Gestural Cues of Discourse Segmentation

Jane Chandlee [1], *Nanette Veilleux*[2]

[1] Department of Linguistics and Cognitive Science, University of Delaware, Newark, Delaware
[2] Department of Computer and Information Sciences, Simmons College, Boston, Massachusetts

janemc@udel.edu, veilleux@simmons.edu

Abstract

Research on discourse segmentation frequently involves the identification of certain cues in the various dimensions of text, speech, and gesture. Advances in automated segmentation models and algorithms have been achieved when these cues are taken into consideration. For gestures in particular, it must be observed that their presence and function as cues for discourse boundaries are both genre- and speaker-dependent. This study uses a recorded lecture to investigate whether speaker gestures can reliably predict the presence or absence of a discourse boundary and whether native speakers are able to make use of such a cue in isolation from others.

Index Terms: discourse segmentation, boundary cues, gestures

1. Introduction

The segmentation of any given discourse can vary depending on the party doing the segmenting – the speaker or the listener – as well as what genre-specific information is available – text, intonation, visuals. This variation creates quite a challenge for automated segmentation models and algorithms. This study narrows in on one genre – the lecture – and two sources of information – the text and speaker gesturing – to assess whether the gesturing behavior between segments is markedly different than that within segments. The hypothesis is that the gestures between segments will be noticeably different. An additional hypothesis is that the speaker will indicate the end of a discourse segment by "de-gesturing", which is a comparative lack of gesture at the point of the penultimate syllable before the start of a new discourse segment (suggested by Shahttuck-Hufnagel, personal communication, February 15, 2009).

In particular, the research questions pursued here are as follows:

1) Will native speakers tend to segment a text with significant agreement?

2) Will native speakers be able to identify discourse boundaries based only on the visual cue of speaker gesturing?

3) Will the speaker consistently signal the approach of a segment boundary with a de-gesture cue?

If the answers to these questions are "yes", we may gain some insight into how a listener can use gestures as cues for discourse segmentation.

2. Related work

Any investigation of discourse segmentation should anticipate disagreement among speakers about the placement of segment boundaries, due to varying mental representations of the segment and the potential for multiple levels of segmentation [1]. In other words, if two people are given a text and one is asked to segment it using only section headings while the other is allowed to use both section headings and subheadings, the resultant segmentations will obviously be different. Now if two more people are given the same task with no limitations, they will each decide for themselves how many levels to apply. So again, the result will inevitably vary.

To go about predicting segmentation, then, first requires a definition of a discourse segment. Grosz and Sidner [2] use intention as the central criterion for such a definition; a segment has a discourse segment purpose (DSP), and a new segment begins when that purpose is fulfilled. This conceptualization underlies much work that relies on or requires segmenting. Other studies have tried to get at the problem by looking not just at what a segment *is,* but what cues exist to signal the boundaries. Litman and Passonneau [3, 4] importantly distinguish between definitions of segments that rely on linguistic cues such as lexical items and those that treat segments as independent constructs motivated by focus of attention or rhetorical purpose. Definitions in the latter category are the ones that allow for analyses of the correlation between those segments and linguistic devices. The authors caution against any segmentation strategy that relies on a single linguistic device, as such an approach disregards the variation both across and within speakers in signaling boundaries. For a model's performance to approach humans', it will have to draw on multiple sources of knowledge and adapt dynamically to speakers' strategies.

One such potential source of knowledge for the case of spoken discourse is gesturing. McNeill [5] says that since English lacks markers of discourse structure, such as lexical items whose sole purpose is to indicate the hierarchical position of a statement in a discourse, speakers look to gestures instead to clarify such pragmatic issues. He identifies the following discourse relationships that gestures can assist in creating: succession, voice, point of view, distance, and level. Cassell et al [6] point out that the most obvious gestures tend to accompany the most prominent syllable of the speech at that time, and listeners may pay more attention to these gestures when they need some assistance in disambiguating speech, for example because of noise. The authors examine "posture shifts" in descriptive monologues and dialogues – they code as a "shift" any motion or position shift of a part of the body other than the hands and eyes that is graded by a percentage energy level relative to the speaker's most emphatic movement. They found that posture shifts did occur more frequently at segment boundaries than within segments.

