunction

MANNER-MEANS
- Manner
- Means

TOPIC-COMMENT
- Problem-Solution
- Question-Answer
- Statement-Response
- Topic-Comment
- Comment-Topic
- Rhetorical-Question

SUMMARY
- Summary
- Restatement

TEMPORAL
- Temporal
- Sequence

TOPIC-CHANGE
- Topic-Shift
- Topic-Drift

Figure 2.4: Simplified Rhetorical Structure Theory categories.

Figure 2.4 shows a simplified version of the categorization used in the corpus. As with the PDTB, some of the categories of rhetoric relations in RST Discourse Treebank intersect with the high-level discourse functions defined by Ädel's taxonomy of metadiscourse.

For instance, the category EXAMPLE matches EXEMPLIFYING, DEFINITION matches COMMENTING ON LINGUISTIC FORM/MEANING or MANAGING TERMINOLOGY, and RESTATEMENT matches REFORMULATING and CLARIFYING.

Another follow-up contribution of this work is SPADE (Soricut and Marcu, 2003). SPADE[1] stands for Sentence-level PArsing for DiscoursE and, as the name states, processes one sentence at a time, outputting a discourse parse tree.

2.2.2 Automatic Classification Approaches

There is a limited amount of work found in the literature in what concerns automatic approaches to the classification of metadiscourse. This section describes studies targeting phenomena that are somewhat related to metadiscourse, and that can help set a baseline for the work proposed in this thesis.

The first relevant work is the building of corpora and classification mechanisms for metalanguage (Wilson, 2010, 2012, 2013). Herein, the author approaches metadiscourse from the point of view of *metasemantics*, which as mentioned before can be defined as the use of language to describe and analyze semantics.

More precisely, the author focuses on the *use-mention* paradigm that was introduced by Lyons (1977). This model defines the distinction between the usage of words or phrases in two situations:

- **Use –** use of language in which words are mapped to concepts outside the language;
  - *E.g.*, I watch **football** on weekends.
- **Mention –** use of language where the representation of the word is not the concept it represents, but the word itself.
  - *E.g.*, The term **football** may refer to one of several sports.

[1] http://www.isi.edu/licensed-sw/spade/ (visited November 2017)

Category	Example	#
Proper name	*A strikingly modern piece called "The Pump Room"...*	119
Translation	*The Latin title translates as "a method for finding curved lines... "*	61
Attributed Language	*"I read a chess book of Karpov", the 21-year-old said.*	47
Words as Themselves	*"Submerged forest" is a term used to describe the remains of trees.*	46
Symbols	*He also introduced the modern notation for the trigonometric functions, the letter "e" for the base of the natural logarithm.*	8
Phonetic	*The call of this species is a high pitched "ke-ke-ke".*	2
Spelling	*"James Breckenridge Speed" (middle name sometimes spelled "Breckinridge")...*	2
Abbreviation	*...often abbreviated "MIIT" for "Moscow Institute of Transport Engineers"...*	1

Table 2.1: Wilson's taxonomy of *mentioned language*.

In Wilson (2010), the author annotates one thousand sentences with *metasemantics* occurrences, proposing a taxonomy of *mentioned language*. Table 2.1 shows the categories included in Wilson's approach, along with examples of each one and their counts on the one thousand sentences sample analyzed by the author. Wilson names each category after the element that it is commenting on (translations, phonetics, symbols, *etc*).

In a follow-up study (Wilson, 2012), the author refines the taxonomy and elaborates a rubric for the annotation of *metasemantics* using the *English Wikipedia*[2] corpus. Wilson (2012) used his experience and composed a list of 23 nouns and verbs that are *mention significant*, *i.e.*, can be used as indicators of *mentioned language*:

- **Nouns –** *letter, meaning, name, phrase, pronunciation, sentence, sound, symbol, term, title, word*
- **Verbs –** *ask, call, hear, mean, name, pronounce, refer, say, tell, title, translate, write*

[2] http://en.wikipedia.org/wiki/English_Wikipedia (visited November 2017

The author then used that set of words (expanded with its correspondent *synset*) as *hooks* to retrieve a set of candidate sentences that include *mentioned language*. After this collection process, Wilson classified each sentence into one of the following categories:

- **WORDS AS WORDS (WW) –** the phrase is used to refer to the word or phrase itself (similar to WORD AS THEMSELVES in Table 2.1);
- **NAMES AS NAMES (NN) –** the sentence directly refers to the phrase as a proper name (similar to PROPER NAME in Table 2.1);
- **SPELLING OR PRONUNCIATION (SP) –** the text illustrates spelling or pronunciation (similar to SPELLING in Table 2.1);
- **OTHER MENTION (OM) –** mentioned language that does not fit the above categories;
- **NOT MENTION (XX) –** the candidate phrase is not mentioned language.

Category	**Global frequency**	**Frequency in the 100 sample**	κ
WW	438	17	0.38
NN	117	17	0.72
SP	48	16	0.66
OM	26	4	0.09
XX	1,764	46	0.74
Total	2,393	100	

Table 2.2: Wilson's annotation results.

After classifying each candidate sentence obtained by searching for the *hooks* in the *Wikipedia* articles, the author recruited three expert annotators to label a subset of 100 candidate instances. The additional annotators worked separately and received guidelines for annotation that included the five categories. The results of this annotation task can be found in Table 2.2, showing the frequency of each category in the original annotation by the author, the frequency of each category in the 100 instances sample submitted to the three additional annotators, and the correspondent *Fleiss' kappa* agreement coefficient (κ).

From the 2,393 candidate sentences retrieved by searching for the *hooks*, only about 26% were considered to contain mentioned language (1,764 were assigned to the category NOT MENTION (XX)). The expert annotators were able to reach an agreement of 0.74 in classifying if a given sentence contained an instance of *mentioned language* or not. However, the agreement for the classification of metalanguage according to the proposed categories was lower (κ between 0.09 and 0.72).

These results suggest that, although annotators tend to agree whether a candidate instance is *mentioned language* or not, there is less of a consensus on how to qualify occurrences according to their function.

In another related study, Madnani et al. (2012) explore the topic of *shell language* in argumentative discourse. As *shell language* the authors refer to language used both to express claims and evidence (*e.g. The argument states that...*), and to organize discourse (*e.g. In sum, the conclusion of this argument is not reasonable...*). These two phenomena naturally link to ARGUING and the categories under *Manage Topic* from Ädel's theory. However, the authors do not try to distinguish occurrences according to any model, encapsulating them under the term *shell language*, and focusing solely on the detection of those high-level organizational elements in argumentative discourse. The authors do so by using two distinct models:

- **Rule-based System –** this model uses a set of 25 hand-written regular expression patterns created by computing lists of n-grams ($n = 1, \ldots, 9$) extracted from annotations of essays written by test-takers of a standardized test for graduate admissions. Individuals experienced in scoring persuasive writing carried out the annotations. The rules were manually written to recognize the *shell language* present in the n-gram lists;

- **Supervised Sequence Model –** a probabilistic sequence model based on Conditional Random Fields (CRFs) (Lafferty et al., 2001), that uses a simple set of features based on lexical frequencies.

The authors evaluated the performance of the *shell text* detection methods by comparing token level system predictions to human labels. In this evaluation, the authors do not consider the exact identification of the span of a sequence of shell-related terms, but rather a token-level evaluation (whether each token is part of *shell language* or not). The rule-based system performed with an f-measure of 0.38, and the sequence model system (combined with the rule-based model) achieved an f-measure of 0.55.

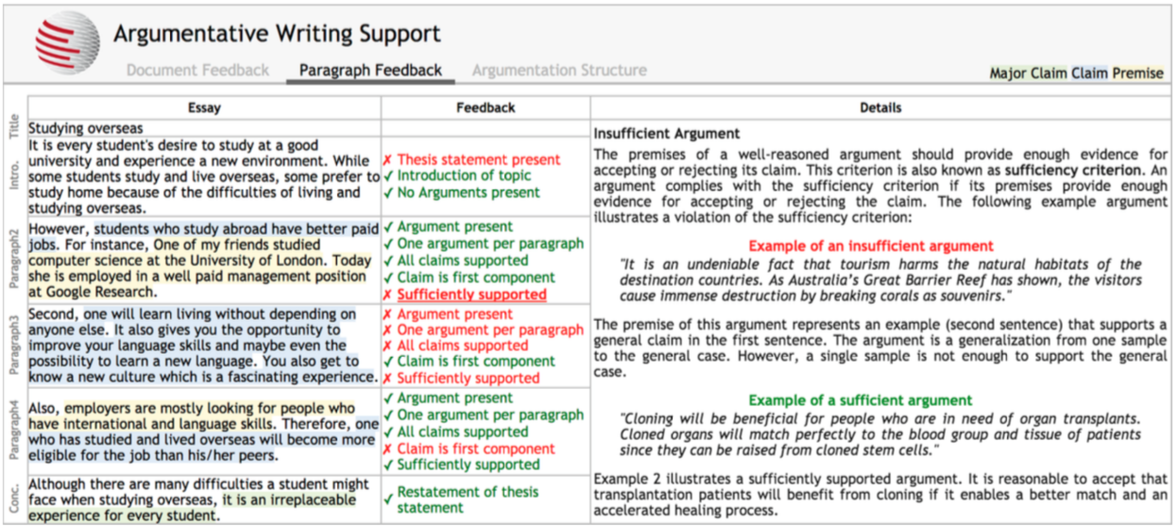
More recently, and still on the topic of detection of argumentative clues, Nguyen and Litman (2015, 2016) explored how to automatically extract argument components and relations in academic and persuasive essays. Here, the authors use a corpus annotated with three categories: *Major Claim* (writer's stance on the topic); *Claim* (statement about the major claim); and *Premise* (underpins the validity of claim), from Stab and Gurevych (2014).

Category	**Baseline**	**LDA 100 ft.**	**LDA 70 ft**
Major Claim	0.54	0.51	0.59
Claim	0.47	0.53	0.56
Premise	0.84	0.84	0.88
None	1.00	1.00	1.00

Table 2.3: Results for classification of argumentation in Nguyen and Litman (2015).

The authors first use traditional features such as word n-grams, POS patterns, lists of discourse markers, and pronouns, to classify the phenomenon. They then enhance the model by establishing a distinction between argument words (such as *"conclude"*, or *"think"*) and domain words (*"art"*, *"life"*), extracted in an unsupervised manner by using LDA (Blei et al., 2003). Table 2.3 summarizes their findings, separating the baseline approach (with the traditional features), and the new setup with 100 and 70 extracted features with the LDA. Their approach generated an overall accuracy of 83%, and F1 measures of 0.59 for *Major Claim*, 0.56 for *Claim*, and 0.88 for *Premise*. The category *None* above represent the non-occurrence of argumentative discourse, for which this work achieved perfect classification.

Suhartono et al. (2016) and Desilia et al. (2017), modeling the same task, introduce word embeddings, reporting an overall accuracy of 79.96%. Concerning each category, they report accuracies of 76.96% for *Major Claim*, 20.52% for *Claim*, and 95.41% for *Premise*.



	Essay	Feedback	Details
Title	Studying overseas		**Insufficient Argument** The premises of a well-reasoned argument should provide enough evidence for accepting or rejecting its claim. This criterion is also known as **sufficiency criterion**. An argument complies with the sufficiency criterion if its premises provide enough evidence for accepting or rejecting the claim. The following example argument illustrates a violation of the sufficiency criterion: Example of an insufficient argument *"It is an undeniable fact that tourism harms the natural habitats of the destination countries. As Australia's Great Barrier Reef has shown, the visitors cause immense destruction by breaking corals as souvenirs."* The premise of this argument represents an example (second sentence) that supports a general claim in the first sentence. The argument is a generalization from one sample to the general case. However, a single sample is not enough to support the general case. Example of a sufficient argument *"Cloning will be beneficial for people who are in need of organ transplants. Cloned organs will match perfectly to the blood group and tissue of patients since they can be raised from cloned stem cells."* Example 2 illustrates a sufficiently supported argument. It is reasonable to accept that transplantation patients will benefit from cloning if it enables a better match and an accelerated healing process.
Intro.	It is every student's desire to study at a good university and experience a new environment. While some students study and live overseas, some prefer to study home because of the difficulties of living and studying overseas.	✗ Thesis statement present ✓ Introduction of topic ✓ No Arguments present	
Paragraph2	However, students who study abroad have better paid jobs. For instance, One of my friends studied computer science at the University of London. Today she is employed in a well paid management position at Google Research.	✓ Argument present ✓ One argument per paragraph ✓ All claims supported ✓ Claim is first component ✗ Sufficiently supported	
Paragraph3	Second, one will learn living without depending on anyone else. It also gives you the opportunity to improve your language skills and maybe even the possibility to learn a new language. You also get to know a new culture which is a fascinating experience.	✗ Argument present ✗ One argument per paragraph ✗ All claims supported ✓ Claim is first component ✗ Sufficiently supported	
Paragraph4	Also, employers are mostly looking for people who have international and language skills. Therefore, one who has studied and lived overseas will become more eligible for the job than his/her peers.	✓ Argument present ✓ One argument per paragraph ✓ All claims supported ✗ Claim is first component ✓ Sufficiently supported	
Conc.	Although there are many difficulties a student might face when studying overseas, it is an irreplaceable experience for every student.	✓ Restatement of thesis statement	

Figure 2.5: User interface for Stab's system of argumentative writing support.

In the same line of research, Stab (2017) developed a system of argumentative writing support capable of providing feedback about structure, reasoning, and presence of opposing arguments. Figure 2.5 shows the user interface of the system. It can provide overall document feedback, particular feedback on a paragraph, and highlight the argumentation structure.

Cotos and Pendar (2016) looked at similar phenomena from the point of view of research papers in scientific conferences, with the ultimate goal of developing a research writing tutor. The theoretical background used in this classification task is the taxonomy of *moves* and *steps*, established by Swales (1990). In it, *moves* are communicative goals, while the *steps* are rhetorical functions that help achieve such goals. This two level categorization contains three *moves* (*Establishing a territory*, *Identifying a niche*, and *Addressing the niche*) and 17 *steps* (including *Outlining the structure of the paper*, *Reviewing previous research*, *Summarizing methods*, or *Clarifying definitions*).

With the theoretical background set up, Cotos and Pendar (2016) build a two-level cascade of classifiers: one for the classification of the *moves* and another for the *steps*. The authors realize such classifiers by using Support Vector Machines (SVM) with lexical features, which include n-grams, information about citations, and special character sequences (HTML, and URLs).

Category	**precision**	**recall**	**F1**
M: *Establishing a territory*	0.73	0.89	0.80
M: *Addressing the niche*	0.78	0.57	0.66
M: *Identifying a niche*	0.59	0.37	0.46
S: *Outlining the structure of the paper*	0.92	0.85	0.88
S: *Reviewing previous research*	0.86	0.85	0.86
S: *Making topic generalizations*	0.70	0.77	0.73
...	...	...	...
S: *Stating the value of the research*	0.39	0.34	0.37
S: *Raising general questions*	0.50	0.28	0.36
S: *Clarifying definitions*	1.00	0.18	0.31

Table 2.4: Results for classification for the *move* level, and top/bottom performant categories in the *steps* level in Cotos and Pendar (2016).

Table 2.4 shows some of the results obtained while automatically classifying discourse with respect to the mentioned taxonomy. With respect to *moves* (first three lines in the table), the classification achieved a maximum F1 of 0.80 (for the category*Establishing a territory*) and a minimum of 0.46 (for *Identifying a niche*). Concerning *steps*, the table shows the top and bottom three performant categories, which achieve a maximum F1 of 0.88 (*Outlining the structure of the paper*) and a minimum of 0.31 (*Clarifying definitions*).

Lastly, the work of Bektik (2017) aims at understanding *"how automated analysis of metadiscourse in student writing can be used to support tutors' essay assessment practices."* Herein, the author investigates a specific discourse analysis tool and discusses how it can be used for classification of metadiscourse in student writing. The analyzed tool is the Xerox Incremental Parser (XIP) (Aït-Mokhtar et al., 2002), which labels rhetorical acts according to 8 categories: *Summary*, *Background*, *Contrast*, *Novelty*, *Emphasis*, *Surprise*, *Open Question*, and *Tendency*.

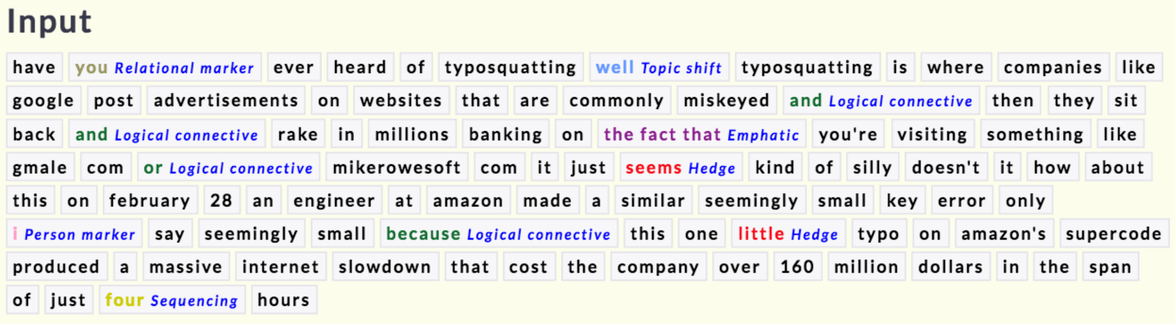
Through interviews with educators and analysis of essays, Bektik concludes that four categories in XIP are suitable to be used for educational purposes: *Background*, *Summary*, *Contrast*, and *Emphasis*.

2.2.3 Annotation Tools

The last set of previous work mentioned herein is composed of two tools that aim at inspecting and annotating metadiscourse. The first is *Text Inspector*[3] (Bax et al., 2013) – a text analysis web tool that provides a vast array of metrics. It is free to use up to 250 words and gives insights about text statistics (word counts, type/token ratios, or syllable information), readability scores, POS tagging, and, more recently, metadiscourse.

With regards to metadiscourse, it can tag a text according to Hyland's theory of metadiscourse, presented briefly at the beginning of the current chapter.

Text Inspector's approach to metadiscourse is to recognize a given list of fixed markers[4] on the submitted segments of text, with no further processing. This tool's website informs users that they need to *"take account of the context to be sure that the term identified by Text Inspector is in fact being used as the discourse marker [they] were expecting."* The web tool further provides the option for users to correct the analysis by changing the tag assigned to a specific token.



Input

have you *Relational marker* ever heard of typosquatting well *Topic shift* typosquatting is where companies like google post advertisements on websites that are commonly miskeyed and *Logical connective* then they sit back and *Logical connective* rake in millions banking on the fact that *Emphatic* you're visiting something like gmale com or *Logical connective* mikerowesoft com it just seems *Hedge* kind of silly doesn't it how about this on february 28 an engineer at amazon made a similar seemingly small key error only i *Person marker* say seemingly small because *Logical connective* this one little *Hedge* typo on amazon's supercode produced a massive internet slowdown that cost the company over 160 million dollars in the span of just four *Sequencing* hours

Figure 2.6: *Text Inspector* results for a TED talk passage.

Figure 2.6 shows the labeling results for a paragraph extracted from a TED talk. *Text Inspector* correctly marked words such as *"and"*, *"or"*, and *"because"* as *Logical Connectives*, and *"you"* and *"I"* as *Relational* and *Person Marker*, respectively. However, looking at the remaining labels brings some questions about the precision of such classification. In the

[3] http://textinspector.com/ (visited in November 2017)

[4] http://textinspector.com/help/?page_id=76 (visited in November 2017)

passage, the word *“well”* does not seem to be performing the function of shifting the topic, nor the phrase *“the fact that”* appears to be an emphatic mechanism.

More recently, Abbas and Shehzad (2017) developed *MetaPak* as an *“exclusive corpus tool for metadiscourse analysis.”* Similarly to *Text Inspector*, *MetaPak* finds instances of metadiscourse with respect to Hyland’s taxonomy, looking for words that are predicted to be associated with each category. It also allows users to correct the annotation and export its results to external files.

The main difference between *MetaPak* and *Text Inspector* is that the former provides the possibility to customize the list of words associated with metadiscourse. In other words, *MetaPak* allows the user to manipulate which words should compose which category.

2.3 Discussion

This section started by describing the existing metadiscursive theories that consider metadiscourse from the spoken variety perspective. Luukka (1992) and Mauranen (2001) focused precisely on the aspect that makes spoken metadiscourse different from the written variety: the immediacy of feedback. As a result, both taxonomies organize metadiscourse according to the individual that is talking and to the number of stakeholders involved in the communication. Consequently, Luukka's and Mauranen's theories focus mostly on form and do not address the function that the phenomena can have in the discourse.

Thompson (2003) first addressed this concern, focusing on discourse organization, and presenting a theory that categorizes the different acts of discourse organization with the level at which they occur (organizing the global topic of the talk, or the various sub-topics). Auría (2006), focusing on speaker intentions, also shows interest regarding the role of metadiscourse. Auría deals with metadiscourse at the level of grammatical units (*Let's, we'll, I'll, etc.*), using pronouns as indicators of the presence of metadiscourse. However, even though both authors address functions of metadiscourse in spoken communications, both taxonomies focus on topic organization only, not considering the full spectrum of functional roles of metadiscourse.

Ädel (2010)'s work stands out from the remaining research with a fully functional and comprehensive approach to metadiscourse. This theory directly aligns with the goal of this thesis, *i.e.*, associating metadiscourse to concepts that, at first glance, seem to be unintelligible to non-experts. For these reasons, the present thesis adopts Ädel's taxonomy as the driving theory for metadiscourse. This taxonomy will be discussed in detail in Chapter 3 while describing the annotation task aimed at building a corpus of metadiscourse for oral presentations.

Concerning existent corpora, there were two projects highlighted: the Penn Discourse Treebank (PDTB) and the RST Discourse Treebank. Even though some intersection was found between these corpora and the discourse functions in Ädel's taxonomy, the categories that compose them are low-level organizational structures, often concerned with sentence-level structure, instead of their role in the full discourse event. Moreover, both resources are built on top of *Wall Street Journal* articles, and therefore contain only instances observed in written text.

The same holds for the Natural Language Processing (NLP) classification tasks found in the literature, as all studies looked at metadiscourse from the written variety perspective only. More particularly, an extensive array of work used academic scientific articles and targeted a particular element of metadiscourse, *i.e.*, argumentative strategies. These used a taxonomy designed explicitly for the phenomenon at hand (composed of three categories), therefore, not addressing the entire spectrum of metadiscourse.

Another common characteristic of all classification tasks described is the use of a supervised approach with expert labeled corpora. On this front Wilson (2012) reported 0.74 overall agreement (κ) on whether a given sentence contained an instance of *mentioned language* or not. However, when considering a taxonomy composed of four categories, the author shows agreements that range from 0.09 to 0.72. In his work, Wilson concluded that while experts can agree on the detection of *mentioned language* in text passages, they have some difficulties in the task of classifying those segments according to their function.

In what concerns classification performance, Madnani et al. (2012) achieved an *f-measure* of 0.55 while identifying *shell language*. On the field of argumentative discourse, Nguyen and Litman (2015) achieved an F1 between 0.56 and 0.88 while distinguishing between three types of argumentative functions. While using a theory with 13 categories, Cotos and Pendar (2016) achieved performances ranged from 0.31 to 0.88. This analysis shows a significant disparity of both annotation and classification performances among items within the same taxonomy.

In sum, little focus on metadiscourse in spoken language was found throughout the literature. Most approaches focused on specific functions, and none addressed the spoken variety. Thus, this thesis aims at filling this gap, by providing a systematic approach to the use of metadiscourse in spoken discourse in the broad spectrum of functions it can perform.

3 Metadiscourse Annotation

As seen in Chapter 2, the available discourse-oriented corpora (PDTB and RST Discourse Treebank) do not directly address metadiscourse. Both corpora also only address written language, which may disregard phenomena that are specific to oral communication. Additionally, the examination of theoretical underpinnings dealing with metadiscourse at spoken language level (Luukka, 1992; Mauranen, 2001; Auría, 2006) revealed a higher focus on form (number of stakeholders involved) than on function. The exception was Ädel's taxonomy, which stands out by unifying previous work using functional concepts to characterize the phenomenon. Such approach suits the goal of representing metadiscourse in a perspective that can be understood by the general public.

Even though Ädel (2010) assembled a corpus of metadiscourse while building the aforementioned taxonomy, it cannot be used herein because of two main limitations. It considers only the pronominal use of metadiscourse, *i.e.*, instances of metadiscourse that contain pronouns (typically *I*, *you*, and *we*), thus ignoring occurrences such as *"This talk is going to be about..."* or *"The take-home message is..."*. Secondly, it is comprised of a small sample of 30 lectures, providing insufficient material to train a classifier such as the one proposed in this thesis.

These two limitations were the motivation for building a new corpus: one that addresses the full spectrum of metadiscursive strategies in spoken language. The new corpus, METATED, was created using crowdsourcing, known to provide expert-comparable quality while using less monetary- and time-related resources (Hsueh et al., 2009; Nowak and Rüger, 2010; Zaidan and Callison-Burch, 2011; Eskenazi et al., 2013). These properties allow for the annotation of a sufficient amount of data that can be used for training purposes. Another advantage of using crowdsourcing for this task is the possibility to investigate how non-experts understand the different functions of metadiscourse in the taxonomy. In practice, and linking to the guiding applications presented in the introductory chapter, this strategy can give insight, for example, on how students would react to these functions if they were to learn them as key concepts in a presentational skills curriculum.

In what concerns the collection of training data for NLP, crowdsourcing has been most commonly used on a limited set of tasks with which workers became progressively familiar with. These include object/text recognition (Von Ahn et al., 2008; Rashtchian et al., 2010; Moyle et al., 2011; Sprugnoli et al., 2017), audio transcription (Parent and Eskenazi, 2010; Eskenazi et al., 2013), and translation (Bloodgood and Callison-Burch, 2010; Ambati et al., 2012; Graça, 2014; Gao et al., 2015), to name a few. However, recent projects challenge the conservative notion of worker and expose the crowd to tasks that are cognitively more demanding and that can often accommodate different opinions. In the area of sentiment analysis, for example, workers are asked to classify a text segment with subjective notions like polarity and emotional state (Brew et al., 2010; Filatova, 2012; Nakov et al., 2013). Another example is Pellow and Eskenazi's (2014) work on simplification, where the task first instructed workers on common simplification techniques, such as lexical simplification, sentence splitting, and reordering, and then asked them to collaborate in a *chat room* to reach a solution.

Similarly, annotating metadiscourse requires workers to grasp a new concept, which depending on the category at hand, may allow for different interpretations. For that reason, such task demands careful instruction design and detailed training sessions. This chapter addresses these considerations, organizing them as follows:

- **Section 3.1** comments on the material to annotate and its metadiscursive coverage;
- **Section 3.2** describes a preliminary annotation that ran on a subset of five categories;
- **Section 3.3** presents METATED, the corpus that resulted from the full crowdsourcing effort, including its validation with experts;
- **Section 3.4** concludes on the process of building METATED and discusses results.

3.1 Source of Spoken Data

Several criteria guided the choice of material to annotate with metadiscourse. The first constraint has to do with the goal of the present research itself: having a broad representation of the different metadiscourse acts, containing strategies for managing the audience, organizing discourse, arguing, *etc.* Other criteria take into consideration one of the possible applications of the end result of this work, *i.e.*, the material suitability to be used in a presentational skills tutor. From this perspective, the literature recognizes the importance of having access to video material, which serves as visual motivation for students (Baggett, 1984; Bandura, 1986; Choi and Johnson, 2005; Wouters et al., 2007; De Grez et al., 2009a), and containing material that spans over a wide range of both topics and language proficiency levels, allowing for individual student adaptation (Brown and Eskenazi, 2005).

Two sources of spoken data that fulfilled these constraints were considered at first: classroom recordings and the online collection of TED Talks. Further analysis, however, proved TED talks be more suitable for the task:

- **Quality –** TED talks are examples of effective public speech and scientific communication, known to illustrate how to rapidly disseminate an idea (Reynolds, 2011; Nicolle et al., 2014). The proficiency quality of lectures would be more difficult to assess;

- **Time span –** each TED talk needs to convey a message in a short span of time, typically between 5 and 20 minutes, *"decreasing the chance of minds wandering or daydreaming about lunch"*[1]. This contrasts with classroom recordings, which are usually longer;

- **Context –** contrarily to lectures, a TED talk is self-contained and not required to be watched in a given sequence;

- **Topic –** TED talks target a broad audience, while lectures may require a significant amount of previous knowledge;

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

- **Communication setting –** when compared to lectures, the format of a TED talk is closer to academic and professional presentation settings, *i.e.*, speakers are typically presenting their own work and the interaction with the audience is limited;
- **Access –** TED talks are available through a `Creative Commons BY-NC-ND` license, and are uniform in what concerns audio and video quality. They are also daily updated and subtitled, providing an excellent source of transcribed material. Classroom recordings, on the other hand, are a more heterogeneous resource, regarding their origin and recording conditions, making them harder to process automatically with the least amount of human intervention possible.

By the time of the preparation of the annotation task, there were 730 TED talks available in English, with subtitles synced at sentence level (180 hours, approximately). It is important to highlight here that subtitles differ from transcripts since they omit disfluencies that typically occur in speech, such as filled pauses, deletions, fragments, or repetitions (Moniz et al., 2012). Literature shows that the removal of disfluencies does not influence comprehension (Jones et al., 2003, 2005), and is a standard procedure when creating textual representations of speech data, such as in automatic speech recognition (Stouten et al., 2006) or summarization (Zhu and Penn, 2006). While the subtitles of the TED talks tend to ignore short disfluencies (fragments of words), they do contain more prolonged events, like abandoned utterances. Such events are known to produce ungrammatical content, which NLP tools that are trained solely on the written form are unable to represent (Hayes et al., 1986).

Having as a primary concern the possibility of non-experts to be distracted by difficult vocabulary, instead of focusing on the rhetorical patterns involved in metadiscourse, the 730 talks were automatically classified according to their lexical level. This was done using the lexical level predictor developed by Collins-Thompson and Callan (2005)[2]. Collins-Thompson and Callan's classifier creates a model of the lexicon for each grade level (between 1st and 12th) and predicts the level of a document using word unigram features (Callan and Eskenazi, 2007; Heilman et al., 2008).

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

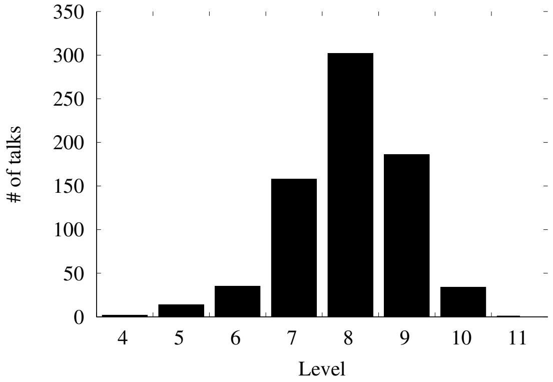


Figure 3.1: Lexical-level distribution of the 730 TED talks.

Figure 3.1 plots the lexical-level distribution of the 730 TED talks according to this method. The vast majority of talks (approx. 85%) correspond to intermediate grade levels (7th, 8th, and 9th grade). These results align with the TED goal of addressing a general audience, with speakers preferring simple vocabulary and refraining from using jargon or technical terms.

Having decided on the taxonomy to explore and on TED talks as a source of presentations, a small preliminary annotation task was carried out to test the suitability of this combination. The goal of this annotation, carried out by the author of this thesis, was to find which metadiscursive categories could be found in the TED talks. Ten TED talks were annotated with the tags from Ädel's taxonomy (see Section 2.1.5 for a detailed description). The ten talks were randomly chosen, spanning different topics and different years.

Three categories of the original taxonomy were excluded *a priori*. The category CONTEXTUALIZING, used to comment on the conditions of the presentation (practical concerns, such as time), was not further considered, since it was vague and therefore difficult to annotate. ENDOPHORIC MARKING, which as Ädel states, *"refers to cases in which it is not clear or relevant whether what is referred to occurs before or after the current point,"* was not considered since it is used to point to elements outside the discourse (such as an image in the presentation). Finally, the generic category OTHER was not annotated for not representing any particular function.

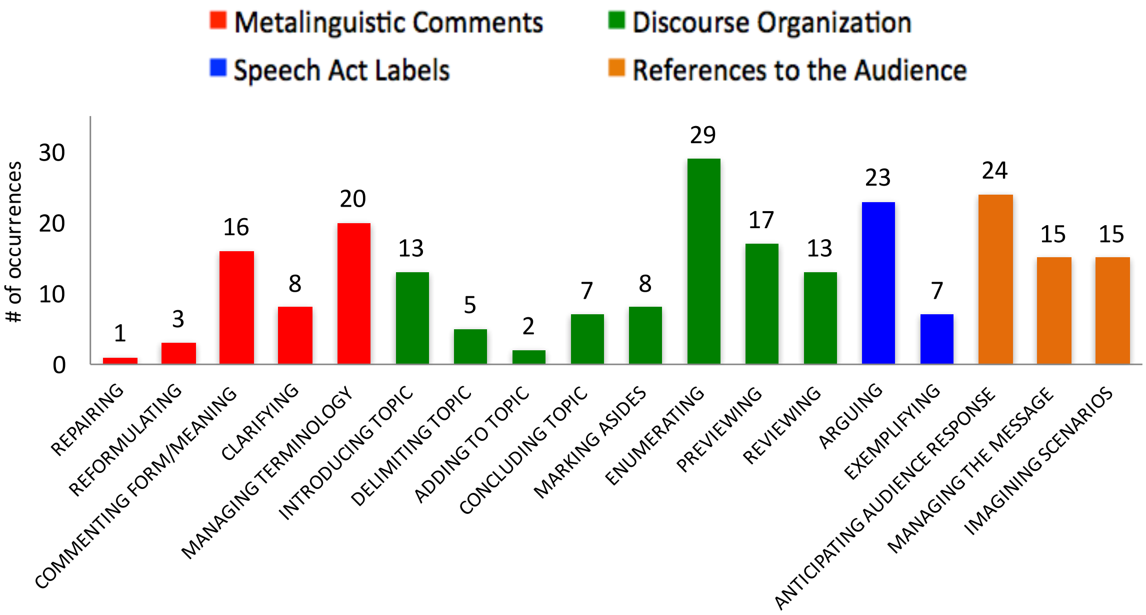


Figure 3.2: Occurrences of metadiscursive acts for the ten TED talks sample.

Figure 3.2 shows the distribution of each category over the ten talk sample. Results highlight the differences in communication setting between TED talks and the material used by Ädel when building the taxonomy (academic lectures). The differences in feedback immediacy between the two contexts affect the way speakers manage audience interaction. In a class, students can contribute to the content by raising their hand and asking a question. Since in TED talks feedback is often limited to nods, applause or laughs, no occurrences were found for MANAGING COMPREHENSION/CHANNEL and MANAGING AUDIENCE DISCIPLINE. Another difference is related to the low amount of instances of REPAIRING and REFORMULATING found. As a TED talk requires a higher degree of preparation (when compared to a lecture[3]), speakers are less likely to use repairing or reformulating strategies.

The following paragraphs, named after the four top-level categories of the taxonomy, detail how each one occurs over the sample, with examples extracted from the talks.

[3] Carmine Gallo states to have practiced his TED talk for three months and mentions another speaker who practiced theirs 200 times. `http://www.forbes.com/sites/carminegallo/2014/03/17/the-one-habit-that-brilliant-ted-speakers-practice-up-to-200-times/` [3 Dec 2015]

Metalinguistic Comments

Metalinguistics refers to the use of language to specifically comment on its form or meaning. Three of the metadiscursive acts that compose this category were found consistently in the sample – COMMENTING ON LINGUISTIC FORM/MEANING (16 times), CLARIFYING (8), and MANAGING TERMINOLOGY (16). The remaining two, REPAIRING (1) and REFORMULATING (3), were found scarcely in the sample. Again, this may be due to the high degree of preparation of each talk and the shortest time allotted to convey the message.

