The Interaction between Information and Intonation Structure: Prosodic Marking of Theme and Rheme

Max M. Louwerse (mlouwers@memphis.edu)[a]
Patrick Jeuniaux (pjeuniau@memphis.edu)[a]
Bin Zhang (bzhang@memphis.edu)[b]
Department of Psychology / Institute for Intelligent Systems[a]
Department of Computer Science / Institute for Intelligent Systems[b]
Memphis, TN 38152 USA

Jie Wu (jie82.wu@gmail.com)[b]
SpeechGear, Inc.
Northfield, MN 55057 USA

Mohammed E. Hoque (mehoque@mit.edu)
Media Lab / MIT
MA 02139 USA

Abstract

Several studies have investigated the relation between information structure and intonation structure. Few studies however have investigated this relationship empirically using natural face-to-face conversations. The current study explores this relation using a large corpus of face-to-face conversations on a map navigation task. In this task dialogue partners sometimes do and sometimes do not have common ground, depending on the differences between their maps. The corpus is therefore ideal to investigate differences between given (theme) and new information (rheme). The current paper presents a technique of automated speech segmentation and transcript time stamping and applies this technique to determine prosodic differences in information structure. Confirming several theoretical studies it shows that the average pitch of the rheme in a turn is significantly higher than the average pitch of the phrasal theme of that turn, showing the relation between information and intonation structure.

Keywords: information structure, intonation structure, theme; rheme; prosody; pitch, multimodal communication.

Introduction

Multimodal communication is comprised of various modalities, both linguistic (intonation and information structure) and non-linguistic (facial expressions, eye gaze and gesture). Despite the deceptively simple appearance of these communicative tools in human-human face-to-face conversation, relatively little is understood about their interaction and alignment. The current paper focuses on the relation between the two linguistic modalities: theme and rheme in language and the prosody in speech.

Knowing the nature of the relation between these modalities can shed light on various areas of cognitive science. From a psychological perspective, an understanding of the interplay of modalities can help us understand language and communication (Clark, 1996). Limited experimental research is available that can help determine whether modalities can be substituted or whether they are complementary (Doherty-Sneddon, et al., 1997).

From an educational perspective, an understanding of modalities can help answer questions regarding student motivation, interest, and confusion, as well as how instructors and tutors can monitor and respond to these cognitive states (Kort, Reilly & Picard, 2001). But with little information available on the conditions under which students use modalities, tapping into students' cognitive states is difficult (Graesser, et al., in press).

From a computational perspective, an understanding of the interplay between modalities can help in the development of animated conversational agents (Louwerse, Graesser, Lu & Mitchell, 2005). These agents maximize the availability of both linguistic (semantics, syntax) and paralinguistic (pragmatic, sociological) features (Cassell & Thórisson, 1999; Massaro & Cohen, 1994; Picard, 1997). But without experimental data on multimodal communication, the guidelines for implementing human-like multimodal behavior in agents are missing (Cassell, et al., 1994).

In an ongoing project on multimodal communication in humans and agents, we are investigating the interaction between dialogue act, speech, eye gaze, facial movements, gesture, and map drawing. The project aims to determine how these modalities are aligned, whether, and if so when, these modalities are observed, and whether the correct use of these channels actually aids comprehension.

Due to the inherent complexity of multimodal communication, controlling for genre, topic, and goals during unscripted dialogue is crucial. With these concerns in mind, we used the Map Task scenario (Anderson, et al., 1991), a restricted-domain, route-communication task. In the Map Task scenario it is possible for experimenters to determine exactly what each participant knows at any given time. In this scenario, the Instruction Giver (IG) coaches the Instruction Follower (IF) through a route on the map.

