PROSODIC CLASSIFICATION OF DISCOURSE MARKERS

Vera Cabarrão[1, 2], Helena Moniz[1, 2], Jaime Ferreira[1], Fernando Batista[1, 3],
Isabel Trancoso[1, 4], Ana Isabel Mata[2], Sérgio Curto[1]

[1] L2F, INESC-ID, [2]FLUL/CLUL, [3]ISCTE, [4]IST – Universidade de Lisboa
{veracabarrao, jaimeferreira.90, scurto}@gmail.com, {Helena.Moniz, fernando.batista, isabel.trancoso}@ inesc-id.pt,
aim@letras.ulisboa.pt

ABSTRACT

The first contribution of this study is the description of the prosodic behavior of discourse markers present in two speech corpora of European Portuguese (EP) in different domains (university lectures, and map-task dialogues). The second contribution is a multiclass classification to verify, given their prosodic features, which words in both corpora are classified as discourse markers, which are disfluencies, and which correspond to words that are neither markers nor disfluencies (chunks). Our goal is to automatically predict discourse markers and include them in rich transcripts, along with other structural metadata events (e.g., disfluencies and punctuation marks) that are already encompassed in the language models of our in-house speech recognizer. Results show that the automatic classification of discourse markers is better for the lectures corpus (87%) than for the dialogue corpus (84%). Nonetheless, in both corpora, discourse markers are more easily confused with chunks than with disfluencies.

Keywords: Discourse markers, Prosody, Lectures, Dialogues, Structural Metadata Events

1. INTRODUCTION

This study aims at describing the prosodic behavior of discourse markers in two speech corpora of different domains in EP, namely university lectures and map-task dialogues. The main motivation of our work is to perform a multiclass automatic classification based on prosodic features to verify which words in both corpora are classified as discourse markers, which are disfluencies and which are words that are neither markers nor disfluencies (chunks). This will allow us, in a near future, to include discourse markers in rich transcription models of our in-house speech recognizer ([2], [16]). The improvement of the output of the speech recognition system was already verified when rich transcriptions, especially metadata events, such as disfluencies and punctuation marks, were encompassed in the language models of the speech recognizer. The recent availability of the large speech corpora used in this study made it possible to analyze the metadata event missing, i.e., discourse markers ([14], [17]).

The importance of discourse markers in automatic speech processing systems was already verified when [11] showed that words occurring in the beginning of utterances or as discourse markers have higher error rates when automatically recognized. The authors divided words into three classes with high error rates: (i) an open class (names and verbs); (ii) a closed class (prepositions and articles), and (iii) discourse markers. The fact that these structures stand as an independent class shows that they are hard to recognize and classify. Therefore, the need to linguistically characterize these structures is crucial to produce richer transcripts.

The study of discourse markers in spontaneous speech is also important if we consider the emerging need to translate these structures to other languages, especially if we intend to use automatic translation tools. The idiomatic nature of discourse markers in spontaneous speech makes it hard to find exact equivalents, which justifies an inventory of these structures and a description of their function in discourse.

In EP, like in several other languages, discourse markers account for different linguistic structures, such as adverbs, conjunctions, interjections, and are classified according to their function in a given text. The different pragmatic functions that can be associated with the same discourse marker ([7], [22]) represent an additional challenge for their classification. Moreover, most of the studies available for EP are based in written texts, and still do not account for all the discourse markers present in oral communication.

In this study, we aim at contributing to the description and classification of discourse markers in spontaneous speech. Here they are produced almost like fixed expressions, which are used to start the utterances, but appear to be deployed of any semantic content ([5]), being rather used as interjections and/or conversational fillers. Under the scope of prosodic analysis, we aim at predicting these types of discourse markers automatically, so that we can include them in the language models for

EP already trained with other structural metadata events. This will also result in enriched automatic transcriptions.

2. RELATED WORK

In EP, as in English and many other languages, authors still do not agree on which structures should be included in a discourse markers category ([8] vs. [21]), and what designation that class should have ([3]), namely, discourse markers, connective markers, phatic markers, conversational markers, among others. The fact that the same word can be classified both as a discourse marker, when it plays a pragmatic or discursive role, and as a non-discursive marker, having a sentential/textual role, makes it more difficult to disambiguate and, therefore, to classify. [19] pointed out that the high frequency of discourse markers in spontaneous speech, the fact that they tend to occur in the beginning of utterances and the fact that they do not necessarily carry any syntactic or semantic information, led some authors to compare them to disfluencies, namely lexicalized filled pauses, and to remove them in the automatic processing of speech. However, authors like [13] and [20] argued that the presence of discourse markers can be used to infer the structure of discourse, and should, therefore, be encompassed in the automatic processing.

