Perception of Spontaneous Narrative Structure

Miguel Oliveira, Jr. [1], *Regina Cruz* [2], *Ebson Wilkerson Silva* [1]

[1] Faculdade de Letras, Universidade Federal de Alagoas, Brazil
[2] Faculdade de Letras, Universidade Federal do Pará, Brazil
miguel@fale.ufal.br, regina@ufpa.br, ebswillk@gmail.com

Abstract

The present paper reports the results of a perceptual experiment that was conducted in order to find out to what extent naïve, untrained labelers can benefit from prosodic information present in spontaneous narratives so as to derive an underlying structure associated with them. Results from inter-rater agreements suggest that even when the lexical, syntactic, and semantic information of a narrative is blocked, labelers can still identify its underlying structure, which clearly indicates that listeners make substantial use of prosodic information in the task of identifying the structure of spontaneous narratives.

Index Terms: spontaneous narrative, structure, prosody, perception

1. Introduction

It´s often postulated that any type of discourse is formed by sets of utterances that have a coherent semantic relationship among them. In this view, discourse is considered to be a structure composed by arranged entities that preserve a similar orientation. The structure of written discourse is most often clear, owing to the use of typographical conventions, such as punctuation, and the organization of the text into paragraphs. Spoken discourse makes use of other mechanisms to signal its structure[1].

There is a vast scholarly work suggesting that one of the most important structuring devices in spoken discourse is prosody. Variation in pitch range ([1], [10], [23], [26]), pausal duration ([5], [9], [25]), speech rate ([8], [11], [15], [21]), and amplitude ([1], [9], [10]) have all been studied, with some success, as potential correlates of discourse structure in speech.

In a large study that used spontaneous narratives told in Brazilian Portuguese as material for analysis, [17] demonstrated that some prosodic variables play a crucial role in structuring discourse, segmenting it into sections that are semantically independent.

Although there are already a considerable number of studies suggesting that speakers often use prosody to structure the flow of information in discourse, there is still very scarce evidence that such procedure is relevant from the point of view of perception. Most of the publications related to this type of research are focused on the Dutch and English languages. [4], for example, based on the analysis of a corpus of sentences read by three different speakers, tried to relate certain prosodic variables with the notion finality from the perspective of perception. Using an empirical unit called Perceptual Boundary Strength (or PBS), which is a measure used by the participants of the experiment to designate how strong they felt the juncture at each sentence boundary to be, he demonstrated that the higher values of the PBS, as judged by the participants on a 10-point scale, the greater the number of prosodic cues associated with that PBS.

[24] demonstrated that listeners not only systematically identify the end of a larger discourse unit, but are also able to predict when there is only one sentence to come or there is much more to come in a monologue. These results would suggest that the proposal that prosody is mainly used to indicate finality or continuity in discourse, as it has been often proposed ([1], [28]) is, perhaps too simplistic.

In order to test whether listeners´ judgments in perceptual experiments of prosody as a structuring device is influenced by lexical, syntactic and semantic information, [19] compared the results of an experiment conducted with normal, read-aloud speech with those derived from an experiment with delexicalized versions of the same speech material. The high correlation between the values of PBS in both experiments indicated that the perception of discourse structure is not biased by lexical, syntactic and semantic information available to the listeners. Prosody would thus be sufficient to derive the underlying structure spoken discourse.

The influence of prosody in the perception of discourse structure was also examined in [23], which, instead of using filtered speech for purposes of comparison, contrasted the results of experiments in different conditions. Two groups of participants were assembled: one that had access only to the transcript of the text under analysis and another that not only had access to the transcript, but also to the audio version of the text. The results of the comparison between these two conditions indicated that access to the speech material incurs into a greater agreement among participants with regard the structure of the text. This would imply that the prosodic information contained in the text makes its structure more transparent from a perceptual viewpoint.