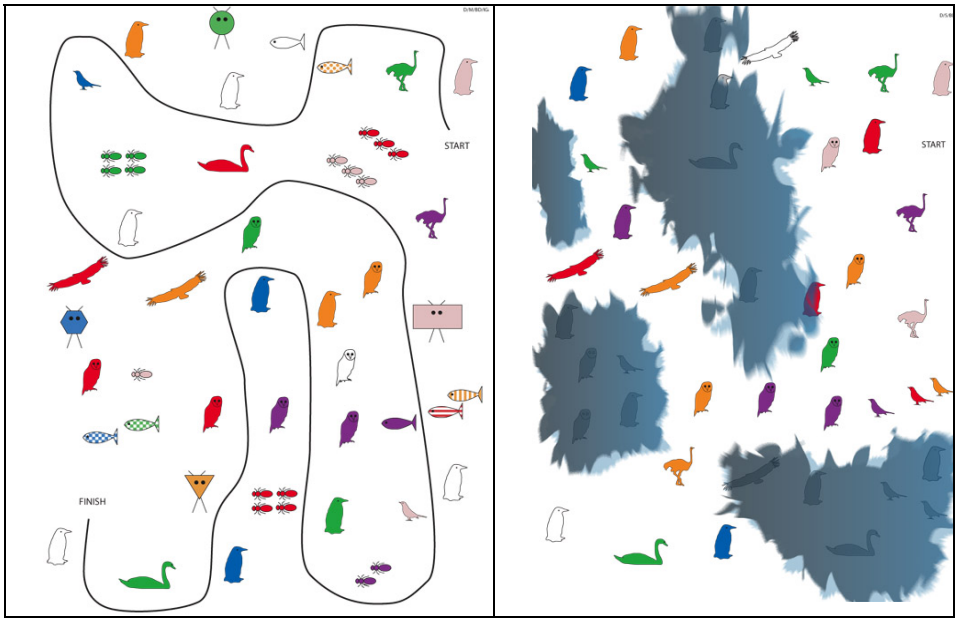

Figure 1. Examples maps for the IG (left) and the IF (right)

By way of instruction, participants are told that they and their interlocutors have maps of the same location, but drawn by different explorers, and so are potentially different in detail.

Sixteen different maps were used, each varying according to the presentation of landmarks, route shape, and method of distortion in the IF map. For instance, IF's maps were distorted with blurred out portions of the map, as shown in Figure 1. The goal of these differences between maps was to elicit dialogue between the participants in a controlled environment whereby dialogue partners sometimes do and sometimes do not have common ground, depending on the differences between their maps. These discrepancies in common ground can be resolved through multimodal communication. Dialogue partners can maintain common ground by using different modalities including eye gaze, facial expressions, gestures, content information or intonation. Elsewhere (Louwerse et al., 2006; 2007) we have reported on the relation between both linguistic and non-linguistic modalities. The current paper investigates the relationship between these two modalities and tests whether information structure can predict patterns in intonation structure.

Information and Intonation Structure

Several studies have discussed the relationship between information and intonation structure (Halliday, 1967; Pierrehumbert & Hirschberg, 1990). In Steedman's (2000) Combinatory Categorial Grammar (CCG), theme and rheme are defined as the basic elements of information structure. Steedman distinguishes the shared topic between interlocutors as the theme and the new information introduced into the dialogue as the rheme. Theme and rheme can next be divided into focus and background. The focus (or contrast) provides alternatives that distinguish the referent of a referring expression from the alternatives that the context affords. The background is everything else. The following exchange, taken from the multimodal communication corpus (Louwerse, et al., 2006; 2007), illustrates these concepts (Example 1). The IG starts speaking and the IF's reply is analyzed in terms of theme/rheme, background/focus.

Example 1

IG: then you're gonna- ok. Then you're gonna stop. OK and now you're gonna start curving down and when you go down, do you see a purple rectangular alien to the left?

IF: uh… is it right above a blue rectangular alien?

background focus background

theme rheme

Steedman (2003) made the claim that theme and rheme can be discriminated in terms of pitch accents, and adds that theme and rheme expose a particular intonation pattern dependent on the common grounds between the speakers. The common ground can vary for instance in function of the agreement between participants. Table 1 illustrates Steedman's proposal.

