INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Measuring Turn-Taking Offsets in Human-Human Dialogues

Rebecca Lunsford[1], *Peter A Heeman*[1], *Emma Rennie*[1,2]

[1]Center for Spoken Language Understanding, OHSU, Portland OR, USA
[2]Reed College, Portland OR, USA

lunsforr@ohsu.edu, heemanp@ohsu.edu

Abstract

This paper examines the pauses, gaps and overlaps associated with turn-taking in order to better understand how people engage in this activity, which should lead to more natural and effective spoken dialogue systems. This paper makes three advances in studying these durations. First, we take into account the type of turn-taking event, carefully treating interruptions, dual starts, and delayed backchannels, as these can make it appear that turn-taking is more disorderly than it really is. Second, we do not view turn-transitions in isolation, but consider turn-transitions and turn-continuations together, as equal alternatives of what could have occurred. Third, we use the distributions of turn-transition and turn-continuation offsets (gaps, overlaps, and pauses) to shed light on the extent to which turn-taking is negotiated by the two conversants versus controlled by the current speaker.

Index Terms: turn-taking, overlapping speech, turn-fights

1. Introduction

As we start employing spoken dialogue systems for more and more complex tasks, it will be essential that the system and user can naturally and efficiently work together. An important component of this is how turn-taking operates, as this regulates how each conversant can contribute to a conversation. By better understanding how turn-taking works in human-human conversations, we will have better guiding principles for how to build spoken dialogue systems (SDS).

There is a large body of work exploring how quickly speakers take the turn. In much of that work an odd phenomenon has been observed, that of a large number of lengthy overlaps [10]. However, recent work by Levinson et al. [13] suggests that these excessive overlaps might be due to how overlaps are identified and measured. Thus, we explore whether a revised measure, one that better accounts for speaker intent, results in offsets that better reflect the generally orderly nature of conversation.

A second motivation for this work is to explore the underlying mechanisms that are used in turn-taking. Many SDSs assume a rigid model of turn-taking, where one person keeps the turn until they decide to release it, which we refer to as the *speaker-control* model. This approach stems from the work of Sacks et al. [14], in which they proposed that the current speaker decides when to allow someone else to take the turn. However, there is evidence that human turn-taking is more flexible than the speaker-control model. For example, Duncan et al. [4] proposed that people bid for the turn. One of the cues could be how quickly each conversant starts to speak [15]. To explore this negotiative view of turn-taking, we will examine turn-transitions together with turn-continuations, as equal alternatives.

This work was funded by the National Science Foundation under grant IIS-1321146. The third author was funded with an REU supplement (Research Experience for Undergraduates).

2. Related Work

For the analysis of turn offsets, a researcher must first decide how to manage segmentation, annotation, and measurement. Here we discuss how these decisions were made in previous work, and will discuss the decisions made in support of our work in Sections 3, 4, and 6.

Segmentation: To analyze pauses and gaps, the speech must first be segmented. This can be performed either algorithmically or manually. Algorithmic segmentation is done using a speech analysis tool to identify silent regions. Examples include the work of Heldner et al. [10, 11] and Kane [12]. To avoid over-segmenting a given speaker's utterance, a minimum pause length is set, typically 200 ms or less, and speech surrounding this small pause is bridged and treated as one utterance. These silence-based approaches to segmenting speech have the advantage of ease, especially as compared to manually segmented speech, but lack insight into speaker behavior or intent.

Alternatively, the speech can be manually segmented. Examples of this include the segmentation performed on Switchboard prior to annotation using DAMSL [1, 3], or the *functional segments* described by Geertzen et al. [5] and in the ISO 24617-2 standard [2]. In these segmentation schemes, the speech is segmented in preparation for the annotation of dialogue acts (DAs). Thus silences are less salient, except in that they may inform the segmenter of the potential completion of the act. These DA-based approaches capture aspects of the speaker's behavior missed in silence-based segmentation, but may not capture locations where the turn might have changed, but did not.

Annotation: Once segmented, annotating the segment's function in the conversation can provide further insights into turn-taking. Although it is possible to annotate a conversation that has been algorithmically segmented, this is unusual. Thus here we will focus on turn-taking related annotation of manually segmented conversations.

In the DAMSL [1, 3] annotation scheme, utterances that are functionally related to a previous utterance are annotated as having a *backward function* – these include responses such as answers, agreement, and back-channels. However, as the annotation did not include an identifier of the related previous utterance, it can be unclear to which utterance the response was directed. The ISO 24617-2 standard [2] addresses this issue by specifying that *functional and feedback dependencies* should be explicitly identified and indicate to which previous segment the current segment relates.

