Dialogue & Discourse 9(1) 79-106 doi: 10.5087/dad.2018.103

Cross-domain analysis of discourse markers in European Portuguese

Vera Cabarrão VERACABARRAO@GMAIL.COM
L2F, INESC-ID
FLUL/CLUL

Helena Moniz HELENA.MONIZ@INESC-ID.PT
L2F, INESC-ID
FLUL/CLUL

Fernando Batista FERNANDO.BATISTA@INESC-ID.PT
L2F, INESC-ID
Instituto Universitário de Lisboa (ISCTE-IUL)

Jaime Ferreira JAIMEFERREIRA90@GMAIL.COM
L2F, INESC-ID

Isabel Trancoso ISABEL.TRANCOSO@INESC-ID.PT
L2F, INESC-ID
IST

Ana Isabel Mata AIM@LETRAS.ULISBOA.PT
L2F, INESC-ID
FLUL/CLUL

Editor: Amanda Stent

Submitted 05/2016; Accepted 07/2017; Published online 06/2018

Abstract

This paper presents an analysis of discourse markers in two spontaneous speech corpora for European Portuguese - university lectures and map-task dialogues - and also in a collection of tweets, aiming at contributing to their categorization, scarcely existent for European Portuguese. Our results show that the selection of discourse markers is domain and speaker dependent. We also found that the most frequent discourse markers are similar in all three corpora, despite tweets containing discourse markers not found in the other two corpora. In this multidisciplinary study, comprising both a linguistic perspective and a computational approach, discourse markers are also automatically discriminated from other structural metadata events, namely sentence-like units and disfluencies. Our results show that discourse markers and disfluencies tend to co-occur in the

dialogue corpus, but have a complementary distribution in the university lectures. We used three acoustic-prosodic feature sets and machine learning to automatically distinguish between discourse markers, disfluencies and sentence-like units. Our in-domain experiments achieved an accuracy of about 87% in university lectures and 84% in dialogues, in line with our previous results. The eGeMAPS features, commonly used for other paralinguistic tasks, achieved a considerable performance on our data, especially considering the small size of the feature set. Our results suggest that turn-initial discourse markers are usually easier to classify than disfluencies, a result also previously reported in the literature. We conducted a cross-domain evaluation in order to evaluate the robustness of the models across domains. The results achieved are about 11%-12% lower, but we conclude that data from one domain can still be used to classify the same events in the other. Overall, despite the complexity of this task, these are very encouraging state-of-the-art results. Ultimately, using exclusively acoustic-prosodic cues, discourse markers can be fairly discriminated from disfluencies and SUs. In order to better understand the contribution of each feature, we have also reported the impact of the features in both the dialogues and the university lectures. Pitch features are the most relevant ones for the distinction between discourse markers and disfluencies, namely pitch slopes. These features are in line with the wide pitch range of discourse markers, in a continuum from a very compressed pitch range to a very wide one, expressed by total deaccented material or H+L* L* contours, with upstep H tones.

Keywords: Prosody, Speech Processing, European Portuguese, Structural Metadata Events.

1 Introduction

The goals of this paper are twofold: (i) a linguistically oriented goal, to describe the acoustic--prosodic properties of discourse markers in European Portuguese (EP), such as *portanto* ('ok'), *pronto* ('ok') or *bom* ('well'); and (ii) a machine learning goal, to classify and discriminate between metadata events, *i.e.*, discourse markers, disfluencies (such as lexicalized filled pauses, like *aam* or *mm*, deletions, substitutions), and sentence-like units. First, the work aims at describing the distribution of discourse markers in different corpora in EP, namely university lectures, map-task dialogues, and tweets. Secondly, for the lectures and dialogues, we also to perform an automatic multiclass classification to verify which segments are classified as discourse markers, which are disfluencies, and which are sentence-like units (SUs), given exclusively their acoustic-prosodic features. This work can be seen as the basis for, in the near future, predicting the acoustic-prosodic features of discourse markers and including them in rich transcription models of a speech recognizer system (Batista, 2011; Moniz, 2013) and also predicting and modeling the suitable integration of discourse markers and disfluencies in spoken dialogue systems (Batista et al., 2016).

Rich transcriptions encompass metadata events, such as disfluencies and punctuation marks (SUs), which promote the legibility of the output of a speech recognizer, enriching the raw sequence of words (Liu et al., 2006, Ostendorf et al., 2008). In this prior work, when metadata events were included in the language models of the recognizer, there was an improvement of the output of the system, expressed by a decrease of the Word Error Rate (WER). No study for EP, however, targeted the inclusion of discourse markers. Given the recent availability of large spontaneous corpora, it was possible to analyze discourse markers (Liu et al., 2006; Ostendorf et al., 2008), the topic of the current paper. Building upon previous results for EP, we expect that incorporating discourse markers will contribute to: (i) improving the output of the recognizer; (ii) discriminating between different metadata events which can share similar acoustic-prosodic properties; and (iii) shedding light on the prosodic patterns of discourse markers in EP, thus contributing to their identification and modeling in spoken dialogue systems.

In the literature, discourse markers are presented as a heterogeneous class that comprises words and expressions from different morphological categories. These carry a specific function in discourse, being mainly used as a metalinguistic strategy to maintain cohesion between utterances

(e.g., Schiffrin, 1987; Fraser, 1990; Schourup, 1999; Aijmer, 2013). The present work follows the definition of discourse markers given by Liu et al. (2006), namely a word or phrase that functions primarily as a structuring unit of spoken language (e.g., *actually, now, anyway, so, I mean, well*), also signaling to the listener the speaker's intention to mark a boundary in discourse, including a change in the dominant speaker or the beginning of a new topic, amongst other functions. The work of Liu et al. (2006) encompasses the discourse markers that are syntactically detachable in the sense of Schiffrin (1987) and for which prosodic cues are the main source of information. In this work, we classify discourse markers in spontaneous speech based exclusively on acoustic-prosodic features, since the lexical output of a speech recognizer has a high WER for spontaneous speech, and thus is not a reliable source of information. Moreover, targeting exclusively acoustic-prosodic features may contribute to the state-of-the-art discussion on the status of their cross-language (in)dependent acoustic-prosodic behavior.

Liu et al. (2006) developed a system that recovers structural metadata events, including sentence boundaries, disfluencies and discourse markers, in order to enrich speech recognition output. The authors considered the identification and recovery of these phenomena crucial to automatically process speech. Goldwater et al. (2010) also showed that discourse markers, along with other words that occur mainly turn-initial, are hard to classify, presenting higher error rates when automatically recognized. The authors divided words into three classes with high error rates, namely an open class (names and verbs), a closed class (prepositions and articles), and discourse markers. We focus mainly in the discourse markers that can both: (a) occur turn-initial, the most frequent location (65% of the total discourse markers found in the dialogues and 34% in the lectures) in our speech corpora; and (b) be extracted from the discourse without compromising its propositional content (like *pronto*/'ok' and *bem*/'well'); so that we can linguistically characterize these units and understand what their most prominent acoustic-prosodic features are. In this study, we consider a turn as a major and/or an intermediate phrasal boundary, with a break index of 3 or 4, respectively, meaning temporal or melodic breaks between words, usually silence and/or distinct pitch contours (Pierrehumbert & Hirschberg, 1990)[1], equivalent to Sentence-like Units. Moreover, there are also discourse markers at the beginning of a turn uttered in co-articulation with the following words, with no silent pause and mostly deaccented. Henceforth, the term *turn-initial* will be used covering the three prosodic patterns described.

Discourse markers also raise problematic issues in the field of translation, both human and automatic. These are characterized by their idiomatic nature, which poses a problem in finding exact equivalents in different languages. Therefore, building an inventory of these words and expressions, as well as a description of their function in discourse, is crucial for several tasks. In Lopes et al. (2015), the authors showed that, out of the list of 60 discourse markers selected for the present study (corpus of university lectures and map-task dialogues), only 18 were available in the Europarl corpus (Koehn, 2005) and had an English translation. The authors present as a possible explanation the register used in the three corpora (dialogues, lectures, and speeches given in the European Parliament), a representation of a continuum encompassing informal to very formal domains, respectively.

In EP, as in several other languages, discourse markers correspond to distinct linguistic categories, namely adverbs, conjunctions and interjections. The fact that the same discourse marker can also be associated with a different pragmatic function, being therefore multifunctional or polysemic (Schiffrin, 1987; Fischer, 1998, 2000; Soares da Silva, 2006), represents an additional challenge for their classification. Moreover, most of the studies available for EP are

[1]For a more detailed description of phrasing in EP, see Frota (2000; 2014), Viana et al. (2007), and Mata & Moniz (2016).

based on written text (e.g., Mendes, 2013; Lopes, 2016), and do not account for all the discourse markers present in speech.

Therefore, our goal with this study is to contribute to the acoustic-prosodic description of the discourse markers that are mainly used in spontaneous speech or in datasets that may represent the continuum between speech and written modalities, like tweets - written products of a rich intersection between speech idiosyncratic features and text representations of those. Although tweets may not faithfully represent speech *per se*, they truly represent a very fuzzy frontier between speech and writing modalities. In that sense, tweets can be seen as a closer scenario and a worthwhile one to study the influences of speech on written modalities, in general, and also to study the frequency and selection of specific discourse markers, in particular. To accomplish these goals, we used a data-driven approach to identify the discourse markers present in the three corpora, followed by an automatic classification task, using acoustic-prosodic features, to differentiate discourse markers and disfluencies (Shriberg, 1994; Liu et al., 2006; Ostendorf et al., 2008; Moniz, 2013) from each other and from SUs. The acoustic-prosodic discrimination between structural metadata events still poses several challenges, due mostly to the prosodic distribution of such events and to the correspondent acoustic-prosodic features of such prosodic contexts. So far in EP, we only manage to discriminate between disfluencies and distinct punctuation marks or SUs, without any information on discourse markers. It is, therefore, expected that the inclusion of discourse markers in the language models for EP, already trained with other structural metadata events, *i.e.*, disfluencies and punctuation marks, will improve the available enriched automatic transcription models and will contribute to the understanding of the structural metadata events as a whole.

This paper is organized as follows: Section 2 is an overview of related work. Section 3 describes the data, the selection of discourse markers, and the automatic segmentation process used. Section 4 describes the distributional patterns of discourse markers in the corpora and their complementary distribution with disfluencies. Section 5 describes the acoustic-prosodic parameters and the machine learning methods used to perform the multiclass classification task. It also presents the results obtained both in-domain and cross-domain, and discusses the most relevant acoustic-prosodic features for discriminating the three classes: discourse markers, disfluencies, and SUs. Finally, Section 6 presents conclusions and future work.

2 Related work

Several authors have showed the importance of discourse markers in different languages, given their high frequency and multifunctionality, both in speech and in written text. There has been a growing effort to define the set of words that can be considered as a discourse marker (Fraser, 1990, *vs.* Schiffrin 1987). Consequently, there are also a variety of definitions to encompass the heterogeneous classes, namely, discourse markers (Schiffrin, 1987; Schourup, 1999), connective markers (Fraser, 1988), pragmatic particles (Beeching, 2002), phatic markers (Fraser, 2009), pragmatic markers (Redeker, 1990; Aijmer & Simon-Vandenbergen, 2006; Denke, 2009) and conversational markers (Urbano, 2003; Borreguerro & López, 2010), among others.

According to Fraser (1999), discourse markers have to be linguistic structures that derive from the morpho-syntactic classes of conjunctions, adverbs, and prepositional phrases. In this work, interjections are not considered as a discourse marker. The author proposes that discourse markers are a pragmatic class, given the fact that they contribute to the interpretation of utterances, but not to their propositional content. Building upon this distinction, Fraser (1999) defines two types of markers, namely those that relate messages and those that relate topics. The main premise of this work is that discourse markers always signal a relationship between the interpretation of the segment they introduce, defined by the author as S2, and a prior segment, S1, that does not necessarily need to be adjacent.