An inventory with the distribution and relative frequency of the distinct discourse markers is still not available for EP, neither the linguistic features that best describe them. The studies available are mainly focused in the behavior of one specific discourse marker and the taxonomy proposals are mostly based on written texts. [5], by adapting the classifications of [9] and [1], made an attempt to describe discourse markers according to the type of text in which they occur, using context to disambiguate its different functions. Authors like [15], [10] and [22] presented studies regarding only a specific discourse marker, namely *então* (so), *portanto* (so, like), and *pronto* (that's it, ok), respectively, and its different functions according to context. Recent studies in EP also analyzed some discourse markers but from a second language acquisition perspective ([18]).

Despite this 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.

3. CORPORA

In this study, we analyzed two spontaneous speech corpora in different domains in EP, namely university lectures (LECTRA corpus), and map-task dialogues (CORAL corpus). Both corpora are available through ELRA. The university lectures corpus (ELRA – S0366), collected within the LECTRA project ([24]), aimed at producing multimedia contents for e-learning applications, mostly for hearing-impaired students. The corpus encompasses 7 courses, 6 recorded in the presence of students, and 1 recorded only with the teacher targeting an Internet audience. The 7 speakers (6 male and 1 female) are all native Portuguese speakers. LECTRA has a total of 75h of speech, of which 33h were orthographically transcribed, totaling 155k words. The CORAL corpus (ELRA – S0367) ([23]), comprising 64 dialogues in map-task format between 32 speakers, has a total of 7 hours orthographically transcribed (61k words). The dialogues occur between two speakers, the giver and 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. The corpus is balanced in terms of gender and of role played by the speaker (giver and follower).

In both corpora, [4] found about 70 discourse markers, corresponding to words and expressions that occur mainly in spontaneous speech, being rare in written texts (e.g. '*pronto'* and '*bem'*/well). This selection was build according to the prospection of the data, and included variations of the same marker, like *então* (so, then) and *mas então* (but then). This data-driven approach allowed studying discourse markers that, despite being very frequent in spontaneous speech, were still not described in the Portuguese grammar. Overall, discourse markers, both in LECTRA and in CORAL, account for 3% of the total number of words in each corpus. The LECTRA corpus has a total of 5,103 vs. 1,719 in CORAL. [4] also showed that the most frequent discourse markers are similar in both corpora, even though their selection is domain and speaker dependent. In the same corpus, there were speakers that tended to use the same discourse marker and those who varied among several structures.

For the multiclass classification task, we only used the discourse markers that occur turn-initially or as isolated utterances. They account, in LECTRA, for 36% of the total discourse markers found, and in CORAL, for about 67%. These results were

expectable considering the nature of the data, CORAL being a corpus characterized by short utterances and sequenced interactions between speakers, whereas LECTRA only encompasses the speech of the teachers.

4. MULTICLASS CLASSIFICATION TASK

For the multiclass classification task, we extracted about 6300 prosodic and acoustic features of the discourse markers in the beginning of utterances for both corpora, using OpenSMILE ([6]). The same set of different features was already available for disfluencies studies. We used several machine learning methods, (e.g., Classification and Decision Trees – CARTs; Linear Regression Support Vector Machines, using the Sequential Minimal Optimization - SMO algorithm; Naïve Bayes), using the toolkit WEKA ([12]). The Naïve Bayes method presented results of correctly classified instances under 70% for CORAL and about 73% for LECTRA. The experiments with CARTs, due to memory restrictions, were only possible for balanced data for both corpora, and the results achieved were around 70% for CORAL, and 78% for LECTRA. Considering that the best results were achieved using SMO, we will only show the results achieved with this method.

As for the selection of the parameter C (complexity), we used the value 0.01, given that, in a previous work for disfluencies, this one proved to be the best value to deal with the amount of features extracted from OpenSmile. As for the number of folds for cross-validation, we conducted experiments with both 5 and 10 folds. The accuracy was very similar for both folds, but considering that the time to compute the experiments for 5 folds is much lower than for 10 (CORAL: 534.35 seconds vs. 542.23; LECTRA: 2,034.38 seconds vs. 3,269.32), we will present only the results with a cross-validation of 5.

As for the type of data, we conducted experiments both with balanced and unbalanced corpora. Regarding the total number of instances, in CORAL, there are 6,381 chunks, 1,834 disfluencies, and 723 discourse markers. As for LECTRA, we have a total of 15,068 chunks, 6,066 disfluencies, and 1,286 discourse markers. Since the discourse markers class has a low number of instances, the number of chunks and disfluencies was randomly selected to total 2,000 in CORAL and 4,000 in LECTRA so that the experiments could be made with more balanced corpora. In order to achieve a baseline for further experiments, we also applied a ZeroR method from Weka, which selects the most frequent class.