Table 1 Pitch Accent Patterns

	Agree	Disagree
Theme	L+H*	L*+H
Rheme	H* or (H*+L)	L* or (H+L*)

L, H, H*, L* are the transcription conventions for intonation and prosody as described in Pierrehumbert (1990). "H" and "L" represent "high" and "low" tone, and "*"

denotes that the tone is aligned with a stressed syllable. "+" is a "followed-by" notation. As for the interplay between focus and background, Steedman's prediction is that focus is marked by prominence in pitch compared to the background and is also emphasized. On the other hand, background is usually unaccented and can even be omitted entirely from conversations. In other words, the theme/rheme partitioning determines the overall intonation pattern, whereas the focus/background partitioning determines the placement of pitch accents.

Despite the fact that there are a number of studies making the link between intonation and information structure at a theoretical level, there is relatively little research that has investigated this link empirically using naturally occurring speech outside of an experimental setting. An exception is Calhoun (2006) who conducted a series of production and perception experiments, showing that differences in pitch mapped onto differences in theme and rheme and extended this conclusion with evidence from corpus linguistic data using the Switchboard corpus. More specifically, Calhoun showed focus is signaled through the alignment of words with prosodic structure.

For the purpose of the current paper, we will however not discriminate between focus/background, because the utterances of interest are phrases which the focus is part of. Take for instance Example 1. Instead of saying "uh is it right above a blue rectangular alien?" the IF could say "blue rectangular alien" where "blue" and "rectangular sign" are similar to a theme/rheme pair.

After a manual inspection of a sample of conversations from the Multimodal Map Task corpus, Guhe, Steedman, Bard and Louwerse (2006) observed that, on average, rheme has a higher pitch than the theme (see Example 2).

Example 2
IG: OK. Do you have a black triangular sign?
IF: No, I have a *red* triangular sign

In this example the common ground between the speakers is confined to a *triangular sign*, which is the theme of the dialogue. However, it happens that the speakers don't agree on its color *red*. Within the IF's utterance *red* is found to conceive a higher pitch than "*triangular sign*".

The current study extends Guhe et al.'s study by taking the same corpus, automatically segmenting the turns and words in the speech, and automatically identifying theme and rheme in the transcripts in order to test whether they differ in terms of prosody in natural face-to-face communication.

Turn Segmentation

Various spoken cues have been used over the years to segment turns, including pitch ranges, preceding pauses, speaking rate, amplitude and pitch contour (Brown, 1983; Grosz & Hirschberg, 1992; Swerts & Ostendorf, 1995).

In this paper, we have used pauses as the initial parameter to detect the beginning and end of a turn in a natural conversation. In the data collection, we used the Marantz PMD670 recorder which enables recording of speech of IG and IF on separate audio channels. Pauses were analyzed using the upper intensity limit and minimum duration of silences. In measurement of intensity, minimum pitch specifies the minimum periodicity frequency in any signal. In our case, 75 Hz for minimum pitch yielded a sharp contour for the intensity. Audio segments with intensity values less than its mean intensity were classified as pauses. We thereby used mean intensity for each channel rather than a pre-set threshold. This enabled our pause detection system to properly adapt to the diverse set of voice properties of the participants. Any audio segment with silences more than .4 second was denoted as pauses. However, the extracted turns were manually inspected to account for different kinds of pauses in the speech signal (e.g. hesitations vs. end of turn). The speech processing software *Praat* (Boersma & Weenink, 2006) was used to perform all calculations to identify these pause regions.

The pause detection algorithm was used separately on the right and left channels of each audio file to detect time-stamp information of turns for both IG and IF. Two audio channels contain separate information for IG and IF, respectively. Using the pause detection algorithm, the beginning and ending time of each turn for both IG and IF are stored separately. Later, the time stamp information for both IG and IF are merged into one file to potentially detect and discard segments where two participants speak at the same time (overlapping speech being difficult to analyze). Examples are given below (Case 1 and 2).