By contrast, Schiffrin (1987) defines discourse markers as multifunctional devices that delimit units of speech, are syntactically detachable, and have a range of prosodic contours. These structures are linguistic expressions, such as conjunctions, interjections, adverbs, and lexicalized phrases, as well as non-verbal devices, such as gestures or paralinguistic features. Contrary to Fraser, Schiffrin (1987) also shows that discourse markers display not only local, but also global relations, *i.e.*, relations between adjacent utterances and across the discourse, respectively. Ultimately, Schiffrin describes the role of discourse markers in structuring a coherent discourse as a whole, beyond the exclusive use of such devices as indexing propositional relations in a text.

Aijmer & Simon-Vandenbergen (2006) distinguish between discourse markers and pragmatic markers, preferring the latter designation. Their distinction is that a pragmatic marker is a word or expression that does not contribute to the propositional, truth-functional content of an utterance, *vs.* a discourse marker which signals coherence relations. Although the authors distinguish between those structures, the frontier between them may not be clear.

The fact that discourse markers are multifunctional (Schiffrin, 1987) or polysemic (Fischer, 1998, 2000), meaning that the same item can have different functions or semantic interpretations in discourse, makes it more difficult to disambiguate and, therefore to classify, them. According to Beeching (2014), pragmatic markers have different functions depending on the nature of the interaction. In spontaneous speech, they allow for hesitations, backtracking, repairs and repetitions, and also occur as turn-taking strategies. Moreover, in a social perspective, pragmatic markers may be sociolinguistically marked, are often associated with naturalness, friendliness and warmth, and may also be a mark of politeness.

We employ the most common designation, "discourse marker", to allow for a broader comparison with the literature. Our work is based on structural metadata events defined by Liu et al. (2006) and Ostendorf et al. (2008) encompassing: (i) Sentence-like Units (SUs), corresponding to a sentence boundary, as either a comma, a full-stop or a question mark; (ii) disfluencies, meaning repetitions, deletions, fragments, filled pauses, substitutions, insertions, and editing terms; (iii) discourse markers (such as *actually, now, anyway, so, I mean, well, like*, *okay*). The authors claim that the automatic detection of these events can enrich speech recognition outputs, by decreasing WER and facilitating subsequent natural language processing techniques.

The automatic detection of disfluencies and punctuation marks was already encompassed in our in-house speech recognizer (Batista, 2011; Moniz, 2013), resulting in enriched transcriptions and improvement of the output of the system. Discourse markers, as other structural metadata events, are very challenging to automatically detect, due to several reasons: (i) they occur mostly turn-initial, locations for higher WER of automatic speech recognition systems (Goldwater et al., 2010); (ii) they share the same location with disfluencies, and (iii) also behave as a prosodic constituent *per se* (Goldwater et al., 2010; Moniz, 2013).

This paper builds upon our previous work, using the same spontaneous speech corpora, namely university lectures (the LECTRA corpus) and map-task dialogues (the CORAL corpus). In Cabarrão et al. (2015), we performed the first experiments with an automatic classification task to discriminate between discourse markers, disfluencies, and SUs, given exclusively their acoustic-prosodic features, extracted from the toolkit openSMILE (Eyben et al., 2010). The results showed that the use of acoustic-prosodic features improved the classification of each task up to 20%. Accuracy was higher for the lectures corpus (87%) than for dialogues (84%). Based on these encouraging results, we now aim at expanding the classification to other sets of acoustic-prosodic features, towards a better understanding of the discriminative properties of each class. We also compare the distributional patterns of disfluencies with discourse markers, in order to verify the differences and/or similarities between those two structural metadata events. In this study, we expanded our datasets to encompass tweets, to compare the distribution and occurrence

of discourse markers in different domains, representative of a continuum from speech to written modalities. We will discuss the most prominent acoustic--prosodic features and relate those to information structure and prosodic patterns of the three classes.

Discourse markers and disfluencies can be processed in two distinct ways (Popescu-Bellis & Zufferey 2011): either to remove these structures from automatic transcripts or to include them to enrich the transcripts with (meta/para)linguistic information. Lease and Johnson (2006) compare discourse markers to disfluencies, namely lexicalized filled pauses, and remove them from automatic transcripts. Hirschberg and Litman (1993), and Samuel (1999) argue that the presence of discourse markers can be used to infer the discourse structure, and for resolving anaphoric references, and should, therefore, be encompassed in the automatic processing. As for their frequency, Jucker and Smith (1998) found a total of 31.8% of discourse markers (reception markers and presentation markers) *per* topic unit (5 minutes). The authors showed that the frequency of both types of markers varied according to the relationship between the speakers (friends or strangers), although the total number of discourse markers was similar in both pairs (32.8 *vs.* 30.7). In line with findings by Shriberg (2001), disfluencies present an interval of 5% to 10% in human-human conversations.

The study by Heeman and Allen (1999) shows that the tasks of segmenting turns and resolving repairs are crucial in studying the speaker's intentions in dialogues along with the identification of discourse markers. The authors follow the definition of discourse marker presented by Schiffrin (1987) and Hirschberg and Litman (1993) and add a definition of *utterance unit*, which they consider to be a block of spoken dialogue where discourse markers operate to relate the current utterance to the discourse context or to signal a repair. In this study, Heeman and Allen (1999) present a statistical language model that aims at repairing speech recognition problems, including the identification of POS tags, discourse markers, speech repairs, and intonational phrases. The corpus is human-human task oriented dialogues (the Trains Corpus), where a human pretends to be an assistant in a computer aided system in a setting as close to human-computer interaction as possible, although both humans know they are talking to a person. In this corpus, discourse markers account for 14% of the total number of words, occurring as part of the editing term or of the alteration in 40% of the fresh starts and in 14% of the modification repairs. Therefore, the authors claim that identifying a word as a discourse marker facilitates the automatic detection of repairs. In this study, the results achieved also showed that modeling repairs, intonational phrases, and discourse markers jointly, rather than as distinct events, helps to improve the performance of the automatic speech recognizer and reduces the error rate in recognizing each one of these events.

In EP, in order to automatically identify discourse markers based exclusively on acoustic-prosodic features, we need to establish, first, what words can be classified as discourse markers and, subsequently, what are their acoustic-prosodic features. For EP, the studies on discourse markers in spontaneous speech are scarce; to the best of our knowledge, our work is the first to consider acoustic-prosodic features.

Coutinho (2009), by adapting the classifications of Fraser (1999) and Adam (2008), describes discourse markers according to the context and the specificity of the text, in order to disambiguate the different discursive uses of discourse markers. Studies by Lopes (1997), Freitas and Ramilo (2003), and Soares da Silva (2006) presented analysis for a singular discourse marker, namely *então* ('so'), *portanto* ('so, like'), and *pronto* ('that's it, ok'), respectively, and its different functions according to context. The study of Pimentel (2012) analyzed a set of discourse markers in a second language acquisition perspective, Lopes and Amaral (2006) analyzed the uses of *agora*/'now' and *então*/'then' both as deictic and anaphoric temporal adverbs and discourse markers, and the work of Lopes (2014a) studied the diachronic behavior of the marker *aliás* ('moreover'). Lopes and colleagues (2006, 2014b) and Lopes (2016) in a recent study have been working on creating a pragmatic categorization of discourse markers, focusing mostly on their connective function in structuring texts. To the best of our knowledge, Lopes (2016) established

the first typology for European Portuguese that relates classes of DMs with discourse relations. However, because our DMs are not being well recognized automatically, this fined-grained typology is still not applicable to our data. We use instead acoustic-prosodic information to distinguish discourse markers from disfluencies and SUs.

Even though there has been a growing effort to describe and classify discourse markers in EP, these structures are still understudied in our language, especially in what concerns their idiosyncratic properties in spontaneous speech. Therefore, the contributions of our work are threefold to: (i) add to previous work the discrimination of discourse markers into the structural metadata events; (ii) describe the most salient acoustic-prosodic properties of discourse markers in spontaneous speech and, thus, contribute to their prosodic characterization, and (iii) contribute to enrich automatic speech transcripts.

3 Data description and selection of discourse markers

In this section, we first describe the data used, namely two spontaneous speech corpora of university lectures and map-task dialogues and a written corpus of tweets. Then, we describe the process used to collect the discourse markers in the three corpora, and, finally, we explain the methods used to perform automatic classification.

3.1 Corpora

This study uses two corpora of spontaneous speech, namely university lectures (the LECTRA corpus) and map-task dialogues (the CORAL corpus). Both corpora are available through ELRA. For comparison, we also analyzed a corpus of geolocated tweets, mostly characterized by a very informal register. The three corpora have different domains and levels of spontaneity: (i) the lectures are oral productions, but previously prepared by the teachers; (ii) in the dialogues, the participants have a specific task to perform, but no previous preparation; and (iii) the tweets correspond to written texts, but tend to represent certain structures of speech.

The university lectures corpus, collected within the LECTRA national project (Trancoso et al., 2008), aimed at producing multimedia contents for e-learning applications, and for hearing--impaired students. The LECTRA corpus (ISLRN 298-379-572-530-5) has a total of 7 courses, 6 recorded in the presence of students, and 1 recorded with the teacher targeting an Internet audience. All the lecturers (6 male and 1 female) are native Portuguese speakers. LECTRA has a total of 75h of speech, of which 32h were orthographically transcribed, totaling 155k words. The transcription process was made by three annotators, with the same linguistic background, following the guidelines described in Moniz (2006) and Trancoso et al. (2008), for disfluencies, and the punctuation summary in Duarte (2000). The inter-transcriber agreement was evaluated (Moniz, 2013) and the results showed an almost perfect agreement between one annotator (A1) and the remaining two (A2 and A3, with a slot accuracy of 0.82 and 0.79, respectively), and a substantial agreement between A2 and A3, with a slot accuracy of 0.69. The corpus was divided into 3 different sets: train (78%), development (11%), and test (11%).

The CORAL corpus (ISLRN 499-311-025-331- 2) (Trancoso et al., 1998) comprises 64 dialogues in map-task format between 32 speakers. The dialogues occur between two speakers with different roles, one the giver of information and the other the follower. The giver has a map with a route drawn and some landmarks and his/her task is to provide information and directions for the follower to reconstruct the same route in his/her incomplete map. There are also several inconsistencies between the names and places of the landmarks to elicit conversation. CORAL is balanced in terms of gender and of role played by the speaker. The corpus has 7 hours orthographically transcribed, totaling 61k words, and was divided into train and test sets.

The corpus of geolocated tweets comprises a total of 307K tweets (with a total of 1,525,437 words), produced in Portuguese regions by about 11k different users in an 8-day period. Brogueira et al. (2014) showed that the tweets collected were produced mainly by teenagers and young adults that used Twitter as a way to share their personal feelings and ideas regarding family, school, and friends. Even though tweets are coded in a written form and can be considered a genre on their own, they also present some characteristics of spontaneous speech, namely contractions and reductions mimicking speech structures. In the current work, we assume that the use of discourse markers typically associated with spontaneous speech can also be a cue to approximate tweets to oral communications.

3.2 Manual selection of discourse markers

Due to the lack of a suitable discourse marker inventory based on spontaneous speech, our first task was to identify possible candidate discourse markers displayed in the corpora. Several criteria guided our selection: previous orthographic guidelines; *stimuli* audition; syntactic detachment of the structures and metalinguistic function; expert agreement in the selected structures; and decisions on ambiguous cases.