5. RESULTS

The classification task using SMO is better for LECTRA than for CORAL, both with balanced and unbalanced corpora (Table 1). These results were expectable considering that the LECTRA corpus has fewer speakers than CORAL (7 vs. 20 speakers), which corresponds to less variation.

Results show that the use of acoustic-prosodic features allow for very significant improvements relatively to the baseline, namely: 20% for the university lectures and 13% for dialogues, with unbalanced data, and 16% and 8% for balanced data. Looking at both types of data, results also show that unbalanced data performs better with a higher accuracy. However, we can also see that the kappa value is slightly lower. The confusion matrix for CORAL unbalanced corpora (Table 2) shows that the class better classified is chunks, followed by disfluencies and, finally, discourse markers. This has the lowest recall (0.617), but a precision (0.772) higher than disfluencies (0.769), which shows that, even though there is a high number of structures that are not being considered as discourse markers, those that are have a correct classification. As for LECTRA (in Table 3) results show that a high number of discourse markers were correctly classified, being the disfluency class the one with more ambiguity with the class chunks. Looking at the accuracy measures, discourse markers show a precision of 0.748 and a recall of 0.625. These results show that discourse markers are a difficult class to identify and classify, and that their structures are more easily confused with chunks than with disfluencies.

Corpora	**Unbalanced**			**Balanced**	
	Kappa	Accuracy		Kappa	Accuracy
		ZeroR	SMO		
Dialogues	0.59	71%	**84%**	0.65	79%
Lectures	0.72	67%	**87%**	0.73	83%

Table 1: SMO classification results for balanced and unbalanced corpora.

Unbalanced CORAL			
Classified as	Chunk	Disfluency	DiscMarker
Chunk	6032	270	79
Disfluency	761	1020	53
DiscMarker	240	37	446

Table 2: Confusion matrix for the CORAL corpus.

Unbalanced LECTRA			
Classified as	Chunk	Disfluency	DiscMarker
Chunk	14216	704	148
Disfluency	1422	4521	123
DiscMarker	272	210	804

Table 3: Confusion matrix for the LECTRA corpus.

Another experiment conducted during this work was a cross-domain evaluation for both corpora (balanced and unbalanced), using a different training set. This was made as an attempt to verify how robust our classification is across domains. To test the CORAL corpus, we used the training set of unbalanced corpora of LECTRA (to account for a higher number of instances), and for the LECTRA corpus, we trained the models with the CORAL corpus also unbalanced.

Corpora	**Unbalanced**			**Balanced**	
	Kappa	Accuracy		Kappa	Accuracy
		ZeroR	SMO		
LECTRA to CORAL	0.41	71%	**75%**	0.4	70%
CORAL to LECTRA	0.55	67%	**80%**	0.52	75%

Table 4: SMO classification results for balanced and unbalanced corpora with cross-domain training set.

LECTRA to CORAL			
Classified as	Chunk	Disfluency	DiscMarker
Chunk	5470	844	67
Disfluency	685	1100	49
DiscMarker	461	131	131

Table 5: Confusion matrix for CORAL with the LECTRA unbalanced corpus as training set.

CORAL to LECTRA			
Classified as	Chunk	Disfluency	DiscMarker
Chunk	13642	1396	30
Disfluency	1891	4151	24
DiscMarker	651	497	138

Table 6: Confusion matrix for LECTRA with the CORAL unbalanced corpus as training set.

The results of the cross-domain experiment were similar to those obtained when the train and test set were randomly selected from the same corpus. Even though the accuracy is a little lower (see Table 1), the best results are still found for the LECTRA unbalanced corpus (80%). Looking at the confusion matrices, we see, however, that the discourse markers class is the one with more classification problems in both corpora. The F-measure of discourse markers is equally low in both CORAL (0.270) and LECTRA (0.187). Nevertheless, also in both corpora, even though there are a fewer number of structures identified as discourse markers (the recall in CORAL is of 0.181, and in LECTRA of 0.107), those that are identified show a correct classification (precision in CORAL is of 0.530, and in LECTRA of 0.719). These results are very promising since they allow us to hypothesize that our classification will be possible in different out-of-domain test sets.