- REPAIRING

  – *“Just for reference, this is a sleep tracking system from just a few years ago –* **I mean, really** *until now.”*

- REFORMULATING

  – *“And that kind of gave me the inspiration –* **or rather to be precise,** *it gave my wife the inspiration.”*

- COMMENTING ON LINGUISTIC FORM/MEANING

  – *“[...] we’re advising seven or eight different countries, or political groups,* **depending on how you wish to define them.***”*

- CLARIFYING

  – ‘**I’m not saying that** *fiction has the magnitude of an earthquake.”*

  – ‘**It doesn’t mean that** *if you are a Republican that I’m trying to convince you to be a Democrat.”*

- MANAGING TERMINOLOGY

  – *“Carbon capture and sequestration –* **that’s what** *CCS* **stands for** *– is likely to become the killer app [...].”*

  – *“This is a wheat bread [...] and it’s made with a new technique [...] which,* **for a lack of better name, we call** *the epoxy method.”*

Discourse Organization

Regarding the first subcategory of *Discourse Organization*, which encompasses topic managing structures, three functions were consistently found – Introducing Topic (13 times), Concluding topic (7) and Marking Asides (8). These structures allow the speaker to manage the content of the talk. The short time frame assigned to each talk and the fact that the audience comes from a broad set of areas, requires the speakers to wisely structure their discourse in order to convey their message effectively. The remaining two functions were less frequent – Delimiting Topic (5) and Adding to Topic (2) – which may be caused by the fact that, aside from being extensively prepared, TED talks have fixed well-defined topics. Speakers tend to focus on what they want to talk about and go straight to the relevant points. Regarding *Manage Phorics*, the other subcategory under *Discourse Organization*, all three categories were found – Enumerating (29), Previewing (17), and Reviewing (13).

- Introducing Topic
  - “**I'm going to talk about** *how they are useful when we [...] want to improve.”*
  - *“[...]* **please allow me to share with you** *glimpses of my personal story.”*
- Delimiting Topic
  - *“The third reason,* **I won't go into.***”*
  - *“But I was thinking that since I planned to make a lifelong habit of coming back to TED, that* **maybe I could talk about that another time.***”*
- Adding to Topic
  - “**I must quickly add** *that this tendency [...] doesn't solely come from the West.”*
- Concluding topic
  - *“So* **to conclude***. You're supposed to read this cartoon, and, being a sophisticated person, say, Ah! What does this fish know?”*
  - “**I've just described to you** *the one story behind that rectangular area in the middle, the Phoenix Islands [...]”*

- MARKING ASIDES
  - *“I want to say –* **just a little autobiographical moment** *– that I actually am married to a wife, and she’s really quite wonderful.”*
  - *“***By the way***, what’s the Hebrew word for clay? Adam.”*
- ENUMERATING
  - *“***I want to start with** *what I call the official dogma.”*
  - *“So* **the second story that I’d like to tell** *is [...]”*
- PREVIEWING
  - *“And* **I’m going to tell you** *that story* **here in a moment***.”*
  - *“***I’ll get into** *why that is* **in just a minute***.”*
- REVIEWING
  - *“Steve Levitt* **talked to you yesterday about** *how these [...] seats don’t help.”*
  - *“[...] these women* **that I told you about** *are dancing every single day.”*

Speech Act Labels

In Ädel’s taxonomy of metadiscourse, two discourse functions compose the category *Speech Act Labels*: ARGUING and EXEMPLIFYING. Both roles were found in significant numbers throughout the ten talk sample (23 and 7 times respectively).

- ARGUING
  - *“***I’m pretty confident that** *we have long since passed the point where options improve our welfare.”*
  - *“But* **my point is** *perhaps* **that** *elusive space is what writers [...] need most.”*
- EXEMPLIFYING
  - *“***I’ll give you some examples of** *what modern progress has made possible.”*
  - *“Or* **another analogy would be** *a caterpillar has been turned into a butterfly.”*

Audience References

Contrary to academic lectures, the audience of TED talks is not entirely present at the moment of the presentation. This means that speakers have to convey the message without direct interaction with the audience. For these reasons, as previously mentioned, the tags MANAGING COMPREHENSION/CHANNEL (check if the audience is in synch with the content of the presentation) and MANAGING AUDIENCE DISCIPLINE (adjusting the channel asking for less noise, for example), were not found in the sample. On the other hand, the categories that allow the speaker to acknowledge the presence of the audience without interacting directly with it were found consistently in the talks – ANTICIPATING THE AUDIENCE'S RESPONSE (24), MANAGING THE MESSAGE (15), and IMAGINING SCENARIOS (15).

- ANTICIPATING THE AUDIENCE'S RESPONSE
  - *"And of course, describing all this,* ***any of you who know*** *politics* ***will think this is*** *incredibly difficult, and I entirely agree with you."*
  - *"Low-cost family restaurant chain,* ***for those of you who don't know it.****"*
- MANAGING THE MESSAGE
  - *"But,* **what's interesting is** *the incredibly detailed information that you can get from just one sensor like this."*
  - *"I said the other night, and* **I'll repeat now***: this is not a political issue."*
- IMAGINING SCENARIOS
  - *"But* **what I want you to do right now is imagine yourself** *400 feet underwater, with all this high-tech gear on your back,* **you're in a** *remote reef off Papua, New Guinea, thousands of miles from the nearest decompression chamber, and* **you're** *completely surrounded by sharks."*
  - *"I was a British diplomat in New York City;* **you can imagine what** *that might have meant."*

3.2 Preliminary Annotation

Before investing resources in annotating the complete set of categories of metadiscourse, a preliminary annotation was set up (Correia et al., 2014a,b). The goal was to build up the annotation interface, fine-tune its parameters, and evaluate its success in a controlled environment.

The criteria to choose which categories to deal with in a first approach took into consideration **(a)** the functions most frequently found in the literature on metadiscourse and discourse in general, **(b)** an empirical opinion of which concepts could be better explained to non-experts, and **(c)** the input from Carnegie Mellon University International Communications Center (entity that holds presentation skills workshops and administrates tests for non-native speakers applying for teaching assistant positions). Out of the 18 categories that were found on the ten TED talk sample (as seen in Figure 3.2), INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING, MANAGING THE MESSAGE, and IMAGINING SCENARIOS were chosen for annotation at this stage.

In line with the goal of representing metadiscourse functionally, the occurrences of IMAGINING SCENARIOS were dealt with as belonging to the category EXEMPLIFYING, since both acts represent the same concept, varying only in form (if it involves mentions to the audience or not). Additionally, for simplification and easier comprehension, MANAGING THE MESSAGE (in Ädel's work, *"used to emphasize the core message in what is being conveyed"*) was renamed and will be further referred to as EMPHASIZING. The remainder of this section deals with the annotation of these five categories, and is organized as follows:

- **Section 3.2.1** describes the conditions in which the annotation was carried out, including interface, instructions, training sessions, and payment;

- **Section 3.2.2** shows the results obtained regarding quantity and quality;

3.2.1 Setup

In this annotation task, workers on Amazon Mechanical Turk (AMT)[4] were asked to annotate the transcripts of 730 TED talks with the five categories of metadiscourse previously mentioned – INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING, and EMPHASIZING. Herein, the goal is to submit for annotation the entire talks and let workers spot the occurrences of metadiscourse in them. This strategy contrasts with Ädel (2010) and Wilson (2012) for example, who used a predefined set of words to retrieve an initial set of candidate sentences that were later labeled. In Ädel's case this set was composed of personal pronouns, and in Wilson's work, of *"mentioned significant"* expressions, such as *meaning*, *term*, or *say*, to name a few.

Contrarily to experts, with whom the annotation environment is more easily controlled, using the crowd requires the set up of training and quality assurance mechanisms to avoid unwanted noise in the answers. It is also necessary to approach tasks differently, dividing complex jobs into subtasks to reduce cognitive load (Le et al., 2010; Eskenazi et al., 2013). As discussed before, the designing phase is particularly relevant for the annotation of metadiscourse since workers are likely unfamiliar with the concepts at hand. It is important to highlight that the considerations presented below were not only used in this preliminary task but constitute the basis for the full annotation effort described ahead in Section 3.3.

The first decision concerns the amount of text to annotate in each Human Intelligence Task (HIT) – the smallest unit of work someone has to complete to receive payment. Each HIT should be simple and allow workers to solve it in the fastest way possible. Knowing that metadiscursive phenomena are not local, usually requiring the understanding of the surrounding context, the decision was to use segments of approximately 300 words (adjusted for each segment to consider full sentences). The 300-word limit was influenced by the design of the interface of the annotation task, taking into consideration that all the text should be visible to the workers at a given point (scrolling increases time-on-task thus influencing answer rate). Additionally, to pay an amount of money that is worthwhile for workers to choose

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

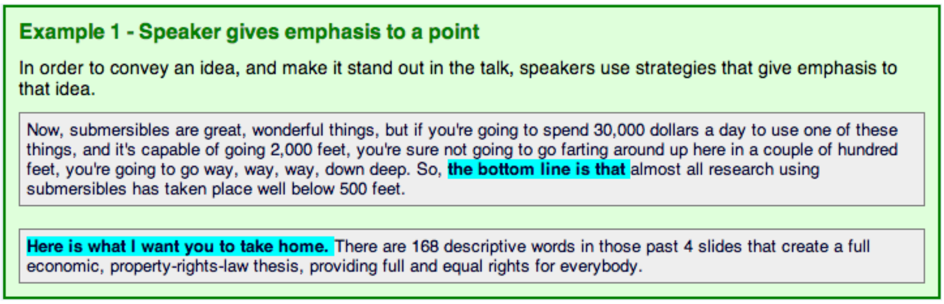
the task, each HIT was composed of four segments in a 4×4 matrix (randomly grouped previously).

The second consideration concerns the design of the instructions. Knowing that metadiscourse is a concept which most workers have probably never heard of, each HIT was designed to target one category at a time, instead of requiring the identification of all five categories in each segment in one single pass. The resulting configuration (300 words per segment, and four segments per HIT) generated a total of 2,461 HITs (or 9,844 segments) per category. Additionally, for quality assessment and agreement report, each HIT was presented to three different workers.

The instructions for each category contained a first paragraph, common to all categories, which revealed the high-level purpose of the work, motivating workers and increasing their sense of contribution, and a second paragraph, with the definition of the category and a short list of examples. As an example, instructions for the category EMPHASIZING 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...".*

After the instructions, there was a section of examples (Figure 3.3) and counterexamples (Figure 3.4), each accompanied with a description of why they were considered (or not) an example of the metadiscursive marker at hand (EMPHASIZING, in the figures). Finally, before showing the segment to annotate (Figure 3.5), there was a succinct description of the set of steps that explained the interface and how to use it to annotate the passages.



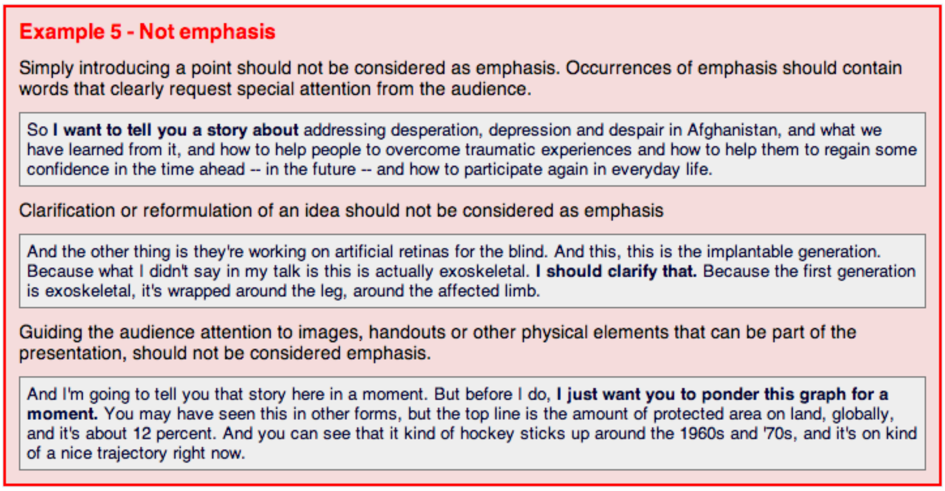
Example 1 - Speaker gives emphasis to a point

In order to convey an idea, and make it stand out in the talk, speakers use strategies that give emphasis to that idea.

Now, submersibles are great, wonderful things, but if you're going to spend 30,000 dollars a day to use one of these things, and it's capable of going 2,000 feet, you're sure not going to go farting around up here in a couple of hundred feet, you're going to go way, way, way, down deep. So, **the bottom line is that** almost all research using submersibles has taken place well below 500 feet.

Here is what I want you to take home. There are 168 descriptive words in those past 4 slides that create a full economic, property-rights-law thesis, providing full and equal rights for everybody.

Figure 3.3: Positive examples for the task of identifying occurrences of EMPHASIZING.



Example 5 - Not emphasis

Simply introducing a point should not be considered as emphasis. Occurrences of emphasis should contain words that clearly request special attention from the audience.

So **I want to tell you a story about** addressing desperation, depression and despair in Afghanistan, and what we have learned from it, and how to help people to overcome traumatic experiences and how to help them to regain some confidence in the time ahead -- in the future -- and how to participate again in everyday life.

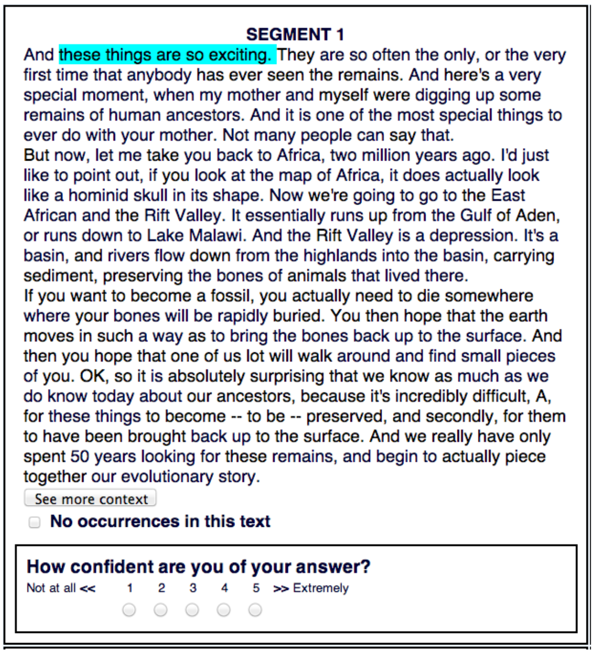
Clarification or reformulation of an idea should not be considered as emphasis

And the other thing is they're working on artificial retinas for the blind. And this, this is the implantable generation. Because what I didn't say in my talk is this is actually exoskeletal. **I should clarify that.** Because the first generation is exoskeletal, it's wrapped around the leg, around the affected limb.

Guiding the audience attention to images, handouts or other physical elements that can be part of the presentation, should not be considered emphasis.

And I'm going to tell you that story here in a moment. But before I do, **I just want you to ponder this graph for a moment.** You may have seen this in other forms, but the top line is the amount of protected area on land, globally, and it's about 12 percent. And you can see that it kind of hockey sticks up around the 1960s and '70s, and it's on kind of a nice trajectory right now.

Figure 3.4: Negative examples for the task of identifying occurrences of EMPHASIZING.



SEGMENT 1

And these things are so exciting. They are so often the only, or the very first time that anybody has ever seen the remains. And here's a very special moment, when my mother and myself were digging up some remains of human ancestors. And it is one of the most special things to ever do with your mother. Not many people can say that.

But now, let me take you back to Africa, two million years ago. I'd just like to point out, if you look at the map of Africa, it does actually look like a hominid skull in its shape. Now we're going to go to the East African and the Rift Valley. It essentially runs up from the Gulf of Aden, or runs down to Lake Malawi. And the Rift Valley is a depression. It's a basin, and rivers flow down from the highlands into the basin, carrying sediment, preserving the bones of animals that lived there.

If you want to become a fossil, you actually need to die somewhere where your bones will be rapidly buried. You then hope that the earth moves in such a way as to bring the bones back up to the surface. And then you hope that one of us lot will walk around and find small pieces of you. OK, so it is absolutely surprising that we know as much as we do know today about our ancestors, because it's incredibly difficult, A, for these things to become -- to be -- preserved, and secondly, for them to have been brought back up to the surface. And we really have only spent 50 years looking for these remains, and begin to actually piece together our evolutionary story.

□ **No occurrences in this text**

How confident are you of your answer?

Not at all << 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ >> Extremely

Figure 3.5: Interface of one segment in a HIT.

Again, for the category EMPHASIZING, each HIT read as follows:

1. For each of the extracts below, click on EVERY word that the speaker uses to EMPHASIZE A POINT. There may be zero, one or more instances in each extract.
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.
3. If you need more information to support your decision, you can click "*See more context*" below the segment to see its surrounding context in the talk.
4. If the speaker does not emphasize any point in the extract, select the "*No occurrences in this text*" checkbox below the text.
5. Rate your confidence level on your answer and click the SUBMIT button afterward.

The set of steps enumerated above are in sync with the interface of the HIT shown in Figure 3.5. Three mechanisms of the HIT are highlight-worthy. First, there is a button that, when clicked on, shows workers the text surrounding the current segment, in the event of needing additional context (step 3 above). More precisely, it redirects workers to a view of the talk with both the previous and next segment of the talk shown in context with the current one. Workers had to intentionally close this view to continue working on the HIT.

Secondly, they were required to explicitly signal the absence of occurrences of metadiscourse in the segment by selecting a checkbox (step 4). The interface did not allow the submission of answers with both no selection of words in the segment or an unchecked box.

Lastly, workers were required to rate their overall confidence in a 5-point Likert scale (step 5), answering the question *"How confident are you of your answer?"*, where 1 corresponded to *"Not at all"*, and 5 to *"Extremely"*.

The last set of design considerations address quality control. AMT provides a prerequisite feature to filter for workers with particular characteristics and demographics. In this case, only native-speakers of English with a reliability rate of at least 95% were allowed to participate. The reliability rate refers to the percentage of the worker's HITs that were accepted before, in the pool of all work they submitted through AMT.

Workers who satisfied these two prerequisites and accepted the HIT were then guided through a category-specific training session. This session was designed to test if the worker read the instructions and examples carefully, and consisted of four sequential segments. Each of the segments was prepared manually to provide targeted feedback depending on the type of problem in the answer (or a success message otherwise). Common problems included **(a)** missing an existent occurrence, **(b)** selecting an occurrence that was not to be considered, and **(c)** boundary selection problems. As an example of the latter criterium, please note the sentence *“Now, I want to emphasize that not every autistic kid is going to be a visual thinker,”* for which the most correct answer is the expression *“I want to emphasize that.”* The interface would warn the users if, for instance, they selected the entire sentence, or on the other hand, just the word *“emphasize”*. The design of each training segment allowed for small selection variation (for example, if the worker did no select the word *“that”*, or if he/she selected the word *“Now”*). Nonetheless, workers always saw the expected answer after passing each segment. Only upon completion of the four training segments were the workers allowed to access real HITs in the corresponding category.

While training is an effective strategy to filter out *bots*, it does not prevent bad-intentioned workers from complying with it just to give random answers to the real HITs afterward. For that reason, and in line with good crowdsourcing practices (Hsueh et al., 2009; Eskenazi et al., 2013), a gold standard was defined for each of the five metadiscursive categories. For every four HITs, the workers saw one previously annotated segment

The gold standard segments were very similar to the examples provided. Given their simplicity, failing one of them flagged the worker as a potential spammer. The decision to accept or reject the work of flagged workers was made by analyzing each case separately.

In this first experiment with crowdsourcing, the choice was to run each category sequentially, *i.e.*, no two categories were *online* at the same time. This configuration allowed for close and detailed inspection of the work, which included giving individual and personalized feedback encouraging workers who were performing well, and warning workers who were in constant disagreement.

3.2.2 Results

Category	**Workers in Agreement**				**%**	**conf**	**avg**	**agr**	κ
	1	**2**	**3**	**2+3**	**exp**	**(stdev)**	**time**	**(%)**	
INTRODUCING TOPIC	732	556	600	1,156	1.32	3.95 (0.98)	03:43	99.36	0.64
CONCLUDING TOPIC	397	346	265	611	37.09	4.00 (0.77)	03:31	99.64	0.60
MARKING ASIDES	N/A	N/A	N/A	N/A	5.52	3.60 (1.54)	10:04	N/A	N/A
EXEMPLIFYING	387	613	674	1,287	4.81	3.94 (0.70)	06:12	99.55	0.72
EMPHASIZING	806	1843	631	2,474	1.14	3.99 (1.02)	06:19	98.89	0.58

Table 3.1: Annotation results in terms of number of instances, behavior, and agremeent.

Table 3.1 summarizes the annotation results regarding the number of instances, annotation statistics, and inter-annotator agreement. The first four columns represent the number of sentences in which workers identified metadiscourse. This information is organized by how many workers agreed on each instance. For example, for the category INTRODUCING TOPIC, there were 600 occurrences selected by all three workers, 556 occurrences selected by two of the workers, and 732 occurrences marked by one worker only. The column *2+3* shows majority vote, *i.e.*, the number of sentences that were signaled by at least two workers. The remaining columns in Table 3.1 address, respectively, the percentage of segments for which workers expanded context (% exp), average of self-reported confidence on a 5-point Likert scale (*conf*), average time spent per HIT in minutes (*avg time*), and inter-annotator agreement, both overall agreement (*agr*) and Fleiss' kappa (Joseph, 1971) (κ). Inter-annotator agreement considered that annotators agree if for each sentence there is at least one selected word in common between the annotators, *i.e.*, if the intersection of the words selected by the annotators is not empty.

The first noticeable result in Table 3.1 is the lack of information for the number of occurrences and agreement for the category MARKING ASIDES. This absence is due to workers showing signs of being unable to deal with this category, which ultimately led to abandoning its annotation. The first failure indicator was the slow response rate: in one week span, less than 10% of the HITs were completed, in contrast with above 30% for the remaining four categories in the same span of time. Another sign was the amount of time spent on the task: 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). Additionally, the comments left by the workers while annotating MARKING ASIDES reflect their discomfort and lack of confidence: *"I am nervous that I am not doing these correctly *at all*"*; *"I hope that this is what you are looking for"*; *"Hope I'm doing well"*; and *"a little difficult."*

Another variable that may have added to the failure of the annotation is the low frequency of MARKING ASIDES. During the 10-talk sample annotation (described in Section 3.1) only eight instances of this category were found. It may be that without regular exposure, workers end up not finding any instances of the phenomenon and get the sense that they are not contributing. The fear of having their work rejected, which besides denying payment also negatively affects their statistics on the platform, is likely to cause them to abandon the task.

Workers were able to complete the remaining four categories. It is important to stress again that there was close control of the answers in the process and the agreement reported does not encompass this rejected work. EXEMPLIFYING was the category with the best performance ($\kappa = 0.72$). As previously mentioned, this category collapses two metadiscursive acts, as defined in Ädel's taxonomy: EXEMPLIFYING and IMAGINING SCENARIOS. Despite the collapse of tags, the fact that in this category annotators reached the highest agreement seems to corroborate the decision to unify both categories under the same functional concept. On the other hand, EMPHASIZING was the category where workers achieved the lowest inter-annotator agreement ($\kappa = 0.58$). This result may occur because this category is the only one that admits a scale of intensity, *i.e.*, different workers have different thresholds for considering that the speaker is emphasizing.

Similarly to what happened before, the number of instances of INTRODUCING TOPIC is larger than the number of talks (13 occurrences in 10 talks vs. 1,156 in 720 talks). This observation occurs since speakers introduce several topics throughout a single talk. The reverse trend occurs in CONCLUDING TOPIC, with the number of occurrences (around 600) being lower than the total amount of talks (and approximately half the number of instances of INTRODUCING TOPIC). This relation shows that speakers do not explicitly conclude every topic in the talk. This discourse function can be performed, for instance, by directly introducing a new topic, which implicitly ends the previous one.

Another significant result concerns the expansion of context. CONCLUDING TOPIC registers a significant difference compared to the remaining categories, with annotators asking to see the surrounding context 37% of the time. This result indicates that conclusions are less local, and people need a broader context to identify them.

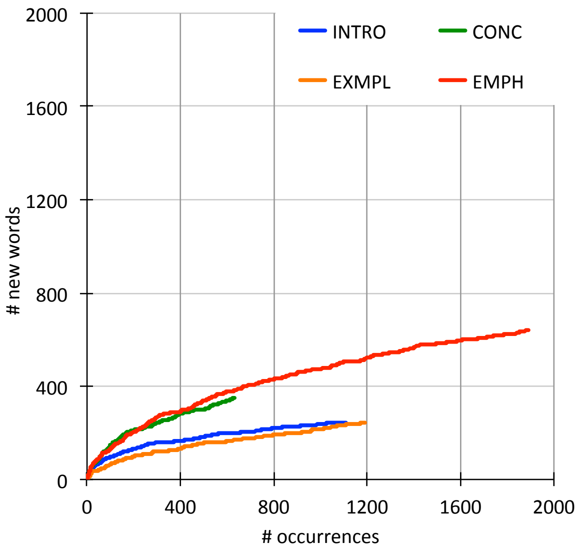


Figure 3.6: Type-token curves for INTRODUCING TOPIC, CONCLUDING TOPIC, EXEMPLIFYING and EMPHASIZING.

Figure 3.6 shows the type-token curves for the four categories annotated. For each occurrence of metadiscourse signaled by at least two workers (x-axis), it plots how many words are newly discovered (y-axis). It is possible to see that, while the rate of new word discovery for the categories EXEMPLIFYING and INTRODUCING TOPIC is stabilizing (around the 200 words threshold), CONCLUDING TOPIC and EMPHASIZING show a linear growth rate towards the right side of the figure. Interestingly, these latter two categories were also the ones where workers agreed the least. These results suggest speakers use a more considerable amount of distinct strategies to conclude a topic and to emphasize. Therefore, to better represent the phenomena that these two categories aim at representing, more annotation would be necessary.

Looking into the actual words that the workers selected (top n-grams), patterns start to appear. For INTRODUCING TOPIC, the verbs *"talk"*, *"show"* and *"tell"* often appear in constructions such as *"I am going to talk about"* and *"I want to show you"*. For CONCLUDING TOPIC, the verbs *"leave"* and *"conclude"* start ranking higher, along with expressions such as *"the last thing"*. For EXEMPLIFYING, as expected, the unigrams *"example"* and *"imagine"* rank first and third (respectively). Finally, regarding EMPHASIZING, the word *"important"* and expressions such as *"I want you to remember"* and *"the bottom line"* show to be the most relevant. Appendix A.1 contains a ranked list of the top n-grams for each of the four categories, accompanied by the same information for the entire set of TED talks, for reference.

3.3 Building of metaTED

The results obtained during the preliminary annotation supported the decision to extend the task to the remainder categories in Ädel's taxonomy of metadiscourse. This section describes the building of METATED (Correia et al., 2016), a freely available corpus of metadiscursive acts in spoken language collected via crowdsourcing.

It takes into consideration the lessons learned during the first annotation trials (described above), scaling the task to the full spectrum of metadiscourse. While most of the interface, instructions, and workings of the annotation remain unchanged, some tuning and revisions took place. The current section is organized as follows:

- **Section 3.3.1** addresses the aforementioned modifications to the annotation setup;
- **Section 3.3.2** presents the results of the annotation task, with detailed statistics on the quality of the collected data;
- **Section 3.3.3** presents a validation task with experts designed to further assess the quality of the material collected.

3.3.1 Annotation Setup

The first adaptation refers back to the unsuccessful annotation of MARKING ASIDES. The hypothesis that low exposure to positive instances discourages participation motivated the grouping of some of the categories. This union only occurred for categories related in some way and, thus, possible to explain together by taking advantage of their similarities. As a result, the categories MARKING ASIDES and ADDING TO TOPIC were merged into a new category – ADDING INFORMATION. Similarly, the categories REPAIRING and REFORMULATING were consolidated into a single category – REPAIR & REFORMULATING.

In contrast, experience with the data showed that REVIEWING, having reasonable representativity, could be further divided into two categories easier to explain separately: RECAPIT-

ULATING, used to go over or to summarize a point (as in *"Let me go through what we've seen so far."*, and REFERRING TO PREVIOUS IDEA, a mechanism used to refer to something mentioned previously, (as in *"The girl that I told you about. . . "*).

Similarly to what happened to MANAGING THE MESSAGE (renamed to EMPHASIZING), some category names were changed, in the hopes of simplifying and aiding understanding. This process of adaptation of the original taxonomy generated a set of 16 categories of metadiscourse: the final group of metadiscursive markers that compose METATED.

A summary of the modifications of Ädel's taxonomy is shown below. For each one, instructions were built (with examples and counterexamples), gold standards defined and training sessions prepared, with the same purpose as in the preliminary annotation. The accompanying acronyms will be frequently used in the remainder of the document when space restrictions apply.

- REPAIR & REFORMULATING (R&R) – union between REPAIRING and REFORMULATING
- COMMENTING ON LINGUISTIC FORM/MEANING (COM)
- CLARIFYING (CLAR)
- DEFINING (DEF) – originally MANAGING TERMINOLOGY
- INTRODUCING TOPIC (INTRO)
- DELIMITING TOPIC (DELIM)
- CONCLUDING TOPIC (CONC)
- ENUMERATING (ENUM)
- POSTPONING TOPIC (POST) – originally PREVIEWING
- DEFENDING IDEA (DEFND) – originally ARGUING
- ANTICIPATING THE AUDIENCE'S RESPONSE (ANT)
- EMPHASIZING (EMPH) – originally MANAGING THE MESSAGE
- ADDING INFORMATION (ADD) – union between ADDING TO TOPIC and MARKING ASIDES
- EXEMPLIFYING (EXMPL) – union between EXEMPLIFYING and IMAGINING SCENARIOS
- RECAPITULATING (RECAP) – subdivision of the original REVIEWING
- REFERRING TO PREVIOUS IDEA (REFER) – subdivision of the original REVIEWING

In response to the slow answer rate and concerns expressed in the comments section about the amount of text to annotate, the workings of the task were redesigned. Workers now saw a single 500-word segment per HIT (instead of the previous four 300-word segments) and received a higher payment for each HIT. As previously, to reduce the impact of lack of exposure to positive instances, gold standards were presented every four HITs.

These set of modifications strongly increased the monetary cost of the task. For that reason, instead of the complete set of 730 talks annotated previously, METATED is composed of a subset of 180 talks, randomly chosen (742 HITs per category), totaling 23,348 sentences and 418,368 tokens. Again, as previously, three different workers annotated each segment.

Finally, for time management reasons, all categories were annotated simultaneously. This setup allowed for much less control, with no manual answer validation, using gold standard agreement as the only accept/reject criteria. Given the substantial differences between setups, especially in what concerns manual quality control, annotation was repeated for the set of four categories of the preliminary task, *i.e.*, INTRODUCING TOPIC, CONCLUDING TOPIC, EXEMPLIFYING, and EMPHASIZING.

3.3.2 Annotation Results

Table 3.2 shows the results of the crowdsourcing task regarding number of instances, annotation statistics, and inter-annotator agreement. As before, the first four columns represent the number of sentences where workers identified metadiscourse. This information is organized by how many workers agreed on each instance. For example, for the category REPAIR & REFORMULATING, there were 46 occurrences selected by all three workers, 233 occurrences selected by two of the workers, and 1,493 occurrences marked by one worker only. The column *2+3* shows majority vote (the number of sentences that were signaled by at least two workers). There was a high disparity amongst all three settings of workers in agreement for all categories, with a low percentage of instances annotated by all three workers. CLARIFYING and ADDING INFORMATION are the most extreme cases, with the number of instances selected by one worker only being 6 to 7 times higher than those marked by the majority.

Cat	Workers in Agreement				%	conf	avg	agr	α	α	κ
	1	2	3	2+3	exp	(stdev)	time	(%)	2+3		
R&R	1,493	233	46	279	3.35	3.57 (0.96)	02:23	95.03	0.61	0.16	0.16
COM	738	271	85	356	1.39	3.10 (0.76)	02:08	97.19	0.66	0.34	0.33
CLAR	1,975	283	58	341	3.46	3.82 (0.90)	02:27	93.57	0.62	0.15	0.15
DEF	836	189	68	257	5.62	4.04 (0.85)	02:27	97.13	0.67	0.28	0.29
INTRO	732	239	131	370	5.08	3.40 (1.17)	01:33	97.31	0.73	0.39	0.40
DELIM	132	28	12	40	1.85	4.21 (0.79)	01:53	99.58	0.70	0.30	0.31
ADD	923	102	33	135	3.51	3.88 (1.10)	01:55	97.14	0.65	0.16	0.15
CONC	153	52	34	86	18.67	4.36 (0.78)	01:12	99.42	0.75	0.43	0.44
ENUM	1,067	368	346	714	2.50	3.74 (0.70)	02:01	95.95	0.79	0.49	0.49
POST	184	23	24	47	4.20	4.17 (0.69)	01:55	99.45	0.80	0.32	0.32
RECAP	202	32	4	36	11.78	3.33 (0.76)	02:15	99.35	0.58	0.16	0.18
REFER	411	83	42	125	3.16	3.93 (0.54)	01:50	98.63	0.72	0.29	0.32
DEFND	1,538	322	223	545	4.56	3.51 (1.18)	02:02	94.77	0.72	0.31	0.32
EXMPL	771	195	140	335	2.50	3.62 (0.72)	01:58	97.31	0.77	0.39	0.38
ANT	1,426	356	100	456	3.80	3.61 (1.02)	01:56	95.07	0.65	0.24	0.24
EMPH	2,023	336	80	446	4.16	3.31 (0.98)	02:17	93.41	0.52	0.18	0.18

Table 3.2: Annotation results in terms of quantity and quality.

Regarding the percentage of times workers asked for additional context (column *(%) exp*), the categories CONCLUDING TOPIC and RECAPITULATING significantly differ from the remaining categories, with workers asking for more context 19% and 12% of the time, respectively. The category CONC had already displayed similar behavior in the preliminary annotation. On the other hand, the categories that seem to be more local, not needing additional context to be identified (besides the 500 words) were COMMENTING ON LINGUISTIC FORM/MEANING and DELIMITING TOPIC, where workers asked for additional context less than 2% of the time.

The next column in Table 3.2 shows the average self-reported confidence (on a 5-point Likert scale) and corresponding standard deviation. This metric, due to a technical fault, was only registered for a subset of 100 HITs. All categories score above the middle of the scale (3), with workers showing less confidence for COM. On the other hand, workers showed to be the most confident when annotating instances of CONCLUDING TOPIC, DELIMITING TOPIC, and POSTPONING TOPIC: three categories that signal change of topic in a talk.

Regarding time-on-task, no significant variations were observed. Most categories required about 2 minutes per segment. The only exception is CONCLUDING TOPIC, taking only about one minute per segment. Interestingly, this was the category where workers most expanded context and achieved the second-best inter-annotator agreement.

The last four columns of Table 3.2 report different measures of agreement. The observed agreement represents the percentage of times workers concur. Concordance includes all instances where workers do not mark any occurrence in a sentence, hence the high values. Krippendorf's α is introduced here since it covers some of the limitations of Fleiss' κ, namely not being able to deal with a variable number of annotators, missing information and not adjusting well to small or unequal samples. The last two columns represent the same information in both measures, showing how tightly connected they are. Using Krippendorf's α allows reporting statistics such as the one in column α *2+3*, which reports inter-annotator agreement if ignoring instances selected only by a single worker.

