Automatic Classification of Metadiscourse for Presentation Skills Instruction

Rui Pedro dos Santos Correia
(M.Sc.)

Ph.D. Thesis Proposal
Information Systems and Computer Engineering

Thesis Advisory Committee

Advisors:  Prof. Maxine Eskenazi  
Prof. Nuno João Neves Mamede

Jury:  Prof. Jorge Manuel Baptista  
Prof. Jaime Carbonell  
Prof. Robert E. Frederking  
Prof. Diane J. Litman  
Prof. Isabel Maria Martins Trancoso

July 2013
Abstract

In this thesis we approach the topic of metadiscourse in a manner that lends itself to presentation skills instruction. More precisely, we address issues related to the function of metadiscursive acts in spoken language. We present existing theory on spoken metadiscourse, focusing on one taxonomy that defines metadiscursive concepts in a fully functional manner, i.e., that assigns a discourse function to occurrences of metadiscourse rather than commenting on its form.

We set up a crowdsourced annotation task with two main goals: (a) use the crowd as a reflection of future students, to assess the non-experts understanding of the different functions in the taxonomy, and (b) build a corpus of metadiscursive acts. Results show that not all metadiscourse acts in the same taxonomy can be labeled and understood by the crowd.

With the collected annotations, we train Decision Trees to identify and classify metadiscourse along four different discourse functions, using simple grammatical and lexical features (n-grams of parts-of-speech, lemmas and words). This strategy performs with classification accuracies between 80% and 90% for the task of identifying sentences in presentations that contain occurrences of metadiscourse used to introduce/conclude a topic, give an example or emphasize a point.

We propose the expansion of the current work with the addition of new categories of metadiscursive functions, the improvement of the current classification methods, and the exploring of metadiscourse in European Portuguese. As a final goal, we aim at packaging this technology in a language learning application, making students aware of the strategies used by professional speakers while presenting a topic.
Keywords

Presentation Skills
Language Learning
Metadiscourse
Discourse Function
Natural Language Processing
# Contents

1 Introduction 1
   1.1 Motivation ................................. 1
   1.2 Goals ........................................ 2
   1.3 Structure of this Document ................. 7

2 Background 9
   2.1 Metadiscursive Theory ........................ 10
      2.1.1 Luukka (1992) ........................... 12
      2.1.2 Mauranen (2001) ........................ 12
      2.1.3 Thompson (2003) ........................ 13
      2.1.4 Auría (2006) ............................ 14
      2.1.5 Ådel (2010) ............................. 15
   2.2 NLP Approaches to Metadiscourse ............... 19
      2.2.1 Penn Discourse Treebank (PDTB) ........... 20
      2.2.2 RST Discourse Treebank .................... 21
      2.2.3 Wilson (2012) ............................ 23
      2.2.4 Madnani et al. (2012) ...................... 25
   2.3 Discussion ..................................... 26

3 A Corpus of Metadiscourse 29
   3.1 Source of Spoken Data ......................... 30
   3.2 A Preliminary Annotation Task .................. 32
   3.3 Annotation using Crowdsourcing ................ 37
   3.4 Annotation Results ............................ 42
   3.5 Discussion ..................................... 47
4 Automatically Classifying Metadiscourse 51
   4.1 Experimental Setup ........................................ 52
   4.2 Classification Results ..................................... 59
   4.3 Discussion ................................................... 66

5 Conclusion & Proposed Work 69
   5.1 Additional Categories ..................................... 71
   5.2 Improving Classification .................................. 72
      5.2.1 Features ................................................. 72
      5.2.2 Algorithms ............................................. 74
   5.3 Metadiscourse in European Portuguese .................... 75
   5.4 Presentation Skills Instruction Tool ....................... 76
   5.5 Timetable .................................................... 78

6 Program Progress Details 81
   6.1 Curricular Plan .............................................. 81
   6.2 Teaching Assistance ........................................ 81
   6.3 Published Work .............................................. 81

ii
### List of Figures

1.1 TED Ed Interface of a Multiple-Choice Question. ........................................ 4
1.2 New York Times interface of the vice-presidential debate. ............................. 6

2.1 Wilbur Schramm’s extension of Shannon-Weaver model of communication . . 10
2.2 ˚Adel’s taxonomy of metadiscourse. .......................................................... 17
2.3 Hierarchy of senses in Penn Discourse Treebank (PDTB) ............................. 21
2.4 Simplified Rhetorical Structure Theory categories. ...................................... 22

3.1 Readability-level distribution of the 730 TED talks. .................................. 32
3.2 Occurrences of the most frequent tags in the 10 TED talks sample. ............... 37
3.3 Interface of one segment in a HIT .............................................................. 39
3.4 Positive examples for the task of identifying occurrences of Emphasizing .. . 40
3.5 Negative examples for the task of identifying occurrences of Emphasizing . 41

4.1 Decision tree for the category Exemplifying using the word bigram model. . 62
4.2 Occurrences of organizational metadiscourse in the beginning of a TED talk. 65
List of Tables

2.1 Wilson’s taxonomy of mentioned language. ................................. 24
2.2 Wilson’s annotation results. ...................................................... 25

3.1 Annotation results in terms of workers’ behavior. ......................... 43
3.2 Annotation results in terms of quantity and quality. ..................... 43
3.3 Inter-annotator agreement by level. ........................................... 46
3.4 Percentage of sentences with metadiscourse occurrences per level. .... 47

4.1 Top $n$-grams in the TED talks compared to annotations of the four metadiscursive categories. .................................................... 57
4.2 Two training data points for the category INTRODUCING TOPIC and the unigram words feature setting. ................................. 58
4.3 Classifiers accuracy for the unigrams, bigrams and trigrams of POS, Lemma and Word Features. ..................................................... 60
4.4 Classifiers accuracy for the combinations of POS, Lemma and Word Features. .......................................................... 63
4.5 Classifiers accuracy when added Positioning & Length features. ....... 65
1

Introduction

Rhetoric is the art of ruling the minds of men.
– Plato

1.1 Motivation

Being able to present an idea, to prove a point, and to convince an audience are essential skills to master in both academic and professional settings. Effective use of presentational skills is even harder for students who are required to make a presentation in a language different from their mother tongue. In addition to the language barrier, individuals have to take into account that the expectations for formal oral presentations, and discourse in general, change from culture to culture (Wierzbicka, 1985; Clyne, 1987; Nelson, 1997; Dahl, 2004).

Presentation skills encompass several components, such as the message itself, body language, eye contact, voice projection, and visuals (De Grez et al., 2009b). For each of these components, there are techniques that can help the speaker to convey his message. In this thesis we focus on the discourse component of presentations, proposing the development of a language learning application that exposes students to good presentations and good speakers. More precisely, we focus on one particular function of language – metadiscourse –, following the consensual opinion in the literature that mastering metadiscourse can help expressing and defending a point of view.

Metadiscourse (also metalanguage, signposting language, or text-referral) is one of the basic functions of language. Commonly referred to as discourse about discourse, it is composed of rhetorical acts and patterns used to make the discourse structure explicit, acting as a way to
CHAPTER 1. INTRODUCTION

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 above quote summarizes the three defining properties of metadiscourse:

1. it occurs in both written and spoken discourse;

2. it does not contribute to the content of the discourse;

3. and it is used by the speaker/writer to guide the audience during the communication event.

Metadiscourse allows the speaker or writer to explicitly refer to events that are situated in the realm of the discourse. Common 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 to note that such events still exist in the discourse even if not explicitly mentioned. For example, a Ph.D. candidate writing his thesis proposal may use bold faced font to emphasize the main concepts, instead of using a metadiscourse strategy such as “It is important to note that. . .”, or he may surround examples with parenthesis instead of explicitly mention that he is exemplifying (“For instance, . . ”).

1.2 Goals

In this thesis we propose to investigate how professional speakers use metadiscursive acts to make the structure of their talks explicit, and package our findings under a language learning framework. Additionally, and given the fact that this thesis is included in the CMU/Portugal program, we aim at doing so for both English and European Portuguese, exploring and comparing the nature of metadiscourse for both languages.
1.2. GOALS

This work can be described as the articulation of two perspectives to metadiscourse. The first is that of Presentation Skills Instruction. Being language learning the end goal of our approach, we look at metadiscourse in a way that lends itself to the teaching purpose. The second is the Natural Language Processing (NLP) perspective, which involves considerations related to the extent to which we can automatically find the rhetorical acts and patterns from which metadiscourse is composed. The next two sections address these two perspectives and their respective contributions in defining the scope of this thesis.

Presentation Skills Instruction

Across the literature authors agree that metadiscourse should be part of the students curriculum regarding the acquisition of academic competencies, since the intentional and prepared use of the linguistic patterns involved in it can constitute an aid to build clearer and more organized presentations (Lyons, 1977; Auría, 2006; Ædel, 2010). However, this set of skills is not usually included in academic curricula.

Traditional learning activities (such as reading a book or attending a class) require planning and scheduling and, for those reasons, are often reserved for core skills improvement. Nowadays, soft-skills instruction (such as the ability to present a topic) is often approached in a learning in the wild philosophy, where the learning opportunity is strategically introduced in everyday activities. An example of such approach is the TED Ed project\(^1\) (Figure 1.1), where users create lectures about the content of TED talks\(^2\) and share them with the TED community. In this example, the educational value of a TED talk is being augmented by explicit instruction that highlights the main points in the presentation.

In this thesis we propose to create similar learning opportunities that, instead of focusing on the content, augment the user experience with explicit presentation skills instruction. This pedagogical aspect of our work imposes two constraints in the scope:

\(^1\)http://ed.ted.com/
\(^2\)http://www.ted.com/talks
Spoken Discourse — this study will focus on the use of metadiscourse in spoken communication only. In comparison to the written form, this specific setting introduces discourse elements that have an impact on how and for what metadiscourse is used, such as the lack of time for planing or revision or direct interaction with the audience;

Functional Approach to Metadiscourse — herein metadiscourse is classified functionally, as opposed to formally. We are interested in identifying the rhetorical functions associated with each metadiscourse occurrence (introductions, conclusions, examples, etc), rather than analyzing metadiscourse according to its form or intrinsic properties (pronominal or non-pronominal, formal or informal). In other words, we keep in mind that the concepts to teach are meant to be understood by non-experts, i.e., students who want to learn how to present.

Associated with the perspective of metadiscourse for presentation skills instruction is our first research question:

*Can non-experts understand the concept of metadiscourse and its distinct functions in discourse?*
1.2. GOALS

In order to create a learning tool that uses metadiscursive acts as instructional goals, we first need to prove that potential users understand the concepts we are trying to teach. Therefore, with this question we aim at understanding if non-experts, as reflection of future students, are able to (a) identify occurrences of metadiscourse as used in spoken language and (b) assign a speaker intention to each occurrence of the phenomenon.

Natural Language Processing (NLP)

The second perspective from which we look at metadiscourse is from the NLP point of view. More than just automatically identifying occurrences of metadiscourse, we want to be able to assign a discourse function to it (constrain imposed by our language learning goal).

This formulation situates our approach in the area of discourse structuring and segmentation. Segmentation deals with the separation of discourse into cohesive segments, going beyond the concept of sentence as unit. Statistical approaches address the issue by tracking dramatic modification of vocabulary trends within the same document (Hearst, 1997; Blei and Moreno, 2001; Mohri et al., 2010). However, earlier approaches to segmentation identified lexical patterns as indicators of discourse structure – Grosz and Sidner (1986) with the theory of attention and intention and Hirschberg and Litman (1993) with the analysis of cue phrases (now, so, well, etc).

Metadiscourse sits on the most explicit and intentional side of the spectrum of use of language as discourse organizer. When a speaker uses metadiscourse, he/she is intentionally making the audience aware of specific discourse elements. It is these explicit mentions of high-level discursive elements that we want to automatically identify and ultimately teach.

Figure 1.2 shows a concrete example of discourse segmentation – the New York Times interface of the vice-presidential debate between Biden and Ryan\(^3\). In the figure we can see that the video recording of this debate was enriched with segmentation information (top-right), transcript (bottom-left) and additional content (bottom-right).

CHAPTER 1. INTRODUCTION

Figure 1.2: New York Times interface of the vice-presidential debate.

When we look at the transcript of the debate, we can find occurrences of metadiscourse that, when assigned to a function, can serve the purpose of dividing the discourse into meaningful segments. Below are three passages extracted from the transcript that contain instances of metadiscourse:

- **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. Thank you, gentlemen. And that coin toss, again, has Vice President Biden starting with a closing statement.”

By identifying and classifying metadiscourse according to its function, we can automatically generate similar content to the one in Figure 1.2, enhance the interactive experience of the viewer, and ultimately package it in a learning framework for presentational skills.
1.3. STRUCTURE OF THIS DOCUMENT

When looking at metadiscourse through the NLP perspective we formulate our second research question:

To which extent can we automate the identification and functional classification of metadiscourse?