Case 1 depicts the ideal cases where one of the participants is silent while other participant is speaking. Case 2 introduces the challenge of segmenting a conversation as two people speak at the same time. Due to a few cases of both of the participants speaking at the same time, it was not possible to attain 100% accuracy in segmenting the audio files in turn level. The chosen audio files, each containing a little more than 80 turns, were processed using the proposed turn detection framework based on pauses. For each audio file, our system was able to map a turn as defined in the transcript into the corresponding audio segment more than 90% of the time with an average of 93% accuracy rate for all the speech files.

Case 1:
IF: is it right above…….a blue rectangular alien

Start of a turn *pause* *continuation of the turn*

IG: …… (pause)……...

Segmented turn:
IF: *is it right above (...) a blue rectangular alien*

Case 2:

IG: Go right……...okay…..........then.. go straight

turn1 | *pause* | *turn2* | *pause* | *turn3*

IF: ……..okay……ummm…okay…

pause |*turn1*| *pause* | *noisy data*

Segmented turns:
IG: *Go right*
IF: *Okay*
IG: *Okay (...) then. go straight.*

Word Segmentation

In order to segment the words of each turn, the Lumenvox's (www.lumenvox.com) Speech Recognition Engine was used, a flexible API that performs speech recognition on audio data from any audio source. One of the strengths of the Lumenvox system is that it is speaker-independent. Spontaneous speech can thus be segmented and recognized based on an acoustic model and a language model. The system provides an API to identify the starting and ending time for every recognized speech unit in the output. This suggests that we have the necessary information to identify the starting and ending times of the leaf node (i.e. the word) of the parse tree induced from the grammar.

A significantly small number of words are used in IF's turns. Indeed, 70% of these turns only contained less than three words, which is a reasonable representation of the conversational nature of the Map Task corpus, given the fact that IFs are generally waiting for instructions and acknowledging the information (Louwerse & Crossley, 2006). The IGs, on the other hand, used longer sentences with around 50% of the IG's turns consisting of more than 10 words. The Lumenvox ASR performs well on shorter streams of speech, but like any other ASR systems, lacks in performance on longer streams of speech, even when the verbatim transcript is available and used for the only purpose of "recognizing" this specific stream. Average performance for turns with more than 10 words was low at 18.51% accuracy and satisfactory for turns ranging between 1-9 words 67.2%. IF turns typically fell in the latter range.

Contrast Marking

As described earlier, Guhe et al. (2006) observed that theme and rheme can be distinguished by their pitch features with which the corresponding words are realized. Guhe et al. therefore predicted that rheme has a significantly higher pitch than theme. A small sample of turns marking contrast confirms this prediction. In the current study we are using an automated approach to extract contrasts from the multimodal Map Task transcripts, and the segmentation techniques proposed above are thereby used to help identify the speech units from the corpora. Contrastive cases were selected using the following algorithm:

1) Adjacent IG and IF turn pairs were selected.
2) Two windows of size *N* were chosen in the turns (see Example 3).
3) These two windows shifted for the two whole turns. Within the window it was determined whether there was a match of *N*-1 words. If this was the case, the pair was considered as a potential theme/rheme pair.

Example 3

IG: We're drawing parallel to the bottom of the page again almost. Uh. there are three purple bugs.

IF: I see three white bugs.

In the current experiment, we set the window size to N=3. In Example 3, the two turns will be chosen because "*three purple bugs*" and "*three white bugs*" have two words in common (i.e. *N*-1=2). This algorithm narrowed down the 25,000 turns of the 258 conversations to 458 turns, all of which were potential candidates for theme-rheme pairs.

In order to precisely derive the pitch information, we needed to filter out noise in the speech data. The pitch for human vocals typically ranges from 100 Hz to 150 Hz for men, and from 170 Hz to 220 Hz for women. We conservatively filtered out the sound information outside the [75-300] range that was caused by noise or non-speech related sounds.