These agreement results show that non-experts have the most trouble while identifying CLAR, ADD, RECAP, R&R, and EMPH, all with $\alpha < 0.20$. The categories CONCLUDING TOPIC and ENUMERATING, on the other hand, show the highest levels of agreement. The four acts annotated during the first crowdsourcing attempt suffered a drastic decrease in agreement. In fact, the values under the column α *2+3* are the ones that seem to be in the order of what was seen previously – INTRO (before: 0.64; after: 0.73), CONC (0.60; 0.75), EXMPL (0.73; 0.77), and EMPH (0.58; 0.52). Such observation confirms that the amount of control performed during the preliminary annotation study strongly influenced the results, leaving the question of whether what was found previously can be considered a reflection of the crowd.

For better comprehension of the annotated material, Appendix A.2 contains a ranked list of the top 10 uni-, bi-, and trigrams for each of the 16 categories. Additionally, the list below, compiled after a qualitative analysis of the collected material, enumerates the primary sources of disagreement:

- **Variance in interpretation –** In categories such as EMPHASIZING and DEFENDING IDEA, it was possible to observe that workers approached the annotation from different standpoints. For instance, regarding EMPHASIZING, some workers signaled occurrences where the emphasis was very subtle (such as *"An important result. . . "*), while others marked only much more explicit cues (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 CLARIFYING 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 into two different 500-word segments). By 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 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 at the beginning of this section, some categories with lower representation were unified under a broader concept, such as the case of the categories ADDING INFORMATION, EXEMPLIFYING, and REPAIR & REFORMULATING. These unions, however, may add to the workers' cognitive load and hinder 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 REPAIR & 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 lower levels of agreement;

- **Lack of attention –** Despite workers' answers were compared to a golden standard and removed if continually missed, there were still some definite occurrences of metadiscourse that ended up not been caught. For example, workers did not always spot the pattern *"by the way"*, a mark of making an aside, even though explicitly taught during training.

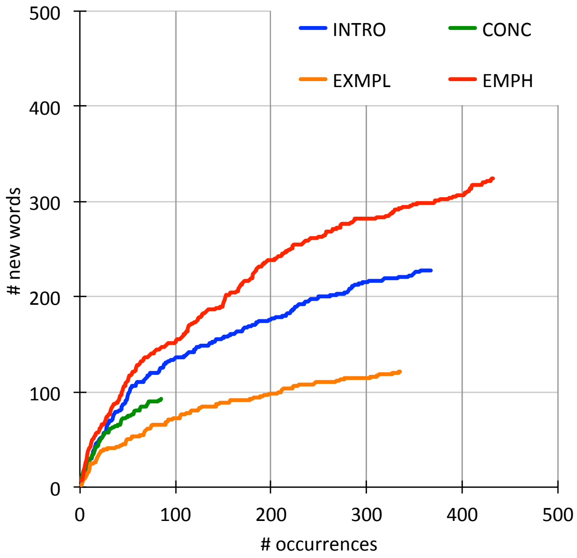


Figure 3.7: Type-token curves for INTRODUCING TOPIC, CONCLUDING TOPIC, EXEMPLIFYING and EMPHASIZING in METATED.

Category	occurr (major vote)	Avg. occurr per talk	New word per occurr	New word per talk
R&R	279	1.55	0.18	0.28
COM	356	1.98	0.26	0.51
CLAR	341	1.89	0.12	0.23
DEF	257	1.43	0.16	0.23
INTRO	370	2.06	0.40	0.82
DELIM	40	0.22	1.14	0.25
ADD	135	0.75	0.48	0.36
CONC	86	0.48	0.23	0.11
ENUM	714	3.97	0.27	1.07
POST	47	0.24	1.10	0.26
RECAP	36	0.20	1.18	0.24
REFER	125	0.69	0.65	0.45
DEFND	545	3.03	0.11	0.33
EXMPL	335	1.86	0.21	0.39
ANT	456	2.53	0.34	0.86
EMPH	446	2.48	0.33	0.82

Table 3.3: Predicted rate of new word discovery after the annotation of 180 talks.

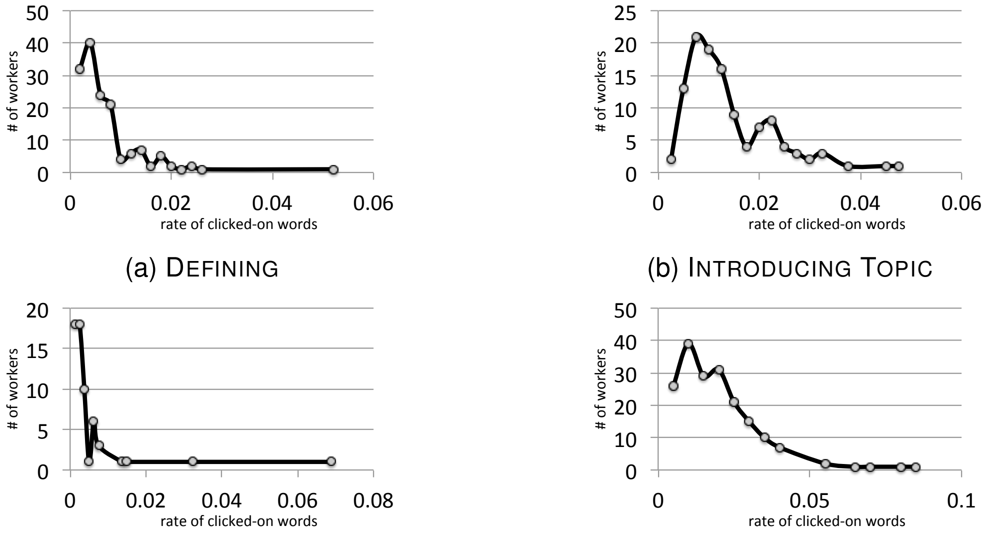
Similarly to what was done for the preliminary annotation (see Figure 3.6), Figure 3.7 plots the type-token curves for the categories INTRO, CONC, EXMPL, and EMPH. It is possible to see that, as new occurrences are analyzed, the rate of new words being discovered decreases. Out of the four categories represented, if the annotation continues, EMPHASIZING is the category in which more new words are expected to be found.

The type-token curves for all 16 functions are shown in Appendix B. While most categories show some decay in the rate of new words as more observations are analyzed, the categories DELIMITING TOPIC, POSTPONING TOPIC, and RECAPITULATING show no signs of stabilizing.

Table 3.3 shows the rate of new words being discovered more quantitatively. The column *occurr (major vote)* corresponds to the number of occurrences selected by at least two workers during annotation (Table 3.2 – column *2+3*); *Avg. occurr per talk* corresponds to the expected number of occurrences per TED talk (previous column divided by 180 talks); the column *New word per occurr* shows the expected number of unseen words if a new occurrence of the corresponding category is found[5]; and *New word per talk* shows how many new words are expected to be found if another talk is annotated.

The *New word per occurr* column highlights which categories are less stable in terms of annotation, by showing how many unseen words are expected to be found in a new occurrence of the phenomenon. Similarly to what happened in the corresponding type-token curves, the categories RECAPITULATING (1.18 new words), DELIMITING TOPIC (1.14), and POSTPONING TOPIC (1.10) are the ones that have the poorest representativity after the annotation of 180 talks. These results are directly related to the number of occurrences of these categories (which were the lowest of all annotations). On the other hand, the categories DEFENDING IDEA (0.11), CLARIFYING (0.12), DEFINING (0.16), and REPAIR & REFORMULATING (0.18) seem to be more stabilized (a new word associated with the concept is only expected to be found if additional 5-10 occurrences are discovered).

[5] Computed based on the difference between the number of words in the last occurrence and the number of words in the $(last - 10)^{th}$ occurrence. The value reported is the average of a 10-fold strategy, each randomizing the order of the annotations.



(a) Defining

(b) Introducing Topic

(c) Postponing Topic

(d) Emphasizing

Figure 3.8: Distribution of selected-words rate between annotators for DEF, INTRO, POST, and EMPH.

Given the annotation agreement observed, several approaches were explored to filter out unwanted answers. The first hypothesis was that some annotators' work could be discarded based on the number of words they clicked on during the tasks. Figure 3.8 shows how annotators behaved for four of the metadiscursive categories, plotting number of workers (y-axis) against the percentage of words clicked on (x-axis). Two different clicking behaviors arise: for Defining and Postponing Topic the vast majority of workers selected only a few words ($< 0.02\%$); for Introducing Topic and Emphasizing, on the other hand, the distribution of workers is more spread out across the clicking rate dimension.

While, intuitively, discarding answers from workers at both extremes of the scale can have a positive impact on quality, this criterion did not improve agreement. Most workers who never selected any words had a perfect agreement.

These were workers who came into the task, did one HIT that did not have any occurrences (which is highly probable), abandoning the task afterward. If the remaining two workers annotating the same segment confirm the absence of occurrences, the first worker has a

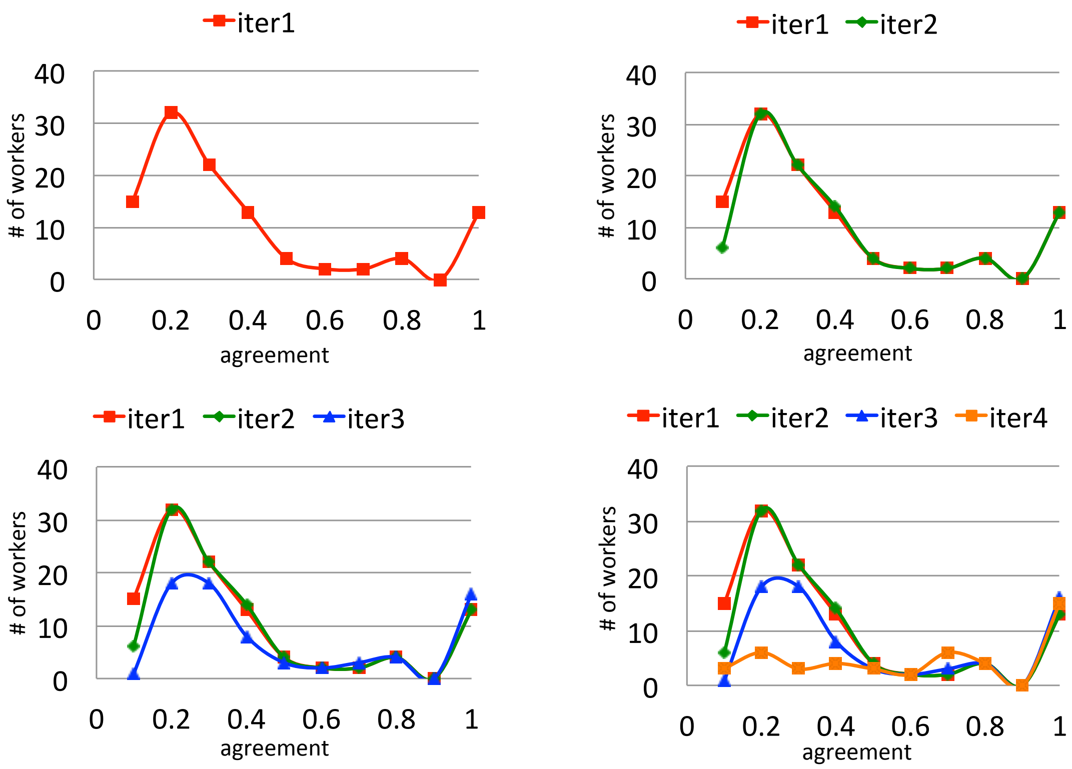


Figure 3.9: Distribution of agreement between annotators, for four iterations of the filter strategy for category `ADD`.

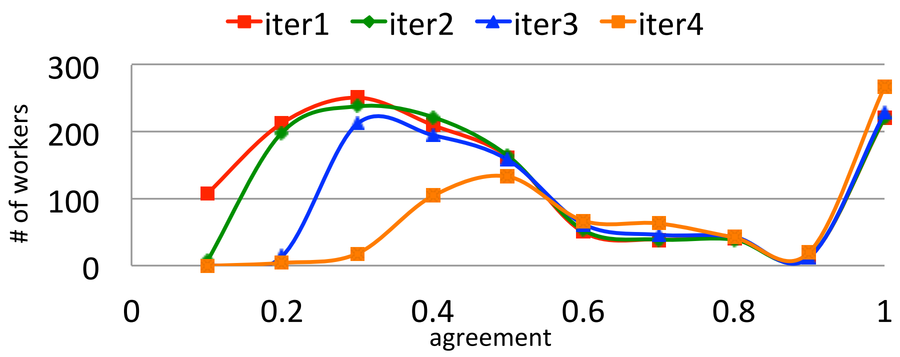


Figure 3.10: Distribution of agreement between annotators, for iteration 4 of the filter strategy considering all 16 categories.

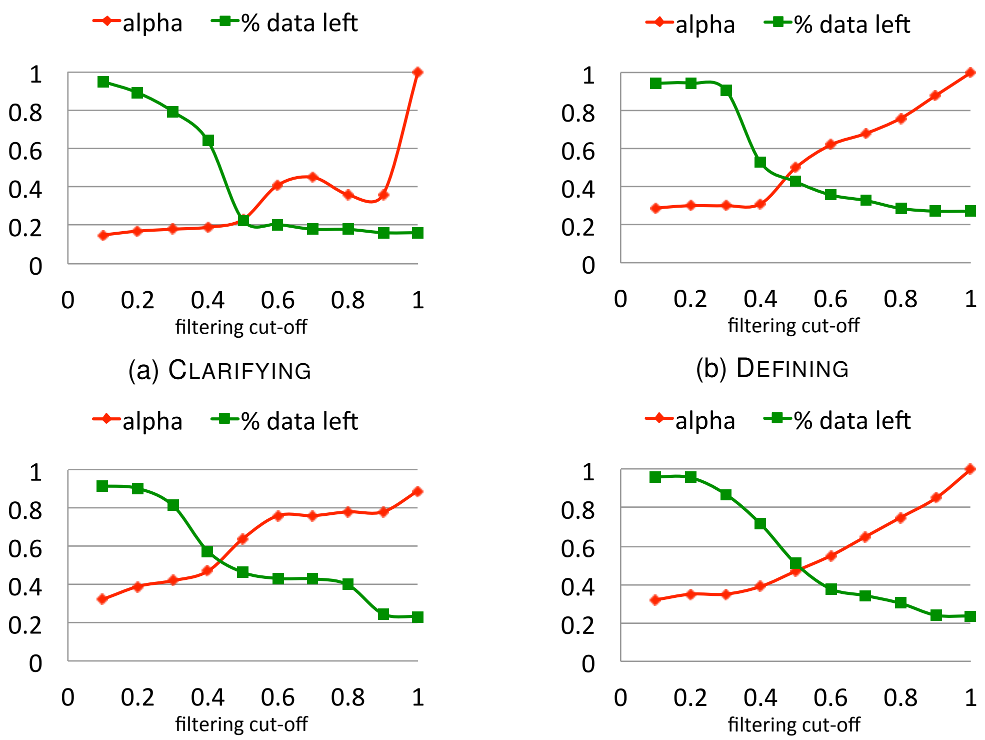
perfect agreement of 1. On the other end of the spectrum, there were only a handful of outliers. By looking closely at their responses, it was possible to observe they were not selecting new occurrences. Instead, these were workers selecting longer passages. As a reminder, two workers are considered to agree if the intersection of their answers is not empty.

Therefore, selecting a longer passage of the same occurrence is not penalized. These observations show that workers performance cannot be judged by merely measuring clicking rate. Appendix C shows the distribution of selected-words rate between annotators for all 16 categories.

The second hypothesis for filtering out answers was to directly discard workers based on their agreement. This filtering was done in incremental steps of 0.1 in α: first discarding workers who had agreement below 0.1 and seeing how that affected overall agreement (iteration 1); then discarding workers who had less than 0.2 agreement, and so forth.

Figure 3.9 plots four iterations of this strategy for the category ADDING INFORMATION. For iteration 2 (green line), which discards work from annotators with $\alpha < 0.2$, there was no improvement in agreement for the remaining workers (red and green line are practically on top of each other). It is only in iteration four that an increase of agreement is seen, with a higher amount of workers performing at a α of 0.7.

The same procedure was done for all 16 categories, and all showed similar behavior (Appendix D). Figure 3.10 shows iteration 4 for the combination of all categories. As previously, all lines are mostly on top of each other, leading to the conclusion that removing workers with lower agreement does not improve the remaining workers' performance.



(a) CLARIFYING

(b) DEFINING

(c) POSTPONING TOPIC

(d) REFERRING TO PREVIOUS IDEA

Figure 3.11: Tradeoff between discarding work based on agreement and percentage of data loss for the categories CLAR, DEF, POST, and REFER.

Figure 3.11 shows the percentage of occurrences of metadiscourse that remain as the filtering strategy in Figure 3.9 is applied (green line), and the corresponding improvement in agreement (red line) for CLARIFYING, DEFINING, POSTPONING TOPIC, and REFERRING TO PREVIOUS IDEA. Appendix E contains the corresponding plots for all 16 categories.

For all four acts shown, overall agreement only improves significantly after iteration five (even later for DEFINING), which means discarding work already at reasonable agreement levels. Regarding the amount of discarded data, CLARIFYING and DEFINING show 40% loss of positive data points in one single step. For POSTPONING TOPIC and REFERRING TO PREVIOUS IDEA the loss of data is less drastic. However, bringing annotation to levels of agreement of 0.5 and above still implies losing about 60% of the examples. These results show that filtering out workers based on low agreement affects answers that contain occurrences of metadiscourse, instead of discarding the workers who skipped those instances while annotating.

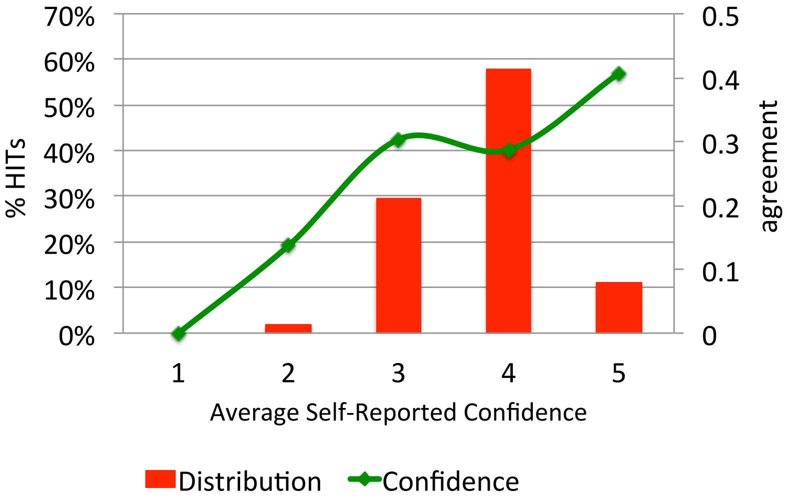


Figure 3.12: Positive examples for the task of identifying occurrences of EMPHASIZING.

The final hypothesis to filter out unwanted answers was that there might be some segments that are intrinsically difficult to annotate, for which workers have trouble identifying metadiscourse. Figure 3.12 shows how the workers' self-reported confidence relates to their agreement. The x-axis corresponds to the average confidence of a given HIT (average of the confidences reported by the three workers who solved that specific HIT). The y-axis on the left is related to the red bars and indicates the % of HITs at each confidence level, while the y-axis on the right is related with the green line, showing the agreement achieved on HITs at different levels of confidence.

The green line shows a significant difference in agreement between HITs with an average confidence of two and five ($\alpha = 0.11$ and $\alpha = 0.39$, respectively). However, the data distribution shows that these two values of confidence occur scarcely (2.23% for the confidence level of 2 and 9.68% for the confidence of 5). The larger bulk of the data was assigned to an average confidence level of 3 (31.39%) and of 4 (56.70%). For these two levels, however, no substantial difference in the inter-annotator agreement was found (variation of 0.01), with the higher level of confidence (4) counterintuitively registering a lower kappa than that of level 3.

3.3.3 Expert Validation

The variation between instances marked by one, two or all three workers (see Table 3.2) served as motivation to validate the data with experts, and thus have further insight on the annotations: how many of the cases selected by one worker only are indeed 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 if it corresponded to the category at hand or not. Experts validated a sample of 300 occurrences of each category (or the maximum number available for Concluding topic, Delimiting Topic, Postponing Topic, and Recapitulating, where total occurrences did not meet 300).

For occurrences marked by more than one worker (columns *2* and *3* in Table 3.2), experts were presented with the union of all workers answers, not knowing how many signaled them. They were also asked to focus on the existence or nonexistence of the function at hand, being permissive about the boundaries of the selection. Two experts revised each occurrence, and in case of disagreement, a third expert's opinion classified the instance.

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

Table 3.4: Results of the expert revision task in terms of agreement (α), occurrence number (#) and true positive rate (TP).

Table 3.4 shows, for each category, the total inter-annotator agreement achieved by the experts, the number of instances evaluated and corresponding true positive rate. 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), showing that the instructions for annotation of these categories, or the category itself, may not have been sufficiently clear. 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 CONCLUDING TOPIC (0.72).

The remaining columns on Table 3.4 show how experts evaluated the crowd's decisions. As previously, results are organized in terms of the 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 columns *2* and *3* should be 1 (experts validate all occurrences marked by at least two workers).

As expected, for most categories, there is a growing trend of TP rate concerning the number of workers in agreement, *i.e.*, the more workers who agree on a given occurrence, the more likely it is for experts to accept it. Exceptions are DEFINING and RECAPITULATING, 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 a perfect agreement ($TP = 1$) for the categories ADD and EXMPL.

For the cases that were selected by precisely two workers (column *2*), experts validate more than half the occurrences for 9 of the categories, with EXEMPLIFYING and REFERRING TO PREVIOUS IDEA reaching TP rates of above 80%. Below the 50% threshold are the categories ADD, CLAR, COM, DEF, POST, RECAP, 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 DELIMITING TOPIC, however, experts accepted more than half (51%) of the instances.

3.4 Discussion

This chapter described the building of METATED – a corpus of metadiscourse use in presentations annotated by the crowd. METATED is composed of 180 TED talks and 16 categories of metadiscourse, adapted from Ädel (2010). The corpus is freely available through LRE Map[6] in the form of 16 XML files, one per category, with all the metadata associated with the annotation (annotator ID, time-on-task, expansion information, self-reported confidence).

The final annotation results show that not all acts are understood by the crowd in the same manner, with agreement varying between 0.15 and 0.49. As a rare phenomenon, the probability of agreeing by chance on the same occurrence is low. This measure, taken into account for agreement, severely penalizes the case where one worker selects an occurrence and others do not. Previous annotation attempts on similar phenomena, such as Wilson's (2012) work on metalanguage, also show similar agreement values for sparser acts (0.09; 0.39), even when annotated by experts and considering only four categories.

When validating the crowd's work, experts behaved similarly regarding which categories had better and worst performance. They also confirmed that the amount of workers agreeing on a given instance is a good indicator of correctness.

The combination of the opinion of both workers on AMT and experts raises some reservations regarding some of the categories. Both non-experts and experts showed the lowest agreement for the categories CLARIFYING (non-experts=0.15; experts=0.28), RECAPITULATING (ne=0.18; e=0.18), and ADDING TO TOPIC (ne=0.15; e=0.40). Such values strongly hint that these categories and corresponding instructions were deficient, and the occurrences for these categories may not reflect the concept for which they were devised.

Additionally, the amount of data for DELIMITING TOPIC (40 occurrences if considering majority vote), CONCLUDING TOPIC (86), POSTPONING TOPIC (47), and RECAPITULATING (36) raises some concerns regarding the suitability of METATED for classification purposes, since they may lack sufficient examples required for training.

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

Given all the properties of the corpus just mentioned, METATED cannot be interpreted nor used as ground truth for metadiscourse. It should always be referred to as non-expert opinion on the phenomena. This idiosyncrasy affects the goal of the current thesis in the sense that it cannot be used in a traditional setting for machine learning. Not all occurrences in the corpus can be interpreted in the same manner, as they include different opinions on the same data points. The automatic classification proposed, besides exploring different features and algorithms, is now dependent on an additional variable: how to better take advantage of the data available, and combine the different opinions.

At this point, as a side experiment, METATED was used in an attempt to add to the understanding of metadiscourse in spoken language (Correia et al., 2015). Such experiment aimed at exploring if metadiscursive acts are used independently of vocabulary complexity and, if so, which ones are used more frequently in more lexically demanding talks. These questions were also raised in Crismore (1984), who stressed out that *"[r]esearchers need to ask about the optimum level of metadiscourse: How much of which type is needed by which students for which tasks under what conditions."*

Briefly put, from the data, it was possible to conclude that some but not all categories correlate with vocabulary level. More specifically, strategies of topic management (DELIMITING TOPIC, INTRODUCING TOPIC, POSTPONING TOPIC) and broadly used functions (EXEMPLIFYING, EMPHASIZING, ENUMERATING) occurred at the same rate in all levels, thus not correlating with lexical complexity. Functions related to paraphrasing were more frequent in higher level talks, but not necessarily in segments containing the highest level vocabulary. This shift in correlation's 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 (COMMENTING ON LINGUISTIC FORM/MEANING and DEFINING) seemed to appear in the context of the vocabulary they address.

Going back to the suitability of using METATED as training data for metadiscursive classifiers, Table 3.5 provides a high-level judgment of the quality of the corpus assembled, in terms of quantity of data and agreement. The first column demonstrates for which categories the rate

Category	rate new	$>$ 200	worker	expert
	words < 1	occurr.	$\alpha \geq 0.2$	$\alpha \geq 0.4$
R&R	✓	✓	✗	✓
COM	✓	✓	✓	✓
CLAR	✓	✓	✗	✗
DEF	✓	✓	✓	✓
INTRO	✓	✓	✓	✓
DELIM	✗	✗	✓	✓
ADD	✓	✗	✗	✗
CONC	✓	✗	✓	✓
ENUM	✓	✓	✓	✓
POST	✗	✗	✓	✓
RECAP	✗	✗	✗	✗
REFER	✓	✗	✓	✓
DEFND	✓	✓	✓	✓
EXMPL	✓	✓	✓	✓
ANT	✓	✓	✓	✓
EMPH	✓	✓	✗	✓

Table 3.5: metaTED high-level judgment by category, regarding the quantity of annotation and annotator agreement.

of new word discovery is more stabilized (less than one unseen word per new occurrence). The second column informs for which categories there are at least 200 occurrences where there was a consensus among non-experts. Ten metadiscursive acts fulfill this criterion, which serves as an indicator of the suitability of using the data in NLP-related tasks. The last two columns in Table 4.6 provide a representation of the reliability of the data in METATED by category. The categories Adding Information, Clarifying, and Recapitulating have severe consensus problems, for both the crowd and for the experts. On the other end of the spectrum are the categories Concluding topic and Enumerating, where the agreement was the highest for both non-experts and experts.

From Table 3.5 it is possible to extract a set of problematic categories (highlighted in red):

- Clarifying – shows signs of not being understood by both the crowd and the experts (lowest levels of agreement in both tasks);
- Delimiting Topic – a low amount of occurrences raises concerns about representa-

tivity. Additionally, the analysis of the type-token curves shows that the expected rate of unseen words when finding a new occurrence is above one, further hinting at representativity issues (contrary to CONCLUDING TOPIC and REFERRING TO PREVIOUS IDEA which seem to be more stable);

- ADDING TO TOPIC – similar to CLARIFYING, showing signs of not being understood by both the crowd and the experts;
- POSTPONING TOPIC – same phenomenon as DELIMITING TOPIC;
- RECAPITULATING – this category triggers none of the positive indicators summarized by the table.

For the current thesis purposes, given reasons mentioned above, it was decided not to proceed with classification attempts for these five categories.

4 Automatically Classifying Metadiscourse

Chapter 2 presented the state of the art Natural Language Processing (NLP) studies that dealt with phenomena related to metadiscourse. Wilson (2012) described a data collection task where experts identified and classified occurrences of *metalanguage* in *Wikipedia* articles. Other studies focused used student essays to analyze argumentative cues (Nguyen and Litman, 2016; Desilia et al., 2017). Madnani et al. (2012) in particular, focused on the identification of *shell text* in argumentative essays.

As mentioned before, all these studies target metadiscourse as used in written discourse exclusively. Madnani et al. (2012) actually perform a comparison between both varieties, applying strategies that were initially developed to identify *shell language* in students' essays to a corpus of political debates. However, this study only analyzes the results of this domain adaptation qualitatively, since the corpus of political debates was not previously labeled with occurrences of *shell language* itself.

Although one can refer to the work focusing on written form for performance comparison reasons, it has limited applicability concerning the goals established in the introductory chapter. In written language, typically, the author does not have to deal with the immediacy of production or feedback (an exception of this would be an *online chatting* situation). *Wikipedia* articles or student essays do not contain occurrences of repairs, or communication channel management strategies, frequently found in spoken discourse.

Consequently, this chapter aims at filling this already identified gap: addressing metadiscourse in spoken language in an automatic manner. More precisely, it presents a supervised learning setup that takes advantage of the material collected via crowd (Chapter 3) to identify passages of metadiscourse and classify them under eleven functional categories: Repair &

REFORMULATING, COMMENTING ON LINGUISTIC FORM/MEANING, DEFINING, INTRODUCING TOPIC, CONCLUDING TOPIC, ENUMERATING, REFERRING TO PREVIOUS IDEA, DEFENDING IDEA, EXEMPLIFYING, ANTICIPATING THE AUDIENCE'S RESPONSE, and EMPHASIZING.

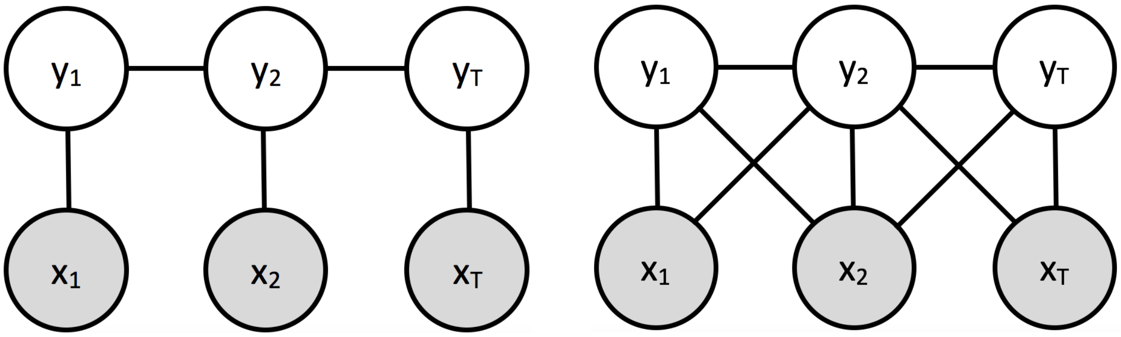
Sections are organized in the following manner:

- **Section 4.1** describes a first classification experiment that constitutes a baseline for the remaining analysis. It discusses the main problems faced when trying to classify metadiscourse, including feature engineering, data balancing considerations, and the search space size;
- **Section 4.2** presents a *divide and conquer* approach to the problem: first considering the problem of identifying which sentences in a talk contain metadiscourse; and secondly, determining the exact words in the candidate sentences that are actually the realization of the metadiscursive acts. This solution encompasses a chain of 2 classifiers, the first acting as a filter to the next phase;
- **Section 4.3** concludes by giving further insight on the relation between the two layers of classification, discussing performance and lessons learned.

4.1 Preliminary Experiments

The first step was to set a baseline system by directly addressing the problem of detecting which words in a given talk encompass each one of the metadiscursive acts.

The approach followed to accomplish this is similar to most common Part-of-Speech (POS) tagging or Named Entity Recognition (NER) tasks: given a set of labels (tags) and observations, learn a model that successfully predicts the act of metadiscourse at hand. A way to formulate this scenario is to consider that a hidden process is generating the observables. The hidden process can be modeled with features that are thought to be relevant for the problem. Such formulation constitutes an undirected graphical model shaped as a linear chain.



(a) (b)

Figure 4.1: Graphical structures of two CRF setups.

A well-known realization of this setup are Conditional Random Fields (CRFs) (Lafferty et al., 2001): a probabilistic model that realizes the segmentation and labeling of sequential data. Figure 4.1 shows two different realizations of a Conditional Random Field (CRF) model: the white circles represent the labels, $y_{1...T}$, while the grey circles represent the observations, $x_{1...T}$. On the left, Figure 4.1a represents a linear chain CRF where labels are linked to observations at the same time step only, *i.e.*, when it is useful to assume that a given label, y_i, depends only on both previous and next labels, y_{i-1} and y_{i+1}, and the current observation x_i. On the other hand, in the model represented by Figure 4.1b, labels are additionally dependent on the observations at the previous and next time steps, x_{i-1} and x_{i+1}.

Both configurations above represent first-order Markov chains since each label depends on the previous and next label only. Higher order chains are also allowed, with each label dependent on a fixed number, n, of previous labels. However, the computational cost of such solutions increases exponentially with n, rendering the problem intractable for large values of n.

A significant advantage of Conditional Random Fields (CRFs) is how easy the adding of new features is. Linking an observation at a given time, t, with labels at a fixed surrounding length is a straightforward process (merely including that information in each of the items).

In more detail, CRFs operate according to the following definitions:

- **Training –** given set of sample observations $\{x_1, ..., x_T\}$ and values for their labels $\{y_1, ..., y_T\}$, model the conditional distribution $P(Y|X)$, subject to:

  $$P(y_i|X, y_j, i \neq j) = P(y_i|X, y_j, i \sim j) \tag{4.1}$$

  where $i \sim j$ stands for i and j being neighbours;

- **Inference –** given a new observation x, find the most likely set of labels y^* for x, *i.e.* compute:

  $$y^* = \arg\max_y P(y|x) \tag{4.2}$$

4.1.1 Experimental Setup

The CRF implementation used to formulate the problem in this experiment follows the work of Okazaki (2007): setting up 1st-order Markov CRFs with state and transition features. State features are combinations of features and labels, and transition features consist of label bigrams.

In this experiment, an independent classifier is trained for each of the eleven categories at hand.

As done in Chapter 3, the way that workers' opinion is combined is through majority vote. This means that a sentence is considered to have metadiscourse if, and only if, at least two workers marked it as such. This also holds true at the token level: if two workers select a different set of words for the act, only the intersection of their answers is considered. For example, if *worker-1* marks *"Today, let me begin with a"*, and *worker-2* selects *"let me begin"*, the tokens associated with an instance of INTRODUCING TOPIC will be `let`, `me`, and `begin`.

This experimental setup uses syntactic and lexical features to support the classification of metadiscourse. Such lexical approach is based on the fact that, in the annotation task described in the previous chapter, non-experts were able to agree on occurrences of metadiscourse while having only access to the subtitles of the TED talks. The practices found in the literature on metadiscourse related-phenomena (discussed in Chapter 2) also sustain the decision of using syntactic and lexical features to support classification.

Similar lexical approaches appear in different research areas such as word sense disambiguation (Pedersen, 2001), sentiment analysis (Pang et al., 2002; Abbasi et al., 2008), or feedback localization (Xiong and Litman, 2010). The prevailing idea behind all these studies is that words can be indicators of the presence of the phenomenon at hand. For instance, in sentiment analysis, some words are associated with positive opinions, others are neutral, while others have negative connotations. The features used in this experiment are:

- **Part-Of-Speech *n*-grams –** presence of POS *n*-grams. The categories considered are the set of 36 POS tags[1] provided by the Stanford Parser (Klein and Manning, 2003);

- **Lemma *n*-grams –** presence of word lemma *n*-grams in the sentence (inflected forms, such as plural and singular, collapsed in a single item, also extracted from the Stanford Parser's output);

- **Word *n*-grams –** presence of word *n*-grams as they are in the transcript (inflected forms of nouns and verbs, *etc*).