With this question we aim at exploring and comparing different techniques that can successfully classify metadiscursive phenomena. More than a simple analysis of performance, we aim at getting insight on the nature of metadiscourse itself, in both English and European Portuguese, i.e., understand which features are representative of the phenomenon and how different algorithms are capable of identifying and functionally classifying occurrences of metadiscourse. Additionally, this question addresses the issues of performance in an area as sensitive as language learning, where the cost of misclassification can be as severe as incorrect instruction.

1.3 Structure of this Document

This proposal is intended to be a proof of concept of the use of metadiscourse as the key concept in a language learning application. For that reason, herein we do not focus on the development of the learning application itself. Instead, we discuss the problem formulation of metadiscourse as a learning goal and the first steps towards automatic identification and classification of the phenomenon in spoken text. We work towards obtaining the first insights to our two main research questions, using our findings to support proposed future work. 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 state of the art approaches to the annotation and classification of metadiscourse-related phenomena;

- **Chapter 3** describes a crowdsourced annotation task aimed at building a corpus of metadiscursive acts. We discuss the choice of material to annotate and the categories
CHAPTER 1. INTRODUCTION

to use in the annotation task, along with considerations related to the instructions and training of non-expert annotators. This section ends with a discussion on the quality of the material collected. The focus of this section is the first research question, *i.e.*, to which extent can non-experts understand the concept of metadiscourse and corresponding functions in the communication event.

- **Chapter 4** presents a first approach to the task of identifying and assigning a function to occurrences of metadiscourse in English. We describe a supervised learning strategy, the features used to support the decision, and the results obtained with such formulation. This section gives the first insights on our second research question, *i.e.*, whether it is possible to automatically identify and classify metadiscourse according to function.

- **Chapter 5** contains a summary and discussion of the work accomplished so far, situating it in the global perspective of the thesis. In this chapter we discuss the proposed work for the remaining two years of the Ph.D. and its alignment with our research questions. These future directions include the annotation of additional metadiscourse functions, the improvement of the current classifiers, the adaptation of the technology to European Portuguese, and the packaging of the technology in a presentation skills learning tool.
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, including 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 own 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 attention of the research community during the mid 80s, with a large body of work focusing on the presence of metadiscursive acts in written academic discourse (Kopple, 1985; Crismore, 1989). Only later, during the 90s and the 00s, the spoken variety of metadiscourse started to be explored and addressed in a systematic and data-driven manner. In this chapter we describe that work focusing in spoken metadiscourse. This discussion is divided in two main sections, related to the two perspectives from which we look at metadiscourse in this study:

- **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 align with the goal of presentation skills instruction;
- **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 produces a signal suitable to be transmitted over the channel (the medium used to transmit 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.

Schramm (1954) expanded Shannon-Weaver’s model incorporating human behavior in the communication process. Schramm’s model is represented in Figure 2.1, where we can see 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.

![Figure 2.1: Wilbur Schramm’s extension of Shannon-Weaver model of communication](image)

Figure 2.1: Wilbur Schramm’s extension of Shannon-Weaver model of communication
2.1. METADISCURSIVE THEORY

By varying the properties of these elements we can define different communication settings. By varying the element channel we distinguish spoken language from its written form. Varying the media in which the message circulates affects other elements in the model. Spoken communications, and oral presentations in particular, can be characterized by both the noise and feedback elements. In spoken language, the immediacy of production affects the noise element, since planning and corrections have to be done 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 element. 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 concrete 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. For example, the speaker might use metadiscourse to manage the audience comprehension ("Can you hear me back there?") or to correct a point ("Sorry, what I meant was..."). For those reasons, in this chapter we focus on research that looks at metadiscourse as used in spoken language. More specifically, in this section we will discuss the existent theories of metadiscourse that either consider only spoken discourse (sections 2.1.3 and 2.1.4) or that adopt an unified approach and discuss both written and spoken varieties of metadiscourse (sections 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 paper: the written version 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 focuses on both written and spoken language in her work, but the author approaches metadiscursive phenomena in 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). MICASE is 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 and dialogic types of spoken discourse, contrasting with Luukka’s five conference paper corpus, which included mostly monologic discourse.
2.1. **METADISCURSIVE THEORY**

The author’s taxonomy is also 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 (similar 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).

By looking at this taxonomy, we can say that 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 can contribute to the message in real-time (the already mentioned immediacy of feedback).

### 2.1.3 Thompson (2003)

Thompson’s work has as a premise the comprehension of lectures in the real world, analyzing how metadiscourse is used in classrooms. The author is focused in showing the misalignment between the curricula for English for Academic Purposes (EAP) courses and the real practices of discourse organization and intonation in real-word communications.

Thompson uses a corpus of six authentic undergraduate university lectures and five EAP published listening skills materials, stressing the mismatch between what is being taught and what is used in real lectures. As a result of this comparison process, the author formulated a taxonomy of metadiscourse in academic lectures, that categorizes metadiscourse in three main groups:
• **Content Markers** – used to give information about the lecture to come;
  
  – *Eg.*: *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;

  – *Eg.*: *I’ll start with water [...] And then I’ll move on to farms [...]*

• **Metastatements** – used to organize the communication event itself (not its content).

  – *Eg.*: *Right. So, with that let me start the lecture.*

Additionally, Thompson further divides each category in 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 different sections/topics of which each lecture is composed.

### 2.1.4 Auría (2006)

Auría focused on the use of spoken metadiscourse in academic settings, comparing it to conversational language and the written register. The author found that metadiscourse is more prominent in the events in which knowledge is being transmitted, since the lecturer, seeking maximum comprehension, explicitly signals his/her communicative intentions. Auría refers to metadiscourse as a powerful linguistic resource in academic speech. Another interesting conclusion from this work is the higher density of metadiscourse acts 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, speakers tend to show greater concern in drawing attention towards discourse organization for a more effective comprehension of the conceptual load of the talk.

The main concept behind Auría’s taxonomy is the lecturer intention. While analyzing the MICASE corpus (described in Section 2.1.2) the author proposes three categories for the
2.1. METADISCURSIVE THEORY

categorization of metadiscourse:

- **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 Ādel (2010)

Ādel is the author of an extensive body of work in metadiscourse in both written and spoken form (Ādel, 2003, 2005, 2006; Ādel and Reppen, 2008; Ādel and Mauranen, 2010). The taxonomy here presented results from the analysis and unification of the existing theories of metadiscourse. Ādel framework encompasses both spoken and written discourse, being 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.

In her work, Ādel stresses the importance of a pedagogically packaged approach to metadiscourse, stating that “*Anyone 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 lead to the creation of a taxonomy that, in contrast with the previously presented theories, fully commits to represent metadiscourse function rather than its form.
Instead of focusing on subtypes of metadiscourse, Ådel relies on discourse functions as a guide to classify metadiscourse. Moreover, Ådel assumes a unifying approach to metadiscourse, presenting a general taxonomy, illustrated with examples from both varieties of discourse.

Figure 2.2 summarizes Ådel’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 Ådel’s taxonomy, illustrating the different categories with examples extracted from the original paper.

Metalinguistic Comments

In this category there are five discourse functions: Repairing, Reformulating, Commenting on Linguistic Form/meaning, Clarifying and Managing Terminology. Repairing refers to alterations that aim at correcting preceding statements. 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 contribute with 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 find 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 introduced by Lyons (1977) and discussed in the beginning of this chapter. Clarifying is a function that 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 . . . ”).
2.1. METADISCURSIVE THEORY

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: Ādel’s taxonomy of metadiscourse.

Discourse Organization

Discourse Organization is further divided in 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 of 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, audiovisual materials, etc 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 to make the action of arguing explicit in speech or writing (like in “I argue that…”), and exemplifying, as the name states, is used to explicitly introduce an example (“I will use the example…”).
2.2. NLP APPROACHES TO METADISCOURSE

**Audience References**

In the last category of Ődel’s taxonomy there are five discourse functions, all related to the interaction between speaker/writer and audience. The first two categories, MANAGE COMPREHENSION and MANAGE DISCIPLINE can only be found in spoken communications. 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?”). The remaining functions in this category are ANTICIPATING RESPONSE, MANAGING THE MESSAGE and IMAGINING SCENARIOS. 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 perspective (e.g. “Suppose you are a researcher.” or “Imagine the following situation.”).

2.2 NLP Approaches to Metadiscourse

As mentioned previously, in this section we will describe work resultant from Natural Language Processing (NLP) research that addressed either discourse functions in general, or metadiscourse in particular. It is important to note that this is not an extensive enumeration of the research that focused on discourse or metadiscourse in English. Instead, we aim at describing relevant resources and state of the art techniques related to our task of identifying metadiscourse in spoken language.

There are roughly two types of works being discussed: in sections 2.2.1 and 2.2.2 we present
existent discourse corpora and discourse parsing tools, and in sections 2.2.3 and 2.2.4 we present state of the art work that addressed the annotation and classification of similar phenomena (more precisely the use-mention paradigm and the detection of shell language in argumentative discourse).

2.2.1 Penn Discourse Treebank (PDTB)

The Penn Discourse Treebank (PDTB) is an ongoing project that started with the contribution of Webber and Joshi (1998). PDTB was 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. Initially, PDTB simply enriched the Penn TreeBank with discourse connectives and respective arguments, organizing them under four categories:

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

More recently, 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 senses 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, we are still able to find some intersection with Ådel’s functional taxonomy described in Section 2.1.5. For instance the category Expansion::Instantiation from PDTB can be compared to Exemplifying in Ådel’s taxonomy, and the category Expansion::Restatement is related to Reformulating and clarifying.
2.2. NLP APPROACHES TO METADISCOURSE

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

2.2.2 RST Discourse Treebank

Marcu (2000) developed a semantics-free theoretical framework of discourse relations. Even though Marcu claims having as a major goal the task of automatic text summarization, the framework developed is intended to be “general enough to be applicable to naturally occurring texts and concise enough to facilitate an algorithmic approach to discourse analysis”.

Marcu’s work on discourse gave origin to the RST Discourse Treebank. Similarly to PDTB, the RST Discourse Treebank is a discourse-annotated corpus intended to be used by the NLP community, based on Wall Street Journal articles extracted from the Penn Treebank. The difference between PDTB and the RST Discourse Treebank is the discourse organization framework, which in the case of the RST Discourse Treebank is the Rhetorical Structure
As with the PDTB, some of the categories of rhetoric relations in RST Discourse Treebank intersect with the high-level discourse functions defined by Ādel’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.

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

\begin{footnotesize}
\begin{center}
\begin{tabular}{lll}
\hline
\textbf{ATTRIBUTION} & \textbf{CONTRAST} & \textbf{JOINT} \\
Attribution & Contrast & List \\
Attribution-Negative & Concession & Disjunction \\
\hline
\textbf{BACKGROUND} & \textbf{ELABORATION} & \textbf{MANNER-MEANS} \\
Background & Elaboration & Manner \\
Circumstance & & Means \\
\hline
\textbf{CAUSE} & \textbf{ENABLEMENT} & \textbf{TOPIC-COMMENT} \\
Cause & Example & Problem-Solution \\
Result & Definition & Question-Answer \\
Consequence & Purpose & Statement-Response \\
\hline
\textbf{COMPARISON} & \textbf{EVALUATION} & \textbf{Comment-Topic} \\
Comparison & Evaluation & Topic-Comment \\
Preference & Interpretation & Rhetorical-Question \\
Analogy & Conclusion & \\
Proportion & & \\
\hline
\textbf{CONDITION} & \textbf{EXPLANATION} & \textbf{SUMMARY} \\
Condition & Comment & Summary \\
Hypothetical & & Restatement \\
Contingency & Evidence & \\
Otherwise & Argumentative & \\
\hline
\end{tabular}
\end{center}
\end{footnotesize}

\begin{footnotesize}
\begin{center}
\begin{tabular}{lll}
\hline
\textbf{CONDITION} & \textbf{EXPLANATION} & \textbf{TEMPORAL} \\
Condition & Evidence & Temporal \\
Hypothetical & Argumentative & Sequence \\
Contingency & Reason & \\
Otherwise & & \\
\hline
\textbf{TOPIC-CHANGE} & & \\
Topic-Shift & & \\
Topic-Drift & & \\
\end{tabular}
\end{center}
\end{footnotesize}

Figure 2.4: Simplified Rhetorical Structure Theory categories.

\textsuperscript{1}http://www.isi.edu/licensed-sw/spade/
2.2. NLP APPROACHES TO METADISCOURSE

2.2.3 Wilson (2012)

Wilson’s approach to metadiscourse sits on the spectrum of metasemantics – the use of language to describe and analyze semantics. The use-mention paradigm was introduced by Lyons (1977), and refers to 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.

As a first approach, Wilson (2010) annotated 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 category and their counts on the one thousand sentences sample analyzed by the author. Each category in Wilson’s taxonomy is named after the element that language is commenting (translations, phonetics, symbols).