[9] and [10] studied the association between the acoustic-prosodic variation and the structure of discourse through the prism of perception. Using an independent model of analysis, these authors examined the relationship between prosodic features and the structure of discourse based on the results of experiments conducted with participants who had access only to the transcription of spoken texts and participants who, in addition to transcription, had access to the recording of the transcribed text. They found a statistically significant association between aspects of the variation in tone, amplitude and temporal variables and the overall structure of the text under analysis. As in [23], it was observed that when participants have access to spoken version of the text, they can derive its underlying structure with greater success than when this material is not available to them, which suggests that

[1] [16] distinguish "discourse" from "text" in terms of the functions each of these concepts convey: the latter is regarded as a "message in its auditory or visual medium," while the former is viewed as an "interpersonal activity." These definitions resemble the common – and misleading – discrimination of linguistic communication between "monologue" and "dialogue." In the present paper, the words "discourse" and "text" will be used interchangeably.

listeners make significant use of prosodic information in the task of identifying the structure of texts (see also [18]).

The purpose of this study is to examine to what extent naïve, untrained labelers can benefit from prosodic information present in spontaneous narrative to derive the underlying structure of this type of discourse. To this end, an experimental study that utilizes spontaneous narratives as stimuli was designed and carried out.

2. Methods

Four spontaneous/non-elicited narratives, told in the course of "spontaneous interviews" ([27]), were selected for this study. These narratives were presented to a total of 48 participants. The participants in this experiment were invited to participate in this study on a voluntary basis. None of them reported any hearing disability.

The task of the participants was to indicate the points in the narratives at which they thought that the speakers intended to mark as boundaries of *communicative units*. Each participant had access to a couple of examples of narratives segmented into communicative units, just as illustrations. No formal definition of what a "communicative unit" was presented and the participants were instructed to judge the boundaries of communication units in a purely subjective way[1].

Each narrative was presented to the participants under four different conditions. In one condition (C1), participants had access only to the transcript of the narrative (a transcription with no punctuation marks and no indication of a paragraph). They were asked to segment the narrative in the transcript, by indicating the communicative unit boundaries with a slash (/). In a second condition (C2), the transcript of the narrative accompanied by its audio was presented to the participants. Similarly, they were asked to segment the narrative in the transcription, by indicating the communicative unit boundaries with a slash (/). In a third condition (C3), only the audio version of the narrative was presented to the participant. After a first hearing, they were asked to indicate the communicative unit boundaries by pressing a key on a computer keyboard. Subjects' answers were registered in the speech annotation tool developed by the Max Planck Institute, ELAN. In a fourth condition (C4), a delexicalized version of the narrative was presented. Following the method employed in other studies ([13], [15], [20], [25]), the original audio files were filtered with a pass Hann band, resulting in unintelligible speech, but with its prosodic information preserved[2]. Like in the third condition, participants were asked to indicate the communicative unit boundaries by pressing a key on a computer keyboard.

The narratives were randomized in such way that all four of them in different conditions appeared at least three times in each order of presentation (4 narratives x 4 conditions x 3 order = 48 unique set of stimuli). In other words, each participant were exposed and responded to 4 stimuli.

Inter-rater agreement in the discourse and dialogue processing community used to be measured as the percentage of the cases on which coders agree ([6]). [2] argued that the Kappa coefficient of agreement ([3], [12]) should be used, because the percentage of times two coders agree with each other is not a meaningful measure, as it is obfuscated by chance agreement. For that reason, the *de facto* standard to evaluate inter-coder agreement in discourse and dialogue processing studies is now considered to be Kappa (K).

[14] propose the following as standards for strength of agreement for the kappa coefficient: 0 = poor, 0.01–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial and 0.81–1 = almost perfect. The dialogue processing community considers K > 0.7 as an indicator of substantial agreement ([7]).

In order to calculate the Kappa values for this study, all the narratives were divided into words; the end of each word was considered to be a potential boundary. So what we wanted to know was whether raters agreed on whether the end of each word corresponded to a narrative section or not.

3. Results

Figure 1 below shows the results of inter-rater agreement for each narrative in condition 1 (C1), where participants had access only to the transcripts of the narratives.

