A Study on the Reliability of Two Discourse Segmentation Models

1 Introduction

The present study describes one of the initial tasks of a project currently being developed with the purpose of investigating how certain prosodic features are used to mark the information structure of spoken discourse and which cues are most relevant for the listeners to identify this structure.

There have been no such studies concerning the Portuguese language so far. This project can thus contribute for a better understanding of the role of prosody in natural language, providing valuable information for computational linguistics. Additionally, it will enable us to compare Portuguese with other languages regarding macro-level prosody.

Following the claims of several authors, we assume that there is a relationship between discourse structure and prosodic features. Crucially, our long term goal is to explain how exactly that relationship holds in European Portuguese.

It has been stated that if we want to identify the role of prosody in the structuring of information, we must compare it with an independently obtained discourse structure, in order to minimize the risks of circularity [1-5]. Previous work on other languages has shown that there is no direct match between syntactic structure and prosodic constituency - see [6] and [7]. Instead, prosody seems to be constrained by semantic and pragmatic aspects. Therefore, we should not rely on syntax for that matter, which would otherwise be the most immediate choice.

In order to have some sort of information structure against which prosody can be confronted, some authors elicit instruction monologues, a method which yields speech with a discourse structure determined a priori [2-3; 5]. Others rely on discourse segmentations resulting from discourse analysis [8-16], whereas still others ask subjects to segment texts according to their idea of paragraph [4]. All these approaches thus assume that spoken discourse exhibits a structure somewhat similar to that of written texts, on what concerns the grouping of sentences into larger units like paragraphs, for instance.

We opted for the second method, which has the advantage of making it possible to study different speech styles. This would be impossible if we were to follow the instruction monologues approach, since it generates a very specific kind of data.

The problem with using the discourse analysis approach is that a priori we do not know whether it will yield more than an individual's intuition of discourse structure. If we are to depend on a discourse segmentation method, we must assure that we are employing one that is reproducible, because the more replicable a discourse segmentation model is, the stronger the evidence that discourse structure does exist.

This paper reports an experiment we have conducted in order to compare two discourse segmentation models. We chose the models of Grosz and Sidner [6] and Passonneau and Litman [16]. These have been widely used and there is extensive research on them, which allows us to compare our results with those obtained in work

done for other languages. Both models produce intention based segmentations. The difference is that while the former generates a hierarchical structure the latter generates only a linear kind of segmentation, and it actually comes very close to asking subjects to segment texts based on an intuitive notion of paragraph.

We will eventually choose one of these models for our future research on prosody. Such choice will be based on a test we carried out with the purpose of evaluating inter-coder agreement.

2 Method

The data used have previously been collected for REDIP [17], a project that aims at collecting and studying the language of Portuguese media, dealing mostly with radio and TV broadcasts.

One of the reasons we are using this corpus is because it contains a large amount of spontaneous speech. The importance of using spontaneous speech in this kind of work has to do with the fact that spontaneous discourse can be prosodically different from prepared or read speech. One of the applications of this sort of work is to make speech technology more natural sounding and more efficient in recognizing natural speech.

For this test, we have selected two excerpts from the corpus, averaging one and a half minute in length. These consist of interviews from the radio, involving both male and female speakers. Using dialogues in this kind of work is novelty, and it will allow comparison to other speech styles. One of our concerns in choosing the dialogue samples was to make sure that they contained speech turns long enough for coders to identify more than one discourse segment within each turn. That way we prevented our participants from placing discourse segment boundaries exclusively at turn boundaries, since our long term interest is in the prosodic means of signaling discourse structure and not in the prosodic strategies used to signal turn taking.

We have asked sixteen naïve coders to annotate these two transcripts using the previously mentioned models.

The participants were split into two different groups according to the model they were asked to work with. They all received an orthographic transcription of the selected texts, but for each model only four of them listened to the original recordings. Since we hypothesize that there is a relation between discourse structure and prosody, we expected the listening and non-listening groups to display a different behavior.