[1] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (June 2013)

For the three sets of features listed above, both unigrams and bigrams were considered. Additionally, at first, both settings with and without stop-words[2] were considered. Stop-word removal is based on the premise that stop-words *"have no meaning"* (Osinski and Weiss, 2005). This strategy is commonly used to decrease model size, filtering data for uninformative words, and therefore improving general performance in areas such as document indexing and retrieval, copy detection and topic modeling (Shivakumar and Garcia-Molina, 1996; Silva and Ribeiro, 2003; Osinski et al., 2004; Wang and McCallum, 2006).

However, considering stop-words has been proven to be successful in areas such as sentiment analysis and word sense disambiguation (Lee and Ng, 2002; Paltoglou and Thelwall, 2010; Maas et al., 2011). In this case, stop-words rank higher in the language models of each metadiscourse act when compared to the language model of the TED talks (Appendix A.2). After a small experiment, the setting considering stop-words outperformed the solution with stop-word removal. For that reason, all further experiments consider stop-words.

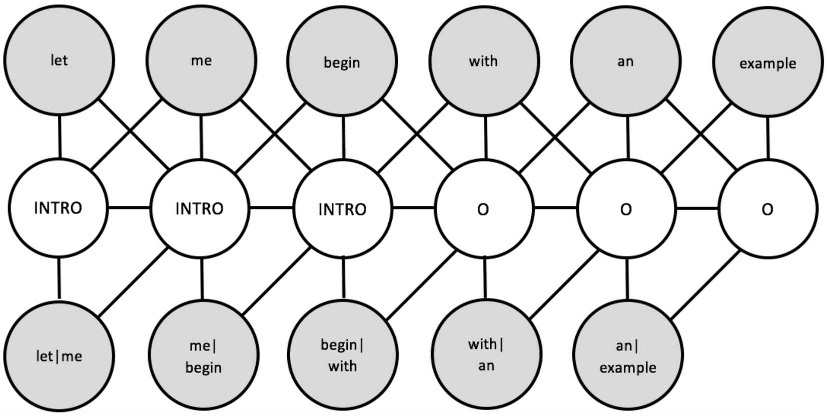


Figure 4.2: Graphical structure for an example of INTRODUCING TOPIC.

Figure 4.2 exemplifies how the model represents an instance of INTRODUCING TOPIC. The white circles in the figure represent the labels – the goal of the classification process –, while the grey circles represent observations – what the model can see and use to classify a given word as being associated with the metadiscursive act at hand. In this model, the labels are simply `INTRO`, which accounts for the current token being part of the act, and `O`, which represents the fact that there is no metadiscourse associated.

[2] http://www.ranks.nl/resources/stopwords.html (June 2013)

For simplification purposes, the figure only presents word unigrams and bigrams. A given label is associated with the previous, current, and next word unigram, and also with the previous and current word bigrams. Again, this also holds true for both lemmas and POS. In the figure, transition features are represented by the lines between two consecutive labels.

```
INTRO  w[0]=let w[1]=me w[0]|w[1]=let|me
INTRO  w[-1]=let w[0]=me w[1]=begin w[-1]|w[0]=let|me w[0]|w[1]=me|begin
INTRO  w[-1]=me w[0]=begin w[1]=with w[-1]|w[0]=me|begin w[0]|w[1]=begin|with
INTRO  w[-1]=begin w[0]=with w[1]=an w[-1]|w[0]=begin|with w[0]|w[1]=with|an
O      w[-1]=with w[0]=an w[1]=example w[-1]|w[0]=with|an w[0]|w[1]=an|example
O      w[-1]=an w[0]=example w[-1]|w[0]=an|example
```

Figure 4.3: Example of feature representation for Introducing Topic with word unigrams and bigrams with a window of one.

Figure 4.3 shows how data is represented in the classifier for the snippet *"let me begin with an example"*. Again, to simplify, only word unigrams and bigrams are present. Each feature is represented by its type (in the figure, `w` stands for *word*), followed by the index to which it refers to (`[-1]`, `[0]`, and `[1]` refer to the previous, current and next item), followed by its value after the = sign. Bigrams are represented by the separation character "|".

In sum, the model will learn the weights between attributes and labels (e.g., if the current item has an attribute `w[0]=begin`, it is likely to have the label `INTRO`). In this approach, inference is made through gradient descent using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method (Liu and Nocedal, 1989), and no cut-off was applied, *i.e.*, it considers all features regardless of their frequency.

To keep the training data manageable, only positive features were considered. This approach contrasts with representing each observation with all features in the model along with the information if the feature is observed at each moment or not. Such strategy, and if considering only bigrams, requires each observation to be described by a set of V^2 features (V being the size of the vocabulary), growth order that quickly becomes unmanageable. Therefore, herein the model is only able to learn from the features that are active at a given point, and not from the absence of specific features.

4.1.2 Experimental Results

The first set of experimental results reported share a common trait: the training data. More precisely, as a first strategy, the training set contains only sentences marked with metadiscourse.

It is important to point out that, since the goal of the classifiers is to identify which words in a sentence are part of a given metadiscursive act, this strategy still provides both positive and negative observations. For instance, in the sentence used above, *"let me begin with an example"*, the words `let`, `me`, and `begin` correspond to positives instances of metadiscourse, while the words `with`, `an`, and `example` correspond to negative observations for the category INTRODUCING TOPIC.

The choice to include, at this point, only sentences that contain metadiscourse in the training set is a naïve strategy to address the problem of sparsity of the phenomenon of metadiscourse. Ignoring the sentences that do not contain metadiscourse while training, balances out the number of positive and negative cases (even if artificially). If the training set contained all sentences (positives and negatives), with no further adjustments or considerations, the algorithm would naturally learn to classify all observations as negative (since they are about 99% of the total number of cases).

With a fixed set of training data, two different experiments were conducted based on the variations of the test set: in the first experiment, the test set is also only comprised of sentences that were signaled to have metadiscourse; on the second experiment, the test set is comprised of all sentences in the corpus. It is clear that the first configuration corresponds to an artificial scenario (where sentences were already known to have metadiscourse or not), while the second scenario encapsulates the ultimate goal – given a talk, identify which words are metadiscursive, and which functions they serve.

$$prec = \frac{TP}{TP + FP} \qquad rec = \frac{TP}{TP + FN} \qquad F1 = \frac{2 \cdot prec \cdot rec}{prec + rec} \tag{4.3}$$

Category	Chance	Test Meta Only			Test All		
		prec	rec	F1	prec	rec	F1
EXMPL	0.0028	0.86	0.69	0.76	0.51	0.69	0.58
DEFND	0.0036	0.83	0.60	0.69	0.26	0.60	0.36
INTRO	0.0055	0.81	0.71	0.76	0.22	0.71	0.33
ENUM	0.0064	0.75	0.54	0.62	0.21	0.54	0.30
COM	0.0028	0.74	0.29	0.42	0.25	0.29	0.27
R&R	0.0016	0.87	0.61	0.71	0.16	0.61	0.25
ANT	0.0051	0.72	0.51	0.59	0.09	0.51	0.16
EMPH	0.0055	0.71	0.51	0.59	0.07	0.51	0.12
DEF	0.0015	0.69	0.34	0.44	0.04	0.34	0.08
REFER	0.0013	0.80	0.53	0.63	0.04	0.53	0.08
CONC	0.0011	0.76	0.44	0.53	0.04	0.44	0.07

Table 4.1: Classification results with only positives vs. all data as test set.

Table 4.1 presents the classification results for the aforementioned settings. The first column shows the probability of correctly classifying an item by chance. In other words, it represents the number of words associated with metadiscourse in each category divided by the total number of words in the corpus. For all categories this probability is below 1%, achieving a maximum of 0.64% for ENUMERATING and a minimum of 0.11% for CONCLUDING TOPIC. These numbers again show the sparsity of the phenomenon and the impact that can have on the task of classification.

The three columns in the middle of the table, under *Test Meta Only*, show precision, recall, and F1 measure (Equation 4.3) for the setup where both train and test sets are composed of sentences that contained metadiscourse only. The categories for which the classification performed best were INTRODUCING TOPIC and EXEMPLIFYING, with $F1 = 0.76$. These two categories were also the ones with the highest recall rates (approximately 0.7). On the other hand, COMMENTING ON LINGUISTIC FORM/MEANING and DEFINING were the categories with the lowest performance concerning both F1 measure (≈ 0.4) and recall (≈ 0.3).

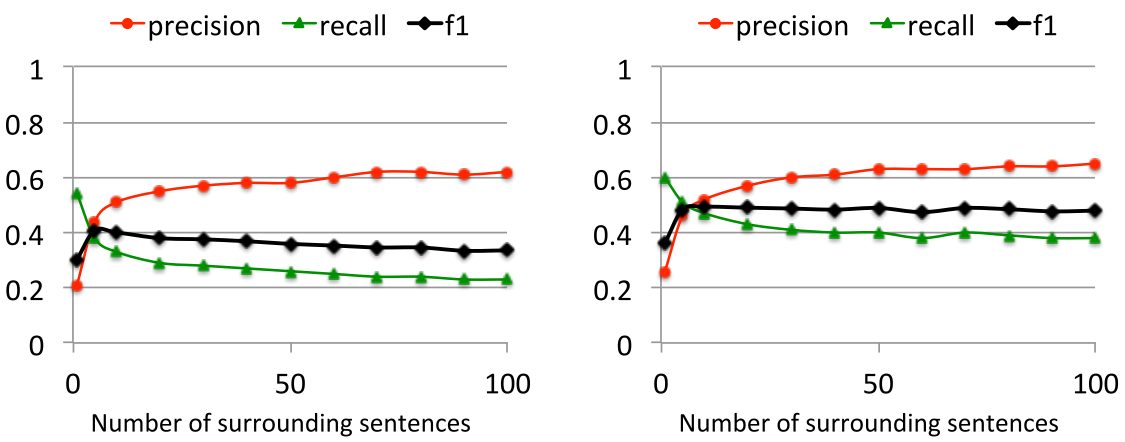
As expected, the values for precision of all the classifiers are high ($0.7 \leq prec \leq 0.9$). By removing all sentences with no metadiscourse from the test set, the classifiers are less prone to produce False Positives (FP), which is one of the factors that impact precision (as can be seen in Equation 4.3).

However, as mentioned previously, this setup does not correspond to a real-world setting, where the system should process the entire text of a talk. The results for such configuration are therefore shown in the last three columns of Table 4.1, under *Test Meta Only*, also showing precision, recall, and F1. The first observation that is possible to draw is that, as expected, recall remains unchanged. It is important to remember that the only difference between the two setups is that the latter has more negative examples, *i.e.*, all sentences that do not contain any traces of metadiscourse. Naturally, testing the classifier with more negative examples does not increase the number of True Positives (TP) nor False Negatives (FN), which are the two factors taken into account to compute recall. Therefore, recall values are the same in both setups.

Contrastingly, values for precision are radically different. Since the training data is composed of sentences that contain metadiscourse only, the real distribution of positives and negatives is skewed. Therefore, when testing the classifiers in data with the real distribution of occurrences, they tend to classify many more instances (mimicking the rates observed during training). Furthermore, merely removing negative examples from the training also contributes to the underrepresentation of the negative class. As a result of these implications (mimic the distribution in the train data and underrepresentation of negative cases), the number of False Positives go up, affecting precision. Consequently, this drop in precision has a significant impact on F1, which in this real setting goes as low as 0.07 for CONCLUDING TOPIC and achieves its maximum for EXEMPLIFYING at 0.58.

Again, these results corroborate the already mentioned problem of unbalanced data. Classifiers trained with the wrong trade-off between positive and negative occurrences cause the under-sampling of negative examples, hindering the representation of the phenomena at hand.

To better understand the problem of training with unbalanced data, a follow-up experiment was carried out. Here, more negative examples were added gradually surrounding each positive example, in order to provide a better representation of the absence of metadiscourse, *i.e.*, negative cases.



(a) Enumerating

(b) Defending Idea

Figure 4.4: Precision, recall and F1 values as more negative examples are provided in the training phase (for `ENUM` and `DEFND`).

Figure 4.4 shows the tradeoff between precision, recall, and F1 for the categories Enumerating and Defending Idea (Appendix F shows the same information for the remainder of the categories). Each graph plots the values of precision, recall, and F1 (y-axis) as more negative examples are added to the training (x-axis).

The leftmost three points in each plot (green, black, and red) correspond to the information on Table 4.1, *i.e.*, the setting where only the sentences that contains metadiscourse is used for training. For Enumerating, this corresponds to $prec = 0.21$, $rec = 0.54$ and $F1 = 0.30$; and for Defending Idea, $prec = 0.26$, $rec = 0.60$ and $F1 = 0.36$.

The second set of points represent the results of classification when using a window of 5 sentences surrounding the positive example in the corresponding talk. These additional sentences naturally introduce more negative examples. The remaining sets of points were obtained in ten-sentence steps, ranging from ten surrounding sentences to one hundred.

For the two categories shown above, it is possible to see that, as more negative examples are added, precision goes up at the cost of recall. This relation reflects correct labeling of more negative examples, at the cost of leaving out positive occurrences of metadiscourse. For Enumerating, this tradeoff penalizes F1 as more examples are added, while for Defending Idea, F1 stays constant.

The F1 peak for both categories is when there is a window of five sentences included in the training. This maximum also holds true DEFINING, INTRODUCING TOPIC, EMPHASIZING, and EXEMPLIFYING. For the remaining categories, the behavior is different. COMMENTING ON LINGUISTIC FORM/MEANING, for example, has its best configuration with a window of zero, while REFERRING TO PREVIOUS IDEA peaks when using a window of 90 sentences. This latter case, however, greatly compromises recall ($window = 0, rec = 0.53$ vs. $window = 90, rec = 0.19$).

In sum, and as expected, as the train data matches the distribution of metadiscourse in the test set, precision goes up, *i.e.*, the classifier has more information to learn what is not to be considered metadiscourse. However, it misses more instances, hence recall goes down, since it is not able to learn positive examples as well as before.

From these experiments, it is possible to draw two conclusions:

- firstly, CRFs seem to perform well in the task of identifying which words in a sentence encompass metadiscursive information, given that said sentence is already known to contain an instance of the phenomena. This conclusion follows from the results shown in the middle columns of Table 4.1, where training and test data were composed of sentences with metadiscursive acts only;
- on the other hand, the unbalance in the training data proved to be a problem. So far the approach was by trial and error, varying the tradeoff between positive and negative instances and looking at what each configuration produces regarding precision, recall, and F1. However, a more robust solution is desirable: one that allows parameters to be tuned to deal with the fact that metadiscourse is a sparse phenomenon and, naturally, there are many more negative examples of it in a talk transcript than positives.

4.2 Classification Chain

The two observations drawn from the preliminary experiment with the classification of metadiscourse guided the decision process to the subsequent phases of the development. This section presents the solution to the problem of classification taking into consideration the lessons learned in past experiments, and building up from their results.

At this point, the decision was to approach the classification of metadiscourse in a more structured manner. Instead of developing a single piece of software that is responsible for all the classification decisions, the objective is to divide the problem into smaller and more manageable tasks that allow allow for more of control.

Knowing that CRFs perform well on the task of identifying which words in a sentence are metadiscursive when a sentence is already expected to have an instance of the phenomenon, there is a need to be able to spot which sentences in a talk contain metadiscourse.

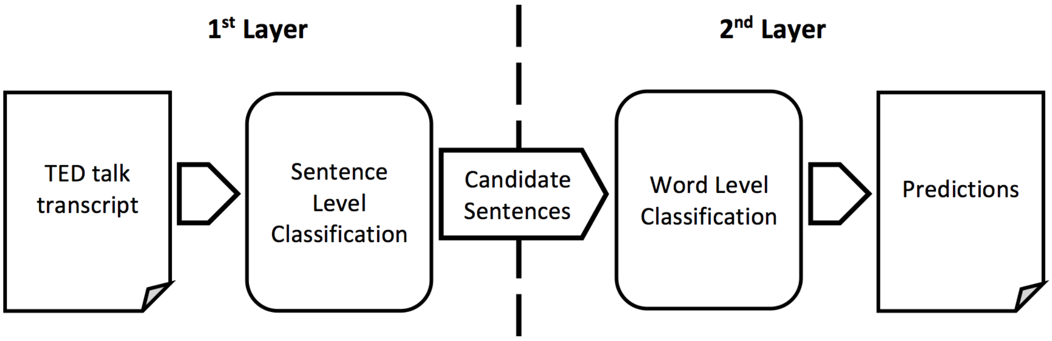


Figure 4.5: Proposed classification chain.

Figure 4.5 presents the proposed solution under a two-level classification chain:

- **Sentence-level classification –** the first layer of the classification chain is responsible for selecting a set of candidate sentences that are expected to contain metadiscourse. Trained on the crowd annotations, given a new talk transcript, it predicts which sentences have an instance os the metadiscursive act at hand. This layer acts as a filter, dealing with the sparsity of the phenomenon and the problem of unbalanced data;

- **Word-level classification –** the second layer, on the other hand, takes as input the candidate sentences that passed the previous step, and predicts, within those sentences, which words materialize the metadiscursive function. This layer performs a similar task to the one in the preliminary experiments, in the sense that it assumes that all sentences submitted to the classifier contain an instance of metadiscourse.

Even though discussed in detail for the remainder of this chapter, it is important to highlight here that the two levels of classifiers are trained on very different sets. At the sentence level, the training data is composed of all data collected via crowdsourcing (*i.e.*, the full TED talk transcripts), while at the word level, the training set comprises only sentences that were marked as metadiscursive by the crowd.

Such formulation allows one to investigate more closely the phenomenon of metadiscourse, focusing on the problem at different levels and facilitating the development of strategies that address specifically the idiosyncrasies of classification in different stages.

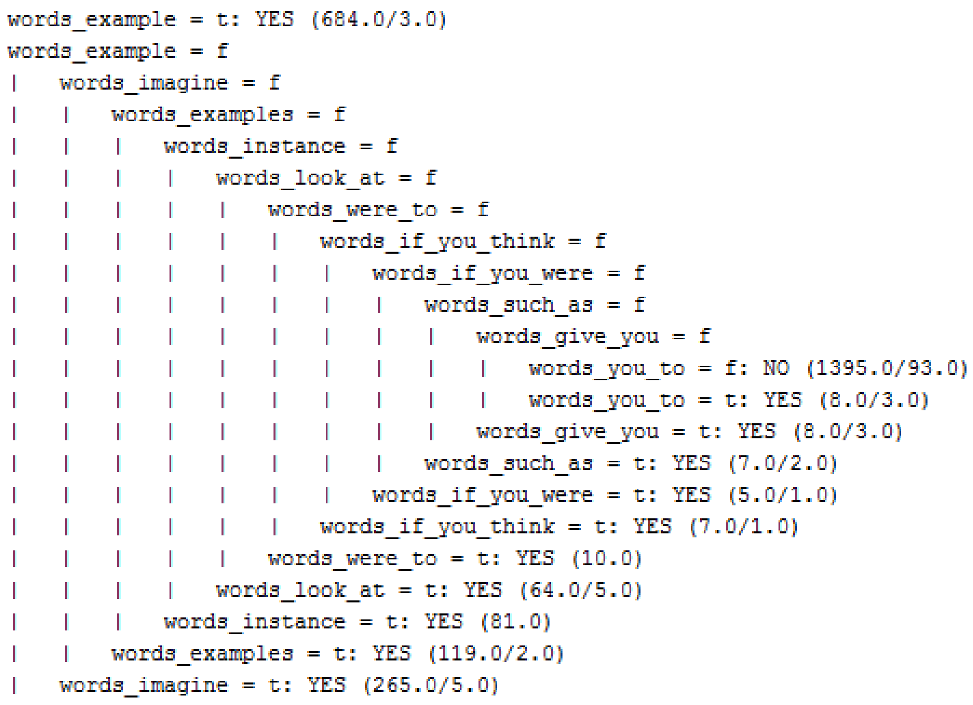
4.2.1 Sentence Level Classification

The goal of the first layer of the classification task can be stated as follows: given a sentence, decide if it has an occurrence of a given metadiscursive act or not. As before, the solution assumes the form of one classifier per function. In other words, each sentence is submitted to different classifiers, each one outputting a binary decision of whether the sentence contains an occurrence of a given metadiscursive act or not.

At this level, the classification units are the sentences. This configuration reduces the number of observations when compared to the preliminary experiments (which tried to classify individual words). With fewer items to train on, it is possible to explore each one at a deeper level, expanding the number of features without jeopardizing training efficiency. Additionally, since sentences are now the classification target, it is possible to include features that are related to sentences, such as its length and position in the talk.

As a proof of concept, this problem was first addressed for a subset of four categories, namely INTRODUCING TOPIC, CONCLUDING TOPIC, EXEMPLIFYING, and EMPHASIZING, which correspond to the first annotation effort described in Section 3.2. This specific experiment used decision trees combined with lexical and syntactic features (n-grams of words and POS tags). By using this particular setup, the goal was to investigate the feasibility of the classification by analyzing the generated output, which for decision trees is a set of rules that is intelligible and can be directly interpreted.

Four test sets of sentences were built (one per category), following a majority vote from the results collected from the crowd. Given the significant disparity between positive and negative cases and the algorithm sensitivity to such setup, data was balanced for each category by randomly choosing negative cases to match the proportion of positive cases. A grid search over type of feature (lemma, word, POS), and n-gram order, achieved classification accuracies of 94.92% for EXEMPLIFYING, 92.71% for INTRODUCING TOPIC, 86.93% for CONCLUDING TOPIC, and 79.77% for EMPHASIZING.



```
words_example = t: YES (684.0/3.0)
words_example = f
|   words_imagine = f
|   |   words_examples = f
|   |   |   words_instance = f
|   |   |   |   words_look_at = f
|   |   |   |   |   words_were_to = f
|   |   |   |   |   |   words_if_you_think = f
|   |   |   |   |   |   |   words_if_you_were = f
|   |   |   |   |   |   |   |   words_such_as = f
|   |   |   |   |   |   |   |   |   words_give_you = f
|   |   |   |   |   |   |   |   |   |   words_you_to = f: NO (1395.0/93.0)
|   |   |   |   |   |   |   |   |   |   words_you_to = t: YES (8.0/3.0)
|   |   |   |   |   |   |   |   |   words_give_you = t: YES (8.0/3.0)
|   |   |   |   |   |   |   |   words_such_as = t: YES (7.0/2.0)
|   |   |   |   |   |   |   words_if_you_were = t: YES (5.0/1.0)
|   |   |   |   |   |   words_if_you_think = t: YES (7.0/1.0)
|   |   |   |   |   words_were_to = t: YES (10.0)
|   |   |   |   words_look_at = t: YES (64.0/5.0)
|   |   |   words_instance = t: YES (81.0)
|   |   words_examples = t: YES (119.0/2.0)
|   words_imagine = t: YES (265.0/5.0)
```

Figure 4.6: Decision tree for the category EXEMPLIFYING using the word bigram model.

Figure 4.6 shows the tree with best performance for the category EMPHASIZING (word bigrams model), composed of only 19 rules (12 leaves). The features *example*, *imagine*, *examples* and *instance* alone account for the correct classification of about 87% of the instances in the corpus.

The high perfomance achieved with this simplistic solution corroborate the hypothesis of looking at metadiscourse at the lexical level. Words alone proved to carry a substantial amount of the information in these metadiscursive items.

Despite the high performance achieved using decision trees to address the sentence level classification, this strategy has several drawbacks:

- Data Imbalance – in this experiment, data was balanced out by randomly choosing some of the negative instances available, given the algorithm's sensitivity to imbalanced data. This formulation made the problem simpler, by making the probability of correctly classifying a sentence by chance 50%. As seen in previous chapters, this is not the case when addressing metadiscourse and is, in fact, one of the major problems that automatic classification faces;
- Unrobustness – decision trees are also known for not being robust upon small variations of data and experiment setup. In fact, in the experiments carried out at this stage, different parameterizations of the trees during training produced drastically different configurations of branches and leaves. Such observations motivate the investigation of options that can generalize better for a greater number of features and conditions;
- Feature Space – finally, feature expansion while using decision trees often lead to overfitting. More specifically, decision trees are more suitable to categorical features making it harder to explore real timed variants, such as word frequency.

Motivated by the performance obtained by this solution and to address the three obstacles pointed out above, a second phase of experiments was carried out, now considering the full set of metadiscursive acts.

In it, linear Support Vector Machines (SVM) were trained with LIBLINEAR[3] (Fan et al., 2008). This decision is based on the premise that, given a large number of both instances and features, a linear classifier gives similar performances to nonlinear solutions, while allowing for faster training (Yuan et al., 2012).

SVMs, in general, and the implementation used in this work in particular, have the advantage of allowing for direct control of the misclassification cost. This property facilitates addressing the issue of imbalanced data, and consequently reduce the impact of under-sampling found during the preliminary experiments.

To fully understand the cost mechanism of SVMs, it is important to look into their definition. In detail, a training set consists of pairs $(\mathbf{x}_i, y_i), i = 1, \ldots, l,$ where $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th training vector and $y_i \in +1, -1$ is the class label of the instance.

During training, SVMs try to minimize:

$$\begin{gathered} f(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \\ \text{subject to } \forall i : y_i(\mathbf{w}^T x_i - b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \end{gathered} \tag{4.4}$$

- where $\mathbf{w}$ is the normal vector to the hyperplane;
- ***b***, the parameter which determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$;
- and ξ_i, a slack variable which predicts misclassification.

The last parameter in Equation 4.4, ***C***, is precisely the cost of misclassification of an item. It follows directly from Equation 4.4 that the higher the value of the cost parameter, the greater the impact of a misclassification. It is this property of Support Vector Machines (SVM) that allows the implementation of a cost-based learning strategy to address the imbalance of the training data.

[3] https://www.csie.ntu.edu.tw/∼cjlin/liblinear/

The first step is to unfold the training error part of Equation 4.4, ($C\sum_i \xi_i$) in two factors:

$$f(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i:y_i=+1} \xi_i + C \sum_{i:y_i=-1} \xi_i \tag{4.5}$$

Equation 4.5 now distinguishes between positive training samples ($i : y_i = +1$) and negative ones ($i : y_i = -1$). For the first parcel, when $i : y_i = +1$, the algorithm is facing a false negative, while in $i : y_i = -1$ it faces a false positive. Given the fact that there are plenty more negative examples in the training data, the error that impacts performance the most is the occurrence of a false negative, *i.e.*, cases where there were instances of metadiscourse that the classifier was unable to capture.

From Equation 4.5, it is now possible to add two weighting constants, j_{pos} and j_{neg}, that perform the role of adjusting the cost of misclassification at different rates depending on what type of error the classification (and the learning process in particular) is incurring in:

$$f(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||^2 + j_{pos}C \sum_{i:y_i=+1} \xi_i + j_{neg}C \sum_{i:y_i=-1} \xi_i \tag{4.6}$$

Following the strategy suggested in He and Garcia (2009), the cost of misclassification for each SVM is then weighted according to the proportion of positive and negative cases in the corpus. Therefore, the cost of missing a positive example, j_{pos}, is set to be the proportion of negative cases (which is high), while the cost of missing a negative example, j_{neg}, is the proportion of positive cases (lower), with $j_{pos} + j_{neg} = 1$. This way, during learning, the algorithm avoids the tendency of classifying every instance as negative, *i.e.*, directly projecting the ratio of positive/negative examples observed in the data.

The set of features used for the first layer of classification encompass most of what was done during the preliminary experiments. However, since sentences are now the units of classification, and consequently there are fewer observations to learn from, it is possible to explore higher order n-grams. Furthermore, it is also possible to take advantage of sentence-specific information.

Below is the list of features considered:

- **Word, Lemma, POS [1-2-3-4]-grams –** similarly to what was done during the preliminary experiment, the features set is heavily composed of lexical and syntactic information. However, instead of using bigrams only (as before), it was possible to represent each sentence with higher order combinations of words, lemmas, and POS tags. Besides, these n-grams were captured in three different modes:
  - **bool –** an n-gram is either inactive or active, *i.e.*, its value is either 0 (when it does not occur in the sentence) or 1 (if it occurs at least once in the sentence);
  - **count –** represents an n-gram by how many times it appears in the sentence. For instance, if the word `we` appears twice in the sentence, the unigrams `word=we` and `lemma=we` are assigned with the value 2, while `pos=PRP` (personal pronoun) will be at least 2;
  - **rate –** represents an n-gram by its normalization with respect to the sentence length, *i.e.*, the appearance of an n-gram multiple times in a sentence has more impact for shorter sentences;
- **Sentence Length and Sentence Position –** includes the length of the sentence itself (number of tokens) and its relative position in the talk, *i.e.*, the order of the sentence in the talk divided by the total number of sentences;
- **Pronouns –** both in **bool** and **count** mode (described above for n-grams), express the existence of pronouns of the first-person singular, second-person singular and first-person plural: *I*, *you*, and *we* (personal, subject); *me*, *you*, and *us* (personal, object); *my*, *your*, and *our* (possessive adjective); *mine*, *yours*, and *ours* (possessive); *myself*, *yourself*, and *ourselves* (reflexive);
- **Reporting Verbs –** both in **bool** and **count** mode, a list[4] of words that are related to arguing and pointing to different sources, such as *tell*, *say*, and *mention*;

[4] http://www.edufind.com/english-grammar/reporting-verbs/

The next two sections present the results for the first layer of classification following the setup that was described above. However, each section will deal with the crowdsourcing data differently:

- **Section 4.2.1.1 –** training is built on the traditional majority vote strategy, *i.e.*, a sentence is considered to have an instance of a given metadiscursive act if at least two (out of three) workers agreed;
- **Section 4.2.1.2 –** training data built considering the idiosyncrasies of the data collection task, addressing the annotation more conservatively, in an attempt to improve the classification performance.

All results reported from this point forward follow a 10-fold cross-validation setup.

4.2.1.1 Majority Vote

As stated above, this section presents the results for the first layer of classification taking into consideration the majority vote for the annotation obtained via crowdsourcing. Therefore, TED talks are segmented into sentences, and each is assigned a binary value representing the occurrence (or non-occurrence) of each act in question. This binary classification is based on the most common opinion among the three annotators who labeled the sentence. In other words, sentences for which at least two workers signaled the existence of metadiscourse are considered positive examples, and all the remaining items are negatives.

The initial set of features tested in this experiment were *words* and *lemmas* (in all three modes described above). It is important to highlight that a setup considering word 4-grams naturally also include lower order n-grams. This setup generates a set of 24 combinations for each classifier (2 types of features × 3 modes × 4 n-gram orders). Additionally, this experiment also varied the SVMs cut-off threshold parameter according to four values ($th \in \{1, 2, 5, 10\}$). Thus, the total number of results for each classifier was 96.

Cat	Chance	Unigram Model			Best Model						
		prec	rec	F1	ft	n	th	md	prec	rec	F1
EXMPL	0.0143	0.68	0.85	0.76	W	1	1	c	0.74	0.86	0.80
DEFND	0.0233	0.36	0.69	0.47	W	3	1	b	0.76	0.58	0.66
INTRO	0.0158	0.31	0.65	0.42	W	3	2	c	0.56	0.54	0.55
COM	0.0152	0.36	0.64	0.46	L	4	10	b	0.43	0.60	0.50
ENUM	0.0306	0.32	0.59	0.41	L	2	5	c	0.45	0.54	0.49
R&R	0.0120	0.16	0.48	0.24	L	2	2	c	0.33	0.44	0.38
REFER	0.0053	0.15	0.44	0.22	W	2	2	b	0.37	0.30	0.33
ANT	0.0195	0.16	0.55	0.25	W	4	2	b	0.38	0.29	0.33
DEF	0.0110	0.15	0.36	0.21	L	2	2	c	0.24	0.27	0.25
EMPH	0.0191	0.12	0.42	0.19	W	2	2	b	0.23	0.26	0.24
CONC	0.0037	0.07	0.44	0.12	W	1	1	c	0.13	0.47	0.20

Table 4.2: Results for the best setting for sentence level classification (and comparison with chance and a straightforward unigram model).

Table 4.2 reports the results of the classification task, dividing them into three main parts. On the left, the probability of selecting a correct instance of metadiscourse by chance (*i.e.*, the number of sentences containing metadiscourse of a given category divided by the total number of sentences). In the middle, labeled as *Unigram Model*, are the results regarding precision, recall, and F1 for the most straightforward model: word unigrams under bool mode. Finally, on the right, under *Best Model*, are the configurations (features used, n-gram order, cut-off threshold, and mode) and classification results for the best settings. Herein, the criterion to choose between settings is the improvement in the F1. In the face of two solutions performing at the same statistically significance level, two additional criteria are considered. First, recall is preferred over precision. This decision reflects the preference for filtering out the least true positives possible, even if by doing so more false positives pass through this layer. Secondly, there is a preference for more straightforward and generic models, more precisely, preferring settings with **(a)** lemmas (over words), **(b)** lower order n-grams, **(c)** higher cutoff thresholds, and **(d)** with the following mode order: boolean $>$ count $>$ rate. The first column in Table 4.2, indicating the probability of selecting a metadiscursive sentence by chance, stresses how rare the phenomenon is, and thus its difficulty to be automatically detected. Only 3% of the sentences contain an instance of ENUMERATING – the most common act –, while the least frequent, CONCLUDING TOPIC, only occurs in 0.4% of the instances.

Another general conclusion is the fact that the mode *rate* never outperformed the other setups. In fact, it performed several times at the same statistical significance level as the mode *count*, demonstrating similarity in the way both represent the data. Given that *rate* mode encompasses a more complex representation of the data, and for the reasons already pointed out, the *count* mode is preferred.

The three categories which achieved better performance on this task were EXEMPLIFYING, DEFENDING IDEA, and INTRODUCING TOPIC ($F1 = 0.80$, 0.66 and 0.55, respectively), while DEFINING, EMPHASIZING and CONCLUDING TOPIC report the worst results ($F1 = 0.25$, 0.24 and 0.20). These values are very similar to the preliminary experiment results, with the top-3 categories being the same, and CONCLUDING TOPIC the worst successful category.

In more detail, and starting from the category with the highest F1 value, the best setting for classifying instances of EXEMPLIFYING only differs from the simplest model in what concerns the mode parameter – the best model also uses word unigrams with a cutoff threshold of one. All three measures (precision, recall and F1) increased simply by representing words with how many times they appear in the sentence (`mode=count`) instead of the binary information of whether they are present or not (`mode=bool`). This change of representation achieved an improvement of 0.04 regarding F1.

The next two top performant categories were DEFENDING IDEA and INTRODUCING TOPIC. For both categories, the best setup was using word 3-grams, improving 0.19 and 0.13 from the simplest model, respectively. Contrarily to what happened with `EXMPL`, there was a drop in recall but a significant improvement in precision ($+0.25$ for `DEFND` and $+0.07$ for `INTRO`).

Regarding COMMENTING ON LINGUISTIC FORM/MEANING, its best setting was one of the two categories, alongside with ANTICIPATING THE AUDIENCE'S RESPONSE, for which the configuration used 4-grams. Such configuration implies that more context is needed to identify these categories. However, the high level of specificity provided by 4-grams is balanced out by the remaining more generic parameters: lemmas (over words), and the value of the cut-off (the only category with `th=10`).

ENUMERATING and REPAIR & REFORMULATING best settings both chose lemma bigrams in count mode, improving F1 by 0.08 and 0.14, respectively, when compared to the unigrams.

REFERRING TO PREVIOUS IDEA and ANTICIPATING THE AUDIENCE'S RESPONSE performed at the same level in what concerns F1, both improving around 0.10 from the simplest model of word unigrams, while using words in bool mode.