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 past 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

The author then used that set of words (expanded with the correspondent synset of each word) as **hooks** to retrieve a set of candidate sentences that included mentioned language.

from the corpus. 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 category **Words as themselves** in Table 2.1);
- **Names as Names (NN)** – the sentence directly refers to the phrase as a proper name (similar to category **Proper name** in Table 2.1);
- **Spelling or Pronunciation (SP)** – the text illustrates spelling or pronunciation (similar to category **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.

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, where we can see 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).

<table>
<thead>
<tr>
<th>Category</th>
<th>Global frequency</th>
<th>Frequency in the 100 sample</th>
<th>κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>WW</td>
<td>438</td>
<td>17</td>
<td>0.38</td>
</tr>
<tr>
<td>NN</td>
<td>117</td>
<td>17</td>
<td>0.72</td>
</tr>
<tr>
<td>SP</td>
<td>48</td>
<td>16</td>
<td>0.66</td>
</tr>
<tr>
<td>OM</td>
<td>26</td>
<td>4</td>
<td>0.09</td>
</tr>
<tr>
<td>XX</td>
<td>1,764</td>
<td>46</td>
<td>0.74</td>
</tr>
<tr>
<td>Total</td>
<td>2,393</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Wilson’s annotation results.

From the 2,393 candidate sentences retrieved by searching for the hooks, only about 26% were actually considered to contain mentioned language (1,764 were assigned to the category NOT MENTION (XX)). The expert annotators were able to reach and 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 positive instances according to their function.

### 2.2.4 Madnani et al. (2012)

In Madnani et al.’s work the authors 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 can be compared to the
ARGUING and the categories under Manage Topic from Ådel’s theory. However, the authors do not try to distinguish occurrences according to any theory, encapsulating them under the term shell language, and focusing simply on the detection of those high-level organizational elements in argumentative discourse. The authors do that using two distinct models:

- **Rule-based System** – this model uses a set of 25 hand-written regular expression patterns. These patterns were 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.

### 2.3 Discussion

In this section we started by describing the existing metadiscursive theories that consider spoken discourse. The first two taxonomies that addressed metadiscourse as used in spoken discourse (Luukka, 1992; Mauranen, 2001) focused exactly on one of the aspects that make spoken metadiscourse different from the written variety: the immediacy of feedback. As a
result, both taxonomies organize metadiscourse according to the individual that is speaking and to the amount of stakeholders involved in the communication. Luukka’s and Mauranen’s theories exclusively focused on classifying the form of metadiscourse in spoken language, and do not address its function.

Thompson (2003), on the other hand, takes the first step in the path of analyzing which discourse functions are associated with metadiscourse. The author focuses on discourse organization, 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 different 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.

˝Adel (2010)’s work stands out from the remaining research with a fully functional and wide approach to metadiscourse. The author theory directly aligns with our goal of presentation skills instructions, associating metadiscourse to concepts that seem to be intelligible to non-experts. For those reasons, we adopt ˝Adel’s taxonomy of metadiscourse in the present work. This taxonomy will be discussed further in Chapter 3 when we describe the crowdsourced annotation task aimed at building a corpus of metadiscourse for oral presentations.

We also looked at two resources that annotate discourse. The Penn Discourse Treebank (PDTB) started by enriching the Penn Treebank with discourse connectives, and was later organized under taxonomy of senses which assigned a low-level discourse functions to the connectives previously identified. The RST Discourse Treebank, on the other hand, enriched the Penn Treebank with discourse concepts adapted from the Rhetorical Structure Theory. Even though we found an intersection between these two corpora and the discourse functions in ˝Adel’s taxonomy, the categories that compose them are still very low-level organizational structures, often concerned with the structure at the sentence-level, instead of the sentences
Finally, we analyzed two state of the art NLP approaches to phenomena related to metadiscourse. Wilson (2012) explored the metasemantics part of the spectrum of metadiscourse, describing an annotation task and making considerations on the agreement achieved when using expert annotators to identify and classify instances of mentioned language. Wilson (2012) concluded that while experts can agree on the detection of metasemantics in text passages, they have some difficulties for the task of classifying those passages according to their function. Madnani et al. (2012) described two techniques used to identify shell language in argumentative discourse, discussing the results and performance achieved. The authors proved the complimentary nature of hand-crafted rules and supervised learning to detect high-level organizational elements in written language.

Throughout the literature we found little focus on metadiscourse in spoken language, and to our knowledge, no NLP approaches to this topic. This fact motivated us to build a corpus of metadiscourse as used in spoken communications (see Chapter 3) and develop automatic classification strategies for the phenomenon (see Chapter 4).
In the process of analyzing the existent theories on spoken metadiscourse, we found a fully functional taxonomy that described metadiscursive acts as high-level organizational elements suitable to be used as instructional goals in a presentation skills instructional tool. However, even though Ādel labeled a corpus of lectures with metadiscourse functions, the author only considered the pronominal use of metadiscourse, i.e., instances that contained pronouns. Ādel’s approach ignores occurrences such as “This talk is going to be about…” or “The take home message is…”, which we do not want to exclude in our work. The absence of a suitable resource that could be used to train a classifier of functions of metadiscourse motivated the decision to build a corpus specifically targeted at metadiscourse in English.

We considered two ways to build such a corpus: hiring expert annotators or use a crowdsourcing platform. While hiring experts requires less effort in terms of task design, we soon realized of a major advantage of using the crowd: the opportunity to answer to our first research question. We chose Ādel’s taxonomy because we found that it described concepts in a sufficiently broad and general manner that could be understood by the general public, and consequently used as key concepts in a language learning application. However, as experts ourselves, it is difficult to judge the understanding of non-experts (Nathan et al., 2001). By using the crowd to annotate metadiscourse, we cannot only build the corpus, but also conclude on the non-experts understanding of the concepts involved in Ādel’s taxonomy. We look at the crowd as a reflection of the future students, in the sense that we instruct workers about metadiscourse and its different functions in presentations, and assess their understanding of the phenomenon (herein measured in annotator agreement).
The next four sections describe the process of building an English corpus of metadiscourse under a crowdsourcing approach:

- **Section 3.1** contains considerations about the material that is going to be annotated with the categories extracted from Ādel’s taxonomy;

- **Section 3.2** describes a preliminary annotation task that aimed at understanding the degree of representation of the categories from Ādel’ theory of metadiscourse along the material of choice;

- **Section 3.3** describes the crowdsourced annotation task itself, with considerations about the instructions and quality insurance mechanisms;

- **Section 3.4** presents the results obtained in the annotation task, including the amount of annotations obtained and the inter-annotator agreement achieved.

### 3.1 Source of Spoken Data

Having a set of metadiscursive acts to annotate, the next step was to select a source of data suitable to be used for language learning. In line with our goals, we restricted the analysis to sources that (a) provided video material, which increases student motivation (Choi and Johnson, 2005; De Grez et al., 2009a), (b) could be found on a wide range of topics, speakers and language proficiency levels allowing individual adaptation (Brown and Eskenazi, 2005), and (c) existed in both English and European Portuguese. Two sources of spoken discourse seemed to fulfill these criteria: classroom recordings and the online collection of presentations TED Talks\(^1\).

The comparison between both resources soon led us to choose TED talks over classroom recordings. First, TED talks are known to be good quality presentations from good presenters. Each talk is rehearsed and carefully prepared beforehand, conveying a message in a short span of time (between 5 and 20 minutes). This contrasts with classroom recordings, which are

\(^1\)http://www.ted.com/
3.1. SOURCE OF SPOKEN DATA

typically longer and where there is a sequence in which classes have to be listened to. Even if
only considering self-contained classes, they are targeted at a very specific audience and the
topics are advanced and require a significant amount of previous knowledge. Additionally,
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.

Secondly, TED talks are uniform in content. They contain high quality audio and video
material and are available in several languages. They are also daily updated and subtitled,
providing a good source of transcribed material. Classroom recordings, on the other hand,
are a more heterogeneous resource, in terms of source and recording conditions, making them
harder to be automatically processed 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). For the European
Portuguese (EP) counterpart, we found 9 TEDx\textsuperscript{2} events, from where we collected 118 talks,
totaling around 29 hours. Contrarily to English, we did not find any subtitles, and to further
use this material we will have to transcribe it first. While the ultimate goal of addressing
metadiscourse in both languages will motivate several decisions throughout this work, in the
remaining of this document we will focus on the set of 730 TED talks in English, which we
propose to annotate and use for training the detection of metadiscourse. Considerations on
porting the technology to EP are described in Chapter 5.

In order to confirm our hypothesis that TED talks are accessible to college level students,
we submitted each talk to a readability detector. Our primary concern is that students are
able to focus on the rhetorical patterns involved in metadiscourse, instead of being distracted
by difficult vocabulary. To do this, we ran a lexical level predictor developed by Collins-
Thompson and Callan (2005), that creates a model of the lexicon for each grade level and
predicts the level of a document using word unigram features (Callan and Eskenazi, 2007;
Heilman et al., 2008). Figure 3.1 shows the readability level distribution of the 730 TED talks

\textsuperscript{2}http://www.ted.com/tedx
in English. As we can see, most talks are situated in an intermediate grade level (8), not containing jargon or other complex vocabulary. This characteristic will allow us to provide learning materials that do not distract students from their learning objectives, keeping them engaged and avoiding frustration (Tucker, 1985; Burns et al., 2006).

![Figure 3.1: Readability-level distribution of the 730 TED talks.](image)

3.2 A Preliminary Annotation Task

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 is to find which metadiscursive categories can be found in the TED talks. Ten TED talks were annotated with the tags from Ådel’s taxonomy (see Section 2.1.5 for a detailed description). The ten talks were randomly chosen, spanning different topics and different years. This annotation task was performed by a single person, which had knowledge of both the metadiscourse phenomena and the goals of the current project. The following paragraphs, each named after the 4 main categories of the taxonomy, describe how each category of metadiscourse is distributed over the sample, showing examples of occurrences of metadiscourse in the TED Talks.
### Metalinguistic Comments

*Metalinguistics* refers to the use of language to specifically comment on its form or meaning. From the five metadiscursive acts that compose this category, only **clarifying** and **managing terminology** were found consistently in the sample. The remaining three metalinguistic acts (**repairing**, **reformulating**, and **commenting on linguistic form/meaning**) were not found in our sample in a representative manner. This fact is due to the high degree of preparation of each talk, when compared with the lectures in the MICASE corpus. Below are examples of the functions **clarifying** and **managing terminology**:

- **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 that will enable us to continue to use fossil fuels in a way that is safe.
  - This is a wheat bread, a whole-wheat bread, and it’s made with a new technique that I’ve been playing around with, and developing and writing about which, *for a lack of better name, we call* the epoxy method.

### Discourse Organization

Regarding the first subcategory of *Discourse Organization*, we consistently found metadiscursive occurrences of the functions **introducing topic**, **concluding topic** and **marking asides**. These structures allow the speaker to manage the content of the talk. The short time frame that is allotted for each talk and the fact that the audience comes from a broad set of areas, requires the speakers to wisely structure their discourse to efficiently convey their
message. The remaining two functions (DELIMITING TOPIC and ADDING TO TOPIC) were not found in our sample. Again, the reason behind this may be the fact that TED talks are extensively prepared, with fixed and well-defined topics. The speakers tend to focus on what they want to talk about, going straight to the relevant points.

Regarding *Manage Phorics*, the other subcategory under *Discourse Organization*, we found significant representation of the functions PREVIEWING, REVIEWING and CONTEXTUALIZING. We decided not to analyze the function ENDOPHORIC MARKING since it refers to elements outside the discourse (such as an image in the presentation). Examples of the discourse organization functions represented in our talks sample are:

- **Introducing Topic**
  - *I’m going to talk about* how they are useful when we reflect, learn, remember, and want to improve.
  - But before I go there, *please allow me to share with you* glimpses of my personal story.

- **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, but every other green patch on that has its own history.

- **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.

- **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.
3.2. A PRELIMINARY ANNOTATION TASK

- **Reviewing**
  
  - *And what is so frustrating and infuriating is this: Steve Levitt talked to you yesterday about how these expensive and difficult to install child seats don’t help.*
  
  - *To start with, these women that I told you about are dancing and singing every single day, and if they can, who are we not to dance.*

- **Contextualizing**

  - *Throughout my talk, you will come across several circles.*
  
  - *I’m going to go through them very quickly and then revisit them.*

**Speech Act Labels**

In Ādel’s taxonomy of metadiscourse, the category *Speech Act Labels* is composed of two discourse functions: **arguing**, and **exemplifying**. Both functions were found in significant numbers throughout our ten talks sample.