Results

The average of pitch across four different types of speech segments was computed: 1) the rheme 2) the head of the phrase that formed the theme (e.g. the head of the NP), 3) the phrase itself (e.g. the NP) and the 4) whole turn containing the theme/rheme pair. In Example 3, the rheme is *white*, the head word of the theme phrase is *bugs*, the theme phrase is *three bugs* and the whole turn is *I see three white bugs*. Following Guhe, et al. (2006) we predicted the average pitch for rheme to be higher than the average pitch computed on the theme segments (head, phrase and turn).

Table 4 presents the results of the analysis. All pitch information showed the expected patterns with the pitch for rheme being higher than the pitch for theme. The difference did not reach significance at the turn level, reached marginal significance in a one-tailed test at the head level (t (45) = 1.42, p = .08) and significance at the phrase level (t (45) = 1.81, p = .04).

Table 4: Mean and SD of pitch of theme and rheme

	Theme		Rheme	
	Mean	SD	Mean	SD
Head	158.02	48.67	165.58	43.75
Phrase	156.80	44.53		
Turn	163.44	39.61		

Conclusion

The current study has explored the relation between information and intonation structure. Several studies have investigated this relation, but few have done this empirically using natural face-to-face conversations. We have used a large corpus of face-to-face conversations on a map navigation task. In this task dialogue partners sometimes do and sometimes do not have common ground, depending on the differences between their maps. The corpus is therefore ideal to investigate differences between given (theme) and new information (rheme).

We presented a technique of automated speech segmentation and transcript time stamping and applied this technique to determine prosodic differences in information structure. Confirming the argument made in a number of theoretical studies the results show that the average pitch of the rheme in a turn is significantly higher than the average pitch of the phrasal theme of that turn in natural face-to-fae communication.

Automated sentence segmentation based on pauses and word segmentation based on automatic speech recognition techniques were employed to help mine the prosodic features of contrasts. Future work includes how to improve the speech segmentation techniques. A proposed method is that, instead of using turns for word segmentation, dialogues acts are used. This would boost the performance of the word segmentation, but would on the other hand, put an extra burden on the dialogue act segmentation. An alternative possibility is to adopt a recent sentence segmentation tool, nailon (Edlund & Heldner, 2006), which segments continuous streams of speech based on the fusion of prosodic features such as pauses, duration of voicing, intensity, pitch, pseudo-syllable durations, and intonation patterns.

The current study focused on the two linguistic modalities of information and intonation structure. Louwerse et al. (2007) provided insight into how eye gaze, facial movements, speech features, map drawings, and dialogue structures correlate with each other and which dialogue acts best predict the expression of a particular modality. Evidence of a mapping between linguistic modalities as well as between non-linguistic modalities is emerging, however, the exact nature of the alignment and whether these modalities add or substitute information remains an open research question.

Acknowledgments

This research was supported by grant NSF-IIS-0416128. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution. We would like to thank Ellen Bard, Art Graesser, Markus Guhe and Mark Steedman for their help on this project and Nick Benesh, Gwyneth Lewis, Divya Vargheese, Shinobu Watanabe and Megan Zirnstein for their help in the data collection and analyses.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, et al. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*, 351-366.

Boersma, P., & Weenink, D. (2006). *Praat: Doing phonetics by computer* (Version 4.4.06) [Computer program]. Retrieved January 30, 2006, from http://www.praat.org/

Brown, G. (1983). Prosodic structures and the Given/New distinction. In D. R. Ladd & A. Cutler (Eds.), *Prosody: Models and measurements* (pp. 67–77). Berlin: Springer

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic research*, *4*, 75-87.

Calhoun, S. (2006). Information structure and the prosodic structure of English: A probabilistic relationship. PhD Dissertation, University of Edinburgh.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., & Stone, M. (1994) Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. *Proceedings of SIGGRAPH '94*, 413-420.