Finally, the three categories with worst performances were DEFINING, EMPHASIZING and CONCLUDING TOPIC. By looking at what happened during the annotation phase (Chapter 3, it is possible to conclude on the reasons for why these acts obtained an F1 lower than 0.30:

- DEFINING was one of the categories with the most unsatisfactory results in the experts' validation task (Section 3.3.3). In sum, experts agreed with the crowd only 36% for instances marked by two workers and 35% for instances marked by all three workers. As mentioned previously, these results show inconsistency in the understanding of the task. These differences in interpretation, which are also expected to have occurred between workers during the annotation on AMT, are generating conflicting training instances and causing problems for classification;

- While validating occurrences of EMPHASIZING, experts' acceptance rates were also further from consensual, agreeing with the crowd about 65% of the times only. More importantly, EMPHASIZING by itself had one of the lowest agreements in the crowd ($\kappa = 0.18$);

- CONCLUDING TOPIC was the scarcest category, with only 0.37% of sentences containing an example of the phenomenon. Therefore, the training data for CONC had the smallest amount of positive examples, making it harder to generalize.

Finally, it is interesting to look at the cut-off parameter in some detail. As mentioned previously, this parameter allows the algorithm to discard features that were seen below a given threshold in order to avoid problems such as overfitting. The results in Table 4.2 were obtained by varying this parameter according to four values: 1, 2, 5, and 10.

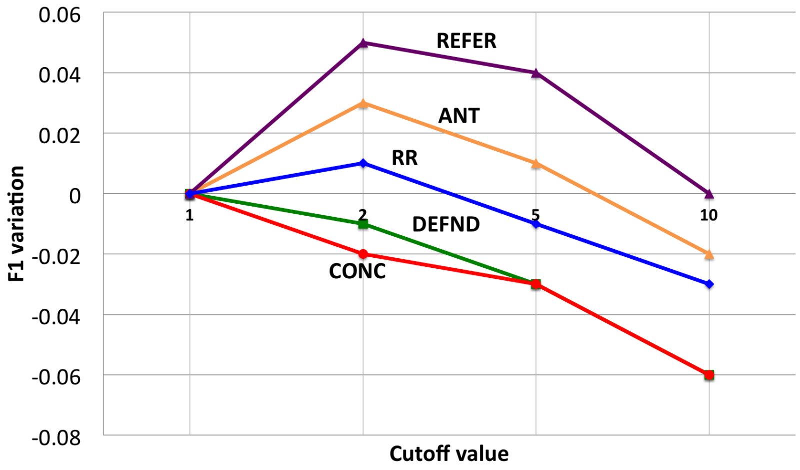


Figure 4.7: Evolution of cutoff parameter.

Figure 4.7 plots the impact of this variation for a set of five categories. In the figure, the x-axis plots the different values of the cut-off that were applied, while the y-axis plots variation in F1 concerning the simplest unigram model. Two trends were observed: first, for some categories pruning any features have a negative impact on performance, such as CONCLUDING TOPIC and DEFENDING IDEA in the figure. This behavior may be due to few training instances (as in CONC), or a high degree of variation inside the category (as in DEFND). The second trend, which happened for more than half of the acts represented, consisted of an improvement in performance when the cut-off value was two, and a loss for higher pruning values. For REFERRING TO PREVIOUS IDEA, merely pruning out all features that only happened once improved the F1 score by 0.05.

4.2.1.2 Beyond Majority Vote

So far, crowd answers were used in a pure majority vote setting: positive examples of metadiscourse are considered to occur when more than fifty percent of the workers agree, and the remaining points are considered to be negatives. However, given the setup of the task (pin-pointing instances of rare phenomena in text segments), it is predictable that some of the occurrences may not be selected. In fact, this was observed in Section 3.3.2, when discussing the primary sources of disagreement, and also during the expert validation task (Section 3.3.3).

In the latter, a team of experts looked at instances that were selected by one worker only, and relabeled them as metadiscourse or not. Results showed that some of these selections contain in fact the metadiscursive acts in question. For ANTICIPATING THE AUDIENCE'S RESPONSE and EXEMPLIFYING, for instance, 36% and 34% (respectively) of such cases were considered to be valid examples. Therefore, these cases used so far in training as negative examples may be hindering learning.

Given this scenario, a different approach to the data was tested. Positive examples are still coming from the agreement of the majority of the workers. However, negative examples will only be included for training if 100% of the workers agreed they were negative. In other words, there is a stricter criterion for considering that a given sentence does not have metadiscourse. Enforcing this rule discards the number of sentences that were selected by precisely one worker during the annotation.

Therefore, this next experiment explores the impact of removing such dubious cases from training. By doing this, two outcomes are expected:

- first, recall is expected to improve. Naturally, by removing negative examples from training, classifiers will be more permissive, allowing for more instances to be classified as occurrences of metadiscourse. Additionally, given the nature of the training points herein removed, the inference will face fewer contradictions, thus improving recall. For instance, and in a simplified manner, if while annotating instances of EXEMPLIFYING, one worker selects the expression *"for example"* and the other two workers miss out on it, this particular case will not count as a negative example;

- on the other hand, removing these cases only from training will hinder precision. This consequence arises given the significant difference between the training and the testing data.

Having in mind these expected outcomes, and taking into consideration that this particular stage of classification is supposed to act as a filter to the next component of the chain, the goal is to be able to achieve configurations that perform at the same level as the best solution so far (equivalent F1 value) while improving recall. This strategy postpones the problems of precision to the next step in the classification chain.

Similarly to what was done before, and since the training data is different, several combinations of parameters were tested: main feature, mode, n-gram order, and cutoff threshold. Again, results reported below consist of a 10-fold cross validation setting. Finally, for solutions that perform at the same statistical significance level ($p < 0.05$), the ones with higher recall rate are preferred (for the reasons mentioned previously).

Cat	Previous Best			− train							− train & test		
	prec	rec	F1	ft	n	th	md	prec	rec	F1	prec	rec	F1
EXMPL	0.74	0.86	0.80	W	1	1	c	0.72	0.87	**0.79**	0.89	0.87	0.88
DEFND	0.76	0.58	0.66	W	4	2	b	0.65	0.65	**0.65**	0.85	0.65	0.74
INTRO	0.56	0.54	0.55	L	4	2	b	0.52	0.61	**0.56**	0.68	0.61	0.64
COM	0.43	0.60	0.50	L	3	10	b	0.40	0.71	**0.51**	0.61	0.71	0.66
ENUM	0.45	0.54	0.49	W	2	2	b	0.47	0.55	**0.51**	0.60	0.55	0.57
R&R	0.33	0.44	0.38	W	3	2	c	0.36	0.45	**0.40**	0.61	0.45	0.52
REFER	0.37	0.30	0.33	W	3	5	c	0.30	0.35	**0.32**	0.35	0.35	0.35
ANT	0.38	0.29	0.33	W	3	5	b	0.24	0.42	**0.31**	0.33	0.42	0.37
DEF	0.24	0.27	0.25	L	2	5	b	0.19	0.36	**0.25**	0.25	0.36	0.30
EMPH	0.23	0.26	0.24	L	2	5	b	0.18	0.39	**0.25**	0.25	0.39	0.30
CONC	0.13	0.47	0.20	W	1	1	c	0.13	0.47	**0.20**	0.14	0.47	0.22

Table 4.3: Results of the best settings by selectively removing negative examples.

Table 4.3 presents the results of the experiment just described, which are divided into three main panes. The leftmost, for comparison purposes, presents the best solution found in the previous experiment where a pure majority vote was taken into account (see Table 4.2). In the middle section are the results in terms of precision, recall and F1 (along with the parameter setting that generated them) for the best solution achieved when removing the dubious annotations from training. Lastly, on the right, are represented the results obtained by also removing these annotations from the test set. The configurations are the same as the ones described in the middle pane.

The first thing worth noting is precisely the difference between the middle and right side panes. In the middle, all instances that were selected by one worker only were excluded from training but were still used in the test data, while on the right they were excluded from both sets. By definition, since all of these examples are considered negative, removing them from the test set have no impact in recall (which deals with the number of positive instances that are retrieved by the classifier).

However, in what concerns precision, there is a significant difference (which also impacts F1 accordingly). It follows logically that this difference in precision between the two configurations (middle and right panes) represents the number of instances selected by one worker only that are being classified as actual instances of metadiscourse.

The three categories with the highest impact in precision are REPAIR & REFORMULATING, COMMENTING ON LINGUISTIC FORM/MEANING, and DEFENDING IDEA, with differences in precision of 25%, 21%, and 20% respectively. The results obtained during the annotation phase for these categories show they had a high number of instances that were selected by one worker only (from Table 3.2, column concerning only one worker in agreement: `R&R` – $1,493$ cases, `COM` – 738, and `DEFND` – $1,538$).

Additionally, by looking at the experts' validation task (Table 3.4), it is expected that 12%, 21%, and 18% of these cases (respectively), are in fact instances of metadiscourse of the corresponding categories. In other words, these were the percentage of cases that were selected by only one worker (on AMT) but counterintuitively experts agreed to have metadiscursive strategies indeed. Interestingly, for `COM` and `DEFND`, these statistics have a close to perfect match with the results obtained in this experiment, more precisely, with the differences in precision between both setups in Table 4.3.

As mentioned before, the goal of this experiment was to be able to improve recall while keeping the overall performance of the classifiers (F1) at the same statistical significance level. By comparing the results on the left and middle panes, it is possible to see that this was true in all cases.

The hypothesis was that recall would go up at the cost of precision. So while that happened for most cases at statistically significant levels ($p < 0.05$), there were some exceptions. For the categories ENUMERATING and REPAIR & REFORMULATING both precision and recall went up by merely removing the one-worker-annotated occurrences from training. For CONCLUDING TOPIC, on the other hand, this strategy left the results unchanged. The best solution consisted of the same configuration of parameters as before, and the values for precision and recall were precisely the same.

The reason for such observation is two-fold: first, CONC is the category with the lowest amount of positive examples; secondly, it is also the category with the least cases of one-worker-annotated occurrences (only 153 instances). Therefore, the modifications introduced in this experiment have less impact on the results.

Finally, the category EXEMPLIFYING had only a non-statistically significant improvement in recall (1%). For this category, the best solution parameters were also the same as in the experiment that followed a pure majority vote setting.

Regarding the remaining categories, a substantial increase in recall was observed. The top three categories regarding improvement in recall were EMPHASIZING ($+13\%$), ANTICIPATING THE AUDIENCE'S RESPONSE ($+12\%$), and COMMENTING ON LINGUISTIC FORM/MEANING ($+11\%$). These categories are amongst the ones with the highest number of cases of one-worker-annotated cases (EMPH had the highest value of all categories – 2,023; ANT – 1,426; and COM – 738).

Finally, regarding the configurations that achieved best solutions in this experiment, it is possible to see that five categories opted for higher order n-grams (DEFND, INTRO, COM, R&R, and ANT) while the remaining were left unchanged. Concerning the cutoff threshold, four categories were able to better generalize by increasing its value (DEFND, ANT, DEF, and EMPH). This result may also be due to the increase of the n-gram order parameter and the consequent need to discard less frequently observed features (more specifically for the categories DEFENDING IDEA and ANTICIPATING THE AUDIENCE'S RESPONSE).

4.2.1.3 Exploring Additional Features

At this point, having stable configurations of classifiers for each category of metadiscourse, the remaining set of features described in the beginning of this section were introduced, one at a time, to test their impact on the classification. As a reminder these features were: part-of-speech, lemmas, sentence length, sentence position, pronouns and reporting verbs.

However, none of the features mentioned above achieved a statistically significant improvement in performance. Only an improvement of 1% in F1 was registered for REFERRING TO PREVIOUS IDEA (when considering reporting verbs), and for ANTICIPATING THE AUDIENCE'S RESPONSE (when using both sentence length and sentence position simultaneously).

In fact, more often the contrary was observed. Including, for example, part-of-speech n-grams (up to 4-grams) significantly hindered the classification, consistently causing worse results across categories (-8% for CONCLUDING TOPIC, -7% for REFERRING TO PREVIOUS IDEA, and -4% for EXEMPLIFYING). These results show the semantic nature of metadiscourse, in the sense that simple syntactic features are not able to capture the phenomenon at the sentence level. For not containing the representative power the task requires, including them generates a situation of overfitting.

4.2.1.4 Qualitative Error Analysis

To better understand the errors resulting from the classification process, this section presents real examples for each of the categories. These excerpts are divided into three groups: *True Positives* contain successful classification examples, *i.e.*, where both classifier and crowd agreed; *False Negatives* are composed of cases labeled as metadiscourse by the crowd, but not automatically classified as such; and *False Positives* are passages that were classified as metadiscourse by the algorithm and not by the crowd.

These examples were selected manually to illustrate the classification process. They were chosen to specifically highlight some of the problems of both classification and annotation.

EXEMPLIFYING

- True Positives

  1. *"They tend to – pigs, for example, are more like dogs."*
  2. *"I'll show you a few examples of this now."*
  3. *"We'll just take one example here: a virus is a natural system, right?"*
  4. *"So you can imagine the scale of this problem."*
  5. *"It's 11 meters in diameter, and we know that it started growing in the year 1584. Imagine that."*

It is important to remember here that this category was a combination of two categories from the original taxonomy – `Exemplifying` and `Imagining Scenarios`. The first three sentences above rely on the hook word *"example"*, but reflect different manners of exemplifying. On the first one, the example is a quick illustration of the matter at hand; a quick side note in the form of an example to help the listener understand what is being discussed. The second and third cases are somewhat different. In a way, they are more metadiscursive in the sense that they signal the listener that what will come next is an example.

The last two examples rely on the word *"imagine"* as a hook for the phenomenon. While they clearly contain metadiscursive phenomena, and even though both the crowd and the classifier selected them, they are less consensual to the category EXEMPLIFYING at hand. It seems that here they may be working more like instances of EMPHASIZING with the speaker highlighting the importance (Example 4) or the magnitude (Example 5) of the subject matter.

- False Negatives

  1. *"But there was one interesting anecdote that I found in Indonesia."*
  2. *"And you can think of this as our satellite view for our map."*
  3. *"And what you're trying to look at is the males have claspers, which kind of dangle out behind the back of the shark."*
  4. *"Look at the waves coming here to shore."*

Above are excerpts from the cases that were selected by the crowd but the classifier was not able to model. Again, while all cases have metadiscourse in them, it is possible to argue whether they are serving as examples or not. In the first one, the speaker is introducing an anecdote which could be more suitable for the category ADDING TO TOPIC. Example 2 seems to be performing a clarification function, where the speaker is creating an analogy to a concept that is probably familiar to the listeners. Finally, examples 3 and 4 above, are performing the function of the category ENDOPHORIC MARKING, establishing a link between the listener and the visual materials displayed on site.

- False Positives

  1. *"Imagine that Earth is at the center of the universe, and surrounding it is the sky projected onto a sphere."*
  2. *"Consider for a moment this quote by Leduc, a hundred years ago, considering a kind of synthetic biology."*
  3. *"Can you imagine the potential offspring applications – environmental detection of pollutants in soils, customs applications, detection of illicit goods in containers and so on."*

The cases above consist of examples that were selected by the classifier but not during the annotation phase by the crowd. The first sentence should have been indeed annotated as an instance of EXEMPLIFYING. In fact, it was selected by one worker only which, in the majority vote technique followed herein, deems it as a negative instance. Even though it was not used for training, it is still considered as negative for testing.

The remaining examples are again less obvious. It seems that the classifier learned to consider quotes as examples, which is acceptable since there was no mention of quotes in the original taxonomy and workers were not given specific instruction on how to handle quotes. Example 3, on the other hand, seems to be performing the same semantic functions as examples 4 and 5 from the *True Positive* group above, *i.e.*, EMPHASIZING, showing some inconsistency during the annotation phase.

DEFENDING IDEA

- True Positives
  1. *"We believe that in the last 10 years companies that we've financed are actually the best media companies in the developing world."*
  2. *"I think one of the most interesting examples of this comes from Australia."*
  3. *"Most important of all, I believe, is working with small groups of women, providing them with opportunities for micro-credit loans."*
  4. *"It's a difficult theory to discount, I think you'll agree."*
  5. *"I mean, the question here is, here we are, arguably the most intelligent being that's ever walked planet Earth [...] and yet we're destroying the only home we have."*

All of the true positive cases listed above for the category DEFENDING IDEA illustrate the speaker giving an opinion, or trying to convince the audience of a certain point, hence being correctly annotated and classified.

- False Negatives
  1. *"Because I thought, I still maintain, that serious and independent media companies are great business."*
  2. *"The answer is it matters quite a lot."*
  3. *"So to conclude, I'd just like to point out, you know, the whales live in an amazing acoustic environment."*
  4. *"We think in terms of war and interstate war."*

Regarding cases that the classifier missed, even though the crowd had selected them, it is possible to see errors introduced in both parts of the process. The first two cases are indeed instances of DEFENDING IDEA (sharing a personal thought and arguing that something matters), wrongly misclassified as not part of the category. Contrarily, in the last two cases, the speaker seems to be stating a fact rather than defending a new idea, belief, or theory.

- False Positives

  1. *"I would think media systems were organizations, which means they should help you."*
  2. *"But the point is, if we ever survive to actually issue them, find enough investors that this can be considered a success, there's nothing stopping the next organization to start to issue bonds next spring."*
  3. *"But I'd also like to point out that the oceans are much more connected than we think."*

This last set of cases for the category DEFENDING IDEA show examples that were automatically classified as part of the category while not identified as such during annotation. In example 1 the speaker is defending a personal belief, using the construction *"I would think"*. However, when looking at the context of the passage above (the previous sentence being *"And what did I know at that time about media systems?"*), it is possible to conclude that this is not an idea that the speaker shares at present.

The second and third examples, on the other hand seem to be correctly classified, since the speaker is trying to prove a point, following a chain of ideas or trying to revert a common misconception.

INTRODUCING TOPIC

- True Positives

  1. *"There's a sense in which it's obvious, and yet, let me tell you a little story."*
  2. *"I want this morning to talk about the biography of one particular object."*
  3. *"I'm going to show you are the astonishing molecular machines that create the living fabric of your body."*

In what concerns INTRODUCING TOPIC, the true positive cases are very consistent and show the speakers signaling the audience what is going to happen next in the speech event.

- False Negatives

  1. *"Here's my approach to it all."*
  2. *"Now, how about incentives?"*
  3. *"Let me begin with an example."*
  4. *"It's a story about lemonade."*

The set of cases above, selected by the crowd but not picked up by the classifier, show some of the difficulties that arise when modeling metadiscourse. The first two cases are used to signal a change of topic, signal what is about to come, but these strategies are very subtle. It is even possible to argue whether example 2 is metadiscursive at all (even though it clearly has the intention of topic introduction). Only the word *"now"* can be interpreted as metadiscourse. In this case, it is mostly the form (that of a rhetorical question) that signals the speaker intention of starting to talk about *"incentives"*.

Example 3 is also complicated since it seems to have other two functions aside from INTRODUCING TOPIC. There is an intention of enumerating, disclosed by the word *"begin"*, and also an intention of exemplifying. Nonetheless, it is an instance of INTRODUCING TOPIC, where the speaker felt the need to introduce an example as a somewhat defined part of the talk.

The last example of false negatives again shows some of the problems of the task at hand. While the word *"story"* situates this sentence in the realm of metadiscourse, it is not straightforward that it should be considered an introduction. In fact, the previous sentence in context is example 1 above, in the true positives group – *"There's a sense in which it's obvious, and yet, let me tell you a little story."* While the latter shows an intention of signaling a new topic, the former, signaled by the crowd only, is already advancing in the content of the discourse, *i.e.*, stating what the *"story"* is about, instead of what the next part of the talk is going to be about.

- False Positives

  1. *"So I talked about the fact that we need to train and support defenders."*
  2. *"Now the final part of the trilogy was I wanted to focus on the body and try to be the healthiest person I could be, the healthiest person alive."*
  3. *"So what we can do is actually tell you about the molecules, but we don't really have a direct way of showing you the molecules."*
  4. *"This is work from a number of years ago, but what I'll show you next is updated science, it's updated technology."*
  5. *"And I want to begin with one episode from that sequence of events that most of you would be very familiar with, Belshazzar's feast – because we're talking about the Iran-Iraq war of 539 BC."*

Regarding cases that the classifier marked as instances of INTRODUCING TOPIC and the crowd did not, it is possible to distinguish three situations from the examples above. Firstly, examples 1 and 2, even if containing some lexical items that might suggest an introduction (*"talk about"* and *"want to focus on"*), are not to be considered as such given the tense of the sentence. Both examples are in the past tense. In fact, the first is an instance of RECAPITULATING, while the second is merely a description of what the speakers did in the past.

Secondly, example 3 is tough to classify because it needs a significant amount of context. To sort this specific example out, it would be necessary to understand if the speaker has been talking about *"molecules"* and is justifying why s/he is not showing them, or if it is a topic that is being introduced at the moment. To make a correct decision, this kind of cases need semantic knowledge of what is happening throughout the entire talk.

Finally, the last two scenarios were wrongly annotated by the crowd since they represent typical cases of introductions to the next topic of the talk (they were actually selected by exactly one worker out of the three who worked on the respective segments).

COMMENTING ON LINGUISTIC FORM/MEANING

- True Positives

  1. *“Well, it may already have occurred to you that Islam means “surrender”.”*
  2. *“So we talk about 300, a batter who bats 300. That means that ballplayer batted safely, hit safely three times out of 10 at-bats.”*
  3. *“What’s a word made of?”*
  4. *“The three words are: Do you remember?”*
  5. *“I personally wrote thousands of lines of code to write this cookbook.”*

The category COMMENTING ON LINGUISTIC FORM/MEANING encapsulates instances of metadiscourse that relate to the *mention* function of the *use-mention* paradigm by Wilson (2012) discussed in the background chapter. It realizes the function of talking about words and terms with respect to their meaning and to how they serve the content of the discourse. The first two examples above specifically show that phenomenon happening, with the speakers talking about the origin of a word’s meaning as well as the use of a given expression in a given context.

The following two examples seem to be less explicit regarding the function that they play in the discourse. Example 3 is embedded in a segment that is meta in itself, where the speaker is talking about words and their meaning. Example 4 appears to be a punchline: the speaker builds up the discourse by referring to *“three important words”*, which eventually results in the statement transcribed herein. This passage seems to be more associated with the EMPHASIZING phenomenon.

The last example was wrongly annotated and classified. The speaker is merely talking about his/her writing process.

- False Negatives

    1. *"Let's rebrand global warming, as many of you have suggested. I like "climate crisis" instead of "climate collapse," but again, those of you who are good at branding, I need your help on this."*
    2. *"And there is a way the world both envisions food, the way the world writes about food and learns about food."*
    3. *"So what I often like to say is that, from a genomic perspective, we are all Africans."*

Both sentences which comprise the first example of the false negative group were annotated by the crowd and missed by the classifier. In fact, they can be seen as instances of COMMENTING ON LINGUISTIC FORM/MEANING, since the speaker is negotiating the use of specific terminology.

Example 2 is somewhat dubious. Even though the speaker is talking about words and the process of writing, this excerpt seems to be part of the discourse and therefore not having metadiscursive mechanisms.

Regarding the last example, it is unclear why a majority of the workers annotated it as an instance of COMMENTING ON LINGUISTIC FORM/MEANING, since it is composed of an opinion or personal idea rather than dealing with terminology.

- False Positives

    1. *"They come up with these very restrictive labels to define us."*
    2. *"But those categories mean even less now than they did before."*
    3. *"I myself am a philosopher, and one of our occupational hazards is that people ask us what the meaning of life is."*
    4. *"I say, "Well, do words exist?"*
    5. *"That horn-shaped region is what we call the sweet spot."*

The first two examples above demonstrating cases selected by the classifier but not annotated by the crowd are indeed instances of COMMENTING ON LINGUISTIC FORM/MEANING. In them, the speaker is talking about *"labels"* and *"categories"* and asserting about their meaning and definition. Examples 3 and 4, on the other hand, show the classifier relying on hooks like *"meaning"* and *"word"*, here not necessarily associated with the metadiscursive act at hand. Example 5 is a clear instance of DEFINING, more specifically an instance of naming (included in the examples of the annotation task upon training).

ENUMERATING

- True Positives
  1. *"And I undertook this for two reasons."*
  2. *"The first thing I did was I got a stack of bibles."*
  3. *"The second type of rule that was difficult to obey was the rules that will get you into a little trouble in twenty-first-century America."*
  4. *"And finally I learned that thou shall pick and choose."*

The true positive cases for the category ENUMERATING show some of the strategies used by speakers to organize discourse. In fact, the cases above are all extracted from one same talk, in which the speaker particularly relies on this strategy to better communicate the message.

- False Negatives
  1. *"The next piece that is going to come up is an example of a kind of machine that is fairly complex."*
  2. *"The first was that I grew up with no religion at all."*
  3. *"Another lesson is that thou shalt give thanks."*

Under false negative cases it is possible to find some strategies that the classifier was not able to generalize, more specifically, the constructions *"next piece"*, *"first was"* (past tense after the word *"first"*), and *"another lesson"*.

- False Positives

    1. *"For the second demo, I have this Wii remote that's actually next to the TV."*
    2. *"And I thought I'd end with just a couple more."*
    3. *"It would be – whatever came across the finish line first would be the winner."*
    4. *"Here's two males fighting."*

Regarding the false positives, there were two cases: situations where the crowd missed some occurrences of ENUMERATING (examples 1 and 2), and cases that the classifier was not able to generalize when speakers used numerals in situations other than enumerations (examples 3 and 4). Example 4 is particularly tricky. As it is, it should be assigned to ENDOPHORIC MARKING, since it is related to an image that the speaker is showing. However, it is the semantics of the object that influences the category to which this instance of metadiscourse belongs. By replacing the words *"males fighting"* with, for instance, *"cases"*, *"examples"*, or *"problems"*, this occurrence would be a clear case of an enumeration.

REPAIR & REFORMULATING

- True Positives

    1. *"I also talked about community-generated data – in fact I edited some."*
    2. *"So I'm highlighting just a few words and saying definitions like that rely on things that are not based on amino acids or leaves or anything that we are used to, but in fact on processes only."*
    3. *"Actually, it turns out that your risk of breast cancer actually increases slightly with every amount of alcohol that you drink."*

By looking at the examples obtained for REPAIR & REFORMULATING, it is possible to see that the crowd heavily relied on the hooks *"in fact"* and *"actually"* to mark occurrences of the phenomenon, not taking into consideration their function. All three examples above seem to be more suitable to be clarifications rather than repairs or reformulations.

- False Negatives

  1. *“There’s the map showing, on the left-hand side, that hospital – actually that’s a hospital ship.”*
  2. *“You know that you’re infected because it actually shows up.”*
  3. *“So how do these guys then monetize those infected computers?”*
  4. *“Well in most cases it never gets this far.”*

Regarding cases identified by the crowd but ignored by the classifier, the same pattern as before appears: *“actually”* being used to identify REPAIR & REFORMULATING even when it does not have that function in the discourse (examples 1 and 2). The last two examples are also confusing on why workers agreed for them to carry reformulation information.

- False Positives

  1. *“So, today I’m back just to show you a few things, to show you, in fact, there is an open data movement afoot, now, around the world.”*
  2. *“When people say fusion is 30 years away, and always will be, I say, “Yeah, but we’ve actually done it.””*

As expected, in this group a lot of unlabeled cases of *“actually”* and *“in fact”* appear, not being connected to the phenomenon at hand. In example 2, for instance, the word *“actually”* is even between quotation marks.

These results and examples align with the fact that this category had the lowest agreement of the ones used for classification. The meaning of the category was not understood and therefore will not be considered further in this thesis.

Referring to Previous Idea

- True Positives

  1. *"And as I told you just now, neural activity can change the connectome."*
  2. *"And as I said earlier, I'm a perfectionist."*
  3. *"You know, I've talked about some of these projects before."*
  4. *"Again, coming back to that thing about the lifestyle, and in a way, the ecological agenda is very much at one with the spirit."*

The examples that were both marked by the crowd and correctly classified show the plenitude of the category Referring to Previous Idea. In the sentences above, the speaker is referring to something mentioned before, heavily using the past tense of verbs like *"tell"*, *"say"*, or *"talk"*. Example 4, however, contains an instance of Referring to Previous Idea using the constructions *"again"* and *"coming back to"*.

- False Negatives

  1. *"And again, we can measure the reduction in terms of energy consumption."*
  2. *"And again, the way in which that works as a building, for those of us who can enjoy the spaces, to live and visit there."*
  3. *"And what had happened was the circle had closed, it had become a circle – and that epiphany I talked about presented itself."*
  4. *"We're being advised by some of these people, as was said, to try and bring all the experience to book."*
  5. *"I also said the other principle that I think we should work on is [...]"*

The first two examples above show that the classifier was not able to associate the word *"again"* (by itself) to instances of this category. The remaining three examples show variations of what was found in the true positives group. These cases were probably not frequent enough to be generalized and, consequently, correctly classified.

- False Positives

  1. *"And if you marry this fact with the incredible abundance of information that we have in our world today, I think you can completely, as I've said, remake politics, remake government, remake your public services."*
  2. *"It needs truth and beauty, and I'm so happy it's been mentioned so much here today."*
  3. *"Okay, let's go back to the slides."*

Regarding cases that the classifier found to be related with REFERRING TO PREVIOUS IDEA, but not marked as so by the crowd, it is possible to see that some of them were mislabeled during the annotation (examples 1 and 2), while others seem to be related to ENDOPHORIC MARKING (example 3).

ANTICIPATING THE AUDIENCE'S RESPONSE

- True Positives

  1. *"And you know what these are."*
  2. *"You think of carbon as black."*
  3. *"Now as you ponder that question, maybe you're thinking about the first day of preschool or kindergarten, the first time that kids are in a classroom with a teacher."*
  4. *"And you might fairly ask: why is that?"*

The presence of the pronoun *"you"* is notorious in most examples of the category ANTICIPATING THE AUDIENCE'S RESPONSE. In the examples above, it appears in the close neighborhood of words such as *"know"*, *"think"*, and *"ask"*.

- False Negatives

1. *"So we've all seen these."*
2. *"It may seem like we're quite remote from other parts of this tree of life, but actually, for the most part, the basic machinery of our cells is pretty much the same."*
3. *"Maybe you've encountered the Zero-to-Three movement, which asserts that the most important years for learning are the earliest ones."*

In what concerns false negatives for the category ANTICIPATING THE AUDIENCE'S RESPONSE, examples 1 and 2 show that the classifier was not able to generalize the phenomenon to instances of the personal pronoun *"we"*. Additionally, example 3 shows the co-occurrence of *"you"* with a less frequent verb *"encounter"* in a sentence that was not captured by the classifier.

- False Positives

1. *"You've seen them all around, especially these days as radars are cheaper."*
2. *"Probably about as many as there are creative people here."*
3. *"You might know that, so far in just the dawn of this revolution, we know that there are perhaps 40,000 unique mutations."*
4. *"It may not be the biggest bamboo building in the world, but many people believe that it's the most beautiful."*

The examples above show that the classifier was able to pick up some of the cases that the crowd missed. All of the four examples above show some sort of pre-assessment of the audience's knowledge.

DEFINING

- True Positives

  1. *"It was here that I had my first encounter with what I call the "representative foreigner.""*
  2. *"The brain is intentionally – by the way, there's a shaft of nerves that joins the two halves of the brain called the corpus callosum."*
  3. *"In fact, creativity – which I define as the process of having original ideas that have value – more often than not comes about through the interaction of different disciplinary ways of seeing things."*
  4. *"That's what the expansion of the universe or space means."*

In the true positive group, different cases were found. Examples 1 and 2 are examples of naming, where the speaker uses the word *"call"* to associate a concept with its name. Example 3 presents an instance of a definition, in this case, a personal definition of *"creativity"*, setting a common ground for the use of the word for the remainder of the presentation. Finally, example 4 shows an example of a mislabel. In it, the word *"means"* is not used to define a term or a concept, but instead to conduct a train of thought.

- False Negatives

  1. *"The Large Hadron Collider, a particle physics accelerator, that we'll be turning on later this year."*
  2. *"By invisible, I mean it doesn't absorb in the electromagnetic spectrum."*
  3. *"So, recently, we have realized that the ordinary matter in the universe – and by ordinary matter, I mean you, me, the planets, the stars, the galaxies – the ordinary matter makes up only a few percent of the content of the universe."*
  4. *"What tribes are, is a very simple concept that goes back 50,000 years."*

A wide array of problematic cases were found under the false negative group. Example 1 contains a definition indeed. However, it is not metadiscursive at all, since no lexical items are signaling that definition. The remaining three examples seem to be more suitable to be classified in other categories: example 2 is a clarification, example 3 is an EXEMPLIFYING strategy, and example 4, where the speaker discusses the origin of a concept, would be more suited to the category COMMENTING ON LINGUISTIC FORM/MEANING.

- False Positives

  1. *"It's called batting average."*
  2. *"We call this gravitational lensing."*
  3. *"And that means that the idea you create, the product you create, the movement you create isn't for everyone, it's not a mass thing."*
  4. *"If it's constant, that means that the stars out here are feeling the gravitational effects of matter that we do not see."*

Under cases that were only picked by the classifier, several instances of naming were found (examples 1 and 2). Another source of problems for the performance of the classifier was the use of the verb *"mean"*, and how it can be used for defining a term or just to mark a chain of thought (as seen in example 4 of the true positives group).

EMPHASIZING

- True Positives

  1. *"The next exercise is probably the most important of all of these, if you just take one thing away."*
  2. *"So what we need to look now is, instead of looking outward, we look inward."*
  3. *"So the whole point of that is not, sort of, to make, like, a circus thing of showing exceptional beings who can jump, or whatever."*

In the successful examples above, the speakers use different strategies to emphasize a point, with constructions like *"the most important"*, *"we need to look"*, and *"the whole point"*.

- False Negatives

  1. *"And so it's very important for those researchers that we've created this resource."*
  2. *"The world I envision for her – how do I want men to be acting and behaving?"*
  3. *"And of course, as I mentioned before, since we can now start to map brain function, we can start to tie these into the individual cells."*
  4. *"So I'm going to leave you with a final note about the complexity of the brain."*

Under the examples that the classifier could not recognize, some cases are clear instances of EMPHASIZING (example 1), but also several cases where the reason why the crowd labeled them as such is not apparent (examples 2, 3 and 4). While example 2 does not have metadiscourse at all, example 3 is an instance of REFERRING TO PREVIOUS IDEA and example 4 is an instance of CONCLUDING TOPIC.

- False Positives

  1. *"Give you one example of that: Intention is very important in sound, in listening."*
  2. *"So let's take a look at the brain."*
  3. *"So let's take a deeper look."*

By looking at the false positive examples, it is possible to see that the classifier generalized the use of the word *"important"* and the word *"so"* at the beginning of a sentence, associating them with the phenomenon. This generalization gave rise to a substantial number of misclassified instances as seen in the cases above.

CONCLUDING TOPIC

- True Positives

  1. *"Finally, what I'm doing now."*
  2. *"The question I leave you with now is which is the ballast you would like to throw?"*
  3. *"If I can leave you with one big idea today, it's that the whole of the data in which we consume is greater than the sum of the parts."*

The true positive cases for the category CONCLUDING TOPIC show the speaker managing the topics in the talk, indicating the last topic that is going to be discussed, and also leaving a final message, highlighting the main idea, and indicating the closure of the discourse.

- False Negatives

  1. *"But I am here to tell you that, based on my experience, people are not where they live, where they sleep, or what their life situation is at any given time."*
  2. *"And the final, sort of, formulation of this "American Idol" format, which has just appeared in Afghanistan, is a new program called "The Candidate.""*
  3. *"And I hope some of you will be inspired for next year to create this, which I really want to see."*

Several of the cases signaled by the crowd but missed by classification, seem dubious. It is not clear how the three examples above are portraying conclusion in any way.