- **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 and artists need most.*

- **EXEMPLIFYING**

  - *I’ll give you some examples of what modern progress has made possible for us.*
  
  - *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 completely present at the moment of the presentation. This means that the message has to be clearly conveyed without direct interaction between speaker and audience. For these reasons, the tags **manage**...
CHAPTER 3. A CORPUS OF METADISCOURSE

COMPREHENSION (check if the audience is in synch with the content of the presentation) and MANAGE DISCIPLINE (adjusting the channel asking for less noise, for example), were not found 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. They are:

• ANTICIPATING 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 incredible 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.

Figure 3.2 summarizes this preliminary annotation task showing the distribution of the tags most frequently found in the ten talks sample. From the resulting fifteen categories, three criteria dictated the final set of metadiscursive acts to annotate. We considered (a) the functions most frequently found in the literature on metadiscourse and discourse in general, (b) the concepts that we felt could be better explained to non-experts, and (c) the input from our university’s International Communications Center (entity that holds presentation skills
3.3. ANNOTATION USING CROWDSOURCING

Figure 3.2: Occurrences of the most frequent tags in the 10 TED talks sample.

workshops and that is responsible for administrating tests for non-native speakers applying for teaching assistant positions).

Therefore, the discourse functions we decided to pursue and label are six: INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING, MANAGING THE MESSAGE and IMAGINING SCENARIOS. We decided to collapse the functions EXEMPLIFYING and IMAGINING SCENARIOS since they both consist of illustrating an idea. Again, we are not interested in the form of the example (if it involves mentions to the audience or not) but in its function. The merging of the two categories will be further referred to as EXEMPLIFYING. Also, for simplification reasons, MANAGING THE MESSAGE (in Ådel’s work, “used to emphasize the core message in what is being conveyed”) will be referred to as EMPHASIZING.

3.3 Annotation using Crowdsourcing

It has been shown that the quality of the results obtained using crowdsourcing can approach that of an expert labeller, 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).
However, this advantage comes at a cost. Unlike experts, using the crowd requires the set up of training and quality assurance mechanisms to eliminate noise in the answers. Additionally, it is necessary to approach problems in a different way, such as dividing complex jobs in subtasks to reduce cognitive load (Le et al., 2010; Eskenazi et al., 2013).

In our case, these intrinsic characteristics of annotation tasks carried out via crowdsourcing constitute an advantage. By designing a task to annotate metadiscourse, we are not only building a corpus of the phenomenon, but also testing how non-experts react to the different functions in discourse. The way we address this task is to ask for workers on Amazon Mechanical Turk (AMT)\(^3\) to annotate the transcripts of the 730 TED Talks with the set of five discourse functions defined in the previous section: INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING, and EMPHASIZING. It is important to say that we are submitting for annotation every sentence that is part of the transcripts of each talk. This contrasts with Ádel (2010) and Wilson (2012) for example, who used a predefined set of words to retrieve an initial set of candidate sentences that were later annotated.

The first decision while setting up this annotation task had to do with the amount of text that workers would annotate in each Human Intelligence Task (HIT) – the smallest unit of work someone has to complete in order to receive payment. Each HIT should be simple and allow workers to solve it in the fastest way possible. However, metadiscursive phenomena are not local, \textit{i.e.}, usually require understanding of the surrounding context to be identified. Having this in mind, we decided 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, influencing the answer rate).

To pay an amount of money that is worthwhile for the worker to choose our task, we included four segments per HIT, shown in a 2-by-2 matrix. Figure 3.3 shows the interface for one of the segments in a HIT. This configuration generated 2,461 HITs (or 9,844 segments) per category.

\(^{3}\)https://www.mturk.com/mturk/welcome
It is also important to notice the presence of the button *See more context* in Figure 3.3. This feature allows the worker to see the surrounding text of the segment in the talk (before and after), in case they need additional context to support their decision.

The second consideration concerns the design of the instructions. Knowing that metadiscourse act is a complex notion which workers may have never heard of, we decided that each HIT would target one category only, instead of requiring the identification of all five categories in each segment in one single pass. This decision lessens the cognitive load for the workers at each point. The 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...”. The passages you mark will be used on a presentation skills virtual tutor, showing students how professional speakers EMPHASIZE a point.

Since the idea is to do one pass over all the segments for each one of the categories of metadiscourse, we designed different sets of instructions depending on each category is being annotated. Common to all categories is the first paragraph on the instructions above. It is important to note that this paragraph was added only after some preliminary trials. Its inclusion was intended to reveal the high level purpose of our work, as a way of motivating workers. And in fact, the inclusion of this paragraph increased the response rate after a first small trial.

After the instructions there is a section of examples (Figure 3.4) and counterexamples (Figure 3.5). Each element was presented to the workers with a description of why they were considered (or not) an example of the metadiscursive marker being labeled.

Figure 3.4: Positive examples for the task of identifying occurrences of EMPHASIZING.
3.3. ANNOTATION USING CROWDSOURCING

Finally, at the bottom of the page, before the segments to annotate, there is a succinct description of the set of steps that explain the interface and how to use it to annotate the passages:

- **STEP 1:** For each of the extracts below, click on EVERY word that the speaker uses to EMPHASIZE A POINT. There may be zero, one or more instances in each extract.

- **STEP 2:** The words you click on will display a light blue background. If you change your mind, you can click on the word again to deselect it.

- **STEP 3:** If you need more information to support your decision, you can click “See more context” below the segment to see the its surrounding context in the talk.

- **STEP 4:** If the speaker does not emphasize any point in the extract, select the “No occurrences in this text” checkbox below the text.

- **STEP 5:** Click the SUBMIT button once you are finished.

The last set of considerations has to do with quality control. We took advantage of the AMT prerequisites feature to filter for workers that were native-speakers of English and who
had a reliability rate of at least 95% (percentage of the worker’s HITs that were accepted). Workers who satisfied the prerequisites and accepted the HIT would then be guided through a training session of four segments. This session tested if the worker read the instructions and examples carefully, giving feedback on each of their decisions. Only upon successful completion of the four training segments were the workers allowed to access real HITs in the category they were just certified on.

While the training strategy described is effective in filtering out bots, it will not prevent bad-intentioned workers from complying with the training session but give random answers to the real HITs. For that reason, and in line with what is done in much of the crowdsourcing community Hsueh et al. (2009); Eskenazi et al. (2013), we defined a gold standard (a set of segments where the answer is known) for each of the five metadiscursive categories. In every four HITs, at least one segment was compared to an expert annotation. The gold standard segments were very similar to the examples provided, and failing one of them flagged the worker as a potential spammer. The decision to accept or reject the work of flagged workers was done manually, by analyzing each case separately. As Figure 3.3 shows, workers also self-assessed their confidence level for each segment on a 5 point Likert scale, which gave us another indicator of the global quality of the worker. A final mechanism to assure quality consisted of submitting the same HIT to 3 different workers, using a majority vote scheme.

Prior to publishing all the HITs in each category, we uploaded a small sample of 100 HITs to test the suitability of the instructions and interface. This trial phase allowed us to modify the instructions and examples for each category if necessary, and to assess the workers’ understanding of the concepts they were asked to annotate.

3.4 Annotation Results

This section presents the results of the crowdsourcing task for each of the metadiscursive acts considered in this study. Tables 3.1 and 3.2 summarize the results for each category. In Table 3.1 we report behavioral data, related with the workers performance while annotating
the transcripts in terms of time spent per HIT (in minutes), self-assessed confidence score (average on a 5-point Likert scale), and percentage of segments in which workers expanded context (by clicking on the See more context button). Table 3.2 reports the quantity and quality of the annotations obtained, showing the number of occurrences of each metadiscourse category and the corresponding inter-annotator agreement ($\kappa$). We used the Fleiss’ kappa (Joseph, 1971) as a measure of annotator agreement, considering that annotators agree if for each sequence of words selected, there is at least one in common between annotators, i.e., if the intersection of the words selected by each annotator is not empty. The occurrences number for each category is the number of sentences that were annotated as belonging to the category by at least two workers.

<table>
<thead>
<tr>
<th>Category</th>
<th>Time-on-task (m)</th>
<th>Self-Reported Confidence</th>
<th>Context Requests (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marking Asides</td>
<td>10</td>
<td>3.60</td>
<td>5.52</td>
</tr>
<tr>
<td>Introducing Topic</td>
<td>3.7</td>
<td>3.95</td>
<td>1.32</td>
</tr>
<tr>
<td>Concluding topic</td>
<td>3.5</td>
<td>4.00</td>
<td>37.09</td>
</tr>
<tr>
<td>Exemplifying</td>
<td>6.2</td>
<td>3.94</td>
<td>4.81</td>
</tr>
<tr>
<td>Emphasizing</td>
<td>6.3</td>
<td>3.99</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Table 3.1: Annotation results in terms of workers’ behavior.

<table>
<thead>
<tr>
<th>Category</th>
<th># occurrences</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introducing Topic</td>
<td>1,159</td>
<td>0.64</td>
</tr>
<tr>
<td>Concluding topic</td>
<td>628</td>
<td>0.60</td>
</tr>
<tr>
<td>Exemplifying</td>
<td>1,327</td>
<td>0.72</td>
</tr>
<tr>
<td>Emphasizing</td>
<td>2,581</td>
<td>0.58</td>
</tr>
</tbody>
</table>

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

Marking Asides

It is important to notice the absence of the category Marking Asides in Table 3.2. All the categories with the exception of Marking Asides produced satisfying results in the trial sample of 100 HITs uploaded prior to submitting the entire set of 730 talks. This fact lead us to discard the asides-related category. We did not proceed with its annotation.

The first indicator of unsuccessful annotations was the slow response rate. The 100 HITs were up for one week, a time frame in which only less than 50% were completed. This slow response
rate could be due to the small amount of HITs that were uploaded, since workers tend to focus on tasks that have a significant amount of HITs online (in the order of thousands), minimizing training time and maximizing payment. However, this hypothesis was rejected since this behavior was not observed in any of the remaining categories, where the 100 samples were fully completed in less than two days.

As we looked further to other indicators, we realized that the crowd could actually be giving us some insight on how appropriate it was to consider MARKING ASIDES as a key concept. Workers were spending 10 minutes on average for each HIT in this category, 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.

Finally, we looked at the comments left by the workers while annotating, which clearly show that this specific task was too hard, justifying the slow response rate and lack of confidence. Workers wrote: “I am nervous that I am not doing these correctly *at all*”; “I hope that this is what you are looking for”; “Hope I’m doing good in my job”; and “little difficult”.

**Introducing Topic**

The task of annotating the category INTRODUCING TOPIC resulted in an inter-annotator agreement of 0.64, considered to be substantial agreement, in the scale proposed by Landis and Koch (1977). Workers took, on average, 3.7 minutes to complete each HIT and identified around 1.2 thousand occurrences of INTRODUCING TOPIC in the entire set of 730 talks. Similarly to what we registered in the preliminary annotation task, the number of instances of INTRODUCING TOPIC is larger than the number of talks (13 occurrences in 10 talks). This is due to the fact that speakers introduce several topics throughout a single talk. A final interesting point was the low amount of times that workers asked for additional context: only in 1.32% of the segments.
3.4. ANNOTATION RESULTS

Concluding topic

The annotation of the category CONCLUDING TOPIC provided analogous results to the previous category: a slightly lower inter-annotator agreement ($\kappa=0.60$), and similar average time-on-task and self-reported confidence. An important different result comes from the percentage of segments for which annotators asked to see the surrounding context: 37% of the time. This result might be an indication that conclusions are less local, and people need a wider context to identify them. It is also important to notice that the number of occurrences of CONCLUDING TOPIC (around 600) is lower than the total amount of talks and approximately half the number of instances of INTRODUCING TOPIC. Again, these results align with what we encountered in the preliminary annotation task (7 conclusions over 10 talks). Speakers do not explicitly conclude every topic in the talk (particularly true for shorter talks). This discourse function can be performed, for instance, by an introduction to a new topic, which implies the closing of the previous one.

Exemplifying

In this category, workers spent on average two more minutes per HIT than while annotating the categories INTRODUCING TOPIC and CONCLUDING TOPIC. This fact might result from the greater quantity of occurrences detected (1,327) and additional context requests (4.8%). The more occurrences a category has, the more time workers will spend clicking on them. While identifying occurrences of the function EXEMPLIFYING, workers showed comparable self-reported confidence to the previous categories (3.94). As previously described, this category collapses two metadiscursive acts, as defined in Adel’s taxonomy: EXEMPLIFYING and IMAGINING SCENARIOS. Despite the collapse of tags, in this category annotators reached the highest agreement ($\kappa=0.72$), which corroborates the decision to unify both categories under the same functional concept.
CHAPTER 3. A CORPUS OF METADISCOURSE

Emphasizing

Regarding the metadiscursive marker **Emphasizing**, the relationship between average time-on-task and number of instances is similar to the one found for the previous category. Workers spent on average 6.3 minutes per HIT and identified over 2.5 thousand occurrences. While identifying emphasis, workers asked for the lowest amount of additional context amongst the five categories (only in 1.14 % of the segments). **Emphasizing** was the category where workers achieved the lowest inter-annotator agreement (0.58), considered to be *moderate agreement* (Landis and Koch, 1977). This result may be due to the fact that this category is the only one in which there is a scale of intensity related to the concept being annotated, *i.e.*, different workers might have different thresholds for considering that the speaker is emphasizing.