Cassell, J., & Thórisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, *13*, 519-538.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Cole, R. A., et al. (1997). *Survey of the state of the art in human language technology*. New York, NY, USA: Cambridge University Press.

Demuynck, K., & Laureys, T. (2002). A comparison of different approaches to automatic speech segmentation. In P. Sojka, I. Kopecek, & K. Pala (Eds.), *Proceedings of the 5th International Conference on Text, Speech and Dialogue (TSD 2002)* (pp. 277-284). New York: Springer.

Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S., & Bruce, V. (1997). Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied, 3*, 105-125.

Edlund, J., & Heldner, M. (2006): /nailon/ - software for online analysis of prosody. In *Proceedings of Interspeech 2006 ICSLP* . Pittsburgh, PA, USA.

Graesser, A.C., D'Mello, S.K., Craig, S.D., Witherspoon, A., Sullins, J., McDaniel, B., & Gholson, B. (in press). The relationship between affect states and dialogue patterns during interactions with AutoTutor. *Journal of Interactive Learning Research.*

Grosz, B. & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing.* ICSLP.

Guhe, M., Steedman, M., Bard, E. G., & Louwerse, M. M. (2006). Prosodic marking of contrasts in information structure. In *Proceedings of BranDial 2006: The 10th Workshop on the Semantics and Pragmatics of Dialogue*,

University of Potsdam, Germany; September 11th-13th 2006.

Halliday, M.A.K. (1967). *Intonation and grammar in British English*. The Hague: Mouton.

Hirschberg, J., & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual meeting* (pp. 286-293). Morristown, NJ, USA: Association for Computational Linguistics.

Jurafsky D., & Martin J. (2000). *Speech and language processing. An introduction to natural language processing, computational linguistics and speech recognition*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings of the International Conference on Advanced Learning Technologies (ICALT 2001)*, Madison Wisconsin, August 2001.

Louwerse, M. M, Jeuniaux, P., Hoque, M. E., Wu, J., Lewis, G. (2006). Multimodal Communication in Computer-Mediated Map Task Scenarios. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1717-1722). Mahwah, NJ: Erlbaum.

Louwerse, M.M., Benesh, N., Hoque, M.E., Jeuniaux, P., Lewis, G. , Wu, J., & Zirnstein, M. (2007). Multimodal communication in face-to-face conversations. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1235-1240). Mahwah, NJ: Erlbaum.

Louwerse, M.M. & Crossley, S.A. (2006). Dialog act classification using n-gram algorithms. In *Proceedings of the 19th International Florida Artificial Intelligence Research Society.*

Louwerse, M.M., Graesser, A.C., Lu, S., & Mitchell, H.H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology, 19*, 1-12.

Massaro, D. W., & Cohen, M. M. (1994). Visual, orthographic, phonological, and lexical influences in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1107- 1128.

Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan & M. Pollarck (Eds.). *Intentions in Communication* (pp. 271-311). Cambridge, MA: MIT Press.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Steedman, M. (2000). *The syntactic process*. Cambridge, MA: MIT Press.

Steedman, M. (2003). Information-Structural Semantics for English Intonation. *LSA Summer Institute Workshop on Topic and Focus*. Santa Barbara, July 2001.

Steedman, M., & Kruijff-Korbayová, I. (2001). Introduction: Two dimensions of information structure in relation to discourse structure and discourse semantics. In I. Kruijff-Korbayová & Steedman (Eds.), *Proceedings of the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*, (pp. 1-6).

Swerts, M., & Ostendorf, M. (1995). Discourse prosody in human-machine interactions. In *Proceedings of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems - Theories and Applications.*

Woodbury, A. C. (1987). Rhetorical structure in a central Alaskan Yupik Eskimo traditional narrative. In J. Sherzer & A. Woodbury (Eds.) *Native American Discourse: poetics and rhetoric* (pp. 176-239). Cambridge, UK: Cambridge University Press.