Considering the high frequency of these structures in the data and the fact that they were often not recognized by the speech recognizer, in the orthographic guidelines it was established, during the annotation process, that this type of word should be followed by a comma or full stop, when occurring turn-initial, or delimited by commas, when occurring in utterance internal positions. Here, the punctuation marks are a way to demark these words, pointing out that a more thorough linguistic analysis of such regions was needed. This analysis was accomplished by listening to all the examples and checking syntactic detachment, contexts, and metalinguistic functions. The data-driven selected inventory was then subject to the evaluation of two experts and both revised the ambiguous cases.

Some structures that are highly ambiguous were disregarded, such as *não é* (“right”), *e* (‘and’), *mas* (‘but’), and *porque* (‘because’). The first one was considered to be more a tag question than a discourse marker. *e* (‘and’) and *mas* (‘but’) were only included when co-occuring with other markers as multiword units. While inspecting the data, we found that *e* (‘and’) and *mas* (‘but’) were being used in two distinct ways: (i) as connectors with propositional content and (ii) as truly discourse markers with interpersonal information, in the sense of Beeching (2014). In the case of *porque* (‘because’), it only occurred in coherence relations with the previous speech turn. In most instances, especially in the lectures corpus, those conjunctions were being used with propositional content. Since we aim at performing a general characterization of discourse markers and not to explore the functions of each marker individually, we decided not to include these structures when they do not co-occur with other discourse markers. Further detailed analyses on both items and adjacent contexts are needed in future work.

On the other hand, we choose to include two frequent connectors, *portanto* (‘ok’) and *agora* (‘now’), given the fact that both, when occuring turn-initial, can be omitted or replaced by a filled pause, for example, without compromising the propositional content of the discourse.

In this selection process, we also included the combination of one marker with other markers, like *então* (‘so’, ‘then’) and *mas então* (‘but then’), to account for as many markers as possible and also to analyze their behavior as possible multiword units.

Since our focus is mainly the identification and classification of discourse markers that tend to occur in spontaneous speech, we excluded connectors, which, by definition, have semantic content. For that reason, the discourse markers collected in all three corpora are devoid of any propositional content and can be detached from the utterance.

The fact that discourse markers occur mostly turn-initial (65% of the total discourse markers found in the dialogues and about 34% in the lectures) was also a very crucial criterion and informed our decision of selecting exclusively structures in this position. The comparatively smaller frequency of turn-initial discourse markers in the lectures is mostly due to the high

frequency of filled pauses occurring in that same position in the lectures, although they do not co-occur (see Section 4.2). Moreover, a single discourse marker (*portanto*) accounts for 26% of the total number of discourse markers in the lectures. Since we want to analyze the acoustic-prosodic behavior of discourse markers overall in EP and not a single marker, we maintained the turn-initial analysis also in lectures. The motivation for this methodological choice is fourfold: (i) turn-initial is the most frequent location for a discourse marker in dialogues; (ii) turn-initial is also the main location for disfluencies, both in the lectures and the dialogues; (iii) turn-initial is the location more prone to trigger automatic speech recognition errors (Goldwater et al., 2010); (iv) turn-initial with adjacent silent pauses corresponds to inter-pausal units (e.g. Levitan et al., 2015; Gravano et al., 2012), a very relevant unit of analysis for the study of discourse markers, in particular, and for the perception and production of communication management in spoken dialogue systems, in general.

List of discourse markers in the university lectures corpus (LECTRA) and in the dialogues corpus (CORAL) with the corresponding percentages[2] of occurrence in each corpus, respectively:

1. *A seguir* / next (1.9%; 2.9%)
2. *Agora* / now (7.2%; 3.8%)
3. *Até agora* / so far (0.2%; 0.0%)
4. *(E) já agora* / (and) by the way (0.2%; 0.1%)
5. *E agora / and now* (2.9%; 1.5%)
6. *Mas agora / but now* (0.2%; 0.0%)
7. *(Mas) eu agora / (but) I now* (0.3%; 0.0%)
8. *Agora é assim* (0.0%; 0.1%)
9. *Atenção / attention* (1.1%; 0.0%)
10. *Tomem lá atenção / Pay attention* (0.1%; 0.0%)
11. *Bem / well* (0.5%; 0.1%)
12. *Bom / well* (1.2%; 0.1%)
13. *De facto / indeed* (0.8%; 0.0%)
14. *Desculpem(a) lá / excuse me* (0.2%; 0.7%)
15. *(Mas) É assim / So* (1.5%; 2.3%)
16. *E depois / and then* (3.5%; 6.8%)
17. *Portanto / ok (*37.9%; 19.1%)
18. *E portanto / ok* (0.6%; 0.0%*)*
19. *Portantos / so* (0.0%; 0.0%)
20. *Pronto / that's it* (2.9%; 1.5%)
21. *E pronto / and that's it* (0.4%; 14.7%)
22. *Mas pronto / but ok* (0.2%; 0.1%)
23. *Prontos / ok (0.0%*; 1.5%)
24. *Eh pa; epá / hey man* (0.4%; 0.1%)
25. *Enfim /anyway* (0.4%; 0.0%)
26. *Mas enfim / but anyway* (0.1%; 0.1%)
27. *Então / so* (8.8%; 21.3%)
28. *Então (isto) agora / so (now) this (0.1%; 0.5%)*
29. *Então a seguir / so next (0.0%; 0.1%)*
30. *Então é assim / so this is it (0.1%; 1.3%)*
31. *Mas então / but so (0.0%; 0.1%)*
32. *Entretanto / meanwhile (0.1%; 0.2%)*
33. *Muito bem / ok; very well (1.0%; 0.0%)*
34. *No entanto / however (0.6%; 0.0%)*
35. *Ó (jovem/pessoal)/ hey (youngman/people) (1.8%; 0.0%)*
36. *Ok / ok (10.0%; 15.3%)*
37. *Olha lá / hey look (0.0%; 0.0%)*
38. *Olha; olhe / hey look (1.2%; 0.3%)*
39. *Ora / well (0.6%; 0.1%)*
40. *Ora bem / well ok (1.3%; 0.4%)*
41. *Ora bom / well ok (0.1%; 0.0%)*
42. *Ora então agora / well now this (0.0%; 0.0%)*
43. *Ora isto / well this (0.0%; 0.0%)*
44. *Ora vejamos / well let's see (0.0%; 0.0%)*
45. *Ou seja /meaning (3.7%; 0.4%)*
46. *Ouve; ouça / listen (0.0%; 0.0%)*
47. *Pá /man (0.2%; 0.2%)*
48. *Pois / ok (0.4%; 3.3%)*
49. *Pois bem / well then (0.0%; 0.0%)*
50. *Pois é / that's right (0.2%; 0.2%)*
51. *Pois então / ok then (0.0%; 0.0%)*
52. *Pois muito bem / ok (0.0%; 0.0%)*
53. *Por acaso / actually (0.6%; 0.1%)*
54. *Quer-se dizer; quer dizer / I mean (0.5%; 0.3%)*
55. *Se calhar / maybe (1.5%; 0.3%)*
56. *Tudo bem / ok (0.5%; 0.2%)*
57. *Vamos lá / let's go (0.0%; 0.1%)*
58. *Vamos lá (a) começar / let's begin (0.0%; 0.2%)*
59. *Vamos lá estar então / let's go then (0.0%; 0.0%)*
60. *Vamos lá ver / let's see that (0.2%; 0.5%)*

[2] There are percentages of 0.0% that correspond exclusively to a single occurrence of a discourse marker in the corpora.

From the corpus of university lectures and dialogues, we selected 60 words (see the list above) as discourse markers occurring turn-initial, even though some of them can also co-occur in other positions in the utterance, and may be extracted from the discourse without compromising its propositional content (like *pronto*/'ok' and *bem*/'well'). These structures also correspond to words or phrases that are very common in spontaneous speech, being rare in more formal written texts[3], can be omitted and do not have coherence relations with the previous speech turn, thus, have no propositional content and are syntactically detached.

This selection was based on the fact that no *a priori* systematic selection of discourse markers was available in EP and our analysis relies in substantial spontaneous material. The criteria of excluding discourse markers based almost exclusively on text materials can be further revised in the future, when a detailed ecosystem of discourse markers in our language will be concluded – a task not conducted so far. The data driven approach ultimately targeted a more homogenous criteria: turn-initial discourse markers, syntactically detachable, with no propositional content, with exclusively metalinguistic function, and very frequent in spontaneous speech.

3.3 Information available for speech corpora

For each of the two speech corpora under study, along with the manual transcripts we have available force time aligned and automatic transcripts, produced by an in-house automatic speech recognition system (ASR) Audimus (Neto et al., 2008). However, the speech recognizer, trained for the broadcast news domain, is totally unsuitable for the domains of university lectures and map-task dialogues. The scarcity of text materials in EP to train language models for these domains has motivated the decision of using the ASR in a forced alignment mode only, in order not to bias the study with bad results obtained with an out-of-domain recognizer. For that reason, our current experiments rely on force aligned transcripts that still contain about 0.9% of unaligned words (mainly due to low energy segments).

The corresponding manual transcripts also provide complementary reference data that are fundamental for speech analysis, supervised training, and automatic evaluation. Instead of creating task dependent links between the force aligned transcripts and the corresponding manual transcripts, the relevant manual annotations, as well as other available elements, were automatically transferred into the automatic transcript itself. Such a self-contained dataset, merging all the relevant information, can be easily dealt with and constitutes a valuable resource to extensively address, study and process speech data. Full reports on the process of extending existing ASR transcripts in order to accommodate relevant reference data coming from manual annotations are described in Batista et al. (2012) and Moniz et al. (2015).

Currently, the two speech corpora are available as self-contained XML files that correspond to enhanced transcripts, integrating information from both manual and force aligned transcripts, enriched with additional prosodic information related to pitch, energy, duration, and other structural metadata (punctuation marks, disfluencies, inspirations, other paralinguistic annotation, etc.). Figure 1. presents a transcript segment, corresponding to the sentence *Portanto o início começa [%aa] do lado direito do cabo.*/'So the beginning is [%aa] on the right side of the cable.'[4] that was automatically enriched with reference data. The example illustrates two important sections: the characterization of the transcript segment and the discrimination of the wordlist that comprises it. Each word element contains the lowercase orthographic form, start time, end time, and confidence level; a discrimination of the focus condition (F1 stands for spontaneous speech without background noise); information about the capitalized form (cap); whether or not it is followed by a punctuation mark (punct=.); and the part-of-speech tag.

[3] This inventory of discourse markers can be further expanded when taking into account ambiguous structures, which were not tackled in this study.