Each participant received a set of instructions which were basically the explanatory texts of [14] and [16] translated with slight modifications. The most significant change we introduced was that people were not restricted to placing segment boundaries at prosodic phrase boundaries previously determined. They could place them between any two words in the text instead. We believe the results obtained this way are more independent from prosody.

3 Results

In order to measure inter-coder agreement, we employed the kappa coefficient, which [18] and [19] consider to be the most adequate for that purpose. Kappa values under 0.6 indicate there is no statistical correlation among coders, whereas results over 0.7 point to replicable coder agreement.

It should be noted that in order to compare these two models we had to discard the hierarchical information that Grosz and Sidner's framework supplies, since Litman and Passonneau's produces linear segmentation. Therefore, the results obtained pertain only to the location of discourse segment boundaries.

As can be seen in the table below, our results show that Passonneau & Litman's discourse segmentation model produces higher inter-coder agreement values, outpacing those of Grosz and Sidner's by almost ten points. This is a significant contrast in terms of reproducibility, with Grosz and Sidner's model below the 0.7 mark and Passonneau and Litman's above it.

Table 1. Observed coder agreement

	Grosz & Sidner's Model	*Passonneau & Litman's Model*
Listening	kappa = 0.59	kappa = 0.74
Non-Listening	kappa = 0.68	kappa = 0.69
Overall	kappa = 0.65	kappa = 0.73

We think that the poorer results of Grosz and Sidner's model might be ascribed to its inherent complexity. The fact that coders had to identify relations between segments caused higher variation among subjects.

Listening to the speech recordings did influence the results, but not quite as we expected, considering that other studies report higher levels of agreement in the listening condition. Our findings show that coders using Grosz and Sidner's model agreed less when listening to the recordings.

The different scores between the listening and the non-listening groups corroborate the hypothesis that discourse structure is reflected in prosody. In Litman and Passonneau's model the effect of hearing the speech shows up in a positive way, suggesting that prosody can make discourse structure more explicit. On the contrary, in Grosz and Sidner's model, access to prosodic information might have caused people to look for prosodic means of signaling hierarchy between segments, resulting in a more disparate segmentation. In fact, some authors comment that it has not been proved if prosody can signal the embeddedness level of discourse segments – [4].

It is important to remember that these results were obtained using spontaneous dialogues. The fact that Litman and Passonneau's model scored well demonstrates that it can be applied to dialogues and suggests that an identifiable discourse structure can be found not only in monologues but also in dialogues.

4 Conclusions

The two models employed in this experiment use speaker intention as a criterion to segment discourse. When participants were instructed to segment discourse, they were also asked to provide a description of the intentions underlying each segment. We want to use that information in a future analysis to check whether different segmentations were caused by discourse ambiguity. This may lead to different results. We are planning on running other statistics too, so that we can compare these results with those presented by other authors in similar experiments.

In addition, we want to find out which model produces segmentation closer to the distribution of certain prosodic variables, since our long term objective is to determine the role of prosodic information as an indicator of discourse structure in European Portuguese.

The results observed so far lead us to choose Passonneau and Litman's model for our future research. As was shown, this method displayed a fair level of inter-coder consensus, well above Grosz and Sidner's. If the level of agreement obtained proves not to be satisfactory for the purpose of our research, we may adapt the chosen model in order for it to produce results further above the 0.7 mark.