Other Results

The two additional results we report use the notion of readability from Collins-Thompson and Callan (2005), which we used in the beginning of this chapter (Section 3.1) to conclude on the vocabulary level of the TED talks. It is important to remember that this readability measure assigns a level to a given document using word unigram features only. Therefore, herein it is used as an indicator of the difficulty of the vocabulary used in the presentations, rather than a predictor of the level of the presentation (for which it would have to be retrained and possibly include features such as speaking rate or grammatical complexity).

The first dimension we analyzed was the relation between the quality of the annotations and the readability level of the transcriptions of the talks. Table 3.3 shows the inter-annotator agreement per level. The correlation between the two variables is positive and negligible (0.2556). We can therefore conclude that vocabulary complexity does not affect the workers’ performance when detecting metadiscursive acts.

<table>
<thead>
<tr>
<th>Level</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>$\kappa$</th>
<th>CORR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$</td>
<td>0.72</td>
<td>0.65</td>
<td>0.65</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.73</td>
<td>1.00</td>
<td>0.2556</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Inter-annotator agreement by level.
3.5. DISCUSSION

Additionally, we looked at the relation between the number of occurrences of each metadiscursive category and the level of the talks. Table 3.4 shows the percentage of sentences in a certain level that contain occurrences of each metadiscursive category, and in the last column ($corr$), the weighted linear correlation between vocabulary level and occurrence percentage. For the categories Introducing Topic and Concluding topic we found no significant correlation between the amount of metadiscourse and the vocabulary level of the talks. On the other hand, we found meaningful correlations for the categories Exemplifying (0.83) and Emphasizing (0.56). The use of more complex vocabulary seems to be correlated with increasing occurrences of Exemplifying and Emphasizing markers.

<table>
<thead>
<tr>
<th>Category</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>$corr$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introducing Topic</td>
<td>0.00</td>
<td>1.03</td>
<td>1.11</td>
<td>1.12</td>
<td>1.10</td>
<td>1.04</td>
<td>1.20</td>
<td>0.68</td>
<td>0.1104</td>
</tr>
<tr>
<td>Concluding topic</td>
<td>0.00</td>
<td>0.88</td>
<td>0.60</td>
<td>0.60</td>
<td>0.61</td>
<td>0.52</td>
<td>0.71</td>
<td>0.68</td>
<td>-0.2404</td>
</tr>
<tr>
<td>Exemplifying</td>
<td>0.00</td>
<td>0.73</td>
<td>0.89</td>
<td>1.05</td>
<td>1.05</td>
<td>1.58</td>
<td>2.52</td>
<td>0.68</td>
<td>0.8141</td>
</tr>
<tr>
<td>Emphasizing</td>
<td>1.27</td>
<td>2.50</td>
<td>2.51</td>
<td>1.93</td>
<td>2.49</td>
<td>2.71</td>
<td>2.47</td>
<td>2.72</td>
<td>0.6425</td>
</tr>
</tbody>
</table>

Table 3.4: Percentage of sentences with metadiscourse occurrences per level.

3.5 Discussion

In this chapter we described an annotation task that took place on Amazon Mechanical Turk (AMT), where workers focused on a predefined set of metadiscourse categories from Ådel (2010) to label TED talks transcripts with metadiscursive data. We started by comparing potential sources of spoken discourse to annotate, and saw how different materials have different coverage of metadiscursive functions. By using a corpus of lectures (MICASE), Ådel found decent amounts of metadiscourse being used for reformulating, repairing, and manage the audience comprehension. However, when we analyzed the intersection of Ådel’s theory with a set of ten TED Talks transcripts, we did not find the full category spectrum.

As we discussed earlier, the immediacy factor of both production and feedback often present in spoken discourse, affects the way metadiscourse is used in this specific setting. But the degree of immediacy in spoken language also varies according to the situation, ending up
conditioning the presence of some discourse functions. For instance, production is affected by presentation preparation. Generally speaking, an event such as a TED talk requires more preparation than a lecture, and as a consequence, the speaker is less prone to use repairing or reformulating strategies. Also, the immediacy of feedback is different between a lecture and a TED talk, impacting the way the speaker manages comprehension and participation. Students attending a lecture can raise their hand and ask a question, directly interfering with the content of the message, while in TED talks the feedback during the time of the discourse event is limited (nods, applause and laughs).

After having decided on using TED talks as the material of choice, we reached a consensus regarding an initial set of concepts to focus on (based on both existent literature in metadiscourse and advice from professional presentational skills instructors). We described the considerations behind the setup of a crowdsourcing task to annotate occurrences of the categories Marking Asides, Introducing Topic, Concluding Topic, Exemplifying and Emphasizing. By approaching the crowd as a reflection of the students, we got not only annotations for four of the categories, but also insight on what can be understood by non-experts and what can be difficult to explain.

The annotation was successfully completed for four of the five categories that we proposed initially. The category Marking Asides was discarded during the trial phase on AMT since workers manifested signs of not understanding the task. This means that not every item in the theoretical framework of metadiscourse can be understood/explained to the same extent. For the remaining four categories we were able to achieve satisfying inter-annotator agreements, situated between 0.68 and 0.72. The lowest agreement (0.68) corresponded to the only category where we can admit different levels of intensity: Emphasizing. Different workers have different thresholds for considering that the speaker is emphasizing. By not establishing a level of intensity of the function in the instructions of the annotation task, we allowed more variability, which ended up reflected in the results. Despite that, these numbers are comparable to what Wilson (2012) reports (see Section 2.2.3). The author had three experts classify sentences according to four categories of metalanguage, reaching agreements
between 0.38 and 0.74.

We also noted variability in respect to the need for additional context when annotating metadiscourse. Workers were able to identify occurrences of Introducing Topic and Emphasizing in a window of 300 words without requesting additional context. On the other hand, while annotating instances of Exemplifying, the crowd expanded the context 5% of the time and an unexpected 37% of the time for the category Concluding topic. These numbers might reflect the necessity to first understand which element is being presented, before deciding that such element is being exemplified or concluded. It is possible that the difference in proportion of context requests (5% vs. 37%) reflects the distance between the introduction of the element and its example/conclusion (an example is closer to the element that is being exemplified, while a conclusion is further away from the point where the element first appeared).

Regarding the influence of the vocabulary complexity in the metadiscursive phenomena, we found a positive correlation between the level of the talks and the occurrences of two categories – Exemplifying and Emphasizing. In other words, when the vocabulary in a presentation is more complex (assigned to higher readability levels), the speakers tend to use more examples and instances of emphasis. On the other hand, the level of the vocabulary of the talks does not influence the number of occurrences of Introducing Topic or Concluding topic. Additionally, we concluded that the presence of complex vocabulary has no effect on the ability to successfully label any of the metadiscursive tags we explored.

So what can we say, for now, about our first research question? By the results obtained in this first annotation task, it seems that non-experts are able to agree when identifying and assigning a function to some metadiscursive events. However, not all the concepts in the same taxonomy can be understood to the same extent. In our case, the crowd (as non-experts) showed discomfort while annotating instances of Marking Asides. It seems that there is a limit for comprehension and we need to explore more in order to generalize. The exploration of additional metadiscursive markers and respective functions is part of the proposed work of this thesis (see Chapter 5).
In Chapter 2, we looked at two 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. Madnani et al. (2012), on the other hand, focused on the identification of *shell text* in argumentative essays. Both studies mainly targeted metadiscourse as used in written discourse. Madnani et al. attempts a comparison between both varieties, applying the algorithms 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 in a qualitative manner, since the corpus of political debates was not previously labeled with occurrences of *shell language*.

Although we can refer to the work focusing on written form (such as Wilson (2012) and Madnani et al. (2012)) for performance comparison reasons, it has limited applicability with respect to the goals we established in the introductory chapter. In written language, typically, the author does not have to deal with 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. It is in this setting of nonexistent previous work on spoken metadiscourse in the NLP area (to our knowledge), that we describe a first approach to automatic detection and categorization of metadiscursive phenomena in spoken discourse.

Herein we take the first step towards our second research question: to what extent can we automate the detection of metadiscourse? What features and traces constitute good detectors of the phenomenon? In this chapter, we describe a supervised learning setup that takes advan-
CHAPTER 4. AUTOMATICALLY CLASSIFYING METADISCUSSION

tage of the material collected via crowd (Chapter 3) to identify passages of metadiscourse and classify them under four functional categories: Introducing Topic, Concluding topic, Exemplifying and Emphasizing.

It is important to point out here again that, in this thesis, we propose to address metadiscourse in both English and European Portuguese. While this proposal focuses only on metadiscourse in English, the ultimate goal of exploring both languages influenced some of the implementation decisions. This fact will be referred to when relevant. This chapter is divided in two sections:

- **Section 4.1** describes the experimental setup, including the formatting of the training data, the set of features explored and the specifics of the algorithm used to automatically identify and classify metadiscourse;

- **Section 4.2** presents the results of the classification task, exploring the different features considered in our approach and discussing how they relate to the performance of the algorithm in general, and to the nature of metadiscourse in English in particular.

### 4.1 Experimental Setup

Having a corpus of annotated material (described in Chapter 3), we use supervised learning to address the task of automatically classifying metadiscourse occurrences according to their function in spoken discourse. We formulate this classification problem as: For each sentence, decide if it contains occurrences of metadiscourse being used to introduce a topic, conclude a topic, exemplifying or emphasizing. We consider one classifier per function, in other words, each sentence is submitted to four classifiers, each classifier outputting a binary decision of whether the sentence contains an occurrence of the respective metadiscourse act or not. It is important to note that, in this experiment, we consider sentences as units of classification. This contrasts with Madnani et al. (2012), for instance, who proposed identification of shell text at token level.
In this experimental setup we use syntactic and lexical features and decision trees to support the classification of metadiscourse. There are two main reasons behind the use of this specific setting. The first has to do with the practices found in the literature of metadiscourse related-phenomena. As we saw in Madnani et al. (2012) and Wilson (2012) the authors rely mostly on lexical information to support their classification decisions. Using previous experience, Wilson defines a set of words that are indicators of metalanguage (such as meaning, sentence, or symbol), while Madnani et al. (2012) asks for experts to create word patterns that match occurrences of shell text in argumentative student essays. Similar lexical approaches are also found 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 common idea behind all these studies is that words can be used as indicators of the presence of the phenomenon being studied. For instance, in sentiment analysis some words are associated with positive opinions, others are neutral and others have negative connotations. In Pedersen (2001) and Xiong and Litman (2010) in particular, the authors use decision trees, inferring word co-occurrence rules that are able to classify word-senses and detect peer-review feedback with reasonable accuracy rates.

The second reason behind choosing this particular strategy has to do with the nature of the output. A decision tree classifier is simply composed of a set of rules that were inferred in the training phase, and their corresponding order of application. These rules are intelligible and can be interpreted by humans. Therefore, the output of a decision tree can be modified and adapted to different domains, or in our specific case, translated to European Portuguese. This necessity has to do with data scarcity and the resulting inability to train models. These considerations are described in detail in the next chapter (Section 5.3).

Algorithm and Implementation Specifics

Decision trees are based on the concepts of Information Gain (IG) and Entropy from Shannon and Weaver’s theory of communication (Shannon and Weaver, 1948). Simply put, Information Gain measures the reduction of uncertainty in a set that results from the occurr-
ence of an event. Formally, the Information Gain obtained when the event $x$ occurs in a set $T$ \((4.1)\) is given by the difference between the original entropy of the set $T$ \((4.2)\) and the entropy of the set after $x$ occurring \((4.3)\).

$$IG(T, x) = H(T) - H(T|x) \quad (4.1)$$

$$H(T) = -\sum_{i=0}^{n} p(t_i) \log_2 p(t_i) \quad (4.2)$$

$$H(T|x) = -\sum_{j=i}^{n} p(y_i|x) \log_2 p(y_i|x) \quad (4.3)$$

In decision trees, the events $x$ are pairs feature-value. In each iteration of the training phase of a decision tree, we are trying to find the event (pair feature-value) that best divide the training set, \textit{i.e.}, the one that reduces uncertainty (entropy) the most. In our specific setup we use Weka’s\(^1\) implementation of the C4.5 algorithm (J48), which briefly operates in three steps:

1. For each pair feature-value, $x$, find the information gain from splitting the training set on $x$. Let $y$ best be the attribute with the highest information gain;

2. Create a decision node that splits the training data on $y$, adding the resulting subsets as children of the node;

3. Recurse for each side of the newly formed node.

In order to use this implementation we need to preprocess the annotations obtained in the annotation task. The first step towards preparing the data for training was to use the Stanford Tokenizer\(^2\) to split each talk into sentences. Then, for each sentence, we looked at the results of the annotation task for each category of metadiscourse, and applied a majority vote rule (two out of three annotators agree) to classify the sentence as containing instances of metadiscourse.