[4] The notation [%aa] corresponds to the orthographic transcription of a filled pause.

```
<TranscriptSegment>
  <TranscriptGUID>6</TranscriptGUID>
  <AudioType start="1734" end="1831" conf="1.000000">Clean</AudioType>
  <Time start="1734" end="1831" reasons="" sns_conf="1.000000"/>
 </TranscriptSegment>
 <TranscriptSegment>
  <TranscriptGUID>7</TranscriptGUID>
  <AudioType start="1832" end="2372" conf="1.000000">Clean</AudioType>
  <Time start="1832" end="2372" reasons="" sns_conf="1.000000"/>
  <Speaker id="1" id_conf="1.000000" name="Unknown" gender="U" gender_conf="1.000000" known="T"/>
  <SpeakerLanguage native="T">PT</SpeakerLanguage>
  <TranscriptWordList>
   <Event name="[l]"/>
   <Word start="1862" end="1906" conf="0.996585" focus="F1" cap="Portanto" pos="Cc" name="portanto"/>
   <Word start="1910" end="1922" conf="0.991042" focus="F1" pos="Td" name="o"/>
   <Word start="1923" end="1982" conf="0.998802" focus="F1" pos="Nc" name="início"/>
   <Word start="1983" end="2066" conf="0.987207" focus="F1" pos="V." name="começa"/>
   <Event name="BEGIN_disf"/>
    <Word start="2099" end="2152" conf="0.005932" status="filled_pause" focus="F1" name="%aa"/>
   <Event name="END_disf"/>
   <Word start="2185" end="2197" conf="0.992067" focus="F1" pos="S." name="do"/>
   <Word start="2198" end="2219" conf="0.987395" focus="F1" pos="Nc" name="lado"/>
   <Word start="2220" end="2250" conf="0.997514" focus="F1" pos="A." name="direito"/>
   <Word start="2251" end="2261" conf="0.994279" focus="F1" pos="S." name="do"/>
   <Word start="2262" end="2289" conf="0.997386" focus="F1" pos="Nc" name="cabo"/>
   <Word start="2290" end="2338" conf="0.998928" focus="F1" punct="." pos="A." name="branco"/>
  </TranscriptWordList>
 </TranscriptSegment>
```

Figure 1. Creating a file that integrates the reference data into the ASR output

The output of forced alignment encompasses phone, syllable, word, and sentence-like unit segmentations, temporally aligned with the audio signal. It is therefore possible to extract acoustic-prosodic features for these different units of analysis. In the present study, the three units selected were segments corresponding to: disfluencies, discourse markers (described in the previous section), and sentence-like units, a generic umbrella covering all the other segments.

4 Distributional patterns of discourse markers

In this section, we first describe the results of the cross-domain analysis of the discourse markers in the three corpora, followed by a speaker-wise analysis. Finally, we describe the distributional patterns of discourse markers and disfluencies in both spontaneous speech corpora.

4.1 Cross-domain analysis

Discourse markers in the initial position account, in the dialogues, for 65% of the total discourse markers found and in the lectures, for about 34%. There are a total of 1719 discourse markers in the dialogues and 4840 in the lectures, which correspond to 3% of the total number of words in each corpus. Even though we focused our analysis on markers that occur turn-initial (see Section 3.2), we also found that the same marker can occur both turn-initial and in any other position in the sentence, even at sentence final position, due to overlapping speech and turn-taking strategies.

Looking only at the most frequent discourse markers (Table 1), results show that the most common in the dialogue corpus are *então* ('so') (14% turn-initial and 7% in other positions),

Discourse Markers	CORAL		LECTRA	
	Turn-initial N (%)	Other positions N (%)	Turn-initial N (%)	Other positions N (%)
(E) Portanto / 'Ok'	**206 (12.0%)**	122 (7.1%)	**612 (12.6%)**	1248 (25.8%)
(E) Agora / '(And) now'	42 (2.4%)	49 (2.9%)	**136 (2.8%)**	351 (7.3%)
Então / 'So'	**245 (14.3%)**	122 (7.1%)	**203 (4.2%)**	223 (4.6%)
(E) Pronto / '(And) that's it'	**197 (11.5%)**	56 (3.3%)	65 (1.3%)	94 (1.9%)
Ok	**232 (13.5%)**	31 (1.8%)	**157 (3.2%)**	347 (7.2%)
E depois / 'And then'	29 (1.7%)	88 (5.1%)	33 (0.7%)	136 (2.8%)
Ou seja / 'Meaning'	3 (0.2%)	4 (0.2%)	27 (0.6%)	150 (3.1%)
Pois / 'Ok'	51 (3.0%)	6 (0.3%)	18 (0.4%)	3 (0.1%)
(Mas) É assim / 'So'	15 (0.9%)	24 (1.4%)	11 (0.2%)	63 (1.3%)
Bom / 'Well'	2 (0.1%)	0 (0.0%)	45 (0.9%)	11 (0.2%)
Olha / 'Look'	1 (0.1%)	5 (0.3%)	6 (0.1%)	50 (1.0%)
Ora bem / 'Well'	6 (0.3%)	1 (0.1%)	63 (1.3%)	0 (0.0%)
Prontos / jargon variation of *pronto*	17 (1.0%)	8 (0.5%)	0 (0.0%)	0 (0.0%)

Table 1 - Most frequent discourse markers in both corpora

portanto ('so') (12% and 7%), *pronto* ('ok') (12% and 3%), and *ok* (14% and 2%). As for the university lectures corpus, the most frequent discourse markers are *portanto* ('so') (13% of occurrences turn-initial, and 26% in other positions in the utterance), *agora* ('now') and ok (both with 3% turn-initial, and about 7% in other positions), and *então* ('so') (4% in all positions). The remaining discourse markers are quite residual in the corpus.

In university lectures, teachers often use these structures to stall time, while they plan the subsequent units and their cohesion nexus. In map-task dialogues, the speakers have shorter interactions in a faster rhythm, structured on the basis of question-answer for clarification and information-seeking purposes.

Discourse Markers	Tweets
(E) agora / ' (And) now'	**6506 (13.4%)**
Bem / 'Well'	**4924 (10.1%)**
Ok	**3943 (8.1%)**
ó/oh / 'Oh'	3311 (6.8%)
Ó pá/opá/opah / variations of 'oh men'	916 (1.9%)
Pronto / 'Ok'	875 (1.8%)
Ya / 'yes'	849 (1.7%)
Tipo / 'like'	817 (1.6%)
Fogo / 'Damn'	812 (1.6%)
Ainda por cima / 'On top of that'	582 (0.6%)
A cena é que/ 'The thing is that'	275 (0.3%)
Well	126 (0.3%)

Table 2 - Examples of discourse markers in the collection of tweets

In the collection of tweets, turn-initial discourse markers also account for about 3% (48,717) of the total of words of the corpus (1,525,437). Since the tweets are limited in length, they tend to resemble a turn. Even though this corpus presents characteristics of oral communications (Brogueira et al., 2014), it is, in fact, written and uses colloquial expressions, which implies that there are word contractions and also reductions, mimicking the frequent vowel/syllable reductions in speech. Sometimes, some of these words can be confused with discourse markers, like in the case of *pá* ('men') and *pá praia* ('to the beach' produced with the contraction of *para* and *a*), which represent an additional challenge in identifying discourse

markers in this corpus. Considering that this is a preliminary analysis to collect and understand what type of words are being used as discourse markers in this domain, we chose not to account here for these ambiguous cases, in line with what was previously said about the discourse markers' inventory and its discriminative properties.

Results show that, out of about 49,000 discourse markers selected, the most frequent ones are *agora* ('now'), with 13%, *bem* ('well'), with 10%, *ok*, with 8% and *ó*, *oh* ('oh'), with 7% (see Table 2). The remaining discourse markers are quite residual in the corpus. The most frequent markers in the tweets are similar to those found in both lectures and dialogues. However, the type of markers found in the collection of tweets also differs from those in the other corpora. Discourse markers like *ya* (form of 'yes'), *tipo* ('like'), and *fogo* ('damn') are very colloquial and not accepted in a more formal context. Another specificity of this corpus is the fact that the authors shift between Portuguese and English, as seen by the co-occurrence of the discourse markers well and *bom* ('well'), both with the same meaning. Here are some examples in order to better understand in what contexts these specific discourse markers occur:

1. *ya, amanhã tenho dos jogos mais importantes e estou aqui* ('yeah, tomorrow I have one of the most important games of my life and I'm here');
2. *tipo nao critikem a nha maneira de escrever aki pk tipo ya este é o meu twitter e eu escrevo o ke kizer como me apetecer kkkkk* ('like don't judge the way I write here, cause like this is my twitter and I'll write whatever I want kkkk');
3. *Fogo hoje tipo que vou ficar todo o dia em casa* :(('Damn today I'm like going to stay all day home');
4. *A cena é que tou cansada*. ('The thing is I'm tired');
5. *Well vou voltar para salvaterra cuz need jantar.* ('Well I'm going back to *salvaterra* cuz need dinner.').

Considering that these examples mimic the speech structures in tweets, we can add a contribution to the findings made by Brogueira et al. (2014) in their characterization of these users. This is clearly a young population that uses Twitter to share their thoughts and opinions. Discourse markers are produced here almost like fixed expressions, which are used to start the utterances, but appear to be devoid of any semantic content, being rather used as interjections, conversational fillers, and/or emphatic expressions, such as *A cena é que*/'The thing is that'.

4.2 Speaker-wise analysis

It is known that disfluencies and sentence types (as tag questions) are speaker and corpus dependent. In this section, we aim at verifying if discourse markers, as part of structural metadata events, can also exhibit speaker and corpus variation.

The overall distribution of discourse markers by speakers reveals that they tend to use similar markers in the same domain, but with different combinations, such as *ora* ('well'), *ora bem* ('well ok') and *ora bom* ('well good') or even *portanto* ('so'), *e portanto* ('and so'), *portanto é assim* ('so this is it'). Results also show that the distribution of discourse markers varies substantially across speakers. For instance, the most common discourse marker in the lectures corpus is not produced by all speakers or in the same proportions: *portanto* ('so') is produced 40% of the times by only one speaker (S7) and 29% by another (S3).

In the dialogues, there is also speaker variation but in a less expressive way than in the lectures. Results show that *então* ('so'), the most common discourse marker, is produced almost 10% of the times by two speakers (S7 and S9), a percentage that drops to less than 8% for all the other 22 speakers. There is also one discourse marker, namely *ok* pronounced as [Okáp6] that is only produced by one speaker (S17) which may suggest, similarly to *portanto* ('ok') in

LECTRA's speaker S7 production, that the selection of discourse markers is influenced by the personal choices of each speaker.

There are discourse markers, like *olhe*, *olha*, *olha lá* ('look'), that mostly occur in the lectures (78% of the times is produced by speaker S6, and 19% by speaker S3) and are very scarce in the dialogues. This marker functions as a deictic to make the student aware of a particular set of information available in the slides or the dashboard. Other discourse markers (*ó jovem, ó pessoal* / 'hey youngman, hey people',) used by the same two teachers (S3 and S6) are vocative expressions to call attention to the content when the students are distracted. These do not occur in the dialogues. Finally, there are discourse markers that occur only in one corpus, like *prontos* ('ok'). In this particular case, the use of the jargon form of *pronto* ('ok') shows that there is a less formal environment between speakers in the dialogues corpus, since they use a more colloquial speech, opposing to the more formal discourse associated with university lectures.

In conclusion, there is a more or less fixed set of discourse markers in the dialogues shared for almost all the speakers, while the lectures present a higher variation per speaker. Previous results for the same dialogue corpus show smaller tempo patterns for disfluencies, their adjacent contexts, silent pauses and turns, evidencing the dynamic temporal characteristics of dialogues (Moniz et al., 2014). Additional to this analysis, selecting a fixed set of discourse markers may contribute to the delivery of information, as if the speaker used a fixed set of discourse markers as a working memory strategy towards temporal fluidity.

4.3 Distributional patterns: discourse markers and disfluencies

Considering that both discourse markers and disfluencies are part of structural metadata events (Liu et al., 2006; Ostendorf et al., 2008; Popescu-Bellis and Zufferey, 2011), the fact that they occur mainly in the beginning of an utterance, and seem to, for instance, share the same function of stalling time or starting a new topic in the conversation, we wanted to find out if these two phenomena are in complementary distribution *vs.* in co-occurrence.

As for the university lectures corpus, filled pauses occur mostly as an intonational unit before a new propositional content in the lectures (Moniz, 2013). There is no register of disfluencies before a discourse marker. However, immediately after, we find about 104 cases of disfluencies, of which 62% correspond to substitutions, 17% to filled pauses, and 10% to repetitions (for a better description of the disfluency typology in these corpora, see Moniz, 2013). This distribution can be explained by the nature of the corpus. We can assume that, because teachers are skilled professionals, after beginning an utterance with a discourse marker, they do not feel the need to produce a disfluency. In this case, we can assume that they are in complementary distribution. These results are not in line with the findings of Heeman and Allen (1999), since the authors found that discourse markers tend to occur mainly after speech repairs or as part of the editing term, being used as a strategy to facilitate the identification of the repair itself.