Measurement: Although it may seem clear how one would measure offsets, differing protocols exist – especially in regard to the classification and measurement of overlaps. In work exploring the duration of pauses, gaps, and overlaps, Heldner et al. [10] included in their analyses overlaps that occurred during speaker transitions, but excluded within-speaker over-

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-1350

laps. In a later work focused on the timing of speaker transitions around *very short utterances*, Heldner et al. [11] included within-speaker overlaps, but measured the overlap as the time from the end of the overlapped speech to the start of the overlapping speech. This measure makes sense if the overlapping speech was a response to the overlapped speech, but misrepresents the overlap in those cases in which the overlapping speech was a response to *earlier* speech.

3. Segmenting into Turn-Taking Units

For this study, we are interested in not only when turn-taking occurred, but also when it could have occurred. This might not be the same as Sacks' definition [14], as we interpret his definition as points where the speaker intends that turn-taking might occur. We feel that there might be points where the current speaker wants to keep talking, but where the listener could speak without it being viewed as rude. It is also possible that there may be points where the speaker has no intentions regarding the turn – essentially points where the speaker has completed his contribution, but it is unclear who should speak next.

Towards this end, we define "turn-interpretable points as any location where a listener might start speaking without appearing to interrupt the current speaker. The specific guidelines for segmentation are as follows:

1. Observe the speech of only one speaker. This should help prevent the annotator from being influenced by whether a turn-transition actually occurred.
2. Segment at every location where there is syntactic, semantic, and intonational completeness (SSI). For those cases where there is continued speech sounds or breath noise after the SSI, mark the following segment with NS (no silence).
3. Take into account only the speech, and any subsequent silence, thus far. This allows for situations where the speech has not achieved SSI but, by remaining silent, the speaker appears to have released the speaking floor.

Our segmentation scheme differs from silence-based segmentation in that we allow within-speaker segmentation based on SSI even if there is no silence, and also allow within-segment silences if it would have seemed an interruption for the listener to start speaking. We differ from dialogue-act segmentation in that we allow for potential turn-transitions even in the middle of a to-be-completed dialogue act.

4. Annotation Scheme

For this work, the primary purpose of our annotation scheme is to clarify any utterance relationships relevant to interpreting turn-taking. Thus, as specified in the ISO 24617-2 standard [2], we identify *functional and feedback dependencies*, but limit ourselves to those that clarify which segment a speaker is responding to. Toward this end, we used DialogueView to annotate the files [17], specifically noting:

Self Talk: Speech that is low-volume and clearly not intended to be part of the dialogue. These segments were excluded from our analyses.

Interrupts: A segment that sounded, in isolation, like the speaker intended to interrupt their interlocutor. Whether the other speaker was in fact interrupted or whether the segment overlapped the preceding speech was not taken into account.

Back-channels: e.g., 'uh-huh', 'yeah', 'okay', etc. These are additionally linked to the segment being acknowledged.

Dual-starts: Pairs of segments in which both interlocutors commenced speaker near simultaneously, without the speakers being aware that the other is speaking, or appearing to interrupt. Dual-starts were cross-linked, with each segment specifying the other.

5. Corpora

Our data includes three corpora: Trains [7], MTD [18], and Switchboard [6]. Trains and MTD are corpora of task-oriented conversations, and Switchboard is conversational speech. In MTD and Switchboard, the two conversants play identical roles. In Trains, however, one person plays the role of the 'user' who has a goal to solve, and the other plays the role of the system, who knows the domain information, but is explicitly told to let the user drive the conversation.

6. Measuring Offsets

For simplicity, we use a single term to describe the amount of time between segments: *offsets*. This includes both when there is a turn-transition and turn-continuation. When there is a turn-transition, if there is overlap between the segments, the offset will be negative (often called a *overlap*); otherwise the offset will be positive (often called a *gap*). When there is a turn-continuation, the offset will be at least 0 (often called a *pause*).

Because of overlapping speech, it is sometimes unclear between which segments to measure the offsets. We first start with a simple definition based strictly on the temporal relationship between segments, similar to that used by Heldner et al. [11]. However, because we segment at locations where the turn might have changed – rather than using silence thresholds – the offset calculation used here can include within-speaker offsets less than 180ms, which are explicitly disallowed in Heldner et al. [11]. We will then refine our offset measurements in the following sections.