---

\(^1\)http://www.cs.waikato.ac.nz/ml/weka/

\(^2\)http://nlp.stanford.edu/software/tokenizer.shtml
of a specific category or not. As we mentioned before, we consider that two annotators agree if the intersection between their annotations is not empty. With this formulation, we ended up with four sets of sentences – one for each category. The amount of positive examples in each category corresponds to the numbers shown in Table 3.2, in Chapter 3.

In the next step we balanced the data so as to have the same amount of positive and negative instances in each category. Using these four sets of balanced examples, we computed the features for each configuration (described below) and proceeded to training the decision tree. The last detail worth mentioning is the adjustment of the parameter \texttt{minNumObj} which imposes a limit on the minimum number of instances per leaf. In order to avoid overfitting, we established that a new rule should only be generated if it applies at least to 5 points in the training set.

**Features**

As we just saw, the execution of the decision tree algorithm depends on the events $x$ tested in each iteration. In decision trees, these events are pairs \texttt{feature-value} that, in the end, constitute the rules inferred by the algorithm. So the set of features to consider should be representative of metadiscursive phenomena. As described in the beginning of the current chapter, in line in what is done in the literature for similar phenomena, our focus in this first attempt goes to lexical indicators.

We base this lexical approach 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. It is important to notice that subtitles differ from expert transcripts, in the sense that they tend to omit disfluencies that typically arise from the immediacy factor present in speech settings, such as filled pauses, deletions, fragments, repetitions or substitutions (Moniz et al., 2012). Literature shows that the removal of disfluencies does not influence comprehension (Jones et al., 2003, 2005), being a common technique when automatically creating textual representations of speech data, such as in the tasks of automatic speech recognition (Stouten et al., 2006) and summarization (Zhu and Penn, 2006).
CHAPTER 4. AUTOMATICALLY CLASSIFYING METADISCOURSE

When looking at the subtitles of the TED talks we collected, we noticed that they tend to ignore short disfluencies, such as fragments composed of individual words, but consider longer events. Such artifacts are known to produce ungrammatical content that language processing tools that are trained on text are unable to deal with (Hayes et al., 1986).

To confirm the suitability of a lexical approach, we looked at the top \(n\)-grams in our annotated corpus for the different categories. Table 4.1 compares the top ten unigrams, bigrams and trigrams, between the entire set of 730 TED Talks, and the annotations given by the crowd for the four categories we are considering. The differences between the highest frequent \(n\)-grams for the total set of talks and the individual categories are clear. INTRODUCING TOPIC ranks higher \(n\)-grams that contain words such as talk, show or tell; CONCLUDING TOPIC contains high frequency of leave and conclude; EXEMPLIFYING, as expected, ranks higher words like example and imagine; and finally, the most frequent \(n\)-grams for the category EMPHASIZING contain words such as important and remember. It is also interesting to see how the categories INTRODUCING TOPIC, CONCLUDING TOPIC and EMPHASIZING share some of the highest frequent \(n\)-grams, particularly in constructs involving the words like as in “I would like...”, and want as in “I want to...”

In this experiment we want to take advantage of this lexical information and see how much representative of spoken metadiscourse it is. Aside from lexical features, we also consider syntax as an indicator of metadiscourse. In sum, we consider four sets of features:

- **Part-Of-Speech \(n\)-grams** – presence of POS unigrams, bigrams and trigrams in the sentence. The categories considered is the set of 36 POS tags\(^3\) provided by the Stanford Parser\(^4\)(Klein and Manning, 2003);

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

---

\(^3\)http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (June 2013)
\(^4\)http://nlp.stanford.edu/software/lex-parser.shtml (June 2013)
4.1. EXPERIMENTAL SETUP

<table>
<thead>
<tr>
<th>n</th>
<th>TED talks</th>
<th>Introducing Topic</th>
<th>Concluding topic</th>
<th>Exemplifying</th>
<th>Emphasizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the, and, to, of, a, that, in, is, I, you</td>
<td>to, you, I, about, going, want, talk, show, tell, like</td>
<td>to, you, I, with, so, and, the, leave, like, that</td>
<td>example, for, imagine, you, of, an, a, is, examples, to</td>
<td>is, the, to, important, I, this, that, you, want, what</td>
</tr>
<tr>
<td>2</td>
<td>of the, in the, this is, and I, going to, and the, to be, to the, on the, is a</td>
<td>going to, I am, am going, want to, I want, to talk, talk about, tell you, show you, like to</td>
<td>leave you, you with, like to, want to, I would, would like, I will, will leave, I want, to conclude</td>
<td>for example, an example, example of, give you, for instance, example for, look at, this is, you an, if you</td>
<td>I want, this is, want to, I think, is that, is important, very important, you to, the most, most important</td>
</tr>
<tr>
<td>3</td>
<td>a lot of, this is a, one of the, I am going, am going to, this is the, I want to, and this is, you can see, going to be</td>
<td>I am going, am going to, I want to, to talk about, to tell you, I would like, would like to, to show you, want to talk, going to talk</td>
<td>leave you with, I would like, I would like to, I will leave, will leave you, I want to, to leave you, the last thing, like to leave, like to conclude</td>
<td>you an example, give you an example, is an example, an example of, I will give, will give you, let me give, me give you, this is an, to give you</td>
<td>I want to, want you to, I want you to, I want you, the most important, this is important, one of the, to point out, the bottom line, you to remember, we need to</td>
</tr>
</tbody>
</table>

Table 4.1: Top n-grams in the TED talks compared to annotations of the four metadiscursive categories.

- **Word n-grams** – presence of word unigrams, bigrams and trigrams as they exist in the transcript (inflected forms of nouns and verbs, etc);

- **Position and Length Features** – contain three additional features:
  - **Length of the sentence** – how many words are in the sentence;
  - **Position of the sentence in the talk** – position of the sentence in the talk (normalized in the [0, 1] interval);
  - **Distance to the last occurrence** – number of sentences between the sentence being currently classified and the last occurrence of the metadiscourse act being considered.
For the first three sets of features (POS, lemmas and words) we considered both setups with and without stop words\(^5\). 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 the size of the models, 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). On the other hand, 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 our case, stop words seem to rank higher in the language models of each metadiscourse act when compared to the language model of the TED talks, as we can see in Table 4.1 for the words to, you and I. Therefore, we hypothesize that occurrences of these words can constitute good indicators of the presence of metadiscourse.

In our settings we represent feature presence rather than frequency. As an example of how we represent our training data, Table 4.2 shows two points for the unigram words setting used to identify instances of INTRODUCING TOPIC. The character \( T \) (true) associated with a word, means that that word is present in the training item, and the character \( F \) (false) means that it is not. The last column \textit{class} contains the decision whether the sentence contains an instance of INTRODUCING TOPIC or not. In this case, both sentences are examples where the speaker is introducing a topic.

<table>
<thead>
<tr>
<th align="left">( I ) am going to talk about it</th>
<th>about</th>
<th>am</th>
<th>art</th>
<th>going</th>
<th>I</th>
<th>it</th>
<th>talk</th>
<th>to</th>
<th>will</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td align="left">( I ) will talk about art</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 4.2: Two training data points for the category INTRODUCING TOPIC and the unigram words feature setting.

A final note goes to feature reduction considerations. The process of inferring a decision tree is time and memory consuming, especially when considering thousands of features, such as all the word trigrams in the training set. In line with what is done in the literature (such as

\(^5\)http://www.ranks.nl/resources/stopwords.html (June 2013)
4.2. **CLASSIFICATION RESULTS**

In Abbasi et al. (2008), for sentiment analysis, we used the Information Gain (IG) metric to discard some of these features. This reduction of the search space is a greedy strategy, since the IG is calculated only on the initial training set, and therefore the choice of threshold can impact the results. In our case, we considered only $n$-grams where $IG > 0.0025$, basing our decision in text feature selection literature (Yang and Pedersen, 1997).

### 4.2 Classification Results

In this section, we show the results in terms of accuracy of classification for four metadiscourse classifiers – **Introducing Topic**, **Concluding topic**, **Exemplifying** and **Emphasizing**. Accuracy is defined as the percentage of correctly classified items. The numbers reported in this section are all result of an evaluation made using a 10 fold cross-validation strategy. It is important to remember that the training set for each category was balanced in terms of positive and negative instances, and therefore the accuracy obtained by chance is 50%.

As mentioned previously, we considered both the inclusion and removal of stop words in our experiments. However, stop word removal proved to hinder the classification for all four categories, producing statistically significant worse results for **Introducing Topic** and **Emphasizing** (decrease in the order of 2% in accuracy). For this reason, in the following sections we report only the results where we consider stop words.

#### n-grams of POS, Lemma and Word

The first setting we tested was using $n$-grams of POS, lemmas, and words individually ($n \in \{1, 2, 3\}$). Table 4.3 reports accuracy for each pair features-category. Naturally, the bigram models also include unigram features, and the trigram models, include both unigrams, and bigrams. Bold face numbers represent the best settings in each category, and results marked with * represent options that are at the same statistically significance level ($p<0.05$) as the best setting.

Before analyzing each category individually, let us first take a look at some general results.
CHAPTER 4. AUTOMATICALLY CLASSIFYING METADISCUSSION

<table>
<thead>
<tr>
<th>Category</th>
<th>POS n-grams</th>
<th>Lemmas n-grams</th>
<th>Words n-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Introducing Topic</td>
<td>79.59</td>
<td>85.11</td>
<td>86.41</td>
</tr>
<tr>
<td>Concluding topic</td>
<td>65.82</td>
<td>66.37</td>
<td>68.13</td>
</tr>
<tr>
<td>Exemplifying</td>
<td>61.82</td>
<td>65.44</td>
<td>68.19</td>
</tr>
<tr>
<td>Emphasizing</td>
<td>67.31</td>
<td>68.15</td>
<td>68.86</td>
</tr>
</tbody>
</table>

Table 4.3: Classifiers accuracy for the unigrams, bigrams and trigrams of POS, Lemma and Word Features.

As expected, the syntactic approach provided the poorest results across the four categories of metadiscourse. All the formulations of lemmas or words as features performed significantly better than part-of-speech approaches in all four categories. However, the POS n-gram setting proved to contain some representation of the metadiscursive phenomenon, achieving accuracies in the order of 60 to 70% for the categories of Concluding topic, Exemplifying and Emphasizing, and an impressive 86% for the category Introducing Topic. Another result shared by all four categories was the statistically equivalent performance of the lemma and word models. All the settings involving lemma and word n-grams performed at the same significance level, with the exception of unigrams of lemmas which was always surpassed by at least one other model, and the word unigram setting for the category Concluding topic.

Introducing Topic Amongst the four categories, Introducing Topic was the one where the models using syntactic information achieved highest accuracy, with the POS trigram model performing at 86.4%. Despite this considerably high performance, all models based on lexical information (words and lemmas) outperformed this solution with an increase of about 15% in accuracy. The bigram and trigram word models achieved the best results for this category, with an equal accuracy of 92.71%. In fact, both models produced exactly the same output, which means that no trigrams of words were included as a rule in the final tree. These two setups were the only ones that significantly outperformed the lemma unigram model. However, their accuracy is not statistically better than any of the other
models involving words or lemmas. Also, within the same feature type (lemmas and words) unigrams, bigrams and trigrams performed at the same level, i.e., lower-order n-grams proved to have the same representative power than higher ones. Regarding the false positive/negative rate we registered a greater amount of false negatives (8.8%) than false positives (5.8%).

**Concluding topic** The model that better predicted occurrences of **Concluding topic** was the lemma bigram model (86.9% accuracy). However, the difference to the lemma trigram model and both models of word bigrams and trigrams is not statistically significant. Similarly to what saw in the previous category, moving from syntactic to lexical features improved 15% of prediction accuracy. For the category **Concluding topic**, unigram models are consistently outperformed by the respective higher-order models within the same feature category. As for **Introducing Topic**, the performance of detecting instances of **Concluding topic** is mostly being affected by the amount of false negatives, which correspond to a percentage of 19.8% of the negative examples, contrasting with 6.4% of false positive occurrences.

**Exemplifying** In this category the impact of the transition from syntactic to lexical features was the most accentuated, with an increase of 25% of overall performance. The best setting for the category **Exemplifying** was the word trigram model (94.9%), the only one who significantly surpassed the performance of the lemma unigram model. However, this best setting is statistically similar to any of the other configurations involving words or lemmas. **Exemplifying** was the category that achieved best overall performance amongst the four discourse functions analyzed. Trigrams of words were enough to correctly classify 95% of the data, with a false positive rate of 2.4% and a false negative rate of 7.8%. Figure 4.1 shows the tree for the word bigrams model, composed of only 19 rules (12 leaves). We can see that the features *example, imagine, examples and instance* alone account for the correct classification of 1,149 positive instances (about 87% of the instances existent in the corpus).