In the dialogues, the number of disfluencies increases both before and after a discourse marker. A total of 41 disfluencies were produced before a discourse marker in the beginning of a sentence and 93 after. 85% of the 41 disfluencies were filled pauses and the remaining ones were deletions, fragments, substitutions, and repetitions. As for disfluencies after the discourse marker, the distribution is more balanced: 35% were filled pauses; 31%, substitutions; 26%, repetitions. The remaining disfluencies are quite residual. Figure 2 shows an example of a sequence of the discourse marker *pronto* ('ok'), corresponding to an independent intonational unit characterized by H+L* L%, a frequent melodic contour for nuclear declaratives in EP, contrasting with a disfluent unit (marked by angular brackets), namely a filled pause with a characteristic plateau contour, H* H%. Again, we can say that the higher number of disfluencies adjacent to discourse markers in the dialogues may be due to the fact that the conversations did not have any pre-planning and that the interactions between participants are faster, quite dynamic, and under temporal constraints.

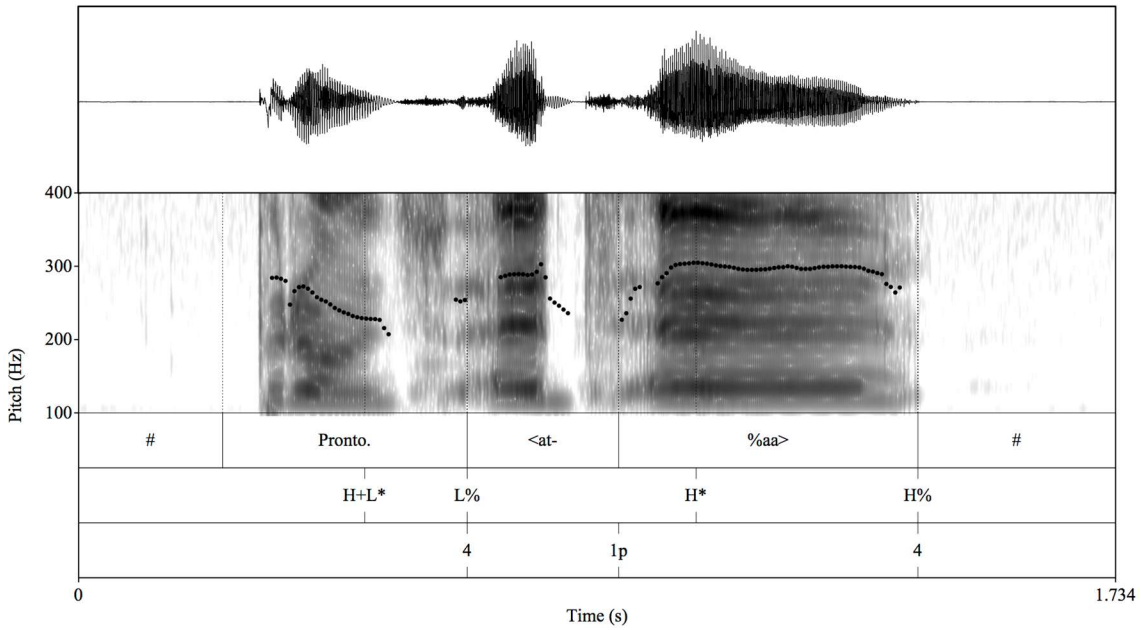

Figure 2. Example of a disfluency following the marker *pronto* ('ok') in the excerpt: *Pronto. <at- %aa>* ('Ok. <at- %@@>') from the dialogue corpus

5 Classification of discourse markers

This section describes a set of machine learning experiments aiming at automatically distinguishing between discourse markers, disfluencies, and segments that are neither discourse markers nor disfluencies (SUs), using acoustic-prosodic features. In order to do so, we have performed intra-domain and cross-domain multiclass classification experiments using three sets of acoustic-prosodic features. We have performed an analysis of the most relevant prosodic features that can be used to characterize the discourse markers, and ranked the most prominent acoustic-prosodic features in the three classes.

5.1 Prosodic parameters and Machine Learning methods

Our experiments are based on three sets of features, automatically extracted using the openSMILE toolkit (Eyben et al., 2010), a toolkit capable of extracting a very wide range of acoustic-prosodic features that has been successfully applied to a number of paralinguistic classification tasks, including disfluency prediction (Schuller et al., 2013).

The first set of features, henceforth referred as IS13, is derived from the Interspeech 2013 Paralinguistic challenge, and corresponds to 6125 speech features calculated by applying segment-level statistics (means, moments, distances) over a set of energy, spectral and voicing related frame-level features. The other two sets of features were recently introduced by Eyben et al. (2016), and correspond to GeMAPS – Geneva Minimalistic Acoustic Parameter Set for Voice Research and Affective Computing, and eGeMAPS – an extended version of GeMAPS. GeMAPS comprises a total of 62 acoustic-prosodic features, whereas its extended version, eGeMAPS, comprises a set of 88 features. These small sets of features are well-known for their usefulness in a wide range of paralinguistic tasks, and are derived from: frequency related parameters (for example, pitch, jitter, and formant 1, 2, and 3 frequency); energy related parameters (such as shimmer, loudness, and Harmonics-to-Noise Ratio – HNR); spectral parameters (like Alpha

Ratio, Spectral Slopes, and Harmonic differences); loudness, pitch (for voiced and unvoiced regions), and temporal features.

In the scope of this work, we have applied a wide range of machine learning methods by means of the open source toolkit Weka (Hall et al., 2009) and the best results were consistently achieved by Logistic Regression (LR) and Support Vector Machines (SVM), the latter using the Sequential Minimal Optimization (SMO) algorithm. For that reason, results reported here are for these two methods. In most cases, our baseline consists of selecting the most frequent class, which was also achieved by applying the ZeroR classification method from Weka.

Experiments reported on the paper are based on 5-fold cross-validation, thus covering all the data both for training and for evaluation. We have also performed 10-fold cross-validation experiments, but the corresponding experiments take about twice the time and results turned out to be very similar. The parameter C (complexity) used in SVMs defines the complexity of the model. The default value (C=1.0) is commonly used in similar experiments, and proved to be a good choice for the two smaller feature sets. However, for the large feature set (IS13), the default value usually leads to poor performance, while taking several days to run, sometimes more than a week. For that reason, based in a previous work with disfluencies, we have set C to 0.01 for this feature set, which also proved to be suitable for the features in use while achieving a good speed performance. We did not tune this parameter in our cross-validation experiments, but we have tested other values of C for all feature sets in order to make sure that the parameter was correctly picked. Results here presented were evaluated using the following standard performance metrics: Precision, Recall, F-measure, and Accuracy (Makhoul et al., 1999):

$$\text{Precision} = \frac{TP}{TP + FP}, \ \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F-measure} = 2 \times \frac{Precision * Recall}{Precision + Recall}, \text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN},$$

where *TP* is the number of hits (true positives), *TN* is the number of correct rejections (true negatives), *FP* is the number of false alarms (false positives), and *FN* is the number of misses (false negatives). Our experiments are also measured in terms of the Kappa statistic, a chance-corrected measure of agreement between the classifications and the true classes. A value close to zero indicates that results could be achieved almost by chance; whereas a value close to 1.0 means that the model is adequate to the problem.

5.2 Intra-domain classification

The number of instances and distribution of discourse markers, disfluencies, and SUs is quite different for both corpora: the dialogue corpus contains 6381 SUs, 1834 disfluencies, and 723 discourse markers, and the lectures corpus contains 16273 SUs, 6618 disfluencies, and 1359 discourse markers. This unbalanced data presents an additional challenge for the classification approaches that may lead to biased results towards the most common class, namely SUs. Table 3 presents the automatic classification results for each corpus, together with the baselines of 71.4% and 67.1%, corresponding to the most frequent class for each corpus. The table shows an overall better performance for the university lectures, despite the baseline being lower, which can be partially explained by the higher number of speakers in the dialogues (7 *vs.* 20 speakers), corresponding to a higher speaker variation. That is also reflected in the kappa values, all of them above 0.60 for the lectures. The acoustic-prosodic features in use allow for very significant improvements relative to the baseline: 20% for the university lectures and 13% for the dialogues. The best results were achieved using the SVMs and the IS13 feature set, achieving accuracies of about 84% and 87%. However, the two other smaller feature sets also proved to be a good choice for discriminating discourse markers, especially when used in combination with LR. On the other

Corpus	Feature set	Classifier	Overall		Discourse Markers			Disfluencies		
			Acc	Kappa	Prec	Rec	F	Prec	Rec	F
Dialogues	Baseline		71.4%	0						
	GeMAPS	SVM (C=1)	77.1%	0.328	1.000	0.001	0.003	0.760	0.395	0.520
		LR	79.1%	0.459	0.573	0.373	0.452	0.711	0.457	0.556
	eGeMAPS	SVM (C=1)	78.0%	0.381	0.660	0.131	0.219	0.749	0.418	0.537
		LR	79.4%	0.478	0.601	0.447	0.513	0.703	0.469	0.562
	IS13	SVM (C=0.01)	84.1%	0.605	0.778	0.625	0.693	0.771	0.557	0.647
		LR	76.7%	0.474	0.567	0.560	0.564	0.544	0.556	0.550
Lectures	Baseline		67.1%	0						
	GeMAPS	SVM (C=1)	83.1%	0.614	0.624	0.364	0.460	0.799	0.655	0.720
		LR	83.3%	0.629	0.599	0.481	0.534	0.788	0.671	0.725
	eGeMAPS	SVM (C=1)	83.5%	0.625	0.605	0.394	0.477	0.812	0.657	0.726
		LR	83.9%	0.643	0.610	0.507	0.554	0.800	0.679	0.735
	IS13	SVM (C=0.01)	86.8%	0.709	0.760	0.623	0.685	0.831	0.734	0.780

Table 3 - Classification results for unbalanced data

hand, LR consistently achieved quite low performance when applied to the IS13 feature set while also taking several days to run. The GeMAPS and their extended version achieved acceptable performance while taking minutes to run, instead of hours and even days. The last 6 columns show individual performance results for discourse markers and disfluencies. With respect to te dialogues, we achieved 85.9% precision and 94.7% recall for SUs, 77.1% precision and 55.7% recall for disfluences, and 77.8% precision and 62.5% recall for discourse markers, revealing that discourse markers are easier to classify than disfluencies in general. That is not the case for the university lectures, where we achieved 17% precision and 62.3% recall for discourse markers and 83.1% precision and 73.4% recall for disfluencies. It is interesting to notice that the model created using GeMAPS and SVMs for dialogues achieved 100% precision with a very low recall. That is because the model classified only one event as a discourse marker and it turned out to be correct.

	Dialogues			**Lectures**		
Classified as =>	SU	Disf	DM	SU	Disf	DM
SU	6041	264	76	15349	773	151
Disfluency (Disf)	759	1022	53	1642	4859	117
Discourse Marker (DM)	231	40	452	297	215	847

Table 4 - Confusion matrix for unbalanced data, achieved using the IS13 feature set and SVMs

As previously mentioned, the data is considerably unbalanced, thus making it more difficult for the machine learning method to classify discourse markers. Table 4 shows the confusion matrix for the two corpora, revealing that SU is, in fact, the class selected more often, followed by disfluencies and, finally, discourse markers.