Kendon [7] describes parsing gestures as those that identify the logical components of a discourse. For example, a certain gesture in the "G-family" (characterized by a "finger bunch") may serve to topicalize an entity in the discourse. Eisenstein, Barzilay, and Davis [8] found that visual features combined with lexical cues can statistically predict segmentation. They extend Hearst's [9, 10] lexical cohesion into their own "gestural cohesion", claiming a consistent use of gestural patterns in segments. They do acknowledge that gesturing is subject to variation by speaker, rather than being predefined, but argue that it is the presence of repeated patterns that leads to semantic coherence in a discourse segment. They identify visual codewords that can be analyzed via changes in distribution, much like lexical items. These codewords are represented as vectors of visual, spatial, and kinematic data. The performance of their segmentation algorithm did improve when gesture was included as a cue in addition to lexical cues. As their dataset was limited to spoken descriptions of physical devices, they conclude with a suggestion that more expressive speakers in other genres may prove a fruitful area for future work.

This study examines one such genre, the academic lecture. Our interest in it stems from its middle-ground status between spontaneous and planned speech. Indeed, Grosz and Sidner [2] claim that even if a discourse is planned beforehand, its intentional structure can still be constructed as it progresses. The speaker in this case has a defined discourse purpose, suggesting an underlying structure and planned direction, but in the course of his delivery personal approaches to fulfilling that purpose – such as intonation cues and gestures – naturally emerge.

3. Method

3.1 Data

The lecture is from a commercially available academic video series. The data is in video, transcript, and sound file form. The video shows the speaker lecturing to a present but unseen (to the viewer of the video) audience, and occasionally graphics (such as charts, diagrams, and photographs) fill the screen and block our view of him. He does refer to notes in front of him, indicating that parts of the lecture were scripted or at least planned, but he also appears to make decisions throughout about how to proceed. For example, a number of times he pauses to consult his notes – even if this is just to recall his place in the lecture it indicates that the text was not memorized and therefore likely changed somewhat in the course of the delivery. One lecture was selected for the study that includes a continuous twelve-minute segment during which the speaker is in view the entire time (no graphics fill the screen) – this was obviously necessary for coding his gestures. In text form, the segment amounts to a total of ninety-two sentences.

3.2 Experiment

3.2.1 Manual text segmentation

Ten native English speakers were given the transcript excerpt as a single paragraph and asked to mark the paragraph boundaries based on only their intuitions. They were not given any formal definition of or criteria for what constitutes a paragraph; they acted only on their own sense of when the speaker was changing topics or points.

The agreement among the coders was calculated using the Kappa statistic [11], specifically the Fleiss' Kappa, [12]. The Fleiss' Kappa indicates the reliability of agreement for a fixed number of raters who assigned categorical ratings to a fixed number of items. Importantly, agreement was gauged among the group of test subjects only, not in comparison to an expert coder or predetermined "correct" segmentation of the text. The Fleiss' Kappa statistic is a number between 0 and 1 that indicates the degree of agreement over that which would be expected by chance. The formula is

$$\kappa = \frac{P - P_e}{1 - P_e} \qquad (1)$$

where $P - P_e$ equals the degree of agreement above chance and $1 - P_e$ equals the degree of agreement above chance that is attainable. Perfect agreement thus is $\kappa = 1$ and zero agreement beyond that expected by chance is $\kappa \leq 0$.

Overall the coders showed substantial agreement (κ = 0.70). Five points in the discourse were identified as the most-agreed upon (i.e. as "majority rules" segments), the criterion being that at least eight of the ten coders selected them. Several previous studies [4, 9] have likewise accepted the boundaries indicated by the majority of test subjects. These points became the basis for the gesture coding and analysis.

3.2.2 Gesture coding

The expectation was that if gestures are a reliable cue for segmentation, then the speaker's gestures at the majority-rules points determined by the textual coders should be somehow distinct from the gestures he uses within a segment. To test that hypothesis, another panel of ten subjects was asked to watch the video segment corresponding to the text excerpt without sound. This second set of coders had no knowledge of the lecture's content and no other familiarity with the speaker. They were given a list of twelve time intervals and instructed to pause the DVD after each interval to select one of the following options:

A. The speaker's gestures during this interval indicate that he is introducing a new topic in the lecture.

B. The speaker's gestures during this interval indicate that he is continuing with the current topic in the lecture.

The intervals they were given included the five majority rules segment boundaries indicated by the text annotators, four points not selected by any of the text annotators, and three randomly-selected filler points. The test subjects were informed that the term "gestures" could refer to any movement of the speaker's body, hands, arms, head, or even eyebrows, as well as a change in facial expression.