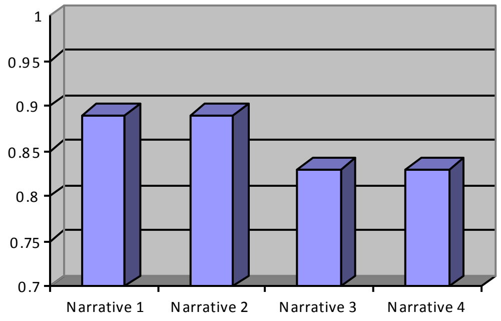

Figure 1: *Kappa values for each narrative in Condition 1 (C1)*

The results in Figure 1 clearly indicate that labelers agree in a statistically significant way on how narratives are segmented in terms of speakers' intentions. These results are in accordance to what has already been reported in the literature.

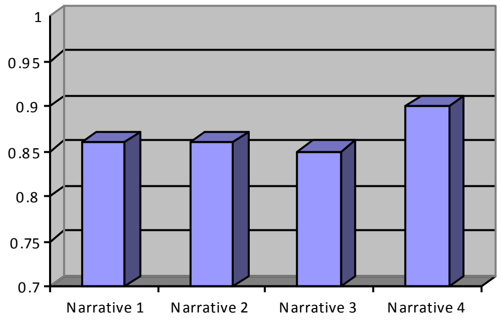

Figure 2: *Kappa values for each narrative in Condition 2 (C2)*

[1] It should be pointed out here that a pilot study was conducted to test the reproducibility of an intention-based discourse segmentation model [22]. The results of this pilot study indicated that coders agree in a statistically significant level with regard to discourse structure as a reflection of speaker intentions, which validates the discourse segmentation model we are presently using in this experiment: [18] (for more details about the pilot study, see [22]).

[2] The filtering was made with Praat. Everything above 400Hz was filtered out.

Figure 2 above shows the results of inter-rater agreement for each narrative in condition 2 (C2), when they had access to both the transcripts and the audio of the narratives.

[22] reported that inter-rater agreement tends to be slightly higher when subjects have access to the audio of the narratives. According to these authors, this happens as the result of the role prosody plays in the perception of narrative structure. The results displayed in Figure 2 do not replicate the tendency that was reported in [22]: two of the narratives actually have a smaller degree of inter-coder agreement in C2 if compared to C1. It should however be pointed out that while the difference between narratives that had a lower degree of inter-rater agreement in C2 differ only in 0,03 points, the difference between narratives that had a higher degree in inter-rater agreement in C2 differ in 0,07 points - a figure relatively higher.

Figure 3 below shows the results of inter-rater agreement for each narrative in condition 3 (C3), when they had access only to the audio of the narratives.

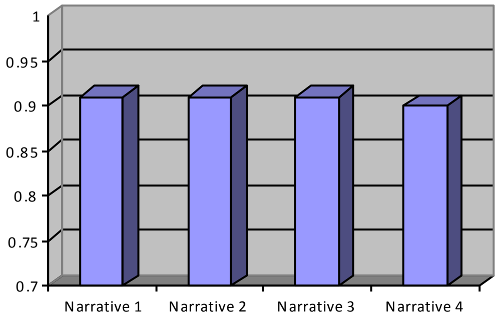

Figure 3: *Kappa values for each narrative in Condition 3 (C3)*

The results in Figure 3 are very homogeneous and show that labelers agree on a very significant level about the structure of spontaneous narratives when only the audio of the narratives are considered.

Finally, Figure 4 shows the results of inter-raters agreement for each narrative in condition 4 (C4), when they had access to a delexicalized version of the original audio of the narratives.

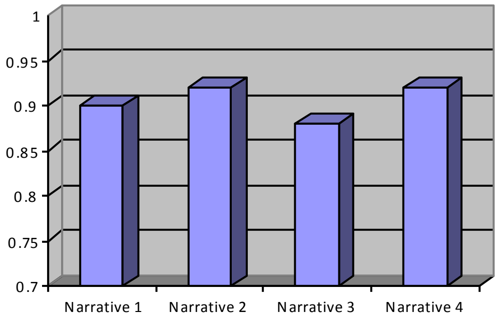

Figure 4: *Kappa values for each narrative in Condition 4 (C4)*

When exposed to a delexicalized version of a narrative, in which only the prosodic information remains intact, labellers agree to a very significant level on how narratives are structured. It should be pointed out that the highest kappa value from this experiment (0,93) was obtained in C4.