In order to prevent biasing the models towards the most frequent class, we conducted the remaining experiments over a balanced version of the data. We considered the frequency of the discourse markers, the least frequent class in each corpus, as the number of samples to select for each class. So, for dialogues we selected 723 samples of each class and for university lectures, this number was extended to 1359. The balanced version of the data was achieved using the filter *SpreadSubsample* available in Weka. The corresponding results are presented in Table 5.

Corpus	Feature set	Classifier	Overall		Discourse Markers			Disfluencies		
			Acc	Kappa	Prec	Rec	F	Prec	Rec	F
Baseline			33.2%	0						
Dialogues	GeMAPS	SVM (C=1)	70.8%	0.562	0.693	0.823	0.752	0.753	0.591	0.662
		LR	69.9%	0.548	0.701	0.791	0.743	0.717	0.624	0.667
	eGeMAPS	SVM (C=1)	71.7%	0.576	0.709	0.835	0.767	0.750	0.620	0.679
		LR	71.0%	0.564	0.729	0.798	0.762	0.700	0.639	0.668
	IS13	SVM (C=0.01)	77.9%	0.668	0.799	0.862	0.829	0.774	0.723	0.748
		LR	49.1%	0.237	0.501	0.527	0.514	0.486	0.480	0.483
Lectures	GeMAPS	SVM (C=1)	78.4%	0.676	0.777	0.872	0.822	0.781	0.654	0.712
		LR	79.3%	0.689	0.799	0.870	0.833	0.777	0.694	0.733
	eGeMAPS	SVM (C=1)	78.8%	0.683	0.786	0.873	0.827	0.791	0.662	0.721
		LR	79.4%	0.692	0.808	0.865	0.836	0.779	0.693	0.734
	IS13	SVM (C=0.01)	82.9%	0.744	0.835	0.918	0.875	0.817	0.732	0.772

Table 5. Classification results for balanced corpora

Once again, SVMs with the IS13 feature set achieve the best performance. Notice that the accuracy baseline is now 33.3% and that these results should not be directly compared with the results from Table 3. Kappa values are high for both corpora, revealing that we achieved good models for the three structures. The last six columns show an impressive F-measure of 82.9%-87.5% for discourse markers. Notice however that discourse markers here being considered are exclusively turn-initial while disfluencies include all possible types of disfluency in distinct positions of the corpus. The discourse marker classification performance is now better than for disfluencies, even on the university lectures.

5.3 Cross-domain classification

In order to verify how robust our classification is across domains, we conducted additional experiments in a cross-domain evaluation scenario, using the training data from one corpus and using the other corpus for testing. Thus, to test if the model trained with the university lectures would generalize for the dialogues, we trained the models with the balanced data from LECTRA and tested on CORAL, and vice-versa.

Table 6 shows the corresponding results. It is known that by using out-of-domain data the performance decreases so, as expected, results are worse than the ones achieved using data from the same domain. In fact, the accuracy performance decreases about 11%-12% absolute. However, the performance achieved suggests that data from university lectures can still be used to classify the events on the dialogues and vice-versa. The eGeMAPS feature set achieved about 71% and 66% accuracy in university lectures and dialogues, respectively, an impressive result when considering the small size of the feature set. The kappa statistic is between 0.666 and 0.717, also suggesting strong and suitable cross-domain models.

An interesting additional result is that, when using balanced data, our results suggest that discourse markers are easier to identify than disfluencies, in both corpora, either using intra--domain or cross-domain models. Disfluencies are recognized as being one of the hardest structures to detect amongst the structural metadata events (Moniz et al., 2016). It is also recognized that the intonation of discourse markers are marker specific and they may share the melodic contours of a neutral declarative in EP, H+L* L%, but they are uttered with a wide pitch range, very distinctive from other units, and do tend to have three main prosodic patterns as previously stated.

Train => Test	Feature set	Classifier	Overall		Discourse Markers			Disfluencies		
			Acc	Kappa	Prec	Rec	F	Prec	Rec	F
Dialogues => Lectures	GeMAPS	SVM (C=1)	69.5%	0.543	0.775	0.615	0.686	0.584	0.768	0.663
		LR	68.5%	0.527	0.770	0.592	0.669	0.588	0.734	0.652
	eGeMAPS	SVM (C=1)	70.6%	0.559	0.736	0.698	0.717	0.638	0.695	0.665
		LR	67.2%	0.509	0.745	0.547	0.631	0.625	0.667	0.646
	IS13	SVM (C=0.01)	68.0%	0.521	0.831	0.468	0.599	0.545	0.800	0.649
		LR	41.3%	0.119	0.378	0.398	0.388	0.398	0.328	0.360
Lectures => Dialogues	GeMAPS	SVM (C=1)	65.9%	0.488	0.671	0.704	0.687	0.702	0.546	0.614
		LR	65.2%	0.479	0.705	0.622	0.661	0.651	0.609	0.629
	eGeMAPS	SVM (C=1)	66.0%	0.490	0.737	0.607	0.666	0.665	0.609	0.635
		LR	63.4%	0.451	0.729	0.548	0.626	0.605	0.617	0.611
	IS13	SVM (C=0.01)	66.1%	0.491	0.758	0.629	0.688	0.609	0.665	0.636
		LR	53.2%	0.298	0.643	0.398	0.492	0.492	0.568	0.527

Table 6. Cross-domain classification results for balanced corpora

Table 7 presents four confusion matrices, achieved with the balanced data, showing the most common correct and incorrect automatic classifications. All the matrices show a strong diagonal, indicating that our models, both inter-corpora and cross-corpora, perform correct classification most of the time. There is also clear evidence that discourse markers are better identified with in-domain data, in line with what was previously said concerning the distributional patterns of such events, *i.e.*, they may occur exclusively in one corpus or be more productive in one corpus than in the other. When using cross-domain models, the tendency is to classify discourse markers as disfluencies and, to a smaller degree, to classify them as SUs. This trend may also be explained by the shared properties between discourse markers and certain types of disfluencies, mostly filled pauses, behaving as vocalic supports with plateau contours (H* H-/%).

Intra-domain

	Dialogues			**Lectures**		
Classified as =>	SU	Disf	DM	SU	Disf	DM
SU	504	105	114	1128	116	115
Disfluency (Disf)	141	448	134	252	899	208
Discourse Marker (DM)	75	44	604	51	121	1187

Cross-Domain

	Lectures => Dialogues			**Dialogues => Lectures**		
Classified as =>	SU	Disf	DM	SU	Disf	DM
SU	553	109	61	985	220	154
Disfluency (Disf)	187	440	96	229	944	186
Discourse Marker (DM)	171	113	439	94	316	949

Table 7. Confusion matrices for intra-domain and cross-domain results achieved with eGeMAPS, balanced data, and SMO with C=1

5.4 Discussion of relevant features and acoustic-prosodic patterns of structural metadata events

The automatic discrimination of structural metadata events is a complex task, due to the diversity of acoustic-prosodic and distributional patterns of such events and also to the prosodic information they may share. However, despite the complexity of such structures, they have discriminative prosodic behaviors captured by acoustic correlates. The literature for Portuguese points out to an array of features relevant for the description of metadata events. The data-driven approaches followed in this work allow us to reach a structured set of basic features towards the disambiguation of such events beyond the established evidences for Portuguese. This study is a contribution to the analysis and discriminative behavior of discourse markers in EP (to the best of our knowledge, completely absent in our literature), and adds to the previous studies on structural metadata events a layer more to the understanding of the acoustic-prosodic behavior of such structures. The automatic discrimination between classes of structural metadata events is feasible because discourse markers have class properties, as will be detailed bellow.

Sentence-like units and disfluencies were previously discriminated (Moniz et al., 2016). Disfluencies are mostly characterized by: two identical contiguous words; both energy and pitch increases in the following word and (mostly) a plateau contour on the preceding word; and a higher confidence level for the following word than for the previous word. This set of features reveals that repetitions are being identified, that repair regions are characterized by prosodic contrast marking (increases in pitch and energy) between disfluency-fluency repair (as in Moniz et al., 2012 and Moniz, 2013), and also that the first word of the repair has a higher confidence score. Since repetitions are more frequent in dialogues (22% of all disfluencies *vs.* 16% in lectures), the feature *identical contiguous words* has a significantly higher impact in dialogues.

Full stops are described by a falling contour in the previous word; a plateau energy slope in the previous word; the duration ratio between the previous and the following words; and previous word higher confidence score. This characterization is the one that most resembles neutral statements in Portuguese, with the canonical contour H+L* L% (Frota, 2001), associated with terminus value.

Question marks in lectures are characterized by two main patterns: a rising contour in the current word and a rising/rising energy slope between previous and following words; and a plateau pitch contour in the previous word and a falling energy slope in the previous word. The rising patterns are not surprising, since interrogatives are cross-language perceived as having a rising contour (Hirst & Di Cristo, 1998). The falling pitch contours have also been ascribed for different types of interrogatives, especially wh- questions in Portuguese.

Commas are the event characterized by the fewest prosodic features, being mostly identified by morpho-syntactic features. However, in dialogues they are better classified. The two most relevant features are: *identical contiguous words* and mostly plateau energy and pitch shapes between words. The first feature is associated with emphatic repetitions, comprising several structures, namely: (i) affirmative or negative backchannels (*sim, sim, sim* 'yes, yes, yes') and (ii) repetition of a syntactic phrase, such as a locative prepositional phrase (*para cima, para cima* 'above, above'). They are used for precise tuning with the follower and for stressing the most important part of the instruction. Emphatic repetitions are annotated with commas separating the repeated item(s) and account for 1% of the total number of words in dialogues. Although not a disfluent structure, if they were accounted as disfluent words, they would represent 16.7% of all disfluent items. As for the features energy and pitch plateau shapes between words, they are linked to lists of enumerated names of the landmarks in a given map, at the end of a dialogue.

Regarding regular words, the most salient features are related to the absence of silent pauses, explained by the fact that, contrary to the other events, regular words within phrases are connected. The presence of a silent pause is a strong cue to the assignment of a structural metadata event.

Adding discourse markers to this characterization, they are described as: having different prosodic patterns: (i) as a major intonational phrase (IP) (see Figure 3); (ii) as an intermediate intonational phrase (ip); (iii) deaccented and functioning as a clitic or an initial vocalic support (see Figure 4), with plateau contours and f_0 values lower than the following prosodic unit, being, therefore, uttered in an intermediate tonal space; (iv) with reduced f_0 slopes relatively to the following prosodic unit (e.g., *então,* 'so' and *pronto,* 'ok')**.** Moreover, the pitch range of discourse markers behaves as a *continuum* from a very compressed to a very wide range.

Feature	Overall		DM vs. Disf.	
	Dialogues	Lectures	Dialogues	Lectures
F3frequency_sma3nz_amean	*****	*****	*****	*****
F2amplitudeLogRelF0_sma3nz_amean	**	*****		*****
F3amplitudeLogRelF0_sma3nz_amean	*****	***		
F0semitoneFrom27.5Hz_sma3nz_amean	**	*****		*****
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	*****	*	*****	
loudness_sma3_percentile50.0		*****		*****
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	***	***	****	****
F2frequency_sma3nz_amean	****		*****	
loudness_sma3_stddevNorm	*	****	**	****
HNRdBACF_sma3nz_amean		****		****
logRelF0-H1-A3_sma3nz_stddevNorm	*****		*	
loudness_sma3_amean	****		*****	
logRelF0-H1-A3_sma3nz_amean	****		****	
F2bandwidth_sma3nz_amean	***	****		
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	****		****	
spectralFluxV_sma3nz_amean		****		****
StddevUnvoicedSegmentLength	**	***	***	***
loudness_sma3_meanRisingSlope	***		****	
spectralFlux_sma3_stddevNorm	***	**	***	
F3bandwidth_sma3nz_amean	****		****	
MeanUnvoicedSegmentLength		****		****
F1amplitudeLogRelF0_sma3nz_amean	***			**
mfcc1V_sma3nz_amean	**		***	
MeanVoicedSegmentLengthSec	*		**	**
alphaRatioUV_sma3nz_amean	*	*		**

Table 8. Top 25 most influent features

Table 8 presents the most relevant features extracted from the set of eGeMAPS features with in-domain data. The choice of this particular set is motivated by: (i) the substantial dimensionality reduction regarding the openSMILE features; and (ii) the faster classification of structural metadata events, with comparable results. The "*" indicates the most informative features, considering their weights extracted with logistic regression models – the higher number of "*" the higher the relevance. The ranking is distributed per corpus, encompassing all the structural metadata events, and also per the most striking differences between discourse markers and disfluencies. We note that the most informative features are the pitch and energy related ones. Considering the prosodic characterization described above, this result may be interpreted as pointing out to the different f_0 slopes in the production of discourse markers, an evidence more of the continuum in the pitch range of such structures.