References

1. Swerts, M. and R. Collier: On the Controlled Elicitation of Spontaneous Speech. Speech Communication 11 (4-5) (1992) 463-468
2. Swerts, M. and R. Geluykens: The Prosody of Information Units in Spontaneous Monologue. Phonetica 50 (1993) 189-196
3. Swerts, M. and R. Geluykens: Prosody as a Marker of Information Flow in Spoken Discourse. Language and Speech 37 (1) (1994) 21-43
4. Swerts, M.: Prosodic Features at Discourse Boundaries of Different Strength. Journal of the Acoustical Society of America 101 (1) (1997) 514-521
5. Swerts, M., R. Collier and J. Terken: Prosodic Predictors of Discourse Finality in Spontaneous Monologues. Speech Communication 15 (1994) 79-90
6. Cutler, A., D. Dahan and W. Donselaar: Prosody in the Comprehension of Spoken Language: A Literature Review. Language and Speech 40 (2) (1997) 141-201
7. Pijper, J. R. and A. A. Sanderman: On the Perceptual Strength of Prosodic Boundaries and its Relation to Suprasegmental Cues. Journal of the Acoustical Society of America 96 (4) (1994) 2037- 2047
8. Grosz, B. and J. Hirschberg: Some Intentional Characteristics of Discourse Structure. Proceeding of the International Conference on Spoken Language Processing (1992) 429-432
9. Grosz, B. J. and C. L. Sidner: Attention, Intention and the Structure of Discourse. Computational Linguistics 12(3) (1986) 175-204
10. Hirschberg, J. and B. Grosz: Intonational Features of Local and Global Discourse Structure. Proceedings of the Workshop on Spoken Language Systems (1992) 441-446
11. Hirschberg, J., C. H. Nakatani and B. J. Grosz: Conveying Discourse Structure through Intonation Variation. Proceeding of the ESCA Workshop on Spoken Dialogue Systems: Theories and Applications, Virgo, Denmark, ESCA (1995)
12. Litman, D. J. and R. Passonneau: Empirical Evidence for Intention-Based Discourse Segmentation. Proc. of the ACL Workshop on Intentionality and Structure in Discourse Relations (1993)

13. Litman, D. J. and R. Passonneau: Combining Multiple Knowledge Sources for Discourse Segmentation. Proc. of 33rd ACL (1995) 108-115
14. Nakatani, C. H., B. J. Grosz and J. Hirschberg: Discourse Structure in Spoken Language: Studies on Speech Corpora. Proceeding of the AAAI Symposium Series: Empirical Methods in Discourse Interpretation and Generation (1995)
15. Nakatani, C. H., B. J. Grosz, D. D. Ahn and J. Hirschberg: Instructions for Annotating Discourses. Technical Report Number TR-21-95. Center for Research in Computing Technology, Harvard University, Cambridge, MA (1995)
16. Passonneau, R. J. and D. J. Litman: Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. Proc. of the ACL (1993)
17. Passonneau, R. J. and D. J. Litman: Discourse Segmentation by Human and Automated Means. Computational Linguistics (1997)
18. Ramilo, M. C. and T. Freitas: A Linguística e a Linguagem dos Média em Portugal: descrição do Projecto REDIP. Paper presented at the XIII International Congress of ALFAL, San José, Costa Rica (2002)
19. Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics 22 (2) (1996) 249-254
20. Flammia, G.: Discourse Segmentation of Spoken Dialogue: An Empirical Approach. Ph.D. thesis, MIT (1998)
21. Beckman, M. E.: A Typology of Spontaneous Speech. In Y. Sagisaka, N. Campbell and N. Higuchi. Computing Prosody: Computational Models for Processing Spontaneous Speech. Springer, New York (1997) 7-26
22. Collier, R.: On the Communicative Function of Prosody: Some Experiments. IPO Annual Progress Report 28 (1993) 67-75
23. Oliveira, M.: Pausing Strategies as Means of Information Processing in Spontaneous Narratives. In: B. Bel and I. Marlien: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France (2002) 539-542
24. Oliveira, M.: Prosodic Features in Spontaneous Narratives. Ph.D. thesis, Simon Fraser University (2000)
25. Oliveira, M.: The Role of Pause Occurrence and Pause Duration in the Signalling of Narrative Structure. In: E. Ranchhod and N. Mamede (eds.): Advances in Natural Language Processing. Third International Conference, PorTAL 2002, Faro, Portugal (2002) 43-51