We define offsets in terms of each segment and its *predecessor* segment. Let u be a segment, and A its speaker. Let u' be the segment by A that precedes u. Let's refer to the other speaker as B and let w be the last segment of B that starts before u starts. Whichever of u' and w ended last is the predecessor of u. Note that every segment (except the initial segment) has a predecessor; however, not all segments have to be a predecessor. For example, a segment entirely embedded inside of another segment will not be a predecessor.

With this definition of a predecessor, we can now define offsets. For segment u with predecessor p, the offset before u is the time from the end of p to the beginning of u. Furthermore, we say u is a turn-transition if u and p are uttered by different speakers, otherwise it is a turn-continuation. In Figure 1, we show an example with segments by two speakers; A and B. Arrows show the relation between each segment and its predecessor. The predecessor of u_2 is u_1, the predecessor of u_3 is u_2, the predecessor of u_5 and of u_6 is u_4 – these are turn-transitions. The predecessor of u_4 is u_3 – this is a turn-continuation. The length of the arrow is the offset, with arrows pointing right being negative amounts. Using the strict temporal definition, u_5 has a large negative offset.

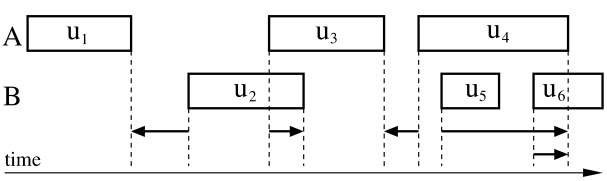

Figure 1: Examples of Offsets

Comparing Switches and Continues: Following Heldner and Edlund [10], we compute a histogram of the delays for turn-transitions, shown in Fig. 2. We used a bin size of 0.1s. A key innovation of this paper is that we contrast the distribution of offsets for turn-transitions (switches) with turn-continuations (continues). As we use do not use silence-based segmentation for determining within-turn segments, delays for turn-continuations start at 0s.

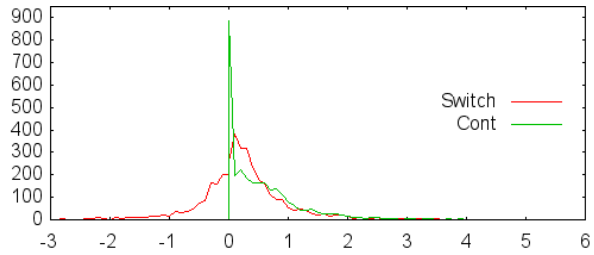

Figure 2: Histogram of delays for turn-transitions

A problem with the histogram, is that it is difficult to determine how many switches versus continues happen within a given amount of time. Hence, in the rest of the paper, we use cumulative distribution curves. Figure 3 shows this curve in terms of the actual number of switches and continues that happened by each time point. For example, we see that the number of switches with negative offsets is just slightly more than the number of continues with 0 offsets. The graph also shows the relative proportion of each and that there are more switches than continues (3573 switches versus 2904 continues).

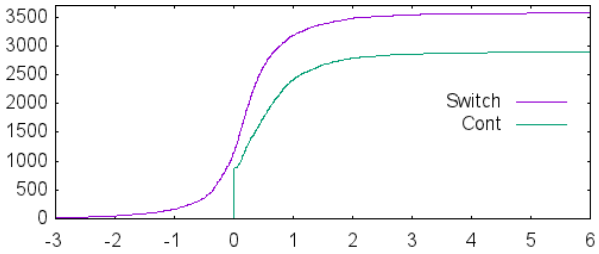

Figure 3: Cumulative distribution

Interrupts: Included in the switches in Fig. 3 are 100 segments that were coded as interrupts. Interrupts are sometimes used to quickly correct a flaw in the other's view of the world or reasoning about how to solve a task [16]. So, we should not expect an orderly progression in turn-taking for them. As computed in Figure 3, their offset is calculated from the end of the speech being interrupted to the start of the interrupt, as shown in Figure 4. These offsets have a median of -0.51s and a mean of -0.68s. This means that when someone interrupts, the interrupted speaker stopped within that amount of time. In our subsequent analyses of switch offsets, interrupts are not included.