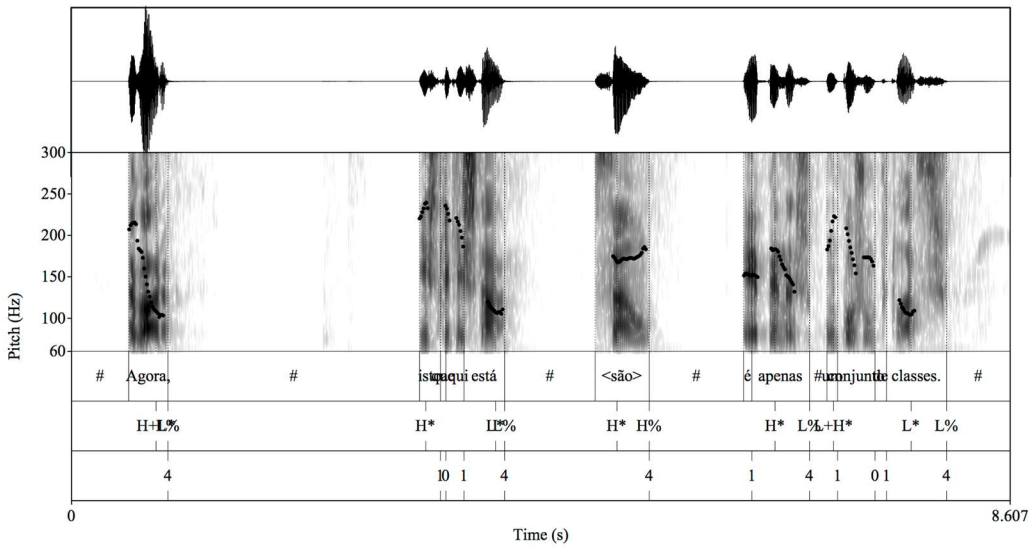

Figure 3: Example of the discourse marker *agora* ('now') in the excerpt: *Agora, isto que aqui está <são> é apenas um conjunto de classes* ('Now, what we have here <are> is just a set of classes'), from the university lectures corpus

The discourse marker *agora* ('now' – Figure 3) tends to be accented, with high f_0 range within the accented syllable, and with similar or higher f_0 values than the adjacent prosodic constituents. This shows that there is an effort to mark this word and to distinguish it from the adjacent contexts. On the other hand, the markers *portanto* ('ok' – Figure 4) and ok are mainly unaccented, present plateau contours, and have f_0 values lower than the following prosodic unit, being, therefore, uttered in an intermediate tonal space, almost as a vocalic support for uttering the following prosodic units. Both *então* ('so'), *pronto* ('ok') and *portanto* (ok) present reduced f_0 slopes, even though the latter tends to be unaccented and the first one is mainly accented.

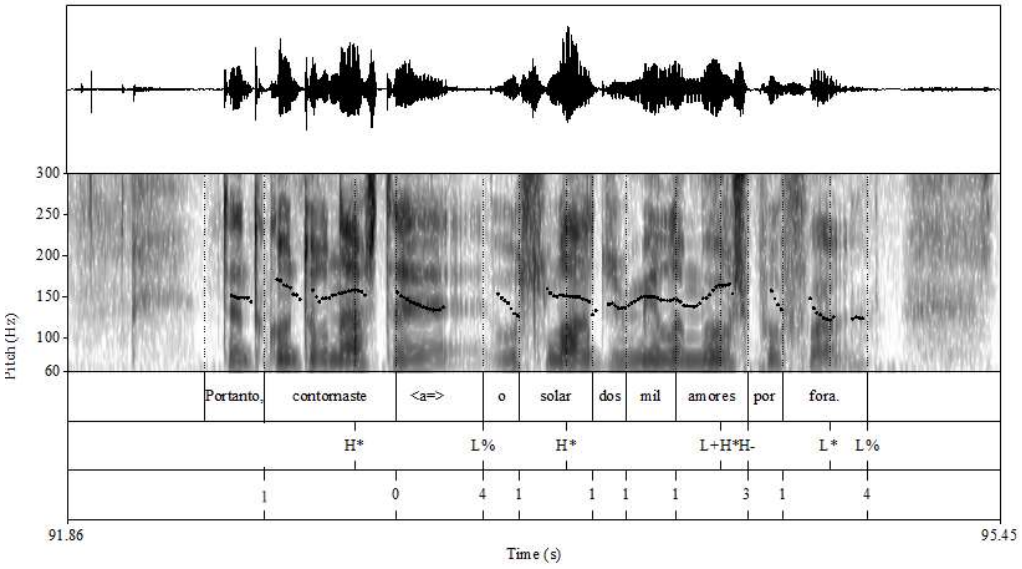

Figure 4: Example of the discourse marker *portanto* (ok) in the excerpt: *Portanto, contornaste <a=> o solar dos mil amores por fora* ('Ok, you've passed the manor of a thousand loves from the outside'), from the dialogues corpus

The prosodic contours can be distinct accordingly to their distribution, *i.e.*, initial, medial or final positions. Since we're targeting the turn-initial ones, the prosodic patterns of such markers highly depend on the marker selected, meaning if it is mostly a deaccented *portanto* ('ok') *vs*. a prominent *agora* ('now'), starting a new topic with a wide range of pitch and energy.

The fact that discourse markers are associated with different pragmatic functions may also be an explanation for the variation found. We can hypothesize that the markers that have a function similar to disfluencies, like stalling, may share with them some prosodic properties, meaning the plateau contours contrasting with the rises in the following prosodic constituents. Other discourse markers, such as *agora* ('now'), introduce a new topic and are prosodically prominent. To confirm this preliminary analysis, we still need to perform a more detailed tonal and acoustic study of the discourse markers and adjacent prosodic contexts.

The acoustic-prosodic characterization of discourse markers, in particular, and structural metadata events, in general, is still a matter of debate. Structural metadata events can be equated to discourse markers as a broad linguistic class, which encompasses both disfluencies and discourse markers[5]. The linguistic literature on discourse markers points out to several strategies when accounting for disfluencies in such a broader class: (i) either consider fillers and reformulation markers as a sub-type of discourse markers, but always stating that they have a distinct nature, although not clarifying truly this nature with empirical evidence; or (ii) consider them as completely different and not part of discourse markers. We should add a third strategy, a flawed or limbo classification, a fuzzy one with no clear criteria for either the inclusion or exclusion of such events into a broader class of discourse markers as a whole. It is our belief that data-driven studies based exclusively on acoustic-prosodic features shed light on the classification of discourse markers, since they are generally defined as syntactically detached structures with no propositional content, and bring to light the discriminative prosodic behavior of disfluencies as a legitimate subtype of discourse markers. In future work, we aim at tackling prosodic parameters per discourse markers subtypes, bridging the acoustic-prosodic properties to the discourse derived sub-categorization of this broad class.

6 Conclusions and future work

This work presented our first attempt to describe discourse markers in three different corpora in EP, namely university lectures, map-task dialogues, and a collection of tweets. Our main goal was to analyze the type of discourse markers used in the various domains. Our results showed that the selection of discourse markers is domain and speaker dependent. Even in the same corpus, there are speakers that tend to use the same discourse marker and there are those who vary amongst several structures. We also found that the most frequent discourse markers are similar in all three corpora. However, in the collection of tweets there are discourse markers that do not occur in the other two corpora. These are discourse markers typically used in more informal contexts, and are clearly produced by teenagers and young adults.

In this multidisciplinary study, comprising both a linguistic perspective and a computational approach, discourse markers are also automatically discriminated from other structural metadata events, namely sentence-like units and disfluencies. As for their distributional patterns, our results showed that markers and disfluencies tend to co-occur in the dialogue corpus, but have a complementary distribution in the university lectures.

[5] In this perspective, punctuation marks are not accounted for, since they are part of sentence-type structures of different illocution values.

We used three acoustic-prosodic feature sets and machine learning to automatically distinguish between discourse markers, disfluencies and SUs. Our in-domain experiments achieved an accuracy of about 87% in university lectures and 84% in dialogues, in line with our previous results. The GeMAPS, and especially the eGeMAPS features, recently introduced for voice research and affective computing and commonly used for other paralinguistic tasks, achieved good performance on our data, while taking minutes to run instead of hours and even days. Our results suggest that turn-initial discourse markers are usually easier to classify than disfluencies, a result also previously reported in the literature.

We replicated the multiclass experiments in a cross-domain evaluation in order to evaluate how robust are the models across domains. The results achieved are about 11%-12% lower than when in-domain data is used, but we have concluded that data from one domain can still be used to classify the same events in the other. These results allow us to hypothesize that it will be possible to classify discourse markers in out-of-domain data. Overall, despite the complexity of this task, these are very encouraging state-of-the-art results for multiclass classification. Both discourse markers and disfluencies are described in the literature as sharing some properties (Goldwater et al., 2010; Liu et al., 2006). Ultimately, using exclusively acoustic--prosodic cues, discourse markers can be fairly discriminated from disfluencies and SUs.

In order to better understand the contribution of each feature, we have also analyzed the impact of the features both in dialogues and university lectures. Pitch features, namely pitch slopes, are the most relevant ones for the distinction between discourse markers and disfluencies. These features are in line with the wide pitch range of discourse markers, in a continuum from a very compressed pitch range to a very wide one, expressed by total deaccented material or H+L* L* contours, with upstep H tones.

In future work, we intend to do a more exhaustive analysis of the most prominent acoustic--prosodic features and also to do a tonal analysis of the discourse markers and adjacent prosodic contexts, to understand the relevance of the discourse marker in the classification process. We also intend to include discourse markers in the language models for EP already trained with other structural metadata events, which will result in enriched automatic transcriptions, and to integrate the classifiers in spoken dialogue systems. Furthermore, we also intend to study morpho-syntactic features of the discourse markers, as well as paralinguistic events, such as laughs, nodding, and hand gestures. We would also tackle the cross-language analysis of discourse markers, in order to verify how specific or universal are their behavior.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and under PhD grant SFRH/BD/96492/2013, and Post-doc grant SFRH/PBD/95849/2013.

References

Jean-Michel Adam (2008). *A lingüística textual. Introdução à análise textual dos discursos*. São Paulo: Cortez Editora.

Karin Aijmer (2013). *Understanding Pragmatic Markers: a Variational Pragmatic Approach.* Edinburgh: Edinburgh University Press.

Karin Aijmer, Ad Foolen and Anne-Marie Simon-Vandenbergen (2006). Pragmatic markers in translation: a methodological proposal. *Approaches to Discourse Particles, 1*:101-114.

Fernando Batista (2011). *Recovering Capitalization and Punctuation Marks on Speech Transcriptions.* PhD thesis, Instituto Superior Técnico, Lisbon.