3.2.3 De-gesturing

To determine whether a de-gesture cue occurred on the penultimate syllable *before* the breaks indicated by the text coders, the authors viewed the video and recorded a description of the speaker's gestures at this point for all of the intervals given to the test subjects. As the intervals the subjects were instructed to code included a few seconds before the suspected break, the subjects did view the time point at which the de-gesturing, if it exists, would have taken place.

4. Results

The Fleiss' Kappa was again calculated for the data provided by the test subjects who viewed the speaker's gestures. The kappa indicates only fair agreement (κ = 0.33) among all twelve intervals. For the five majority rules points, the percentages of coders who believed the gestures indicated a new topic are shown in Table 1.

Majority rules point	% of gestures coders who indicated a change in topic
1	50
2	80
3	100
4	60
5	100

Table 1. Percentage of test subjects (n = 10) who believed the gestures at a point in the discourse indicated a change in topic. These points are the majority rules points.

The gestures that the subjects unanimously agreed were boundary gestures can be classified into two categories. The term "category" is necessary, as opposed to simply "gesture", since it is unlikely that any given speaker will perform an absolutely identical combination of gestures more than once in a discourse. But this particular speaker does have a degree of consistency that makes the descriptions useful enough.

> Category 1: He raises both hands briefly then lowers them to the table again.
>
> Category 2: He pauses, quickly turns to his right and walks a few steps, then comes back.

In addition to examining the presence of gestural cues at discourse boundaries, we should also expect there to be an *absence* of such cues at non-boundary points. If the cue is discovered to be randomly distributed throughout the discourse, meaning it only occurs at boundaries by chance, then we cannot be as confident in asserting its value as a signal to the viewer. Therefore, the intervals presented to the gesture coders included four points in the discourse that no text annotator selected as a break. The hypothesis was that at these points we should find the gestural cue less frequently.

To prevent any kind of "paragraph length" bias, the earliest point in this set was the twentieth sentence in the excerpt. In other words, many points prior to that were not selected by any of the textual annotators, but that is likely because they weren't expecting a new segment to start so early. The behavior at the second sentence, for example, may not be as revealing as that at the twentieth. The percentage agreement for the gestures at the four points not selected by any of the text annotators is shown in Table 2.

Interval	% of gestures coders who indicated a change in topic
1	0
2	30
3	60
4	10

Table 2. Percentage of test subjects (n = 10) who believed the gestures at a point in the discourse indicated a change in topic. These points were not selected by any of the text annotators.

The overall trend is that at the points not selected by anyone as discourse breaks (in the text) the gesture test subjects also did not believe the speaker's gestures indicated a change in topic. The exception is the third interval in Table 2, which was marked by more gesture coders as transitional than not; however, it was a 6/4 split, which is approaching the 5/5 split that equals chance. Therefore this number is not convincing evidence either way.

4.1 De-gesturing

Again looking at the five majority rules points, we noted the following descriptions of the speaker's gestures at the point of the penultimate syllable before the start of a new discourse segment. Again, the intervals the gestures coders were instructed to focus on included a few seconds before the start of the segment. At the penultimate syllable before point one (Table 1), the speaker (who frequently paces or shifts his weight from front to back) is standing still with his hands together in front of him just before he speaks the sentence selected as the start of a new discourse segment. He is in this same posture just before points two and three. Just before point four, he is also still but pinches his chin rather than putting his hands together, and just before point five, he is still and raises and lowers both hands once. Keep in mind that "de-gesturing" is a comparative term, meaning what counts as a lack of gesture may vary from speaker to speaker – the utter absence of movement is not likely to occur during any spoken discourse, but for a speaker who is for the most part extremely animated and constantly moving, as this one is, the comparative stillness can serve as the de-gesture cue.

Thus at all of the five points in question, some manner of de-gesturing was observed just before the break. So although the test subjects who categorized the speaker's gestures may not have had unanimous agreement about the gestures at the start of the segments, their overall agreement combined with the consistent presence of de-gesturing before those segments suggest that de-gesturing may be part of the gestural cue.

5. Discussion

The experience of listening to a lecture involves (at least) the following three dimensions: hearing the words, hearing the speaker's intonation, and seeing the speaker's movements. Separating these three dimensions will always create a more artificial experience. Arguably the least artificial form of that separation is reading the lecture transcript. Reading is extremely natural to literate, native speakers of the language, but a lecture transcript does not entirely resemble, say, an article on the same subject, especially if the transcript preserves discourse markers such as "okay", "now", and "well". Nonetheless, it is not surprising that our strongest agreement statistic came from the first test, in which native speakers segmented the transcript in text form.