**Emphasizing** This category follows the same trend as the categories **Introducing Topic** and **Exemplifying**: the word trigram model contributed for the best performance (79.8%), but only statistically better than the three POS models and the lemma unigram model. **Emphasizing** was the category where we achieved the lowest overall accuracy, which is a sign of
Figure 4.1: Decision tree for the category Exemplifying using the word bigram model.

the greater lexical variability used to emphasize a point in a talk, when compared to strategies used to introduce/conclude a topic or exemplify. This variability and complexity property is also confirmed by the size of the trees generated for this category, which contain around 100 leaf nodes and 200 rules. This means that each rule accounts for the classification of fewer instances, when compared with the category Exemplifying, for instance. Regarding the false positive/negative rates, we registered the same relation than in the previous categories, where the classification errors are mostly due to false negatives (27.1%) than false positives (12.4%).
4.2. CLASSIFICATION RESULTS

Combinations of POS, Lemma and Word Models

Since the results from the previous experiments were inconclusive regarding the better adequacy of the different sets of features to classify sentences as containing metadiscursive acts, we decided to analyze combinations of the previous settings. We wanted to understand if the reason why using lemma and word features produced similar results was due to the fact that they both represent the same information, or on the other hand if they can complement each other, and when combined improve classification accuracy. We also wanted to test if including syntactic features in the lexical models contributed for best performances, or if POS tags do not add any information in what concerns metadiscourse.

To test these hypotheses, we combined the trigrams of POS, lemmas and words and trained new trees. The same feature reduction criterion applies: features with $IC < 0.0025$ are discarded. Table 4.4 shows the results obtained for the different settings. Again, results marked with * represent settings that performed at the same statistical level (p<0.05).

<table>
<thead>
<tr>
<th>Category</th>
<th>Best</th>
<th>Lemmas +POS</th>
<th>Words +POS</th>
<th>Lemmas +Words</th>
<th>Lemmas +Words +POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introducing Topic</td>
<td>92.71*</td>
<td>90.03</td>
<td>89.99</td>
<td>92.53*</td>
<td>90.11</td>
</tr>
<tr>
<td>Concluding topic</td>
<td>86.93*</td>
<td>85.50</td>
<td>87.17*</td>
<td>86.69*</td>
<td>86.69*</td>
</tr>
<tr>
<td>Exemplifying</td>
<td>94.92*</td>
<td>94.20*</td>
<td>95.02*</td>
<td>94.91*</td>
<td>95.02*</td>
</tr>
<tr>
<td>Emphasizing</td>
<td>79.77</td>
<td>78.63</td>
<td>79.90</td>
<td>81.50*</td>
<td>80.59*</td>
</tr>
</tbody>
</table>

Table 4.4: Classifiers accuracy for the combinations of POS, Lemma and Word Features.

The combination of the models did not produce significantly better results for the categories Introducing Topic, Concluding topic or Exemplifying. For Introducing Topic, any combination containing syntactic features resulted in a significant decrease of overall performance, while for the category Concluding topic only the combination Lemmas+POS produced worse results than the standalone model. Exemplifying was the only category
where all the combinations operated at the same accuracy level. These results show that most information between the models based on lemmas and words are sharing a great deal of information. In practical terms, under situations of equivalent performance, we will opt for the models that generalize the most, i.e., preference of lower-order models, and of lemmas instead of words.

The exception to this observation was the category Emphasizing. The combination of the lemma and word trigrams significantly improved the accuracy of the classification. For this category, the standalone model that generated the best results was the word trigram model, and so we looked at the tree generated for the combination of lemmas and words to understand how lemmas were impacting the results. We noticed that while the great majority of the rules were still based on word features, the algorithm created rules with lemmas that were able to generalize certain cases. Typically, the rules with lemmas represent the collapse of number and person variation in nouns and verbs. Some examples are the inclusion of the lemma trigram that_be_really, which collapses the word trigrams that_is_really and that_are_really, or the lemma unigram message, which encompasses both singular and plural forms of the word.

**Positioning and Length**

In this last experiment, we considered features that expressed qualities of the discourse that POS, lemmas and words could not. As mentioned before, the features we explore here are (1) the length of the sentence, (2) the position of the sentence in the talk, and (3) the distance between the sentence being classified and the last occurrence of the metadiscursive function.

The inclusion of these three feature results from the experience obtained during the preliminary annotation task (Section 3.2). Figure 4.2 shows the distribution of metadiscourse in the beginning of one of the TED talks analyzed. We can see a substantial amount of organizational metadiscourse in the beginning of the talk, occurring in consecutive sentences.

Table 4.5 summarizes the results obtained when enriching the best model for each category with the aforementioned features, first individually and then combined. Results show that
none of the features have any impact in the detection of these four categories of metadiscourse (all accuracies are statistically equivalent). This does not mean that they are not able to reflect the properties of these specific metadiscursive acts, but instead, that they do not add information when combined with the previous models. In fact we looked at the resulting decision trees and found that the length and position features were being used at the leaf nodes, which means that they cannot generalize as much as the lexical features.

<table>
<thead>
<tr>
<th>Category</th>
<th>Best</th>
<th>Best + Sentence Length</th>
<th>Best + Sentence Positioning</th>
<th>Best + Distance Previous</th>
<th>Best + All</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCING TOPIC</td>
<td>92.71</td>
<td>92.49</td>
<td>92.79</td>
<td>92.71</td>
<td>92.62</td>
</tr>
<tr>
<td>CONCLUDING TOPIC</td>
<td>86.93</td>
<td>86.69</td>
<td>86.29</td>
<td>86.77</td>
<td>86.85</td>
</tr>
<tr>
<td>EXEMPLIFYING</td>
<td>94.92</td>
<td>94.87</td>
<td>94.79</td>
<td>94.94</td>
<td>94.76</td>
</tr>
<tr>
<td>EMPHASIZING</td>
<td>81.50</td>
<td>81.22</td>
<td>81.42</td>
<td>81.38</td>
<td>81.21</td>
</tr>
</tbody>
</table>

Table 4.5: Classifiers accuracy when added Positioning & Length features.
4.3 Discussion

In this chapter, we described our first attempt at the classification of metadiscourse according to four functions: INTRODUCING TOPIC, CONCLUDING TOPIC, EXEMPLARYING and EMPHASIZING. Under a supervised learning approach, we used the corpus of annotated metadiscourse obtained through crowdsourcing (described in Chapter 3) to train decision trees. The decision to use this algorithm was motivated by similar approaches in the literature and by the output model intelligibility. This latter property is especially important in our case, since it allows the adaptation of the resulting models to different domains (in our case, to European Portuguese). The set of features herein explored were also motivated by existent literature on areas such as sentiment analysis, word sense disambiguation, and feedback localization. We explored syntactic features (POS n-grams), lexical features (lemma and word n-grams), and position and length features, and concluded how each set of features (or their combination) contributes to the task of identifying metadiscourse in oral presentations.

This experimental setting allowed us to start understanding the nature of metadiscourse, giving us some direction on our second research question, related to the properties of metadiscourse that can constitute good detectors of the phenomenon. We started by analyzing the standalone models with POS, lemma or word features, and saw how stop words contribute to the detection of metadiscourse. INTRODUCING TOPIC was the category where syntactic patterns most contributed for the detection, where the part-of-speech trigram model achieved 85% accuracy. However, across the four categories the models based on syntactic information were consistently outperformed by the lexical approaches. The latter contributed for performance increases in the order of 15 to 25%. Even though we only looked at four categories (Adel’s theory is composed of 23), we believe that the representative power of POS in what concerns metadiscourse is limited, and will never attain the performances of the lexical approaches.

For the first three categories, the best solutions were the standalone models of either lemma or word n-grams. For the category INTRODUCING TOPIC, the best configuration was when
4.3. DISCUSSION

using bigrams of words as features, while for Concluding topic the best setting was the lemma bigram model. The category Exemplifying was the one where we achieved the best classification performance – 95% with the word trigrams model. All other models tested in further experiments performed at the same statistical level for these three categories.

A different result was found for the fourth and last category we analyzed - Emphasizing. For this category lemmas and words complemented each other, generating a model that outperformed any other setting tested. We saw how lemmas provide a back-off mechanism, when the representative power of words is too specific. While this trend was only seen here for the category Emphasizing, we believe that it can manifest itself for other categories of metadiscourse where there is a higher variability of strategies.

The last settings we looked at considered the inclusion of position and length features, derived from our experience in the preliminary annotation task. The three features analyzed – sentence length, sentence position, and distance to last occurrence – did not impact the results in any of the four categories. Despite this result, we believe that these positioning and length features might have impact in other discourse functions.

A final note regarding the results goes to the relation between the type of model with the best performance and the size of the training data. Categories with more training data ended up being better classified by models where features are more specific, i.e., higher order n-grams and words instead of features. The category Emphasizing, having the most training data available, was best classified by word and lemma trigrams; Exemplifying with word trigrams; and Introducing topic with word bigrams. Concluding topic having the least data, was better represented by a bigram lemma model.

In sum, in this first classification task we reached performances between 80 and 95% with models based on lexical n-grams. It is important to point out again that these values correspond to the task of classifying metadiscourse at sentence level. As we saw in Madnani et al. (2012), the performance of token-level approaches is naturally lower (f-measure of 0.55).

This tradeoff between granularity of classification and performance needs to be discussed in
the perspective of the goal of this work. In an area as sensitive as learning, the correctness of the material is an imperative requirement. So far with our approach we have no way of measuring confidence of classification. To accomplish that we need to consider probabilistic models and study the threshold of compromise between precision and recall. In language learning, the cost to pay for missing one item of metadiscourse (i.e., a false negative) is missing one learning opportunity, while the cost of wrongly identifying one occurrence (i.e., a false positive) is incorrect learning. These considerations are going to be addressed in the next chapter, where we discuss the future directions of this work.
In this proposal, we worked on the first steps towards understanding the nature of metadiscourse. To do so, we defined two distinct perspectives from which to look at metadiscourse: language learning and NLP. But more than looking at metadiscourse from these two perspectives individually, we showed how they articulate and influence each other.

On the first front, we addressed how metadiscourse could be looked at in a way that lends itself to presentation skills instruction. We started by analyzing different metadiscursive theories focused on spoken discourse and discussed how they align with our ultimate goal. In this process we opted for a taxonomy where the author showed clear concern towards a pedagogical approach to metadiscourse, associating occurrences of metadiscourse with their function in conveying the message.

We then looked at different sources of material that could be used as good models of presentations. Uniformity, broad set of topics, and existence in both English and European Portuguese, were some of the properties that lead us to choose TED talks over classroom recordings. To check the intersection of the chosen theory and the material of choice, we proceeded with a preliminary annotation task. We found that the situational settings in which the presentation occurs determine what type of metadiscourse strategies the speaker uses. With the experience obtained on the preliminary annotation task, along with advice from the Intercultural Communication Center, we came up with a set of five functional categories of metadiscourse to further process: **Introducing Topic, Concluding topic, Marking Asides, Exemplifying and Emphasizing**.

Finally, we set up a crowdsourcing annotation task to label the five categories along the material of choice (TED Talks). This allowed us not only to collect the labels but also to see the
crowd as the reflection of the future students. The crowd was able to annotate four of the
categories, showing apprehension only when labeling the category Marking Asides. This
annotation generated a corpus of metadiscourse for four categories, annotated at sentence
level, with agreement scores comparable to the ones described by Wilson (2012). Addition-
ally, in the process of building this corpus we found some particularities of the nature of
metadiscourse 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.

On the second front, we explored the automatic identification and classification of metadis-
course following NLP techniques. Having as premise what was found by looking at metadis-
course from the learning perspective, we implemented a classifier capable of detecting func-
tional categories of metadiscourse in TED talks transcripts. Having a corpus of annotated
metadiscourse, we were able to look at this problem in a supervised learning strategy. As
a first attempt, we used decision trees and syntactic and lexical feature, motivated by the
intelligibility of the setup and the literature in areas such as word sense disambiguation, sen-
timent analysis and feedback localization. With this strategy, we found that lexical features
are capable of reaching accuracies of 80 to 95%.

The work we developed up to this point serves as a proof of concept of a tool to be used
for presentation skills instruction. From the learning perspective, we saw that non-experts
understand the notion of metadiscourse and the function different strategies place in presen-
tations. This means that we can use the metadiscourse as an instructional goal. From the
NLP perspective, we saw how a simple setting performs considerably well.