Fernando Batista, Helena Moniz, Isabel Trancoso, Nuno Mamede and Ana Isabel Mata (2012). Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. *Journal of Speech Sciences*, *2*:115-138.

Fernando Batista, Pedro dos Santos Lopes Curto, Isabel Trancoso, Alberto Abad, Jaime Rodrigues Ferreira, Eugénio Alves Ribeiro, Helena Moniz, David Martins de Matos, Ricardo Ribeiro (2016). SPA: Web-based Platform for easy Access to Speech Processing Modules. In *LREC, European Language Resources Association (ELRA),* pages 3886-3892, doi: ISBN: 978-2-9517408-9-1, Portoroz, Slovenia.

Kate Beeching (2002). *Gender, Politeness and Pragmatic Particles in French*. Amsterdam: John Benjamins.

Kate Beeching (2014). Just a suggestion - Pragmatic marker just/e in French and English. Oral communication in the International *Workshop - Pragmatic Markers, Discourse Markers and Modal Particles: What do we know and where do we go from here?*, Università dell'Insubria, Como (Italy).

Margarita Borreguero and Araceli López (2010). Los marcadores del discurso y la variación lengua hablada vs. lengua escrita. *Los estudios sobre marcadores del discurso en español, hoy*. Madrid: Arco/Libros: 415-496.

Gaspar Brogueira, Fernando Batista, João Paulo Carvalho, Helena Moniz (2014). Expanding a Database of Portuguese Tweets. In *SLATE'14 3rd Symposium on Languages, Applications and Technologies*, Schloss Dagstuhl, vol. 4569, series OpenAccess Series in Informatics (OASIcs), pages 275-282, Bragança, Portugal.

Vera Cabarrão, Helena Moniz, Jaime Ferreira, Fernando Batista, Isabel Trancoso, Ana Isabel Mata, Sérgio Curto (2015). Prosodic classification of discourse markers. *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences*, pages 14-17, Glasgow, UK: The University of Glasgow.

Maria Antónia Coutinho (2009). Marcadores discursivos e tipos de discurso. *Linguistic Studies* 2:193-210.

Annika Denke (2009). *Nativelike Performance. Pragmatic Markers, Repair and Repetition in Native and Non-native English Speech*. VDM Publishing.

Inês Duarte (2000). *Língua Portuguesa. Instrumentos de Análise*. Lisbon: Universidade Aberta.

Florian Eyben, Martin Wöllmer, Bjorn Schuller (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. *ACM Multimedia (MM),* ACM, Firenze, Italy.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julian Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, no. 2:190-202.

Kerstin Fischer (1998). Discourse particles, turn-taking, and the semantics-pragmatics interface. *Revue de Sémantique et Pragmatique, 8*:111–137.

Kerstin Fischer (2000). *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles.* Mouton de Gruyter, Berlin/New York.

Bruce Fraser (1988). Types of English discourse markers. *Acta Linguistica Hungarica, 38*:19-33.

Bruce Fraser (1990). An approach to discourse markers. *Journal of Pragmatics, 14*:383–395.

Bruce Fraser (1999). What are discourse markers? *Journal of Pragmatics, 31*:931–952.

Bruce Fraser (2009). Topic orientation markers. *Journal of Pragmatics, 41*:892-898

Tiago Freitas, Maria Celeste Ramilo (2003). O actual estatuto da palavra portanto. In *Actas do XVIII Encontro da Associação Portuguesa de Linguística*, pages 357-369, Lisbon.

Sónia Frota (2000). *Prosody and focus in European Portuguese: Phonological phrasing and intonation.* New York, Garland Publishing.

Sónia Frota (2014). The intonational phonology of European Portuguese. In Sun-Ah Jun (ed.) Prosodic Typology II: 6-42. Oxford: Oxford University Press.

Sharon Goldwater, Dan Jurafsky, Christopher D. Manning (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52:181–200.

Agustin Gravano, Julia Hirschberg, and Štefan Beňuš. (2012). Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, *38*(1):1-39.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10-18.

Peter Heeman and James Allen (1999). Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, *25*:527–571.

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

Daniel Hirst and Albert Di Cristo (1998). *Intonation systems: a survey of twenty languages*. Cambridge University Press.

Andreas H. Jucker and Sara W. Smith (1998). And people just you know like "wow": Discourse markers as negotiating strategies. In Jucker, Andreas H. and Yael Ziv (eds). *Discourse Markers: Description and Theory,* 57:171. Amsterdam: John Benjamins.

Philipp Koehn (2005). Europarl: A parallel corpus for statistical machine translation. In *MT SUMMIT*, vol. 5, pages 79-86.

Matthew Lease and Mark Johnson (2006). Early deletion of fillers in processing conversational speech. In *Proceedings of HLT-NAACL 2006 (Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics)*, Companion Volume: Short Papers, pages 73–76. New York, NY.

Rivka Levitan, Stefan Benus, Agustin Gravano, and Julia Hirschberg (2015). Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction, Stanford, CA*.

António Calado Lopes, David Martins de Matos, Vera Cabarrão, Ricardo Ribeiro, Helena Moniz, Isabel Trancoso, Ana Isabel Mata (2015). Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers. http://arxiv.org/abs/1503.0914.

Ana Cristina Macário Lopes (1997). Então: elementos para uma análise semântica e pragmática. *Actas do XII Encontro Nacional da APL*, vol. 1:177-189. Lisbon: Colibri.

Ana Cristina Macário Lopes (2009). Justification: a coherence relation. *Pragmatics,* 19(2): 223-239.

Ana Cristina Macário Lopes (2014a). Aliás: A contribution to the study of a Portuguese discourse marker. In Piera Molinelli & Chiara Ghezzi, (Eds.), *Discourse and Pragmatic Markers from Latin to the Romance Languages,* (9). Oxford University Press.

Ana Cristina Macário Lopes and Patrícia Amaral (2006). From time to discourse monitoring: agora e então in European Portuguese. In Corneille, B. and Delbecque, N. (eds.), Topics in subjectification and moralisation. *Belgian Journal of Linguistics,* (20):3-18.

Ana Cristina Macário Lopes and Sara Sousa (2014b). The discourse connectives ao invés and pelo contrário in European contemporary Portuguese. *Journal of Portuguese Linguistics,* 13(1):3-27.

Ana Cristina Macário Lopes (2016). Discourse Markers 24. In Wetzel, Leo, Menuzzi, Sergio & Costa, João (eds.). *The Handbook of Portuguese Linguistics*: 441-456. New York: Wiley-Blackwell.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf and Mary Harper (2006). Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. In *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n. 5, pages 1526-1540.

John Makhoul, Francis Kubala, Richard Schwartz and Ralph Weischedel (1999). Performance measures for information extraction. In *Proceedings of the DARPA Broadcast News Workshop*, pages 249-252. Herndon, VA.

Ana Isabel Mata and Helena Moniz (2016). Prosódia, variação e processamento automático. In Martins, Ana Maria & Ernestina Carrilho (eds). *Manual de Linguística Portuguesa* (Vol. 16). MRL Series. *De Gruyter.*

Amália Mendes (2013). Organização textual e articulação de orações. In Raposo, Eduardo B. P., M. Fernanda Bacelar do Nascimento, M. Antónia Coelho da Mota, Luísa Segura, Amália Mendes (orgs) *Gramática do Português*, vol. II: 1691-1755. Lisboa: Fundação Calouste Gulbenkian.

Helena Moniz (2006). *Contributo para a Caracterização dos Mecanismos de (Dis)Fluência no Português Europeu.* MA thesis, University of Lisbon.

Helena Moniz (2013). *Processing disfluencies in European Portuguese*. PhD thesis, University of Lisbon.

Helena Moniz, Fernando Batista, Isabel Trancoso, Ana Isabel Mata da Silva (2012). Prosodic context-based analysis of disfluencies. In *Interspeech 2012, ISCA,* pages 1961-1964. Portland, Oregon, U.S.A.

Helena Moniz, Fernando Batista, Ana Isabel Mata, Isabel Trancoso (2014). Speaking style effects in the production of disfluencies. *Speech Communication, vol. 65*:20-35.

Helena Moniz, Jaime Rodrigues Ferreira, Fernando Batista, Isabel Trancoso (2015). Disfluency Detection Across Domains. In *DISS 2015*, Edinburgh, Scotland, U. K.

Helena Moniz, Fernando Batista, Ana Isabel Mata, Isabel Trancoso (2016). Towards automatic language processing and intonational labeling in European Portuguese. In M. Armstrong, N. C. Henriksen, & M. M. Vanrell (Eds.), *Intonational Grammar in Ibero-Romance: Approaches across linguistic subfields*, pages 295-324. Philadelphia, USA: John Benjamins.

João Neto, Hugo Meinedo, Márcio Viveiros, Renato Cassaca, Ciro Martins and Diamantino Caseiro (2008) Broadcast news subtitling system in Portuguese. In *Proceedings of ICASSP'08*, pages1561–1564. Las Vegas, USA.

Mari Ostendorf, Benoît Favre, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Dustin Hillard, Julia Hirschberg, Heng Ji, Jeremy G. Kahn, Yang Liu, Sameer Maskey, Evgeny Matusov, Hermann Ney, Andrew Rosenberg, Elizabeth Shriberg, Wen Wang, and Chuck Wooters (2008). Speech Segmentation and Spoken Document Processing. In *IEEE Signal Processing Magazine*, pages 59-69.

Janet Pierrehumbert and Julia Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in communication*:271-311.

Ana Pimentel (2012). *Os marcadores conversacionais no ensino de Português Língua Estrangeira: um estudo de caso*. MA thesis, Faculty of Letters, University of Oporto.

Andrei Popescu-Bellis, Sandrine Zufferey (2011). Automatic identification of discourse markers in dialogues: An in-depth study of like and well. *Computer Speech & Language 25 (3)*:499-518.

Gisela Redeker (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics, 14*:367-381.

Kenneth Brian Samuel (1999). *Discourse learning: an investigation of dialogue act tagging using transformation-based learning.* PhD thesis, University of Delaware.

Deborah Schiffrin (1987). *Discourse Markers.* Cambridge University Press, Cambridge.

Deborah Schiffrin (2001). Discourse markers: Language meaning and context. In Deborah Schiffrin, Deborah Tannen and Heidi Hamilton (Eds.) *Handbook of Discourse analysis.* Oxford: Basil Blackwell.

Lawrence Schourup (1999). Discourse markers. *Lingua,* 107: 227-65.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan (2013). Paralinguistics in speech and language state-of-the-art and the challenge. *Computer Speech & Language,* 27, no. 1: 4-39.

Elizabeth Shriberg (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of Califórnia.

Elizabeth Shriberg (2001). To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31:153–169.

Augusto Soares da Silva (2006). The polysemy of discourse markers: the case of pronto in Portuguese. *Journal of Pragmatics, 38:*2188-2205.

Isabel Trancoso, Maria do Céu Viana, Inês Duarte, and Gabriela Matos (1998). Corpus de Diálogo CORAL. In *Proceedings of PROPOR'98*, Porto Alegre, Brasil.

Isabel Trancoso , Rui Martins, Helena Moniz, Ana Isabel Mata, and Maria do Céu Viana (2008). The LECTRA Corpus - Classroom Lecture Transcriptions in European Portuguese. In *Proceedings LREC'08*, Marrakech, Morocco.

Hudinilson Urbano (2003). Marcadores conversacionais. In Preti, D. (org.). *Análise de textos orais.* 6a Edição:93-116. São Paulo: Humanitas.

Maria do Céu Viana, Sónia Frota, Isabel Falé, Flaviane Fernandes, Isabel Mascarenhas, Ana Isabel Mata, Helena Moniz & Marina Vigário (2007). Towards a P_ToBI. In http://www.ling.ohio-state.edu/~tobi/.