- False Positives

  1. *"A real airplane that we could finally present."*
  2. *"So with that, I thank you."*
  3. *"Now, personally, I think I'm not the first one who has done this analysis, but I'll leave this to your good judgment."*
  4. *"I'll show you just a couple others."*

The dispersion of annotations, as seen in the false negative group above, along with the small number of examples obtained for CONCLUDING TOPIC, generates a set of false positive cases that are not consistent with the concept at hand. The classifier leans towards picking up sentences with the words *"finally"* and *"leave you"*, but as shown above, it is not always the case that the presence of such expressions is enough to imply the presence of a conclusion.

4.2.2 Word Level Classification

As described at the beginning of this section, the Support Vector Machines (SVM) that make up the first level of the suggested classification chain serve as filters to the next set of classifiers. In sum, at the first stage of classification, they receive an entire talk and output the sentences with a high probability of containing each of the metadiscursive acts considered.

Having now a set of potential sentences that contain metadiscourse, discretized by function, the next step (and ultimately the goal of this system) is to pinpoint the tokens in those sentences that realize the metadiscourse act itself.

More precisely, in this second classification task, Conditional Random Fields (CRFs) will be trained with positive examples only and tested on the candidate sentences that passed the first classifier (SVM). Training only with sentences that were labeled by the crowd as containing metadiscourse allows one to focus precisely on which tokens are indicators of the phenomenon.

For this second and last component of the chain, two primary sets of experiments took place. The difference between these experiments is the way data is represented, *i.e.*, the features used for classification. They are organized in the following manner:

- **Section 4.2.2.1 –** similarly to what was done in the preliminary experiments at the beginning of this section (Section 4.1), the feature space is composed of n-grams (lemmas, words, and POS tags) and other syntactic features thought relevant;
- **Section 4.2.2.2 –** presents an approach based on word embeddings, where words are described as vectors of real numbers, representing the context in which such words appear.

4.2.2.1 N-grams

As just mentioned, the first approach to token-level classification closely follows the setup described in the preliminary experiments (Section 4.1). In sum, the approach consists of training one CRF per category and can be formulated as: given a set of words, distinguish between the ones that make up the metadiscursive act and which are part of the content of the presentation itself.

The difference between the experiment reported below and what was done during the preliminary experiments is the test space. The preliminary approach was trained with sentences which contained metadiscourse but was then tested in the entire talks. This mismatch between train and test severely impacted performance.

Now, however, while the training data is still composed of sentences that contain metadiscourse only, the classifiers are going to be tested on the candidate sentences that passed the previous step of classification. This formulation guarantees a more fair distribution of positive and negative examples between train and test sets.

Besides starting off in a more balanced situation concerning train and test sets, in this new approach to token-level classification both higher-order n-grams and contexts are going to be explored, and analyzed.

Regarding the details of implementation, again, it follows the work of Okazaki (2007) to train CRF using lexical and syntactic features. Given the amount of feature combination sets, the strategy used in this experiment was to explore them in a hill-climbing fashion.

For each category, two classifiers were trained using unigrams with no surrounding context: one with lemmas and another with words (features `L1_1` and `W1_1`). These two baseline results were first expanded augmenting the window size (until a maximum of four), and then varying the order of the n-grams (also until a maximum of four). With these results, models with different orders were combined, and such combinations were kept when an increase of F1 was registered.

After testing all such combinations, POS features were introduced in the same fashion to the best combination of lemma or word n-grams. Finally, other features such as named entities, position in sentence, reporting verbs, and personal pronouns were tested.

Cat	Chance	Unigram Model			Best Model			
		prec	rec	F1	ft	prec	rec	F1
EXMPL	0.1474	0.69	0.56	0.62	W1_2	0.70	0.75	0.72
DEFND	0.1095	0.54	0.74	0.62	L2_3	0.60	0.79	0.68
INTRO	0.2737	0.47	0.81	0.59	W1_2-W2_1-W3_1-P4_1	0.53	0.86	0.66
ENUM	0.1881	0.32	0.35	0.33	L1_4-L2_1	0.42	0.62	0.50
REFER	0.1889	0.35	0.26	0.30	W2_2-P3_2	0.47	0.59	0.48
CONC	0.2444	0.23	0.27	0.25	W1_2-W2_3-ner	0.49	0.30	0.37
COM	0.1370	0.37	0.19	0.25	L1_2	0.34	0.37	0.35
ANT	0.2185	0.17	0.36	0.23	L2_3-L3_4	0.38	0.29	0.33
EMPH	0.2262	0.17	0.33	0.22	W1_2-W2_3	0.21	0.64	0.32
DEF	0.0980	0.21	0.37	0.27	L1_3-L4_2	0.21	0.42	0.28

Table 4.4: Results of the best setting for word-level classification.

Table 4.4 summarizes the results of this experiment. To the left, it shows the probability of a given word, for a given category, being part of the metadiscursive act, *i.e.*, the number of words marked as metadiscourse in the test set divided by the total number of tokens. This probability of correctly classifying a word by chance is here much higher than what was seen in the preliminary experiment (Table 4.1), fact that is due to the filtering process that occurred in the previous step of the classification chain. However, the data is still not balanced and the probability of making a correct guess randomly is low, with categories Introducing Topic, Concluding topic and Emphasizing having the highest chance (27.4%, 24.4%, and 22.6% respectively) and with Defining and Defending Idea at the bottom (9.8% and 10.1%).

The next part of the table, labeled as *Unigram Model*, contains the results for the simplest model: word unigrams with no context (feature W1_1). This feature was in itself enough to provide information for the classification of the phenomena, achieving a maximum of performance for the categories Defending Idea and Exemplifying (which F1 is about 50 percentage points above what was expected to achieve by chance). Interestingly, three categories did not benefit at all from this simplest model: Concluding topic, Anticipating the Audience's Response, and Emphasizing. In other words, the simplest model here presented performed at the same statistically significant level as chance. This result shows that these categories are the most sensitive to the surrounding context and words alone are not representative of the phenomena at hand.

Finally, to the right, the table shows the results for the best setups on this task. It shows the set of features used (column *ft*), and the results regarding precision, recall, and F1. As before, when two setups produced results at the same statistical significance, simpler and more generic models were chosen, *i.e.*, preferring **(a)** lemmas (over words), **(b)** lower n-gram orders, and **(c)** lower window sizes.

As in the previous step in the filter chain (decide which sentences contained metadiscourse), the top three categories were EXEMPLIFYING, DEFENDING IDEA and INTRODUCING TOPIC (with F1 of 72%, 68%, 66%, respectively). At the lower end of the table there were some variations, but DEFINING, EMPHASIZING, and ANTICIPATING THE AUDIENCE'S RESPONSE are still the least successful. The category which most benefited with the hill climbing technique and the addition of new features configuration was ENUMERATING, with an improvement of 17% in F1 when compared to the simple unigram model.

In what concerns the set of features used to achieve the best performance, there are also some relevant considerations. First, the two best performing categories (along with COMMENTING ON LINGUISTIC FORM/MEANING) use the simplest set of features, whether unigrams or bigrams, but not combining them with other configurations. On the other hand, INTRODUCING TOPIC uses four different sets of features: word unigrams, bigrams, trigrams, and part-of-speech 4-grams. Another interesting consideration is, in fact, the selections of POS as features. This preference only happened for INTRODUCING TOPIC and REFERRING TO PREVIOUS IDEA (in the latter improving as much as 9% over the best solution without the feature), suggesting that there are some syntactic clues associated with metadiscourse for these two categories. Another surprising result was the presence of the named entities feature (`ner`) for the category CONCLUDING TOPIC, which by itself brought the F1 measure from 31% to 37%.

Finally, it is also noteworthy that the categories for which the unigram model did not add any performance when compared to chance (`CONC`, `ANT`, and `EMPH`, as mentioned previously) were the ones which explored the window size parameter the most, using the surrounding three or four words to support classification.

4.2.2.2 Word Embeddings

Based on the famous premise by Firth (1957) – *"you shall know a word by the company it keeps"* –, recent research in word representation shows that the context similarity of words related to the semantic relation between them (Collobert et al., 2011; Mikolov et al., 2013).

This strategy, commonly referred as word embeddings, consists of representing of words as vectors of real numbers, with a fixed length. The mapping between a word and its vector is learned by observing the contexts in a training set (typically in the order of billions of words). The main advantage of such approach is the reduction in vocabulary size. I.e., assuming the preservation of the semantic relationship between words, it is possible to reduce the vocabulary size from the number of words in the English language to a controlled parameter (the chosen size of the vectors).

This technique has been used recently and proven successful in several NLP related tasks, such as machine translation (Ling and Ist, 2013; Bojar et al., 2016), sentiment analysis (Tang et al., 2014; Rosenthal et al., 2017), dependency parsing (Chen and Manning, 2014), or topic modeling (Das et al., 2015).

The final set of experiments is therefore intended to investigate in what way can word embeddings provide information to identify which words in a sentence are metadiscursive. This approach contrasts with the previous formulation in the sense that the set of features are now more close to the semantics (instead of lexical and syntactic clues).

For this reason, to focus on the feature change in itself, the machine learning mechanism is left unchanged. Thus, the new features would ideally be introduced directly in the CRF architecture.

Such configuration, however, poses a problem, since CRFs are typically trained with discrete data (and, as mentioned before, words are now represented by vectors of real numbers). Combining CRFs with word embeddings has however been addressed in previous research (Turian et al., 2010; Tao et al., 2017). In the current experiment, this problem was overcome by taking advantage of feature weighting capabilities. More precisely, each dimension of the word vector is a feature weighted by its corresponding value.

```
INTRO   w.0.0:0.6120    w.0.1:0.5592    ...     w.0.49:0.9123
        w.1.0:0.2322    w.1.1:0.5491    ...     w.1.49:-0.0098
        w.-1.0:0.7152   w.-1.1:-0.3822  ...     w.-1.49:0.8353
INTRO   w.0.0:0.2322    w.0.1:0.5491    ...     w.0.49:-0.0098
        w.1.0:-0.1153   w.1.1:0.8204    ...     w.1.49:-0.0021
        w.-1.0:0.6120   w.-1.1:0.5592   ...     w.-1.49:0.9123
```

Figure 4.8: Simplified representation of two words for training the CRF with word embeddings for INTRODUCING TOPIC.

To better illustrate how this proposed setup represents words, Figure 4.8 shows two simplified training instances for the category INTRODUCING TOPIC. The figure represents two words that are encoded in the training data as vectors of real numbers (of size 50). The name of each feature is in the form `w.<window_value>.<vector_index>`. Consequently, for example, `w.0.0` symbolizes the current word and its first dimension in the array, and `w.-1.49` is related to the 49^{th} dimension of the previous word. Therefore, all words have exactly the same features. The differentiation between each item comes from the weight associated with it. In the figure, this weight is located after the semicolon character and takes values from -1 to 1.

Another consideration regarding this experiment is the set of word embeddings to use. There are several pre-trained vectors available such as the most known word2vec (Mikolov et al., 2013), Stanford's GloVe (Pennington et al., 2014), or LexVec (Salle et al., 2016). For the purposes of this work, the GloVe embeddings were preferred to the remaining solutions. This decision has to do with both the form and variety that GloVe provides when compared to the remaining options.

First, GloVe provides vectors based on vocabulary that is uncased. This property is relevant given the task at hand: identifying metadiscourse in transcripts of talks, which are not expected to have casing information. A cased solution (such as word2vec) would require manipulating the vectors to obtain only one vector for the two versions of the same word (either by choosing the one that was uncased or by averaging the dimensions of both versions).

The second reason why GloVe was adopted is that it is composed of different sets, of different vector sizes and different data sources. More precisely it provides four sets:

- Wikipedia 2014[5] + Gigaword 5[6] – trained on a collection of 6B tokens, with vocabulary size of 400K words, uncased, and with vectors of 50, 100, 200, and 300 dimensions;
- Common Crawl[7] (uncased) – trained on a collection of 42B tokens, 1.9M words of vocabulary, and vectors of size 300;
- Common Crawl (cased) – trained on a collection of 840B tokens, with 2.2M words and vectors of 300 dimensions;
- Twitter[8] – trained on 2B tweets (27B tokens), 1.2M words in the vocabulary, cased, and vectors of 25, 50, 100, and 200 dimensions.

For the reasons above, only the two first items of the list were explored (the first one in all versions of vector dimensions). The training data for the metadiscourse classifiers was built by directly looking up in a word/vector file, filing up the weights for each dimension accordingly. In the case of a miss (Out Of Vocabulary), the word is represented by a value of zero in all dimensions.

Finally, aside from which sets to use (with their differences in data source and vector size), the window size around the current word was also varied. All experiments were carried out with 10-fold cross-validation.

Table 4.5 shows the most relevant results of this experiment. More precisely, on the left, it shows the best results from the previous experiment (with lexical and syntactic features), and the results for three experiments varying the size of the vectors (50, 200, and 300 dimensions) with respect to the *Wikipedia 2014 + Gigaword 5* set. The value for the window size is constant and equal to three.

[5] https://corpus.byu.edu/wiki/

[6] https://catalog.ldc.upenn.edu/ldc2011t07

[7] http://commoncrawl.org/

[8] https://twitter.com/

Cat	Previous Best			50_w3			200_w3			300_w3		
	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1
EXMPL	0.70	0.75	0.72	0.54	0.19	0.28	0.59	0.54	0.56	0.59	0.58	0.58
DEFND	0.60	0.79	0.68	0.65	0.36	0.46	0.57	0.52	0.54	0.55	0.51	0.53
INTRO	0.53	0.86	0.66	0.54	0.35	0.42	0.57	0.55	0.56	0.53	0.53	0.53
ENUM	0.42	0.62	0.50	0.50	0.08	0.14	0.44	0.26	0.33	0.43	0.30	0.35
REFER	0.47	0.59	0.48	0.32	0.26	0.29	0.38	0.24	0.29	0.37	0.21	0.27
COM	0.34	0.37	0.35	0.04	0.01	0.02	0.27	0.17	0.21	0.32	0.27	0.29
CONC	0.49	0.30	0.37	0.16	0.08	0.11	0.46	0.29	0.36	0.50	0.37	**0.43**
ANT	0.38	0.29	0.33	0.20	0.01	0.02	0.21	0.13	0.16	0.23	0.21	0.23
EMPH	0.21	0.64	0.32	0.36	0.03	0.06	0.30	0.13	0.18	0.20	0.19	0.19
DEF	0.21	0.42	0.28	0.00	0.00	0.00	0.12	0.08	0.10	0.15	0.08	0.10

Table 4.5: Experiments with embeddings for word-level classification, varying vector size.

It is important to clarify that *Common Crawl* data and different values for window size were also tried out, as mentioned. However, since these experiments did not produce any statistically significant improvement over the previously reported lexical setup, they were omitted here for the sake of simplicity.

The first observation from Table 4.5 is that, in general, word embeddings seem to provide some information to identify metadiscourse, as for the most part, classifiers perform above what could be expected by chance. However, this was not true for one category in particular: EMPHASIZING (where the probability of getting an item correct by chance is 22.6%). It also performed only marginally better than chance (not statistically significant) for ANTICIPATING THE AUDIENCE'S RESPONSE and DEFINING (where chance is 21.9% and 9.8%, respectively).

Even though word embeddings can distinguish between words that are metadiscursive and words that are not, this specific setup only performed better than its lexical counterpart for the category CONCLUDING TOPIC: an improvement of 6% on the setup with 300 dimensions. Interestingly, this was the category that, in the previous experiment, also shown significant improvement when considering a feature related to named entities. This result shows that, for this category, semantic information helps to decide whether a given word materializes the metadiscursive function.

More generally, it is also possible to observe that, with except for three categories, increasing the size of the vectors improved the overall classification. The three exceptions are DEFENDING IDEA, INTRODUCING TOPIC, and REFERRING TO PREVIOUS IDEA, where F1 decreases between the experiment with 200 and 300 dimensions.

It is also interesting to see that the main impact on the F1 is caused by recall. In fact, precision stays about the same throughout the results reported in the table. In sum, it seems that using word embeddings in this configuration to label which words are metadiscursive, suffers from recall.

4.3 Discussion

This chapter described the process of building a set of classifiers to identify and classify metadiscourse as used in spoken language in the context of oral presentations. By taking a supervised approach, such classifiers took advantage of the corpus collected in this thesis – METATED – which was presented and discussed in detail in Chapter 3.

A first preliminary experiment took place where the ultimate goal of identifying which words in a talk have metadiscursive connotation was tackled directly. Soon this approach revealed the problem of dealing with hugely imbalanced training and test sets.

With this realization, a cascade of classifiers was set up. Composed of two levels, this chain of classifiers broke down the process of identifying metadiscursive words in a talk in two different tasks. First, given a talk, produce a set of sentences that are predicted to have each of the metadiscursive acts at hand. Second, given these candidate sentences, select the words in them (if any) that are having a metadiscursive role in the discourse.

By dividing the task into these two steps, it was possible to address some of the problems of the higher-end goal carefully. In the first classification step, the issues of imbalance of data were dealt with, taking advantage of the cost mechanism provided by the SVMs. Still, at the sentence level classification, it was possible to take advantage of the training data particularities (the way that it was obtained via crowdsourcing), going beyond the traditional majority vote technique. This strategy had a significant impact on the recall metric of the solutions achieved which, for a classification step that is acting as a filter for the next level, has a significant impact in the performance of the overall system.

At word-level classification, it was possible to revisit the task of pinpointing the words in a talk that are genuinely conveying metadiscourse, now in a much more controlled environment of already labeled sentences as potentially containing the phenomenon. On this front, it was possible to explore two different sets of features: the classic lexical approach (n-grams), and word embeddings.

Cat	Preliminary			Sentence-Level			Word-Level			Overall		
	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1
EXMPL	0.51	0.69	0.58	0.72	0.87	0.79	0.70	0.75	0.72	0.70	0.65	0.67
DEFND	0.26	0.60	0.36	0.65	0.65	0.65	0.60	0.79	0.68	0.60	0.51	0.55
INTRO	0.22	0.71	0.33	0.52	0.61	0.56	0.53	0.86	0.66	0.53	0.52	0.52
ENUM	0.21	0.54	0.30	0.47	0.55	0.51	0.42	0.62	0.50	0.42	0.34	0.38
REFER	0.04	0.53	0.08	0.30	0.35	0.32	0.47	0.59	0.48	0.47	0.21	0.29
COM	0.25	0.29	0.27	0.40	0.71	0.51	0.34	0.37	0.35	0.34	0.26	0.29
CONC	0.04	0.44	0.07	0.13	0.47	0.20	0.50	0.37	0.43	0.50	0.17	0.25
EMPH	0.07	0.51	0.12	0.18	0.39	0.25	0.21	0.64	0.32	0.21	0.25	0.23
ANT	0.09	0.51	0.16	0.24	0.42	0.31	0.38	0.29	0.33	0.38	0.12	0.18
DEF	0.04	0.34	0.08	0.19	0.36	0.25	0.21	0.42	0.28	0.21	0.15	0.18

Table 4.6: Summary of classification task results.

Table 4.6 summarizes the results of this task. It shows precision, recall, and F1 for four specific situations. First, *Preliminary* corresponds to the results that were obtained during the preliminary experiment, where a CRF was trained on positive examples only and tested on the full talks. As discussed at the time of the experiment, this setup severely impacted precision, as the distribution of positive examples was very different between train and test sets.

In the middle of the table are the performances of the sentence-level (SVM) and word-level (CRF) classifiers described in this chapter. By comparing the performances of both tasks, it is possible to see that the categories that achieve better F1 in one task are also the ones that are best in the other. The same holds true for the categories towards the bottom of the table. In fact, during the different tasks (annotation and classification), the order in which categories performed has generally been the same. This observation shows the impact of the quality of the annotation during classification.

Finally, the last portion of the table combines the filter chain results and presents the expected performance of the full task of detecting and classifying metadiscourse for each category. Here, the value of the precision is given by the precision achieved in the word-level task, since this will ultimately be what the system retrieves. Recall, however, is the combination of the recall values for both levels of the classification chain. This calculation was done to fairly represent the positive instances that are mistakenly filtered out during the sentence-level classification, and the ones that the word-level step also misses. So, the recall of the overall system takes into consideration these both loss steps. Finally, F1 is recomputed with the new recall values.

The categories that achieved best overall results were EXEMPLIFYING, DEFENDING IDEA, and INTRODUCING TOPIC, with F1 of 0.67, 0.55, and 0.52, respectively. On the other side of the table, the worst performant categories were ANTICIPATING THE AUDIENCE'S RESPONSE and DEFINING (both with $F1 = 0.18$). It is very important to highlight here again the sparsity of these metadiscursive instances. The chances of a word being part of a metadiscursive act are below 1%, often even below 0.2%.

Given the novelty of the task accomplished herein, it is difficult to compare these results with previous research. However, referring back to the background chapter, where some studies on related phenomena were presented, it is possible to see that results are in line with what was expected, and in some cases exceed what was reported.

In what concerns approaches that considered token-level classification, Madnani et al. (2012) achieved an f-measure of 0.55 on the task of deciding if a given word was *shell language* or not. This binary classification was based upon a set of expert labeled *Wikipedia* articles and would be equivalent to the task of deciding if a word was metadiscursive or not (with no further comment on its function).

Regarding approaches that do not focus on token-level classification, but do address issues related to metadiscourse, very different results are reported. Nguyen and Litman (2015) work on argumentative discourse achieve performances of between 0.56 and 0.88 in a 3-category taxonomy, while Cotos and Pendar (2016), in a 13-category theory, achieve performances between 0.31 and 0.88.

This disparity in results between categories of the same taxonomy was also observed in this work, with the sentence-level classification results ranging between 0.20 for CONCLUDING TOPIC and 0.79 for EXEMPLIFYING. In the particular case of the work described in this thesis, such disparities started to be observed at the annotation phase and carried on to the classification stage.

5 Conclusion & Future Work

This thesis worked towards understanding the nature of metadiscourse, specifically regarding how it is used in spoken communication. It provided a broad view of the phenomenon, by looking at it from a functional perspective and analyzing it in a data driven manner.

This systematic approach was composed of three main steps: theoretical background discussion, corpora collection, and development of automatic solutions for classification.

On the first front, the existing taxonomies that encompassed the spoken variety of metadiscourse were presented and compared. The adopted taxonomy resulting from this discussion was the one that stated as a goal to both unify the existing approaches on metadiscourse and provide a functional standpoint of the phenomenon.

Secondly, a large-scale annotation of 16 categories of metadiscourse that served two purposes: build training data for classification and make considerations about understanding of the chosen taxonomy. Different sources of presentations that contained metadiscourse were also looked at. Quality, uniformity, and broad set of topics were some of the properties that lead to the choice of TED talks over classroom recordings. A preliminary annotation task checked for the intersection of the chosen theory and the material of choice. As a result, it was shown that the situational settings in which a presentation occurs determine what type of metadiscourse strategies the speakers use.

The full annotation effort generated a corpus of metadiscourse for 16 categories, annotated at token level – metaTED. In the process of building this corpus, some particularities of the nature of metadiscourse were discussed, such as the amount of context needed to identify occurrences of Concluding topic and the relation between the level of the talk and the presence of metadiscourse.

Finally, the last step consisted of building metadiscursive classifiers, exploring different features and algorithms in a structured chain of supervised classification. As a result, ten classifiers capable of detecting functional categories of metadiscourse in TED talks transcripts were implemented and their performance was discussed.

The next two sections summarize the contributions of the current thesis (Section 5.1) and point to future work directions (Section 5.2).

5.1 Contributions

METATED

The first contribution worth highlighting is METATED – a corpus of metadiscursive use in presentations annotated by the crowd. METATED is composed of 180 TED talk transcripts and 16 categories of metadiscourse, adapted from Ädel (2010). The corpus is freely available through LRE Map in the form of 16 XML files, one per category, where each token of the talk is enriched with information on how many workers selected that token as being part of the metadiscursive act (from 0 to 3).

METATED cannot be interpreted nor used as ground truth for metadiscourse. It should always be referred to as non-expert opinion on the phenomena. Details on the quantity and quality of the data were highlighted. The goal of this corpus is to provide insight on the use of metadiscourse in spoken language in the setting of a presentation.

Metadiscourse Classifiers

A set of 10 metadiscursive classifiers were developed in this work. These classifiers are the result of a classification chain where first, using SVMs a list of candidate sentences that have metadiscourse is generated, and then, through CRFs the exact tokens that are part of the metadiscursive act are extracted.

Given the substantial difference between the categories that were analyzed, the classifiers have different performances: whether because of the number of positive examples available or by the quality of the annotation itself.

Metadiscourse Understanding

In general, both the annotation process and the creation of classifiers targeted explicitly at a fixed set of metadiscursive acts allowed for better understanding of how the phenomenon is used in spoken settings.

Through the process of annotation with non-experts, it was possible to conclude what categories could be better understood and which ones, on the other hand, were more difficult to deal with. Further insight was given by additional statistics, such as time on task and inter-annotator agreement. Particularly impressive were the differences in the number of times the workers requested for additional context while annotating specific categories. This difference is an indicator that not all of the categories require the same amount of context to be understood.

During classification, the algorithms, features, and parameters also gave insight on the differences between the acts analyzed. Some seemed to rely only on lexical cues to be identified, while others took advantage of the semantic information. The order of the n-grams and the cutoff parameter were also analyzed, and the best solutions for the different categories were comprised of variations of these two measures.

The building of METATED also allowed to address one area that the literature on the theory of metadiscourse highlights as necessary even though not much work can be found on it, *i.e.*, its relation with the lexical level of the content for which the metadiscourse is used. An analysis of the distribution of metadiscourse across different vocabulary levels of proficiency (Correia et al., 2015) showed that different markers are used in different ways as the level of the discourse is higher or lower. Furthermore, these differences were approached at two levels: whole talk and 500-word segment. This latter analysis allowed to see that some acts are used to refer to something that is in close context, while others refer to something that is further (or back) in the discourse.

5.2 Future Work

Data Enhancement

As mentioned above, during the course of this work, a corpus of metadiscursive phenomena in TED talks was built. However, as discussed in the annotation section, the statistics regarding agreement showed some difficulties for non-experts to execute the task. In this sense, METATED can be build upon by submitting it to an exhaustive validation task (either by experts or non-experts) in order to filter out noisy data points and come up with an overall more robust set of annotations.

This should be a first step towards improving the models developed herein, since the training data contained contradicting instances that were not possible to identify automatically.

Theory Refining

The metadiscursive theory used in this work was chosen given two main criteria: its unification of different taxonomies on the same phenomenon and its functional approach. Throughout the course of the work, the taxonomy was adjusted according to the specific setup for which it was being used, by either merging categories together, separating them into different concepts, or renaming them. These decisions were all supported by the need to reduce the cognitive load of the non-experts while annotating metadiscourse.

For some of these new categories, both annotation and classification performances were lacking. This indicates that, while a taxonomy may comprehensively represent a phenomenon, it might not be suitable to be used in a systematic approach as the one carried out.

Further refining of the taxonomy can help improve data quality and classification performance by continuing to divide or aggregate concepts, leading to less ambiguous classes. As an example, definitions were used as presented in the original taxonomy, while in the automatic classification phase it was identified that, in fact, two concepts were at stake: defining and naming.

Multimodal Approach

When using TED talks, aside from text, one has access to two other dimensions: audio and video. The literature on discourse structure and topic segmentation gives some insights that these dimensions also contain metadiscursive clues.

Cassell et al. (2001) and Hyland and Guinda (2012) for example show how changes in the topic might correspond to changes in physical posture of the speaker or even the audience. Hirschberg and Nakatani (1998) looked at how acoustic indicators can predict topic frontiers, and Passonneau and Litman (1997) concluded how pauses patterns could help in the task of topic segmentation. Purver (2011) summarized these results stating that people tend to pause for longer than usual just before moving to a new segment and that speakers tend to speed up, speak louder and make fewer pauses when starting a new section. These observations may not only apply to topic segmentation (such as the categories INTRODUCING TOPIC and CONCLUDING TOPIC) but can also be indicators of other categories like EXEMPLIFYING or EMPHASIZING. For EMPHASIZING, in particular, studies in the area of speech synthesis manipulate pitch to approximate the synthesized speech to what humans do when emphasizing (Raux and Black, 2003).

Exploring Additional Features

This thesis looked at metadiscourse from a broad perspective, aiming at understanding the phenomenon as a whole and how it plays in the structure of spoken discourse. The drawback of such an approach, however, is that there was no focus to specific acts in particular. Instead, it presented a category independent classification with a general set of features.

Therefore, category specific features should be able leverage the performance of the models. Examples include:

- Verb Tense – looking at the tense being used in the discourse would be particularly suited to strategies related to conclusions and reviews, where speakers tend to use the past;

- TF-IDF – the term frequency and documents frequency can help identify keywords (such as words being defined or introduced) from words that are being used to compose the discourse;
- Intonation – as mentioned before, audio features, and most specifically intonation, can help signal change of topic or emphasis;
- Dependencies – analyzing sentence dependencies can help understand the structure of more complex strategies, such as clarifications.

Interaction with other NLP tasks

It is also left as future work the task of measuring how metadiscourse in general, and its automatic classification strategies in particular, can enhance the performance of other common Natural Language Processing tasks.

The classifiers built as a result of this work can be used to improve summarization tasks, for instance, by removing examples and making sure that parts that are emphasized by the speaker are included. There is also a connection between the use of metadiscourse and topic segmentation, where introductions, conclusions or delimitations give clues to the boundaries of micro topics in the same talk. Finally, the analysis of metadiscourse could be used to improve machine translation, since knowing the existence of an act can generate a lookup in a database of metadiscursive expressions for the same function.

Presentation Skills Instruction Tool

Another future application of the work accomplished in his thesis is the building of a presentation skills tutor. It was shown that non-experts understand the notion of metadiscourse. Therefore, a potential presentation skills tool could use the idea of metadiscourse as learning goals, following the idea that mastering metadiscourse can help expressing and defending a point of view. Students would be able to focus on several categories of metadiscourse, watch professional speakers using them in different contexts, and ultimately create a model that they can use for future presentation opportunities.

The literature on this topic has shown that explicit instruction of presentational skills is needed since students do not intuitively recognize the value of such skills (Börstler and Johansson, 1998; Pittenger et al., 2004). However, few individuals are exposed to courses that specifically target presentational skills. These abilities are often developed simultaneously as the core skills, with students being asked to present course-related topics or results from a class project (Kerby and Romine, 2009). This trial and error instruction of presentation skills has proven to fail when there is no explicitly targeted feedback at the presentation component (De Grez et al., 2009a).

De Grez et al. (2009a) stressed how making discourse concepts explicit could improve presentation skills instruction. The authors found that students, when presented merely with strict rules, do not change their presentations according to the context they are in. Therefore, students should be introduced to adequately explained concepts, allowing them to adapt according to their needs. Presenting these concepts and showing them in different contexts and realizations delegates on the students the responsibility to extrapolate and formulate models tailored to their own reality and needs. Additionally, Haber and Lingard (2001) supports a technologic approach to presentation skills instruction, defending creative control over the contents, activities that integrate text and images, and engagement with different types of media.

The tools which resulted from this work can then be a first step towards the development of presentation skills curricula, by providing efficient ways to tag metadiscourse in large amounts of data. They can be both used to highlight metadiscourse in documents that the students themselves can choose to watch, and to automatically generate exercises on those same documents.

I
Appendices

A Top N-Grams

A.1 Preliminary Annotation

		n-grams	
	1	**2**	**3**
TED talks	the	of the	a lot of
	and	in the	this is a
	to	this is	one of the
	of	and I	I am going
	a	going to	am going to
	that	and the	this is the
	in	to be	I want to
	is	to the	and this is
	I	on the	you can see
	you	is a	going to be
Introducing Topic	to	going to	I am going
	you	I am	am going to
	I	am going	I want to
	about	want to	**to talk about**
	going	I want	**to tell you**
	want	**to talk**	I would like
	talk	**talk about**	would like to
	show	**tell you**	**to show you**
	tell	**show you**	**want to talk**
	like	like to	**going to talk**
Concluding topic	to	**leave you**	**leave you with**
	you	you with	I would like
	I	like to	would like to
	with	want to	**I will leave**
	so	I would	**will leave you**
	and	would like	I want to
	the	I will	**to leave you**
	leave	**will leave**	**the last thing**
	like	I want	**like to leave**
	that	**to conclude**	**like to conclude**

Table A.1: Ranked top 10 n-grams obtained during the preliminary annotation for the categories Introducing Topic and Concluding topic. Top 10 n-grams of the entire set of TED talks provided in the first line for reference.

		n-grams	
	1	**2**	**3**
TED talks	the	of the	a lot of
	and	in the	this is a
	to	this is	one of the
	of	and I	I am going
	a	going to	am going to
	that	and the	this is the
	in	to be	I want to
	is	to the	and this is
	I	on the	you can see
	you	is a	going to be
EXEMPLIFYING	**example**	**for example**	**you an example**
	for	**an example**	give you an
	imagine	**example of**	**is an example**
	you	give you	**an example of**
	of	**for instance**	I will give
	an	**example for**	will give you
	a	**look at**	let me give
	is	this is	me give you
	examples	you an	this is an
	to	if you	to give you
EMPHASIZING	is	I want	I want to
	the	this is	want you to
	to	want to	I want you
	important	I think	**the most important**
	I	is that	**this is important**
	this	**is important**	one of the
	that	**very important**	**to point out**
	you	you to	**the bottom line**
	want	**the most**	**you to remember**
	what	**most important**	**we need to**

Table A.2: Ranked top 10 n-grams obtained during the preliminary annotation for the categories EXEMPLIFYING and EMPHASIZING. Top 10 n-grams of the entire set of TED talks provided in the first line for reference.

A.2 Full Annotation

		n-grams	
	1	**2**	**3**
Repair & Reformulating	in	in fact	in other words
	actually	other words	if you want
	fact	in other	or more precisely
	well	you want	
	other	if you	
	words	i mean	
	then	but actually	
	or	or actually	
	you	or more	
	say	no no	
Commenting on Linguistic Form/Meaning	means	the word	in other words
	word	other words	does it mean
	words	to write	with the word
	the	in other	what does it
	it	that means	what it means
	that	it means	in his writeup
	mean	three words	would mean that
	write	means that	in the words
	what	is a	is a word
	is	what it	there are words
Clarifying	I	I mean	it's not just
	not	it's not	this is not
	mean	is not	what I mean
	but	this is	in other words
	it's	not just	but it's not
	that	I'm not	I don't mean
	means	what I	it's not that
	is	that means	I'm not saying
	it	it means	what that means
	what	but it's	I want to
Defining	is	this is	what that means
	means	which is	what I call
	called	I call	this is called
	which	we call	that means is
	this	that is	in other words
	that	which means	is something called
	it's	that means	this means that
	call	called the	
	what	it means	
	the	what that	

Table A.3: Ranked top 10 n-grams obtained during annotation for Repair & Reformulating, Commenting on Linguistic Form/Meaning, Clarifying, and Defining.

		n-grams	
	1	**2**	**3**
INTRODUCING TOPIC	to	going to	I'm going to
	you	I'm going	I want to
	I'm	want to	to show you
	I	show you	going to show
	going	I want	to tell you
	about	tell you	to talk about
	show	to talk	I'd like to
	want	let me	going to talk
	talk	to show	let me show
	tell	talk about	me show you
DELIMITING TOPIC	I	I'm not	don't have time
	to	go into	I will not
	not	going to	time to go
	I'm	I won't	I don't have
	into	time to	not going to
	time	don't have	not here to
	go	I don't	I'm not going
	don't	I can't	I'm not here
	have	have time	I'm going to
	it	to go	to go into
ADDING INFORMATION	I	the way	by the way
	to	by the	I want to
	the	want to	want to say
	by	I want	I have to
	way	to say	to point out
	say	like to	I should say
	just	let me	I'd like to
	want	tell you	have to say
	and	have to	to tell you
	in	I just	like to point
CONCLUDING TOPIC	to	leave you	leave you with
	I	you with	to leave you
	with	to leave	I want to
	you	let me	I'm going to
	so	the final	want to leave
	the	want to	going to leave
	leave	I want	so that is
	that	to conclude	let me finish
	last	to finish	finish up with
	and	going to	like to finish

Table A.4: Ranked top 10 n-grams obtained during annotation for INTRODUCING TOPIC, DELIMITING TOPIC, ADDING INFORMATION, and CONCLUDING TOPIC.