But there is some promise when it comes to gestures. It is hard to ignore that 100% of the annotators, who had no exposure to the content of the lecture either through reading or hearing, agreed that certain gestures of the speaker indicated that he was starting a new topic in the discourse. The weaker statistical agreement for this portion of the study could only mean that gesturing alone is not a sufficient cue for segmenting discourse; it does not mean it is not helpful. Again, as opposed to reading or listening, watching without sound is not the most natural way of experiencing discourse for a hearing individual (otherwise the game of charades would not be so challenging). So the fact that certain gestures definitively stood out as "boundary gestures" to all the subjects does grab attention.

Also noteworthy, however, is the presence of a gesture cue at a point that no text coder thought represented a new segment. Again, this occurred only once (interval three of Table 2), and only a slight majority of the gestures coders marked it. What does this mean for an analysis that identifies a segmentation cue? Conclusions drawn from an examination of a cue in isolation need to acknowledge the potentially dynamic relationships between multiple cues. Perhaps a cue that occurs in isolation is/should be disregarded, while cues that work in tandem demand more attention. Such a view would allow cues to occur randomly throughout the discourse so long as they occur in isolation from other cues. Thus there is no inherent "boundary" quality or meaning in a boundary gesture, but its potential to signal a boundary can be activated by the presence of other cues, such as textual markers, textual cohesion, or intonation.

6. Conclusion

This study is part of a larger attempt to isolate cues that could aid a listener and/or viewer in the task of discourse segmentation. A good amount of the previous work on segmentation has considered different types of cues individually – cues such as textual coherence, discourse markers, intonation, and gestures. The overall trend is that more of these cues appear at those points that humans intuit as discourse boundaries. But the relationship is not absolute. More work is needed to identify the mechanisms by which different cues work together, as well as the role of speaker and genre variation in this process. Also, the experiment detailed above suggests that more than one cue is needed to activate the ability to signal a discourse boundary, but this claim should be further quantified. Does "more than one cue" mean two? Three? At most four? Again, examining other speakers in this genre, as well as other genres, could shed further light on such questions. The ultimate goal of improving automated segmentation algorithms may be facilitated by first isolating the audio, visual, and textual dimensions of discourse genres.

7. Acknowledgements

This work was partially funded by the Simmons College Fund for Research.

8. References

[1] M. Walker, A. Joshi, and E. Prince, "Centering in naturally occurring discourse: an overview," in *Centering Theory in Discourse*, M. Walker, A. Joshi, and E. Prince, Eds. New York: Oxford UP, 1998, pp. 1-28.

[2] B. J. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics,* vol. 12, pp. 175-204, 1986.

[3] D. Litman and R. J. Passonneau, "Combining multiple knowledge sources for discourse segmentation," in *Proceedings of the 33rd Annual Meeting*, Association for Computational Linguistics, 1995, pp. 108-115.

[4] R. J. Passonneau and D. J. Litman, "Empirical analysis of three dimensions of spoken discourse: segmentation, coherence, and linguistic devices," in *Computational and Conversational Discourse: Burning Issues – an Interdisciplinary Account*, E. H. Hovy and D. R. Scott, Eds. New York: Springer, 1996, pp. 161-194.

[5] D. McNeill, *Hand and Mind: what gestures reveal about thought.* Chicago: The University of Chicago Press, 1992.

[6] J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich, "Non-verbal cues for discourse structure," in *Proceedings of ACL*, 2001, pp. 106-115.

[7] A. Kendon, *Gesture: Visible action as utterance.* Cambridge: Cambridge UP, 2004.

[8] J. Eisenstein, R. Barzilay and R. Davis, "Gestural Cohesion for Topic Segmentation," in *Proceedings of the 46th Annual Meeting*, Association for Computational Linguistics, 2008, pp. 852-860.

[9] M. A. Hearst, "TextTiling: A quantitative approach to discourse segmentation," University of California, Berkeley, Tech. Rep. 93/24, 1993.

[10] M. A. Hearst, "TextTiling: segmenting text into multi-paragraph subtopic passages," *Computational Linguistics,* vol. 23.1, pp. 33-64, 1997.

[11] S. Siegel and N.J. Castellan, Jr., *Nonparametric Statistics for the Behavioral Sciences,* 2nd ed. New York: McGraw-Hill, 1988.

[12] J. L. Fleiss, "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin,* vol. 76.5, pp. 378-382, 1971.