Figure 5 below gives a comparison of mean kappa values in each different condition that narratives were presented to labelers. It is interesting to notice that higher kappa values are associated to C3 and C4, i.e., to judgments that were made based solely on the audio version of the narratives.

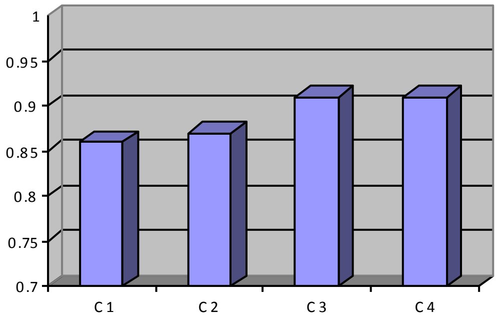

Figure 5: *Mean kappa values in each of the four conditions that stimuli were presented to labelers*

4. Discussion and Conclusions

What the general results from inter-raters agreement on narrative structure reported here suggest is that naïve, untrained labelers can indeed consistently identify a discursive structure in spontaneous narratives. [22] had already demonstrated this with a corpus in which the informational content of its constituent parts was available to the labelers. This study takes one step further and shows that even when the lexical, syntactic, and semantic information of a narrative is blocked, making it unrecognizable, labelers can still identify its underlying structure. This clearly indicates that listeners make substantial use of prosodic information in the task of identifying the structure of spontaneous narratives.

Evidently prosody is not the only resource that people use in order to derive the underlying organization of narrative texts. Results of inter-rater agreement for narratives in Condition 1, which were presented without their corresponding sound version, demonstrate that the information content of narratives is sufficient for labelers to agree on an underlying structure of narratives. Prosody, in principal, would function as additional information to make this agreement even more strong. A number of papers ([4], [19], [23], [24]), including a pilot study that used the same stimuli as in the present study [22], have already demonstrated that this is the case. Although two of the narratives in the present study apparently contradict this finding - in that inter-rater agreement in Condition 1 for them were relatively higher than in Condition 2 -, a trend in this direction is still present in the data, if the results are considered in block (see Figure 5).

Perhaps, the most important finding in the present study is that prosodic information is sufficient for the identification of spontaneous narrative structure. This finding derives from the fact that listeners agree in a very systematic way on the segmentation of narratives when no lexical information is available to them (C4) in the perceptual experiment reported here. Previous works have already demonstrated that the perception of discourse structure is not biased by lexical, syntactic and semantic information available to the listeners ([10], [13], [19], [20], [25]). Most of these works, however, used read-aloud / elicited / non-spontaneous material. Furthermore, most of them dealt with either English or Dutch languages. What justifies the present study is the fact that it utilizes spontaneous, non-elicited narratives told in Brazilian Portuguese. Although there is already an increasing amount of work dealing with prosodic aspects of Brazilian Portuguese and with the functions of prosody in spontaneous discourse, studies on prosody as a structuring device in Brazilian Portuguese spontaneous speech are still very scarce. The present paper stands thus as a contribution to this area of research.

A natural follow-up of this study will be an analysis of the prosodic features that are associated to perceptual boundary strengths (PBS) in different conditions of the narratives. [17], in a production study that used the same material that was utilized here, found that the boundaries occurring between major narratives units (narrative boundaries) are prosodically different from those that occur elsewhere. This was verified in terms of (i) pause occurrence and duration (pauses occur more frequently and are generally longer at narrative boundaries), (ii) pitch reset values (the difference in pitch range values between two adjacent clauses is higher at narrative boundaries) and (iii) boundary tones (low boundary tones usually occur at narrative boundaries).

It is a well-known fact that any study of prosodic aspects of speech should not only consider the production, but also the perception of this phenomenon. The relevance of prosody in the demarcation of discourse structure can only be fully validated after the consideration of their effectiveness from the perspective of perception. The present study is thus a first step toward this pursuit.

5. References

[1] Brown, G., Currie, K. and Kenworthy, J. “Questions of Intonation”. London: Croom Helm, 1980.

[2] Carletta, J. “Assessing agreement on classification tasks: the kappa statistic”. Computational Lingustics, 22(2):249–254, 1996.

[3] Cohen, J. “A coefficient of agreement for nominal scales”. Educational and Psychological Measurement, 20:37–46, 1960.