6. CONCLUSIONS

This paper presented our first attempt to use prosodic features to classify discourse markers in two different corpora in EP, namely university lectures (LECTRA corpus), and map-task dialogues (CORAL corpus). Regarding the type of discourse markers used in both corpora, [4] showed that the selection of discourse markers were domain and speaker dependent. Nevertheless, the most frequent discourse markers were similar in both corpora. Taking into account this overall analysis of how discourse markers occur in our corpora, and considering the fact that both discourse markers and disfluencies are part of structural metadata events ([14]), we wanted to find out in this study if they have distinguishable prosodic features. The main purpose of this work is to contribute to a data-driven description and categorization of discourse markers, in order to both better understand their pragmatic functions and to best predict them automatically.

Results of the automatic classification task showed that the acoustic-prosodic features improved up to 20% the prediction of the events. It also showed that discourse markers are the hardest event to predict, being a class very difficult to identify and classify. Their prosodic properties are more similar to chunks than to disfluencies, which poses an important question for further analysis, since both discourse markers and disfluencies are considered in the literature to share some properties ([11], [14]). Regarding the experiment with a cross-domain evaluation for both data (balanced and unbalanced), using a different training set, we had similar results to those obtained with the train and test sets selected from the same corpus. This allows us to hypothesize that it will be possible to use our classification in test sets from different domains. Overall, despite the complexity of this task, these are very encouraging state-of-the-art results for multiclass classification.

In a future work, we intend to look at the most prominent prosodic and acoustic features, to understand their relevance 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 dialogue systems.

7. ACKOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, under PhD grant SFRH/BD/96492/2013, and Post-doc grant SFRH/PBD/95849/2013, and also by EU project SPEDIAL (FP7 611396).

8. REFERENCES

[1] Adam, JM. (2008). A lingüística textual. Introdução à nálise textual dos discursos. *São Paulo: Cortez Editora*.

[2] Batista, F. (2011). Recovering Capitalization and Punctuation Marks on Speech Transcriptions. PhD thesis, Instituto Superior Técnico.

[3] Blakemore, D. (2002). Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers. *Cambridge University Press*, Cambridge.

[4] Cabarrão, V., Moniz, H., Batista, F., Trancoso, I., Mata, A. I., Curto, S. (2014). Discourse markers in spontaneous speech in European Portuguese: a first approach. Selected communication presented 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), 16-17 October, 2014.

[5] Coutinho, M. A. (2009). "Marcadores discursivos e tipos de discurso". *Estudos Linguísticos/Linguistic Studies 2*, 193-210.

[6] Eyben, F., Wöllmer, M., Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. *ACM Multimedia (MM), ACM*, Firenze, Italy, 25.-29.10.2010.

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

[8] Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics* 14, 383–395.

[9] Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics 31*, 931–952.

[10] Freitas, T., Ramilo, M. C. (2003). "O actual estatuto da palavra *portanto"* In *Actas do XVIII Encontro da Associação Portuguesa de Linguística*, Lisbon.

[11] Goldwater, S., Jurafsky D., Manning, C. (2008). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *In Proceedings of the Joint Meeting Association for Computational Linguistics and Human Language Technology Conference (ACL/HLT).*

[12] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10-18.

[13] Hirschberg, J., Litman, D.J., (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19 (3), 501–530.

[14] Liu, Y., Shriberg, E., Stolcke, A., Dustin, H., Ostendorf, M. and Harper, M. (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, pp. 1526-1540.

[15] Lopes, A. (1997). "Então: elementos para uma análise semântica e pragmática." *Actas do XII Encontro Nacional da APL*, vol. 1. Lisbon: Colibri, 177-189.

[16] Moniz, H. (2013). Processing disfluencies in European Portuguese. PhD thesis, University of Lisbon.

[17] Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J., Liu, Y., Makey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W. and Wooters, C. (2008). Speech Segmentation and Spoken Document Processing. In *IEEE Signal Processing Magazine*, 59-69.

[18] Pimentel, A. (2012). "Os marcadores conversacionais no ensino de Português Língua Estrangeira: um estudo de caso." MA thesis, Faculty of Letters, University of Oporto.

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

[20] Samuel, K., (1999). *Discourse learning: an investigation of dialogue act tagging using transformation-based learning.* Ph.D. Thesis. University of Delaware.

[21] Schiffrin, D. (1987). *Discourse Markers*. Cambridge University Press, Cambridge.

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

[23] Trancoso, I., do Céu Viana, M., Duarte, I., Matos, G. (1998). *Corpus de diálogo CORAL*. In PROPOR'98, Porto Alegre, Brasil.

[24] Trancoso, I., Martins, R., Moniz, H., Mata, A.I., Viana, M.C. (2008). *The LECTRA corpus - classroom lecture transcriptions in European Portuguese.* In LREC 2008 - Language Resources and Evaluation Conference. Marrakesh, Morocco.