In the next two years of the PhD. we propose working towards the learning tool itself. To
accomplish that, we define four main lines of research. The next four sections describe the
future work in detail.
5.1 Additional Categories

We started by submitting five categories of metadiscourse to annotation in order to understand if the crowd, as a reflection of the future students, was able to agree and correctly spot occurrences of metadiscourse for the different functions. Having proved that non-experts can understand the concept of metadiscourse and associate it to a function in the communication event, we propose to explore additional functions from Ådel’s theory. There are two steps involved in this process of expanding the set of categories to consider. These steps should be executed for both English and European Portuguese settings.

Testing Understanding

In our experiment we saw how the category MARKING ASIDES could not be annotated. When we asked the crowd to annotate a small subset of the TED talks with occurrences of MARKING ASIDES, we noticed an increase of time-on-task and decrease in answer rate and self-reported confidence, to an extent that lead us to discarding this category for further consideration. We learned that not all categories in the same taxonomy can be understood at the same level and some categories might be too hard to explain.

Consequently, we propose to explore the remaining categories in Ådel’s taxonomy and conclude on the suitability to use them as key concepts for instruction. This can be done in a similar setup to what was described in Chapter 3, asking the crowd to annotate the categories in a small sample of TED talks and registering the behavior of the crowd in terms of time-on-task, self-reported confidence, and comments.

Defining the Final Set of Key Concepts

At the present moment, we are considering four categories of metadiscourse: INTRODUCING TOPIC, CONCLUDING TOPIC, EXEMPLIFYING and EMPHASIZING. However, in order to build a robust learning tool we need to consider additional concepts also frequently used by professional speakers when presenting.
Based on the small annotation task performed by the crowd (described in the previous section) we can consider an additional set of categories to teach. This includes pursuing with the complete annotation of the selected categories (in the set of 730 TED talks), and setting up a classifier for each new category.

5.2 Improving Classification

The second big area to address in this thesis is the improvement of the classification techniques. In this proposal we used very simple algorithms and features to classify metadiscourse according to its function. We can consider the present solution as a baseline from which we want to improve. We can improve the classification by exploring two different dimensions, discussed in the next sections.

5.2.1 Features

So far we used lexical and syntactic features and saw how the former outperform the latter, reaching accuracies between 80 and 95%. With these results we concluded that metadiscourse is a very lexical phenomenon. There is an additional set of features that we propose to explore, which can not only contribute to the performance of the classification but can also give insights on the nature of metadiscourse, i.e., what is representative and what is not. There are roughly three types of features we want to explore:

- **Dependencies** – So far we looked at how words by themselves are good indicators of the function of metadiscourse. However, this can be improved if we look at the role of the words in the sentence. Dependencies are representations of the grammatical relations between the words in a sentence. By looking at the relations of the words that were selected as rules in the decision tree inference process, we can have a better representation of the phenomenon. For instance, in the sentence “I will talk about art.”, the dependency root(Root-0, talk-3), might be a better indicator of the presence of
5.2. IMPROVING CLASSIFICATION

Introducing Topic than just the word *talk*. Stanford parser provides 53 grammatical relations (de Marneffe and Manning, 2008).

- **Discourse Structure** – The second set of features we want to explore refer to discourse analysis and topic segmentation. The idea is to use fine-grained discourse analysis as indicator of higher level concepts. We can test occurrences of metadiscourse around sub-topic barriers. We believe that some categories are more cohesive with the surrounding context than others. For example, the categories EXEMPLIFYING and EMPHASIZING are deeply related with the surrounding context, while INTRODUCING Topic and CONCLUDING Topic represent a breaking with the topic. To do this, we can use techniques adapted from topic segmentation, that explore dramatic changes in vocabulary or more sophisticated approaches such as using discourse parsing tools like SPADE\(^1\) (presented in Section 2.2.2). Another possibility is to explore how categories interact with each other, by looking at patterns of occurrence of between them. This will only be possible once we collect a substantial amount of annotations for different metadiscursive markers.

- **Audio** – While using TED talks, aside from text, we have access to two additional dimensions: audio and video. Cassell et al. (2001) for example showed how changes in topic might correspond to changes in physical posture of the speaker or even the audience. While video is out of the scope of this thesis, we propose to analyze audio features and conclude if they are representative of metadiscourse. Again, literature on discourse structure and topic segmentation gives us some insight that this dimension might be relevant for the case. Hirschberg and Nakatani (1998) looked at how acoustic indicators are able to predict topic frontiers and Passonneau and Litman (1997) concluded how pauses patterns can 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 pause less when starting a new segment. We believe that these observations do not only apply to topic

\(^1\)http://www.isi.edu/licensed-sw/spade/
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).

5.2.2 Algorithms

The second dimension to address while trying to improve classification is the algorithm itself. While decision trees achieved accuracies between 80% and 95%, they do not allow us to comment on how confident we are in each decision. They simply output rules to be applied. This setup does not allow to make the distinction between occurrences where we are almost sure that they contain an occurrence of metadiscourse of a certain type (and can therefore be used to teach) and occurrences where we are not so sure (and can be discarded). In other words, decision trees provide a binary classification, which does not allow us to study a precision-recall tradeoff that is appropriate to the high standards of a learning application. To be able to do this we need to move to probability models. Such models will also allow us to consider the full set of AMT annotations. Instead of recurring to a majority vote and discarding work where workers did not agree, we can give different credit to different annotations according to the number of annotators who agreed. The down side of such approaches is that we will not have an explicit model that can be interpreted by humans and adapted to European Portuguese (for which we might not have enough data to train a probabilistic classifier).

The other possible direction regarding classification setup refers to fine tuning the task. So far, we considered the task of classifying metadiscourse at sentence level, i.e., for each sentence, decide if it has an occurrence of each of the metadiscursive categories. The reason behind this approach is the fact that the crowd did not provide reasonable agreement at token-level, due to the fact that the cognitive load of identifying sentences with metadiscourse was already high. However, we can resubmit the sentences that were identified as containing metadiscourse to
a second pass in Amazon Mechanical Turk (AMT) and specifically train the workers to find the words that are part of the metadiscursive strategy. With annotation at token level we can then train a classifier to identify the exact terms that are used in metadiscourse. This is similar to what was described in Section 2.2.4 where Madnani et al. (2012) used Conditional Random Fields (CRFs) to identify shell text at word level. Madnani et al. achieved f-measure of around 0.6 for the task of identifying if each word was part of shell text or not, without further specifying the function of the occurrence of shell text. In our case, since we want to assign a function to each occurrence, the task is more complex and the performance achieved might not be sufficient for consideration in a language learning application.

5.3 Metadiscourse in European Portuguese

As we have been discussing throughout this document, analyzing metadiscourse in both English and EP is one of the major goals of our work. However, the task of porting this technology to European Portuguese constitutes a challenge. This porting task encompasses three steps:

- **Data collection** – One of the reasons why we considered TED Talks as a source of good presentations was the fact that they are available in both languages. As we mentioned previously we collected a total of 118 talks, distributed along a set of 9 events, totaling around 29 hours.

- **Transcription** – While the criteria to select a TED talk in English and add it to our corpus was based on the fact that it had subtitles, for European Portuguese none of the talks are subtitled. This means that to further process the TED talks and use them as a source of examples of good presentations, we need to transcribe them. To reduce transcription time and cost we consider the use of L²F automatic speech recognition engine for European Portuguese AUDIMUS (Neto et al., 2008) as a first pass.

- **Annotation** – The last step, similarly to what was done for English, is to test which categories of metadiscourse non-experts can understand, and annotate them in the set
of EP talks.

One of the challenges in executing these steps is the non-existence of a crowdsourcing technology specifically targeted at European Portuguese and the fact that the representation of the language on well-known crowdsourcing platforms (such as Amazon Mechanical Turk or CrowdFlower\(^2\)) is very limited. As a consequence, annotation will require more resources to be executed.

Even if we are able to transcribe and annotate the complete set of 118 talks, we still will not have sufficient data for training an algorithm. As we mentioned previously, this is the reason behind using decision trees. We can take advantage of what we learn for English, and see how it transfers to European Portuguese. This will include translation of the lexical features, and adaptation of the remaining features to the EP formulation (for example, map the dependencies given by the Stanford parser to the ones provided by L\(^2\)F NLP chain – String (Mamede, 2011) ).

Regarding evaluation of the algorithm we can adopt two strategies: report \textit{f-measure} or only precision. If we are not able to annotate the corpus with metadiscourse, we can still apply the decision tree and ask experts to classify the positive instances retrieved. This task takes considerably less time than reviewing all the sentences in each talk, but allows only concluding on precision.

### 5.4 Presentation Skills Instruction Tool

Although we supported our decisions with the presentation skills instructional software goal, in this thesis proposal we focused on proving its concept. We saw that non-experts understand the notion of metadiscourse. Therefore, the tool we propose to develop will use the notion of metadiscourse as learning goals. Students will 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 in future presentation opportunities.

\(^2\)http://crowdflower.com/
5.4. PRESENTATION SKILLS INSTRUCTION TOOL

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 as proven to fail when there is no specifically targeted feedback at the presentation component (De Grez et al., 2009a). And in fact, instructors are often limited to give feedback on the content instead of on the form due to time and cognitive load constraints.

De Grez et al. (2009a) stressed how presentation skills instruction can be improved by making the rules of presentation explicit. The authors found that students, when simply presented to strict rules, do not change their presentations according to the context they are in. Therefore, students should be presented to concepts, which should be properly explained, allowing them to adapt according to their needs. Presenting the concepts and showing them in different contexts and realizations delegates on the students the responsibility to extrapolate and formulate models adapted to their own reality and needs.

De Grez et al. (2009a) shifted the learning paradigm away from a teacher-centered approach. 73 Business Administration freshmen engaged in a computer-based instruction focusing on presentation skills. The authors concluded that the performance of students significantly improved when compared to instruction that makes no use of multimedia, particularly regarding the correct formulation of introductions and conclusions. While evaluating the learning environment, students selected video clips as the top constraint of appreciation, with 100% of the participants completing the multimedia instruction in an efficient way. Haber and Lingard (2001) also support this 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.

When there is an effort to create a course on presentation skills however, it is often presented as an option, and successful completion depends on motivation and is the responsibility of
the student. In fact, in one of the studies analyzed, the authors mention the dropout rate as one of the major problems of the designed course (Börstler and Johansson, 1998).

We propose to develop a tool that addresses all of these issues, following the learning trends of *just-in-time* learning (Romiszowski, 1997) – where the learning experience is situated close to the knowledge application – and *serious games* (Susi et al., 2007) – where learning appears associated with a form of entertainment. The first step towards the instructional software is to develop a visualization software that enriches a TED talk with occurrences of metadiscourse and respective function. The second step includes packaging the visualization tool around the instructional goal, developing definitions of the concepts being illustrated, examples by category, exercises, *etc.* Current plans point to the development of both web and smartphone applications. For European Portuguese, we intend to package the technology around REAP.PT (REAding Practice for PorTuguese), the EP version of REAP (Heilman et al., 2006), developed at Carnegie Mellon University. REAP.PT started with the original notion of learning vocabulary in context, but has been enriched with instruction of additional components of the language, such as listening comprehension (Lopes et al., 2010; Pellegrini et al., 2013) and syntactic exercises (Freitas et al., 2013).

5.5 **Timetable**
<table>
<thead>
<tr>
<th>Year</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2013</td>
<td>Conclude on the final set of categories</td>
</tr>
<tr>
<td></td>
<td>Gather needed resources for European Portuguese</td>
</tr>
<tr>
<td></td>
<td>Visualization Tool</td>
</tr>
<tr>
<td>Spring 2014</td>
<td>Exploring new features and classification techniques</td>
</tr>
<tr>
<td>Fall 2014</td>
<td>Design instructional platform</td>
</tr>
<tr>
<td>Spring 2015</td>
<td>Test instructional platform</td>
</tr>
<tr>
<td></td>
<td>Write thesis</td>
</tr>
<tr>
<td>Summer 2015</td>
<td>Thesis Defense</td>
</tr>
</tbody>
</table>
6.1 Curricular Plan

<table>
<thead>
<tr>
<th>Type</th>
<th>Course</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Natural Language</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Spoken Language Processing</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Statistical Learning</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Portuguese Linguistics II</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Educational Goals, Instruction and Assessment</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Structured Prediction for Language</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>and Other Discrete Data</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Grammar Formalisms</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Language &amp; Statistics</td>
<td>A</td>
</tr>
<tr>
<td>Lab</td>
<td>Information Retrieval</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Natural Language Processing Project</td>
<td>17</td>
</tr>
</tbody>
</table>

6.2 Teaching Assistance

- Independent Studies I, II, III & IV
  - IST Fall 2010 & Spring 2011 (Professor Pedro Barros)
- 08-710 Search Engine Portals & 08-711 Data Mining
  - CMU Spring 2013 (Professor Jaime Carbonell)

6.3 Published Work

Conferences

CHAPTER 6. PROGRAM PROGRESS DETAILS

*Interspeech* (pp. 1629-1632).


**Journal**