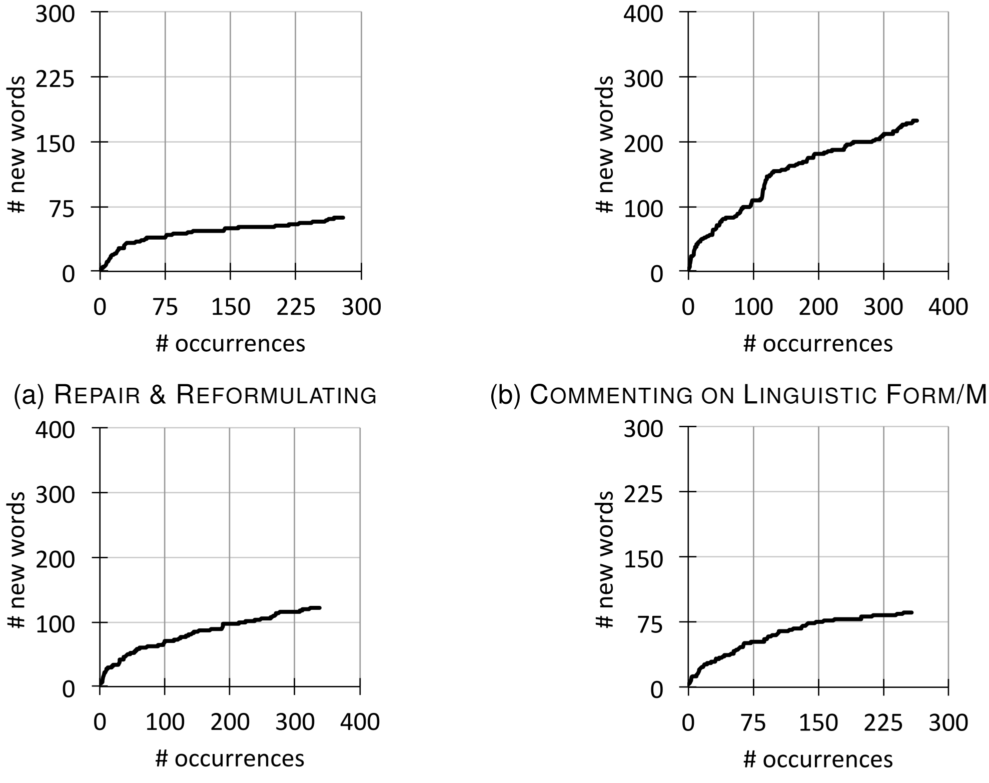
	n-grams		
	1	**2**	**3**
ENUMERATING	the first one two second is to three of and	the first the second of all want to I want first of the next there are let me going to	I want to first of all I'm going to the first is the first thing there are two let me just to talk about the second is I'd like to
POSTPONING TOPIC	to I you that before a but show about in	before I but before in a show you tell you a moment to that a little but first back to	but before I I'm going to I tell you in a moment before I go before I tell I show you for a moment before I show come back to
RECAPITULATING	again I as so to said and back the mentioned	as I and again I said I mentioned so again to the go back	as I said as I mentioned
REFERRING TO PREVIOUS IDEA	I you said as the about that I've to heard	I said as I told you I told I mentioned back to that I I showed showed you we heard	as I said I told you I showed you told you about at the beginning we were talking as I mentioned as we heard said at the just told you

Table A.5: Ranked top 10 n-grams obtained during annotation for ENUMERATING, POSTPONING TOPIC, RECAPITULATING, and REFERRING TO PREVIOUS IDEA.

	n-grams		
	1	**2**	**3**
DEFENDING IDEA	I	I think	what that means
	think	I believe	that means is
	that	I mean	that is why
	is	is that	it means that
	believe	that means	the answer is
	the	I know	I was thinking
	mean	the answer	the truth is
	means	we think	what I mean
	we	we believe	I would say
	to	means that	like to think
EXEMPLIFYING	example	for example	you an example
	for	an example	you can imagine
	imagine	give you	give you an
	you	for instance	is an example
	an	look at	here's an example
	examples	one example	an example of
	of	can imagine	this is an
	can	example of	I'll give you
	instance	you can	me give you
	give	you an	let me give
ANTICIPATING THE AUDIENCE'S RESPONSE	you	you know	you can see
	know	you can	many of you
	think	of you	you can imagine
	of	you might	some of you
	to	can see	you all know
	I	if you	as you can
	can	you think	you're going to
	see	as you	a lot of
	we	you may	you might think
	that	all know	you want to
EMPHASIZING	is	this is	I want to
	the	I want	the most important
	to	want to	one of the
	I	most important	I'm going to
	this	I think	of the most
	important	the most	this is the
	you	is the	want you to
	what	let me	the interesting thing
	that	to understand	a very important
	of	very important	what's interesting is

Table A.6: Ranked top 10 n-grams obtained during annotation for DEFENDING IDEA, EXEMPLIFYING, ANTICIPATING THE AUDIENCE'S RESPONSE, and EMPHASIZING.

B Type-token Curves



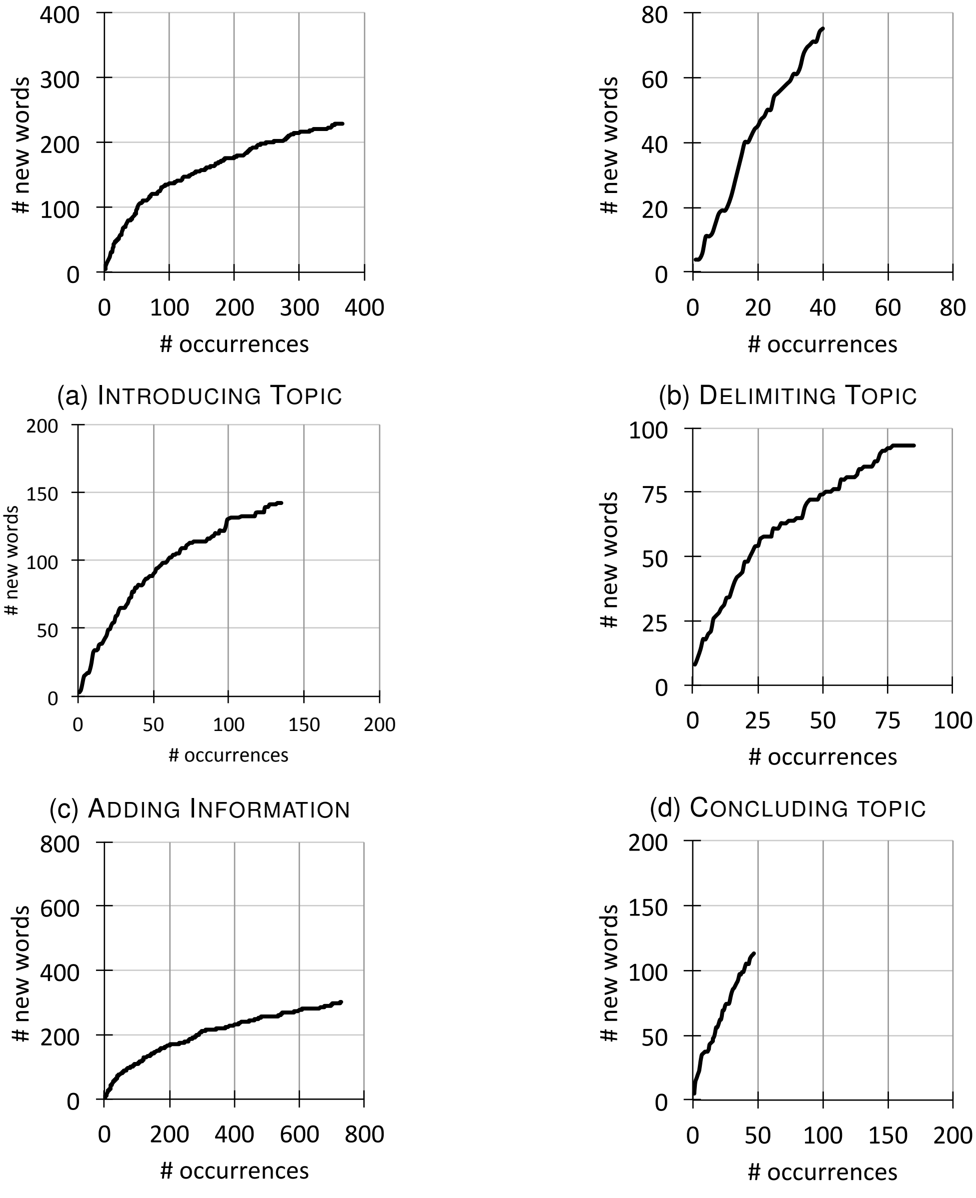
(a) Repair & Reformulating

(b) Commenting on Linguistic Form/Meaning

(c) Clarifying

(d) Defining

Figure B.1: Type-token curves for R&R, COM, CLAR, and DEF.



(a) INTRODUCING TOPIC

(b) DELIMITING TOPIC

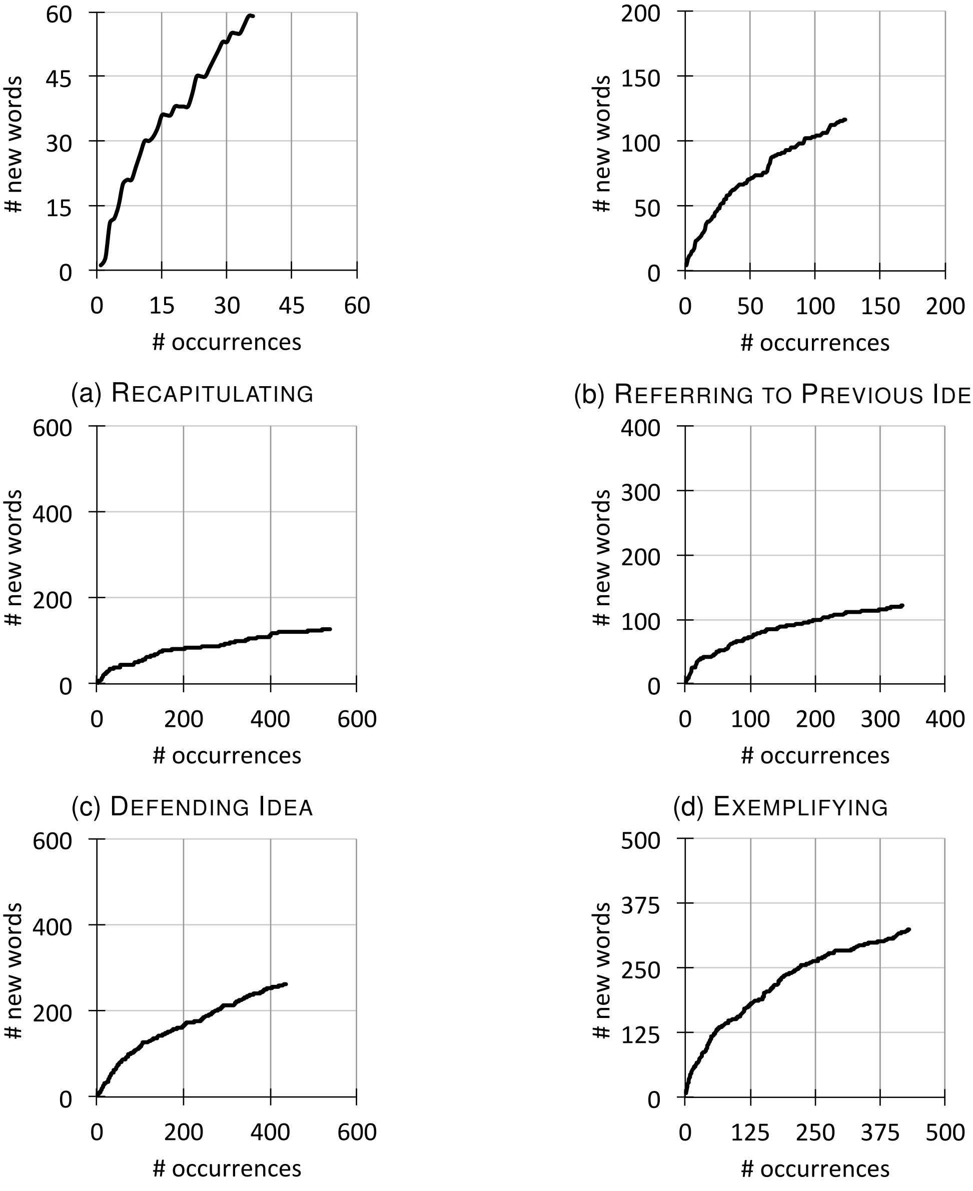
(c) ADDING INFORMATION

(d) CONCLUDING TOPIC

(e) ENUMERATING

(f) POSTPONING TOPIC

Figure B.2: Type-token curves for INTRO, DELIM, ADD, CONC, ENUM, and POST.



(a) RECAPITULATING

(b) REFERRING TO PREVIOUS IDEA

(c) DEFENDING IDEA

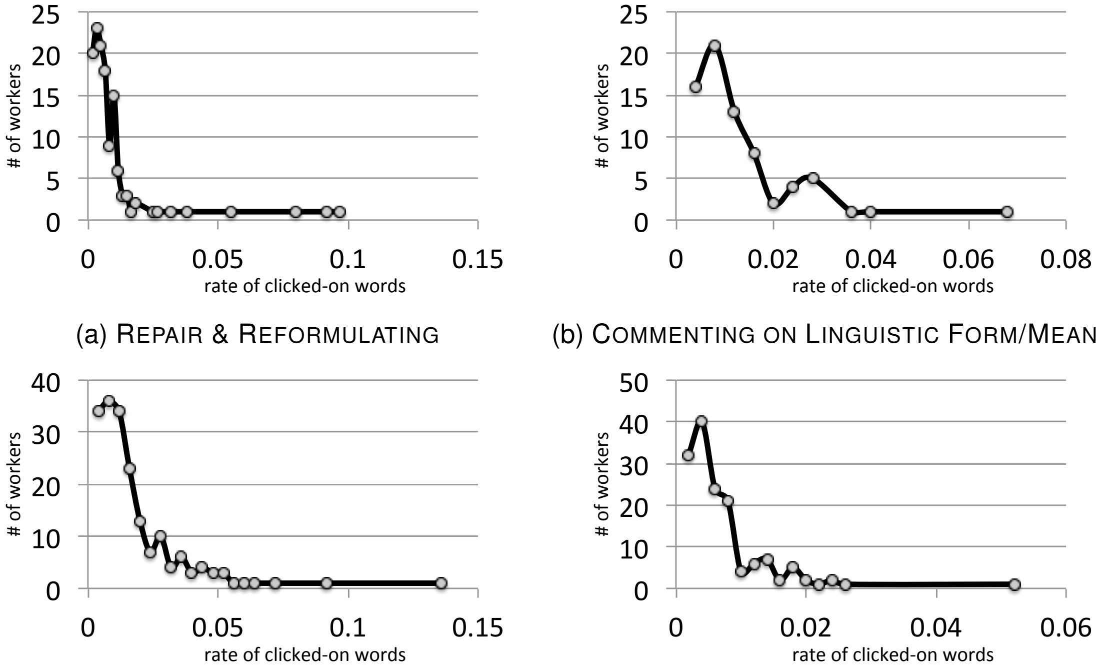
(d) EXEMPLIFYING

(e) ANTICIPATING THE AUDIENCE'S RESPONSE

(f) EMPHASIZING

Figure B.3: Type-token curves for RECAP, REFER, DEFND, EXMPL, ANT, and EMPH.

C Clicking Rate



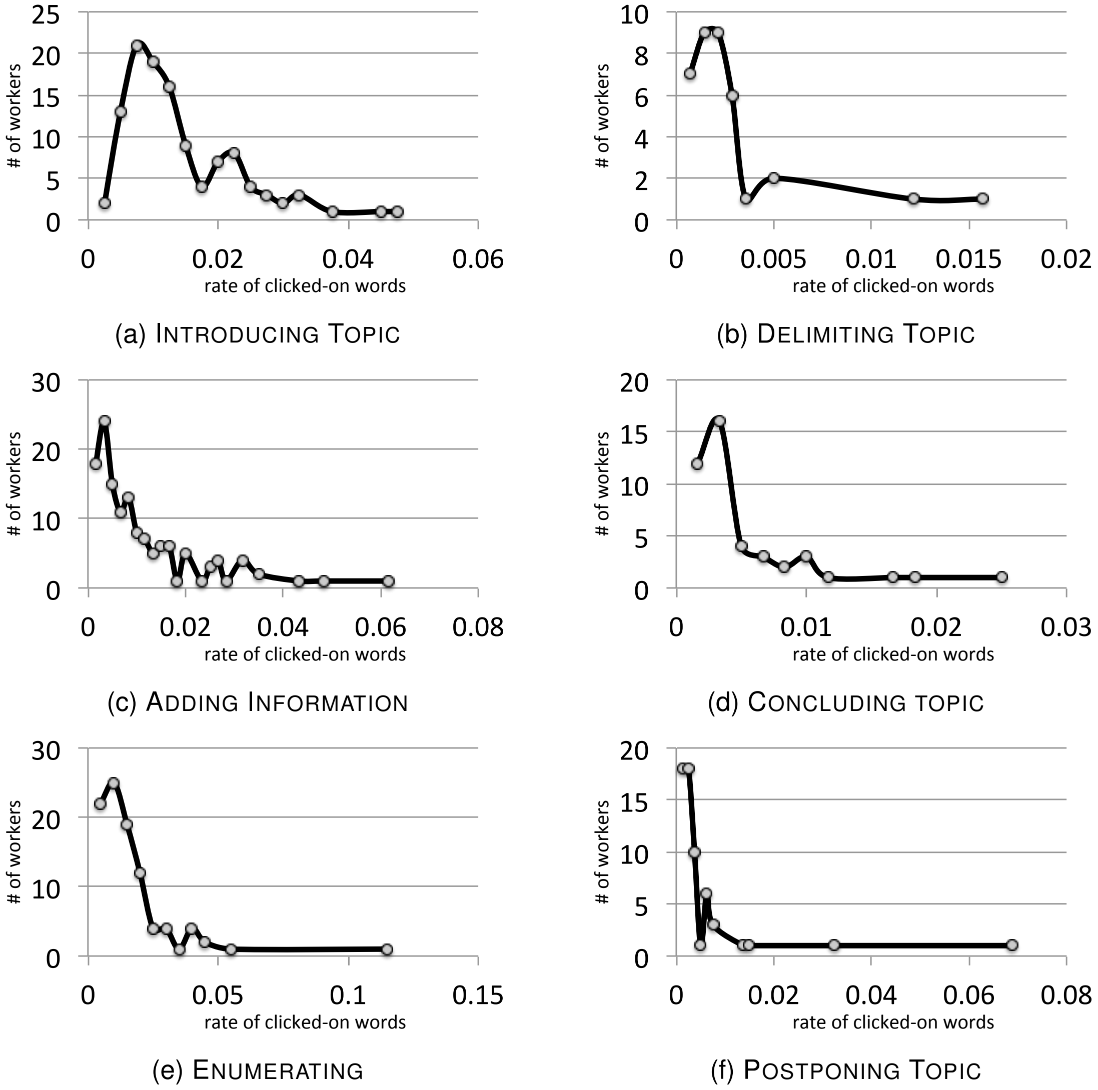
(a) Repair & Reformulating

(b) Commenting on Linguistic Form/Meaning

(c) Clarifying

(d) Defining

Figure C.1: Distribution of selected-words rate between annotators for `R&R`, `COM`, `CLAR`, and `DEF`.



(a) Introducing Topic

(b) Delimiting Topic

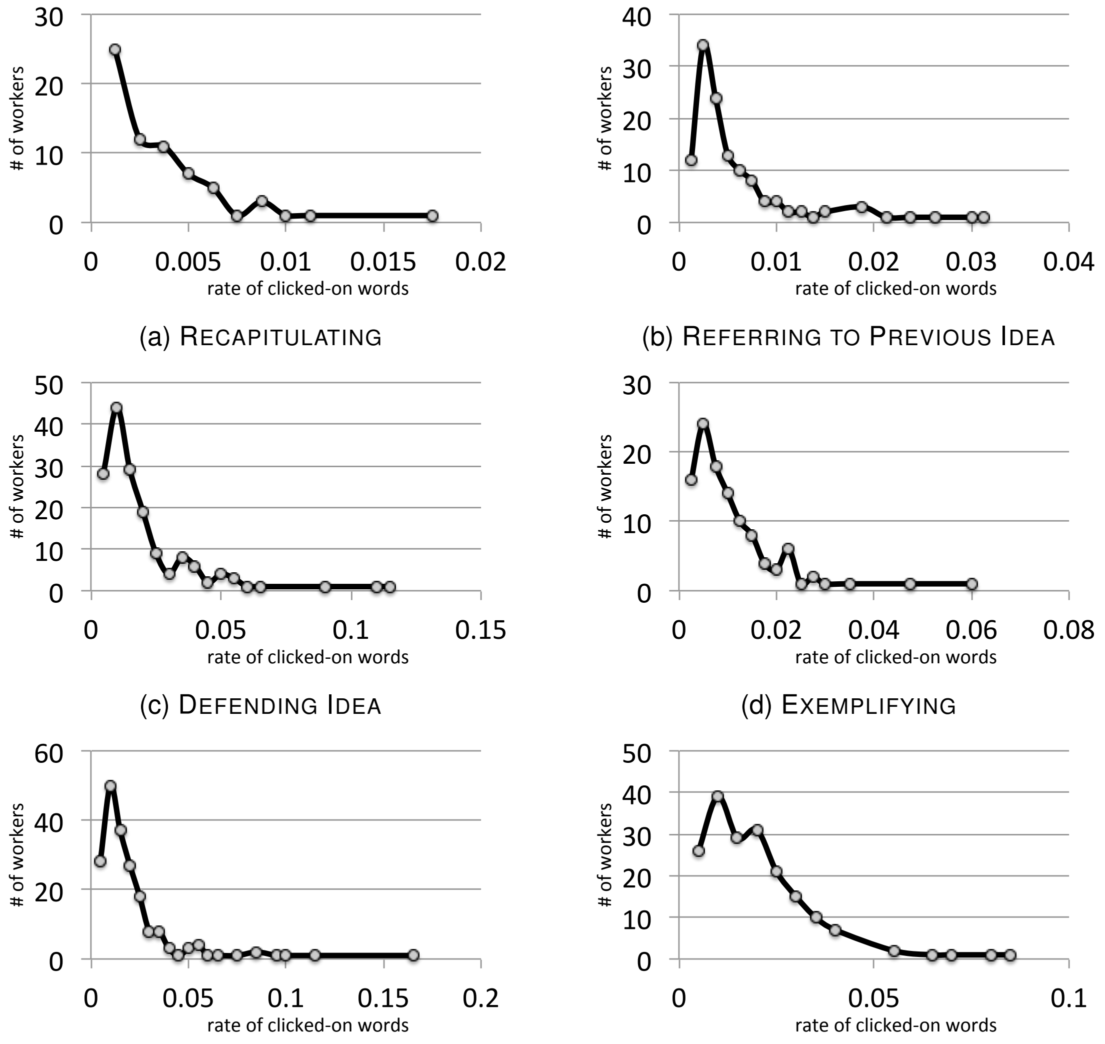
(c) Adding Information

(d) Concluding topic

(e) Enumerating

(f) Postponing Topic

Figure C.2: Distribution of selected-words rate between annotators for `INTRO`, `DELIM`, `ADD`, `CONC`, `ENUM`, and `POST`.



(a) Recapitulating

(b) Referring to Previous Idea

(c) Defending Idea

(d) Exemplifying

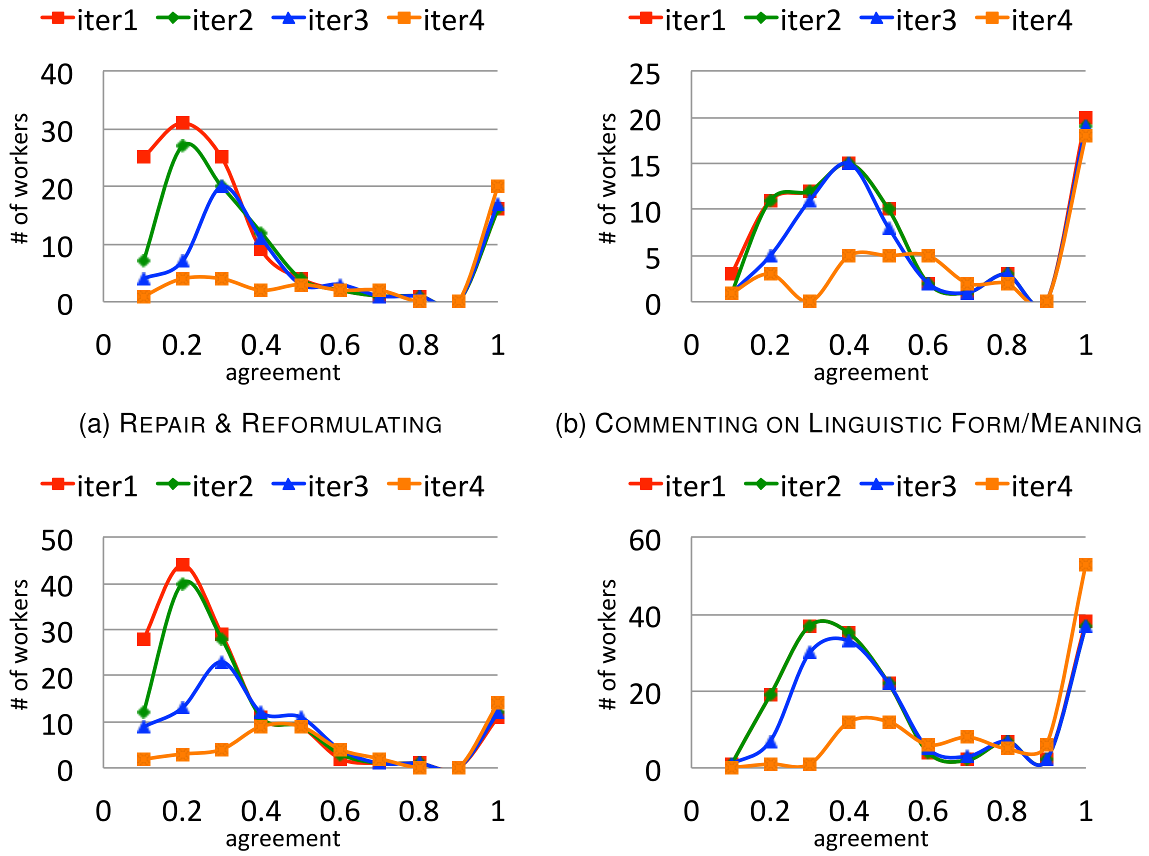
(e) Anticipating the Audience's Response

(f) Emphasizing

Figure C.3: Distribution of selected-words rate between annotators for RECAP, REFER, DEFND, EXMPL, ANT, and EMPH.

D

Agreement Filter



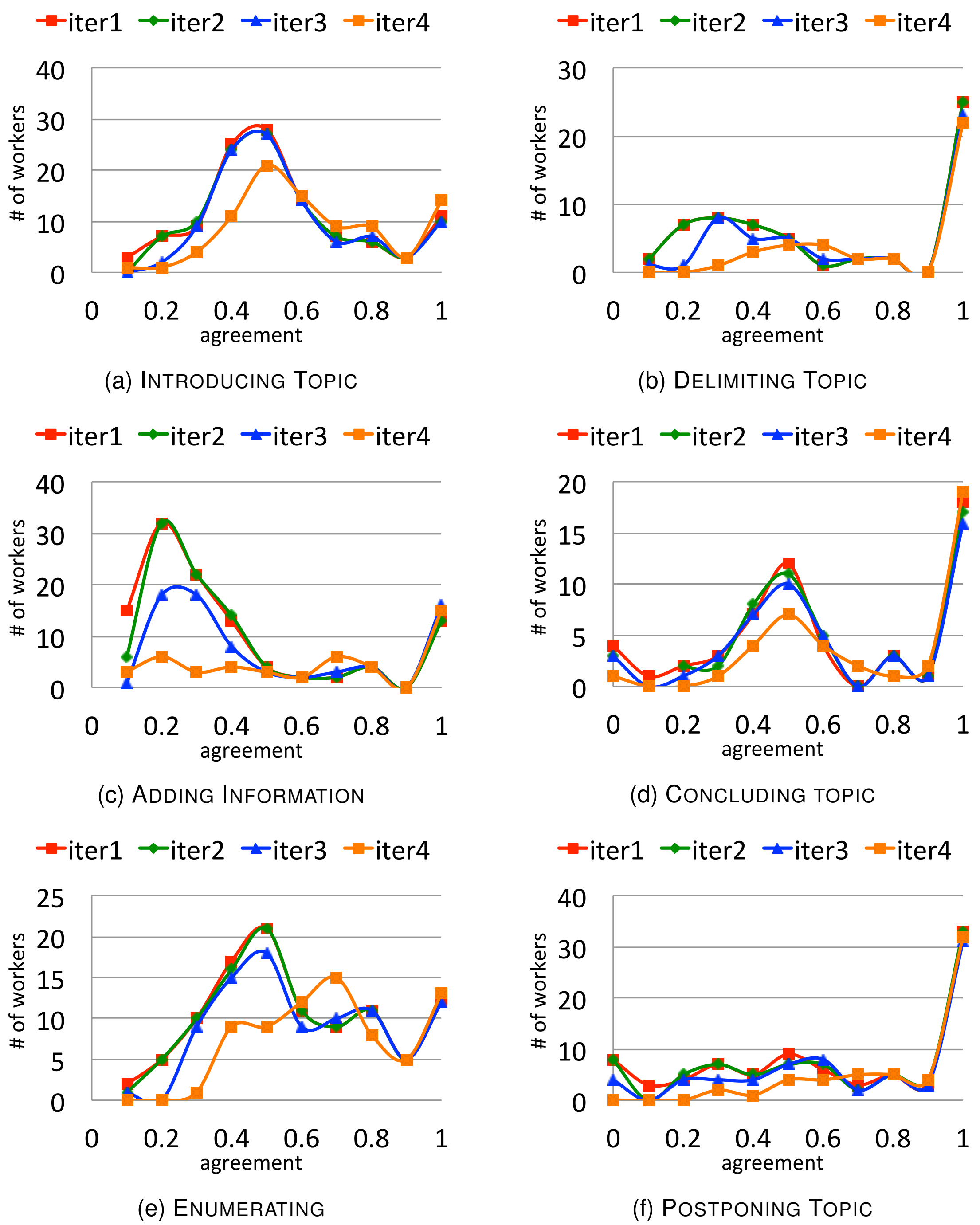
(a) Repair & Reformulating

(b) Commenting on Linguistic Form/Meaning

(c) Clarifying

(d) Defining

Figure D.1: Distribution of agreement between annotators, for four iterations of the filter strategy for `R&R`, `COM`, `CLAR`, and `DEF`.



(a) INTRODUCING TOPIC

(b) DELIMITING TOPIC

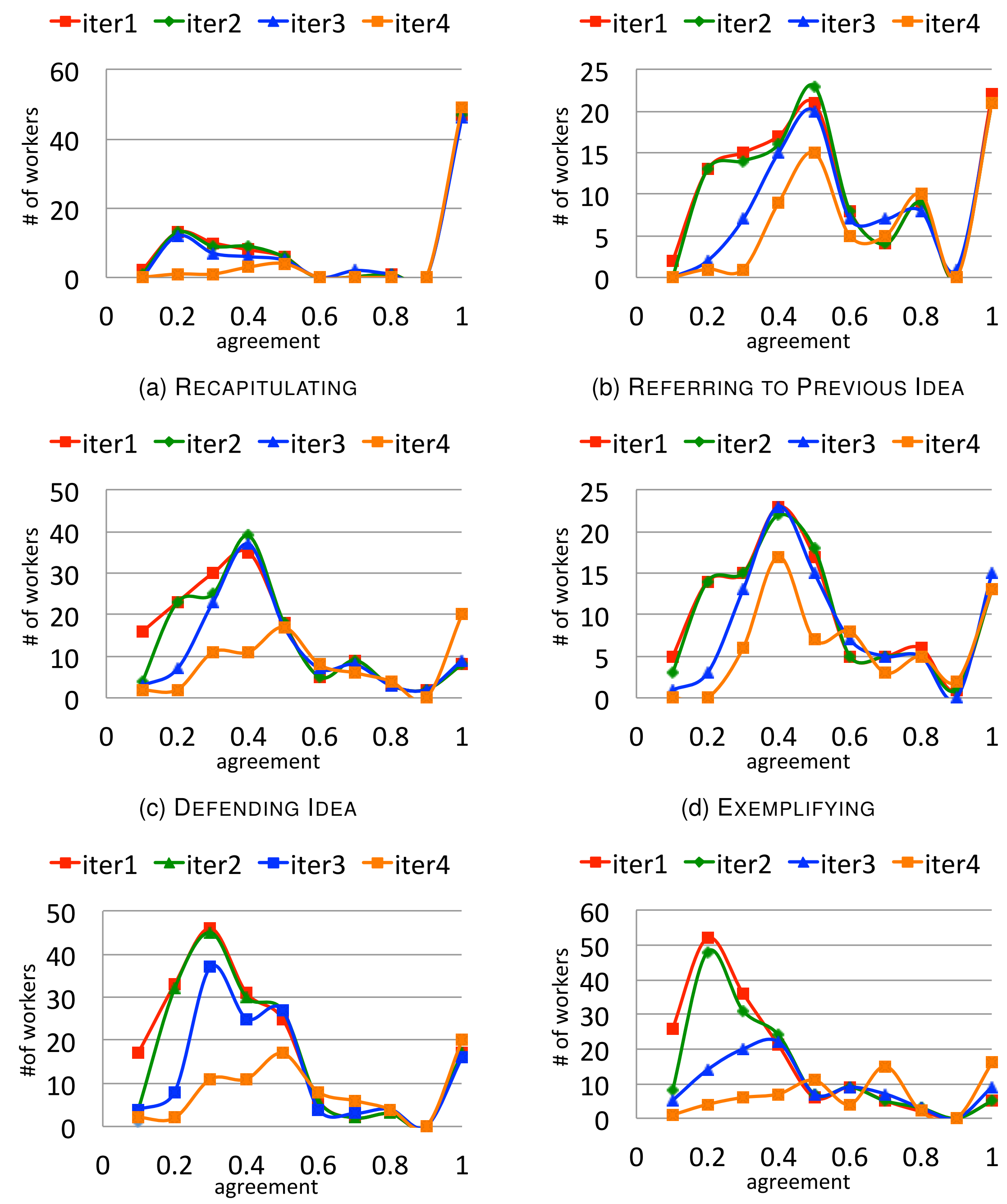
(c) ADDING INFORMATION

(d) CONCLUDING TOPIC

(e) ENUMERATING

(f) POSTPONING TOPIC

Figure D.2: Distribution of agreement between annotators, for four iterations of the filter strategy for `INTRO`, `DELIM`, `ADD`, `CONC`, `ENUM`, and `POST`.