[4] Collier, R. “On the communicative function of prosody: some experiments”. IPO Annual Progress Report, 28:67-75, 1993.

[5] Collier, R., Piyper, J. R. D. and Sanderman, A. “Perceived prosodic boundaries and their phonetic correlates”. Proceeding of the ARPA Workshop on Human Language Technology. Plainsboro, New Jersey, USA: Morgan Kaufman Publishers, 341–345, 1993.

[6] Di Eugenio, B, “On the usage of Kappa to evaluate agreement on coding tasks”. In: Proceedings of the Second International Conference on Language Resources and Evaluation, 441-444, 2000.

[7] Flammia, G. “Discourse Segmentation of Spoken Dialogue: An Empirical Approach”. Ph.D. thesis, MIT, 1998.

[8] Fon, J. “Speech rate as a reflection of variance and invariance in conceptual planning in storytelling”. In: Proceeding of the 14th ICPhS. San Francisco, 663–666, 1999.

[9] Grosz, B. and Hirschberg, J. “Some intonational characteristics of discourse structure”. In: Proceeding of the International Conference on Spoken Language Processing, Banff, 429-432, 1993.

[10] Hirschberg, J. and Grosz, B. “Intonation features of local and global discourse structure”. In: Proceeding of the DARPA Workshop on Spoken Language Systems. Arden House, 1992.

[11] Koopmans-van Beinum, F. J. and Van Donzel, M. E. “Discourse structure and its influence on local speech rate”. In: Proceeding of the International Conference on Spoken Language Processing. Philadelphia, 1724-27, 1996.

[12] Krippendorff, K. “Content Analysis: an Introduction to its Methodology”. Beverly Hills: Sage Publications, 1980.

[13] Kreiman, J. “Perception of sentence and paragraph boundaries in natural conversation”. Journal of Phonetics, 10:163-175, 1982.

[14] Landis, J. R. and Koch, G. G. “The measurement of observer agreement for categorical data” Biometrics, 33:159–174, 1977.

[15] Lehiste, I. “Perception of sentence and paragraph boundaries”. In B. Lindblom & S. Öhman. Frontiers of Speech Communication Research. London: Academic Press, 191-201, 1979.

[16] Leech, G. and Short, M. “Style in Fiction: A Linguistic Introduction to English Fictional Prose”. London: Longman, 1981.

[17] Oliveira, M. “Prosodic Features in Spontaneous Narratives”. Ph.D. Thesis. Vancouver, BC, Simon Fraser University, 2000.

[18] Passonneau, R. J. and Litman, D. J. “Discourse Segmentation by Human and Automated Means”. Computational Linguistics, 23(1):103-139, 1996.

[19] Pijper, J. R. d. and Sanderman, A. A. “On the perceptual strenght of prosodic boundaries and its relation to supresagmental cues”. J. Acoust. Soc. Am. 96(4):2037-2047, 1994.

[20] Schaffer, D. “The role of intonation as a cue to topic management in conversation”. Journal of Phonetics 12:327-344, 1984.

[21] Selting, M. “Prosody in conversational questions. Journal of Pragmatics”, 17:315-345, 1992.

[22] Silva, E. W. and Oliveira Jr., M. “A percepção dos elementos prosódicos como marca de estruturação de narrativas espontâneas”. Anais do Colóquio Brasileiro de Prosódia da Fala, 1, 2011.

[23] Swerts, M. “Prosodic features at discourse boundaries of different strength”. Journal of the Acoustical Society of America, 101(1):514-521, 1997

[24] Swerts, M., Collier, R. and Terken, J. “Prosodic predictors of discourse finality in spontaneous monologues”. Speech Communication, 15:79-90, 1994.

[25] Swerts, M. and Geluykens, R. “Prosody as a marker of information flow in spoken discourse”. Language and Speech, 37:21-43, 1994.

[26] Silverman, K. “Natural prosody for synthetic speech”. Cambridge: Cambridge University Press, 1987

[27] Wolfson, N. “Speech events and natural speech”. Language in Society 5:189-209, 1976.

[28] Yule, G. “Speakers' topics and major paratones”. Lingua 52:33-47, 1980.