(a) Recapitulating

(b) Referring to Previous Idea

(c) Defending Idea

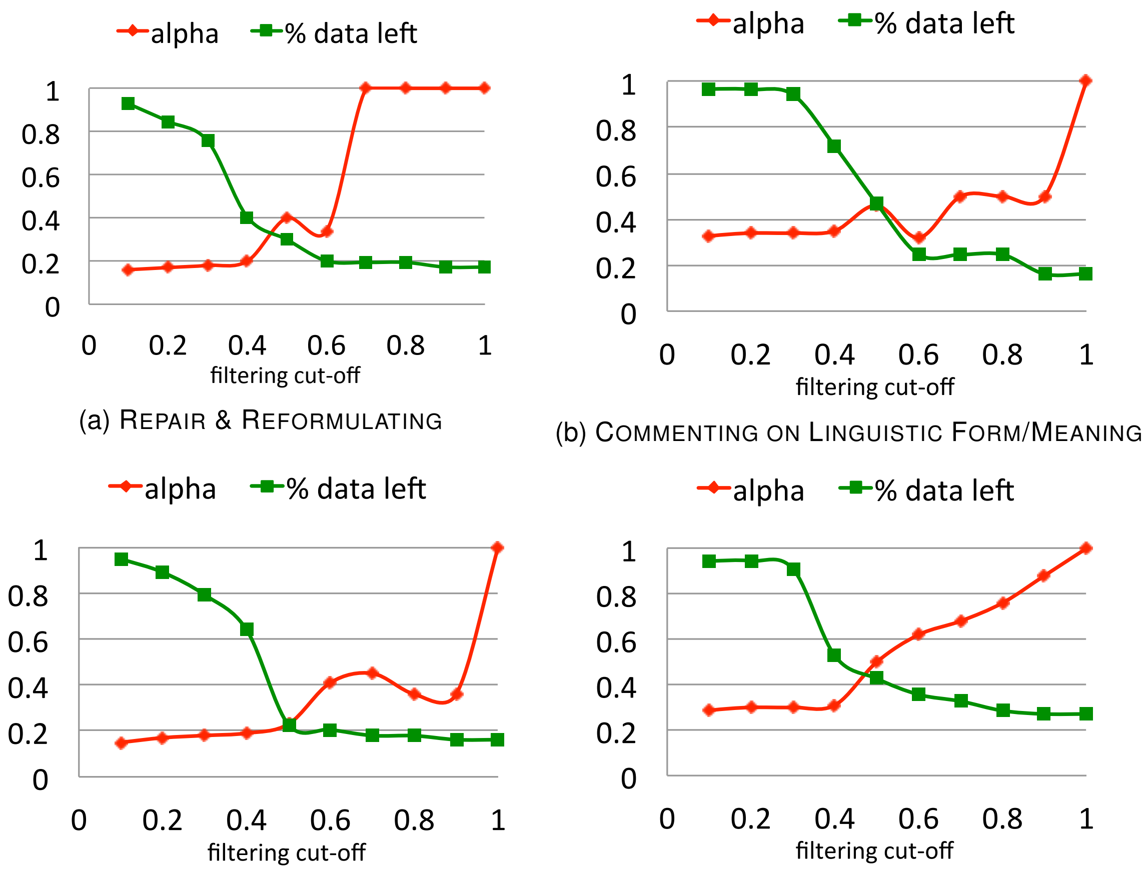
(d) Exemplifying

(e) Anticipating the Audience's Response

(f) Emphasizing

Figure D.3: Distribution of agreement between annotators, for four iterations of the filter strategy for RECAP, REFER, DEFND, EXMPL, ANT, and EMPH.

E Filtering Trade-off



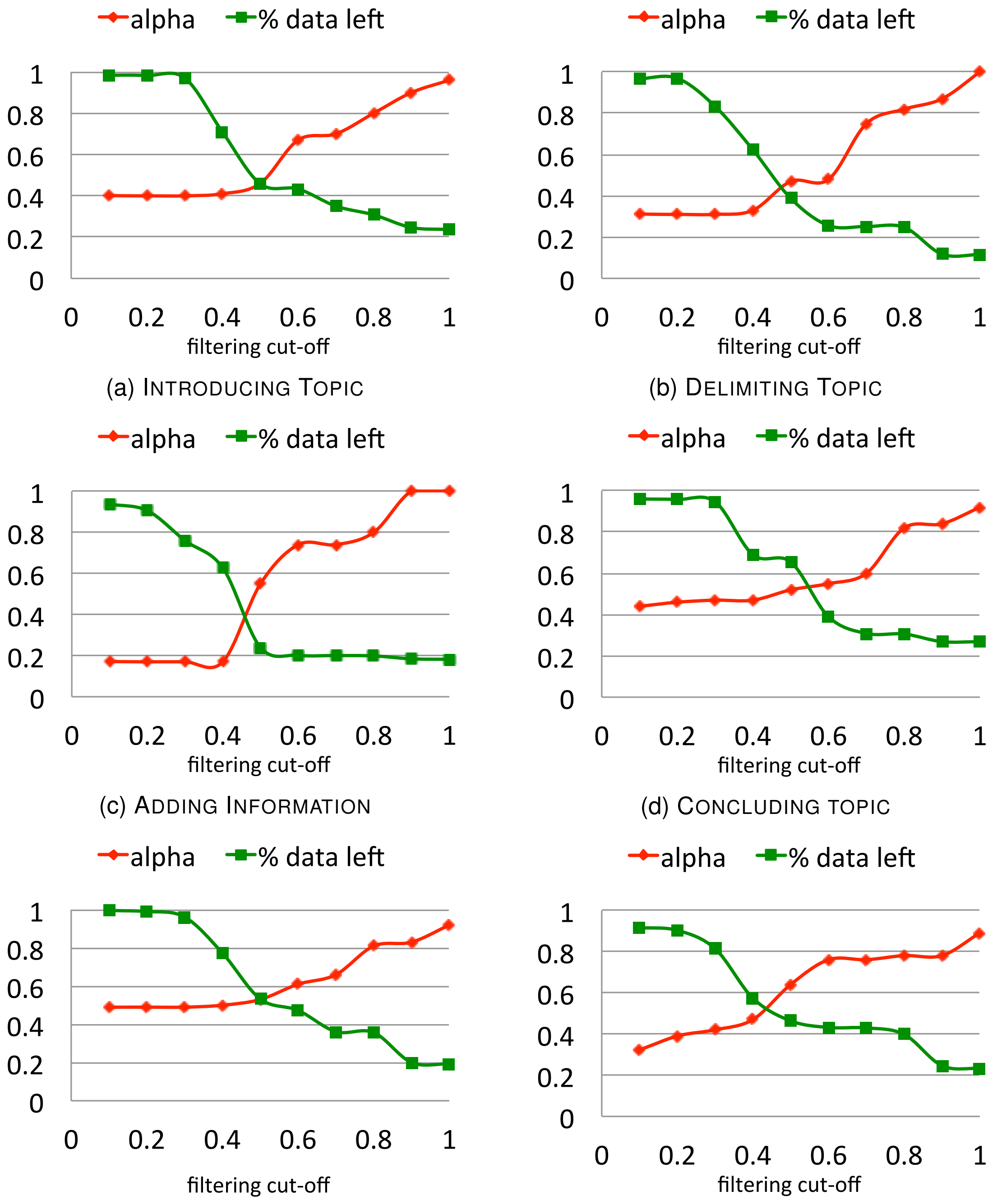
(a) Repair & Reformulating

(b) Commenting on Linguistic Form/Meaning

(c) Clarifying

(d) Defining

Figure E.1: Tradeoff between discarding work based on agreement and percentage of data lost for R&R, COM, CLAR, and DEF.



(a) INTRODUCING TOPIC

(b) DELIMITING TOPIC

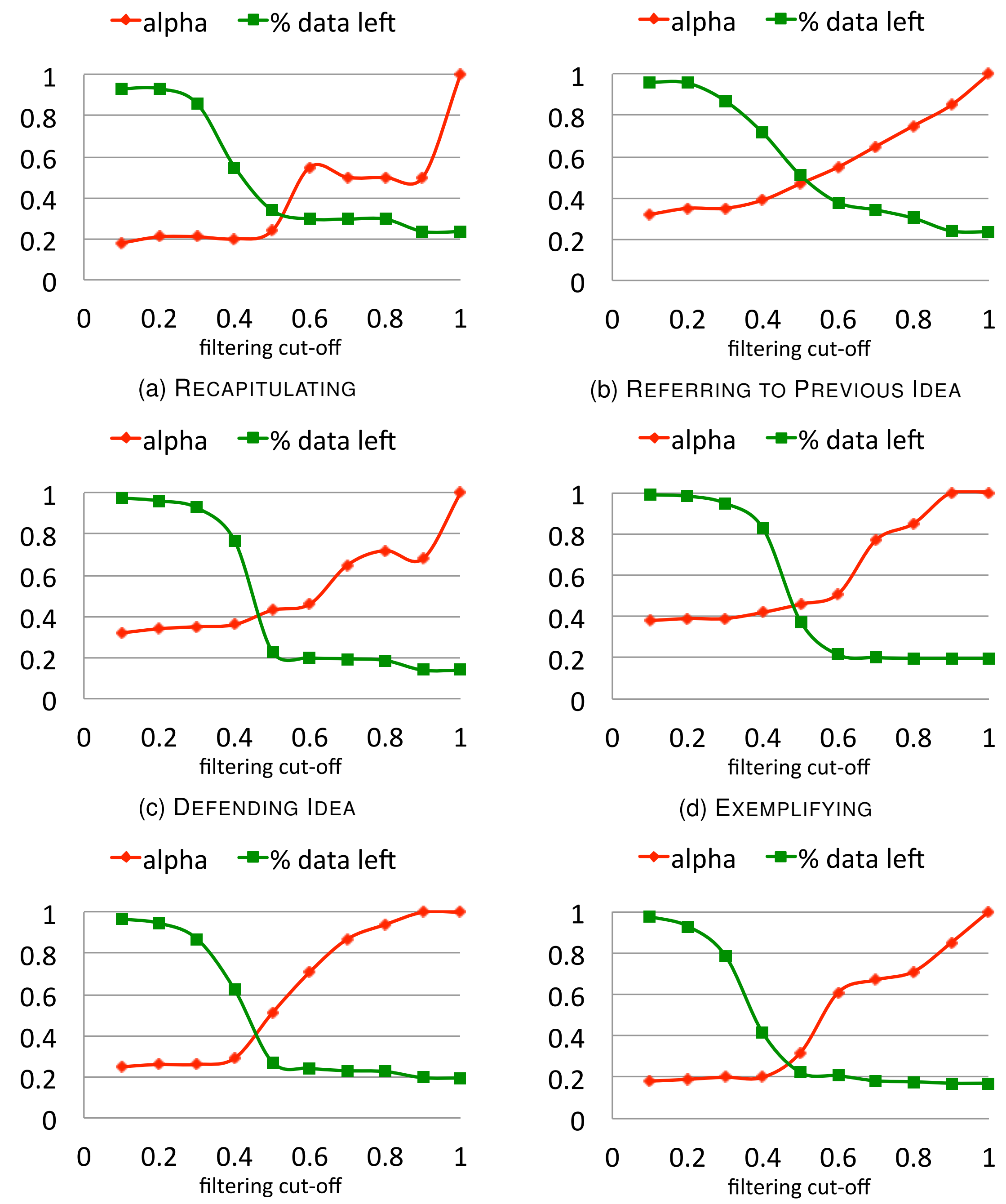
(c) ADDING INFORMATION

(d) CONCLUDING TOPIC

(e) ENUMERATING

(f) POSTPONING TOPIC

Figure E.2: Tradeoff between discarding work based on agreement and percentage of data lost for `INTRO`, `DELIM`, `ADD`, `CONC`, `ENUM`, and `POST`.



(a) RECAPITULATING

(b) REFERRING TO PREVIOUS IDEA

(c) DEFENDING IDEA

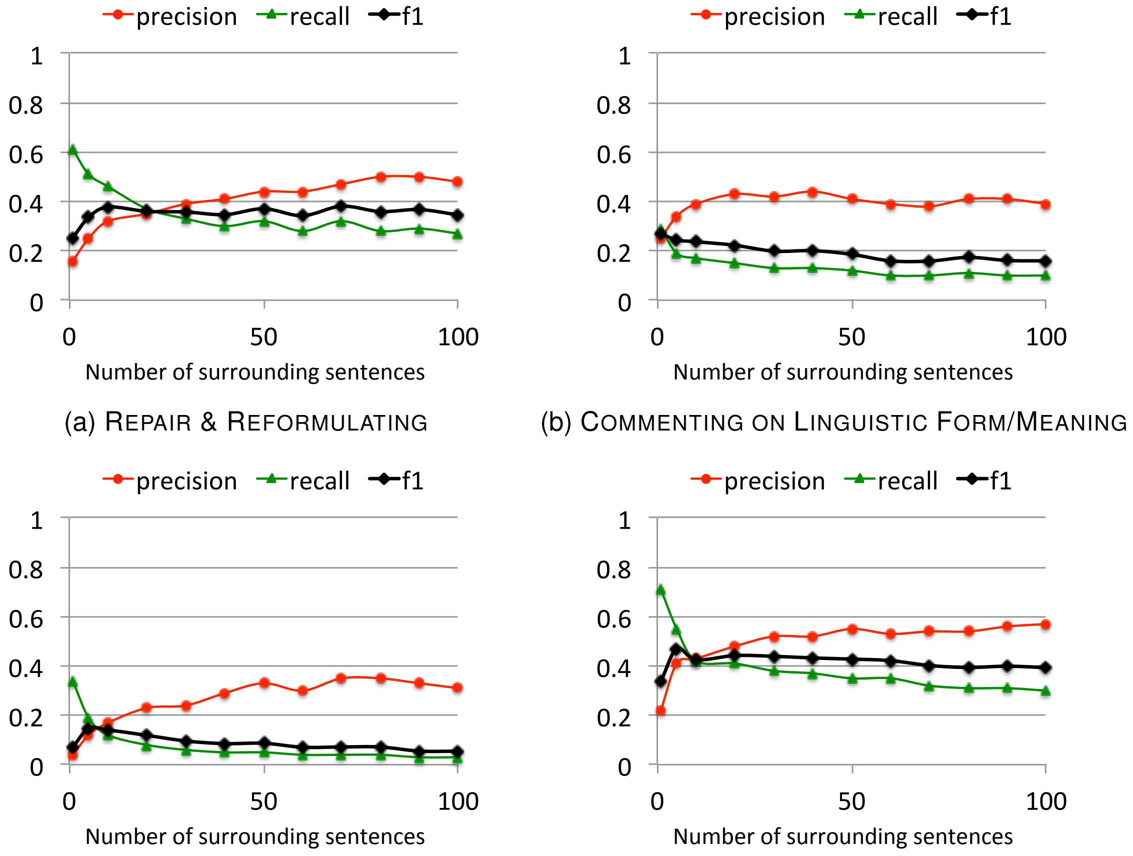
(d) EXEMPLIFYING

(e) ANTICIPATING THE AUDIENCE'S RESPONSE

(f) EMPHASIZING

Figure E.3: Tradeoff between discarding work based on agreement and percentage of data lost for RECAP, REFER, DEFND, EXMPL, ANT, and EMPH.

F Training data balance impact



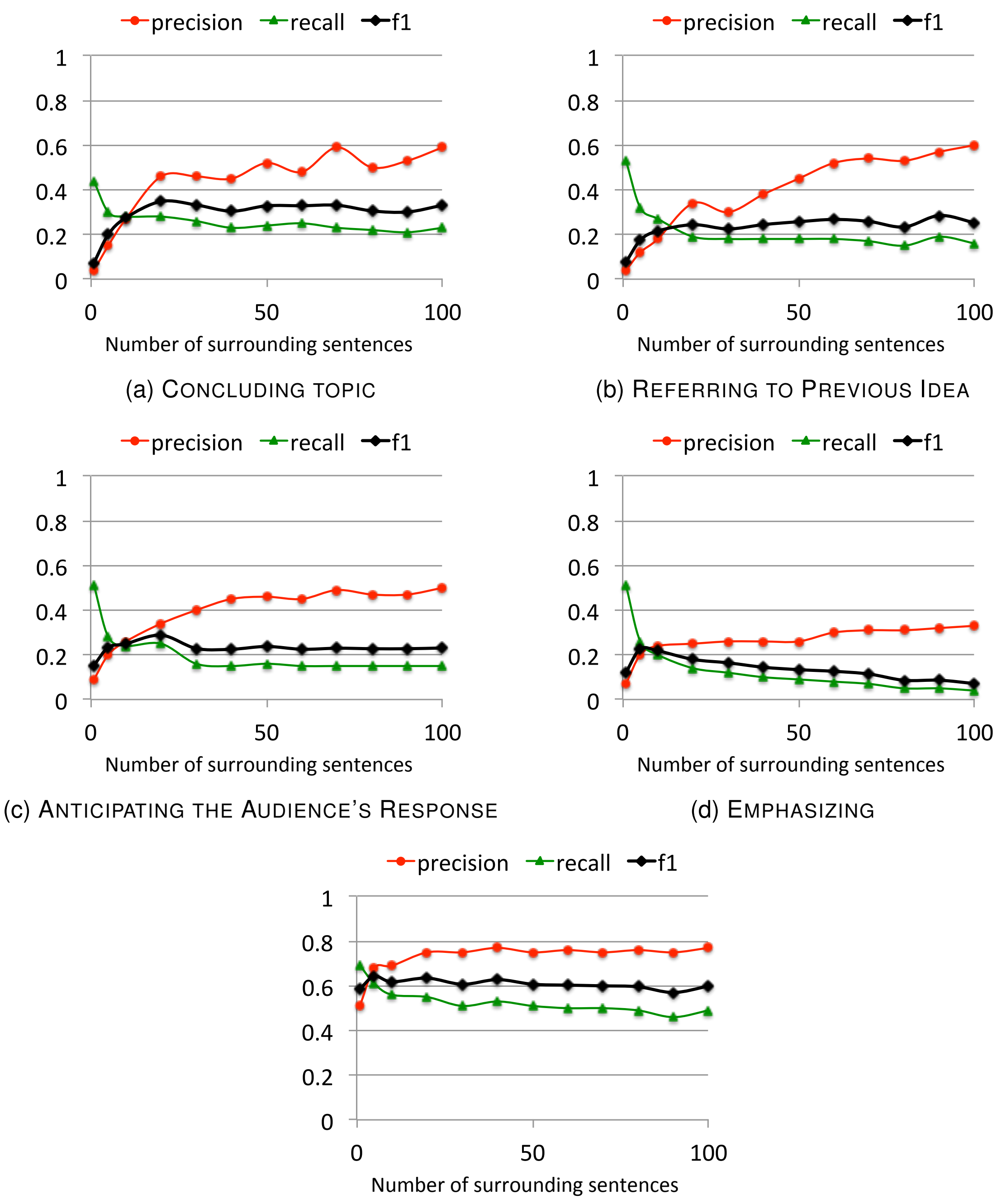
(a) REPAIR & REFORMULATING

(b) COMMENTING ON LINGUISTIC FORM/MEANING

(c) DEFINING

(d) INTRODUCING TOPIC

Figure F.1: Tradeoff between precision, recall and F1 as more negative examples are added to the training data for `R&R`, `COM`, `DEF`, and `INTRO`.



(a) CONCLUDING TOPIC

(b) REFERRING TO PREVIOUS IDEA

(c) ANTICIPATING THE AUDIENCE'S RESPONSE

(d) EMPHASIZING

(e) EXEMPLIFYING

Figure F.2: Tradeoff between precision, recall and F1 as more negative examples are added to the training data for `CONC`, `REFER`, `ANT`, `EMPH`, and `EXMPL`.

Bibliography

Akhtar Abbas and Wasima Shehzad. Metadiscursive Role of Author(s)'s Exclusive Pronouns in Pakistani Research Discourses. *International Journal of English Linguistics*, 8(1): 71, 2017.

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

Annelie Ädel. *The Use of Metadiscourse in Argumentative Writing by Advanced Learners and Native Speakers of English*. PhD thesis, Göteborg, Sweden: Göteborg University, 2003.

Annelie Ädel. On the boundaries between evaluation and metadiscourse. *Strategies in academic discourse*, pages 153–162, 2005.

Annelie Ädel. *Metadiscourse in L1 and L2 English*, volume 24. John Benjamins Publishing, 2006.

Annelie Ädel. 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, 2010.

Annelie Ädel. Remember that your reader cannot read your mind: Problem/solution-oriented metadiscourse in teacher feedback on student writing. *English for Specific Purposes*, 45:54–68, 2017.

Annelie Ädel and Anna Mauranen. Metadiscourse: Diverse and divided perspectives. *Nordic Journal of English Studies*, 9(2):1–11, 2010.

Salah Aït-Mokhtar, J-P Chanod, and Claude Roux. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2-3):121–144, 2002.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Collaborative workflow for crowdsourcing translation. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1191–1194. ACM, 2012.

Carmen Pérez-Llantada Auría. 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, 2006.

Patricia Baggett. Role of temporal overlap of visual and auditory material in forming dual media associations. *Journal of Educational Psychology*, 76(3):408, 1984.

Albert Bandura. *Social foundations of thought and action: A social cognitive theory.* Prentice-Hall, Inc, 1986.

Stephen Bax, F Nataksuhara, and D Waller. Researching metadiscourse markers in candidates' writing at cambridge fce, cae and cpe levels, 2013.

Duygu Bektik. *Learning Analytics for Academic Writing through Automatic Identification of Meta-discourse*. PhD thesis, The Open University, 2017.

Douglas Biber. Spoken and written textual dimensions in english: Resolving the contradictory findings. *Language*, pages 384–414, 1986.

David M. Blei and Pedro J. Moreno. Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in Information Retrieval*, pages 343–348. ACM, 2001.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Michael Bloodgood and Chris Callison-Burch. Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 208–211. Association for Computational Linguistics, 2010.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 131–198, 2016.

Jürgen Börstler and Olof Johansson. The students conference – a tool for the teaching of research, writing, and presentation skills. In *ACM SIGCSE Bulletin*, volume 30, pages 28–31. ACM, 1998.

Anthony Brew, Derek Greene, and Pádraig Cunningham. Using crowdsourcing and active learning to track sentiment in online media. In *ECAI*, pages 145–150, 2010.

Jonathan Brown and Maxine Eskenazi. Student, text and curriculum modeling for reader-specific document retrieval. In *Proceedings of the IASTED International Conference on Human-Computer Interaction. Phoenix, AZ*, 2005.

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

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486. ACM, 2001.

Wallace Chafe and Jane Danielewicz. *Properties of spoken and written language.* Academic Press, 1987.

Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

Hee Jun Choi and Scott D. Johnson. The effect of context-based video instruction on learning and motivation in online courses. *The American Journal of Distance Education*, 19 (4):215–227, 2005.

Michael Clyne. Cultural differences in the organization of academic texts: English and german. *Journal of Pragmatics*, 11(2):211–241, 1987.

Kevyn Collins-Thompson and Jamie Callan. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462, 2005.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Rui Correia, Nuno Mamede, Jorge Baptista, and Maxine Eskenazi. Toward automatic classification of metadiscourse. In *Advances in Natural Language Processing*, pages 262–269. Springer, 2014a.

Rui Correia, Nuno Mamede, Jorge Baptista, and Maxine Eskenazi. Using the crowd to annotate metadiscursive acts. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 102, 2014b.

Rui Correia, Maxine Eskenazi, and Nuno Mamede. Lexical level distribution of metadiscourse in spoken language. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, page 70, 2015.

Rui Correia, Nuno Mamede, Jorge Baptista, and Maxine Eskenazi. metaTED: a Corpus of Metadiscourse for Spoken Language. In *Proceedings 10th Language Resources and Evaluation Conference*, 2016.

Elena Cotos and Nick Pendar. Discourse classification into rhetorical functions for awe feedback. *calico journal*, 33(1):92, 2016.

Avon Crismore. The rhetoric of textbooks: Metadiscourse. *J. Curriculum Studies*, 16(3): 279–296, 1984.

Avon Crismore. *Talking with readers: metadiscourse as rhetorical act*, volume 17. Peter Lang Pub Inc, 1989.

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

Trine Dahl. Textual metadiscourse in research articles: a marker of national culture or of academic discipline? *Journal of Pragmatics*, 36(10):1807–1825, 2004.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *ACL (1)*, pages 795–804, 2015.

Luc De Grez, Martin Valcke, and Irene Roozen. The impact of an innovative instructional intervention on the acquisition of oral presentation skills in higher education. *Computers & Education*, 53(1):112–120, 2009a.

Luc De Grez, Martin Valcke, and Irene Roozen. The impact of goal orientation, self-reflection and personal characteristics on the acquisition of oral presentation skills. *European journal of psychology of education*, 24(3):293–306, 2009b.

Yunda Desilia, Velizya Thasya Utami, Cecilia Arta, and Derwin Suhartono. An attempt to combine features in classifying argument components in persuasive essays. In *17th Workshop on Computational Models of Natural Argument (CMNA)*, 2017.

Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann, editors. *Crowdsourcing for Speech Processing*. John Wiley & Sons, 2013. ISBN 978-1-118-35869-6.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008.

Elena Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, pages 392–398, 2012.

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Mingkun Gao, Wei Xu, and Chris Callison-Burch. Cost optimization in crowdsourcing translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015.

João Graça. Crowdsourcing in MT. the 9th Machine Translation Marathon, Keynote Speech. `http://www.statmt.org/mtm14/uploads/Crowdsourcing_MTM2014.pdf` [Accessed: 02 Dec 2015], 2014.

Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

Richard J. Haber and Lorelei A. Lingard. Learning oral presentation skills. *Journal of General Internal Medicine*, 16(5):308–314, 2001.

Philip J. Hayes, Alexander G. Hauptmann, Jaime G. Carbonell, and Masaru Tomita. Parsing spoken language: a semantic caseframe approach. In *Proceedings of the 11th coference on Computational linguistics*, pages 587–592. Association for Computational Linguistics, 1986.

Haibo He and Edwardo A Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. Retrieval of reading materials for vocabulary and reading practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88. Association for Computational Linguistics, 2008.

Julia Hirschberg and Diane Litman. Empirical studies on the disambiguation of cue phrases. *Computational linguistics*, 19(3):501–530, 1993.

Julia Hirschberg and Christine Nakatani. Acoustic indicators of topic segmentation. In *Proc. ICSLP*, volume 4, pages 1255–1258, 1998.

Charles F. Hockett. The problem of universals in language. *Universals of language*, 2:1–29, 1963.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35. Association for Computational Linguistics, 2009.

Ken Hyland. *Disciplinary discourses, Michigan classics ed.: Social interactions in academic writing*. University of Michigan Press, 2004.

Ken Hyland. *Metadiscourse*. Wiley Online Library, 2005.

Ken Hyland and Carmen Sancho Guinda. *Stance and voice in written academic genres*. Springer, 2012.

Ken Hyland and Polly Tse. Metadiscourse in academic writing: A reappraisal. *Applied linguistics*, 25(2):156–177, 2004.

Douglas A. Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas A. Reynolds, and Marc A. Zissman. Measuring the readability of automatic speech-to-text transcripts. In *Interspeech*, 2003.

Douglas A. Jones, Wade Shen, Elizabeth Shriberg, Andreas Stolcke, Teresa M. Kamm, and Douglas A. Reynolds. Two experiments comparing reading with listening for human processing of conversational telephone speech. In *Interspeech*, pages 1145–1148, 2005.

Fleiss L. Joseph. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382, 1971.

Debra Kerby and Jeff Romine. Develop oral presentation skills through accounting curriculum design and course-embedded assessment. *Journal of Education for Business*, 85(3): 172–179, 2009.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

William J. Vande Kopple. Some exploratory discourse on metadiscourse. *College composition and communication*, pages 82–93, 1985.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann, 2001.

John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 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, 2010.

Yoong Keok Lee and Hwee Tou Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 41–48. Association for Computational Linguistics, 2002.

Stephanie Lindemann and Anna Mauranen. "it's just real messy": the occurrence and function of just in a corpus of academic speech. *English for specific purposes*, 20:459–475, 2001.

Wang Ling and Luísa Coheur Ist. Machine translation in microblogs. *System cybernetics*, 1(2013), 2013.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

John A. Lucy. *Reflexive language: Reported speech and metapragmatics*. Cambridge University Press, 1993.

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

John Lyons. *Semantics. vol. 2*. Cambridge University Press, 1977.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28. Association for Computational Linguistics, 2012.

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

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

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

Anna Mauranen. Reflexive academic talk: Observations from micase. In *Corpus linguistics in North America: selections from the 1999 symposium*, page 165. University of Michigan Press/ESL, 2001.

Anna Mauranen. A good question?: Expressing evaluation in academic speech. *Domain-specific English: Textual practices across communities and classrooms*, pages 115–140, 2002.

Anna Mauranen. " but here's a flawed argument": Socialisation into and through metadiscourse. *Language and Computers*, 46(1):19–34, 2003.

Anna Mauranen. Discourse reflexivity-a discourse universal? the case of elf. *Nordic Journal of English Studies*, 9(2):13–40, 2010.

Anna Mauranen. 'but then when i started to think?': Narrative elements in conference presentations. *Narratives in Academic and Professional Genres*, pages 45–66, 2013a.

Anna Mauranen. Speaking professionally in an l2. *Variation and Change in Spoken and Written Discourse: Perspectives from corpus linguistics*, 21:1, 2013b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. Discriminative topic segmentation of text and speech. In *Proceedings of the 13th international conference on artificial intelligence and statistics (AISTATAS'10)*, 2010.

Helena Moniz, Fernando Batista, Isabel Trancoso, and Ana Isabel Mata. Prosodic contex-based analysis of disfluencies. 2012.

Martin Moyle, Justin Tonra, and Valerie Wallace. Manuscript transcription by crowdsourcing: Transcribe bentham. *Liber Quarterly*, 20(3/4):347–356, 2011.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.

Gayle L. Nelson. How cultural differences affect written and oral communication: The case of peer response groups. *New Directions for Teaching and Learning*, 1997(70):77–84, 1997.

Huy Nguyen and Diane J Litman. Extracting argument and domain words for identifying argument components in texts. In *ArgMining@ HLT-NAACL*, pages 22–28, 2015.

Huy Nguyen and Diane J Litman. Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics. In *FLAIRS Conference*, pages 485–490, 2016.

Eileen Nicolle, Emmanuelle Britton, Praseedha Janakiram, and Pierre-Marc Robichaud. Using ted talks to teach social determinants of health maximize the message with a modern medium. *Canadian Family Physician*, 60(9):777–778, 2014.

Stefanie Nowak and Stefan Rüger. 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, 2010.

Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007. URL `http://www.chokkan.org/software/crfsuite/`.

Stanislaw Osinski and Dawid Weiss. A concept-driven algorithm for clustering search results. *Intelligent Systems, IEEE*, 20(3):48–54, 2005.

Stanislaw Osinski, Jerzy Stefanowski, and Dawid Weiss. Lingo: Search results clustering algorithm based on singular value decomposition. *Proceedings of the international IIS: intelligent information processing and web mining IIPWM*, 4:359–368, 2004.

Georgios Paltoglou and Mike Thelwall. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395. Association for Computational Linguistics, 2010.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

Gabriel Parent and Maxine Eskenazi. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 312–317. IEEE, 2010.

Rebecca J. Passonneau and Diane J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, 1997.

Ted Pedersen. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

David Pellow and Maxine Eskenazi. An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 84–93, 2014.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Khushwant K. S. Pittenger, Mary C. Miller, and Joshua Mott. Using real-world standards to enhance students' presentation skills. *Business Communication Quarterly*, 67(3):327–336, 2004.

Matthew Purver. Topic segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317, 2011.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

Antoine Raux and Alan W. Black. A unit selection approach to f0 modeling and its application to emphasis. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 700–705. IEEE, 2003.

Garr Reynolds. *Presentation Zen: Simple ideas on presentation design and delivery*. New Riders, 2011.

Ute Römer and John M. Swales. The Michigan Corpus of Upper-level Student Papers (MICUSP). *Journal of English for Academic Purposes*, 9(3):249, 2010.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. Matrix factorization using window sampling and negative sampling for improved word representations. *arXiv preprint arXiv:1606.00819*, 2016.

Wilbur Schramm. How communication works. the process and effects of mass communication. *Urbana: University of Illinois Press*, 1954.

Claude Elwood Shannon and Warren Weaver. A mathematical theory of communication, 1948.

Narayanan Shivakumar and Hector Garcia-Molina. Building a scalable and accurate copy detection mechanism. In *Proceedings of the first ACM international conference on Digital libraries*, pages 160–168. ACM, 1996.

Catarina Silva and Bernardete Ribeiro. The importance of stop word removal on recall values in text categorization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 1661–1666. IEEE, 2003.

Michael Silverstein. Shifters, linguistic categories, and cultural description. *Meaning in anthropology*, pages 11–55, 1976.

Rita C. Simpson and John Swales. *Corpus linguistics in North America: Selections from the 1999 symposium*. University of Michigan Press/ESL, 2001.

Radu Soricut and Daniel Marcu. 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, 2003.

Rachele Sprugnoli, Giovanni Moretti, Luisa Bentivogli, and Diego Giuliani. Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation*, 51(2):283–317, 2017.

Christian Stab and Iryna Gurevych. Annotating Argument Components and Relations in Persuasive Essays. In *COLING*, pages 1501–1510, 2014.

Christian Matthias Edwin Stab. *Argumentative Writing Support by means of Natural Language Processing*. PhD thesis, Technische Universität Darmstadt, 2017.

Frederik Stouten, Jacques Duchateau, Jean-Pierre Martens, and Patrick Wambacq. Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Communication*, 48(11):1590–1606, 2006.

Derwin Suhartono, Afif Akbar Iskandar, M Ivan Fanany, and Ruli Manurung. Utilizing word vector representation for classifying argument components in persuasive essays. In *3rd International Conference on Science, Engineering, Built Environment, and Social Science (ICSEBS)*, 2016.

John Swales. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (1)*, pages 1555–1565, 2014.

Carson Tao, Michele Filannino, and Özlem Uzuner. Prescription extraction using crfs and word embeddings. *Journal of Biomedical Informatics*, 72:60–66, 2017.

Susan Elizabeth Thompson. Text-structuring metadiscourse, intonation and the signalling of organisation in academic lectures. *Journal of English for Academic Purposes*, 2(1):5–20, 2003.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321 (5895):1465–1468, 2008.

Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

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

Anna Wierzbicka. Different cultures, different languages, different speech acts: Polish vs. English. *Journal of pragmatics*, 9(2-3):145–178, 1985.

Shomir Wilson. Distinguishing use and mention in natural language. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 29–33. Association for Computational Linguistics, 2010.

Shomir Wilson. The creation of a corpus of English metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 638–646. Association for Computational Linguistics, 2012.

Shomir Wilson. Toward automatic processing of english metalanguage. In *IJCNLP*, pages 760–766, 2013.

Pieter Wouters, Huib K Tabbers, and Fred Paas. Interactivity in video-based models. *Educational Psychology Review*, 19(3):327–342, 2007.

Wenting Xiong and Diane Litman. Identifying problem localization in peer-review feedback. In *Intelligent Tutoring Systems*, pages 429–431. Springer, 2010.

Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. Recent advances of large-scale linear classification. *Proceedings of the IEEE*, 100(9):2584–2603, 2012.

Omar F. Zaidan and Chris Callison-Burch. 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, 2011.

Xiaodan Zhu and Gerald Penn. Summarization of spontaneous conversations. In *INTERSPEECH*, 2006.