Dual-Starts: Dual-starts are also problematic using the simple approach to offsets for switches and continues. A dual-start is where two segments overlap but, unlike an interruption, neither speaker seems to be aware of the other one. Let's refer to the two parts of a dual-start as u_1 and u_2, where u_1 started before u_2, as shown in Fig. 5. Depending on what preceded u_1 and u_2, u_1 will be viewed as a continue or a switch, and its offset computed with respect to the preceding segment. However, u_2 will be always computed as a switch, and its offset will be from the end of u_1 to the start of u_2, as u_1 will have overlapped

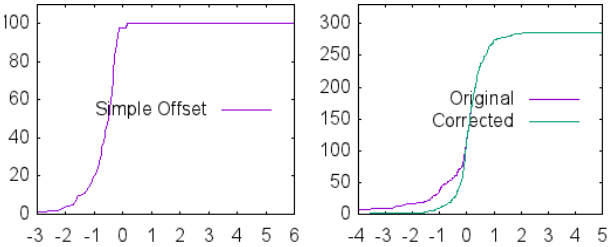

Figure 4: Interrupts (left) and Delayed Back-Channels (right)

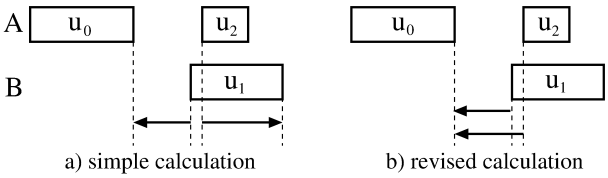

Figure 5: Calculation of offsets for second part of dual-starts

u_2, a negative amount. Figure 5a shows how the offsets are computed, and Fig. 6a shows the distribution of the offsets for the first and second parts.

As argued by Fang and Heeman, dual-starts are probably unintentional collisions. So, the second speaker, in starting to say the second part, was not influenced by the first part. Thus, the offset for the second part should be computed relative to the same segment as the first part. Figure 5b shows how the offsets should be measured, and Fig. 6b shows the corrected offset distributions.

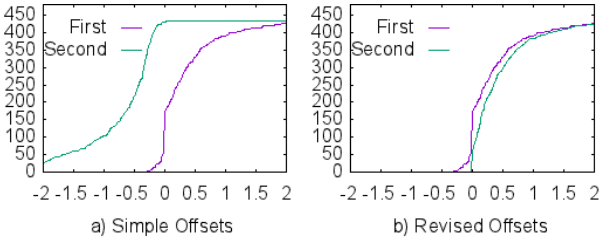

Figure 6: Dual-starts: (a) simple offset and (b) revised offset

In addition to correcting the offsets, this also corrected whether the second part is viewed as a switch or a continue. Before correction, the second part was always viewed as a switch, relative to the first part. Of the 433 dual-starts, 220 of the second parts were by the same conversant who last spoke before the dual-start, and so they were changed to continues.

Delayed Back-Channels: Back-channels are very common in dialogue, where the listener gives the speaker a signal that they are understanding so far, but where both know that the speaker is not finished. Usually, these overlap with the end of the utterance from the speaker, but they might be delayed, and in fact start after the speaker has started their next contribution. These are not coded as a dual-start as back-channels are commonly viewed as not even taking the turn, and often overlap the speech of the other person.

When conversant B makes a back-channel to respond to something that conversant A had said, sometimes A starts speaking before B starts making their back-channel. Using the simple definition of offsets, switches, and continues, the back-channel would then be viewed as having a large overlap with A's second segment, just as dual-starts have with the simple definition. Hence, for delayed back-channels, we compute their

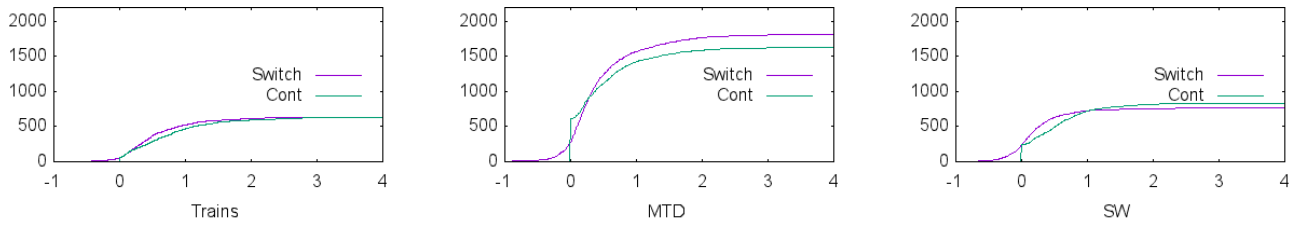

Figure 7: Comparison of two task-based corpora, Trains and MTD, and one conversational corpus, Switchboard (SW)

offset relative to the segment they are responding to, which we annotated. Figure 4 shows the offsets for all back-channels, both the original offsets from the simple definition, and our corrected offsets. Of the 285 back-channels, we corrected the offsets of 62 of them. Of those, 7 changed from switch to continue; these were cases where a speaker made two back-channels in a row, where the second one happened to overlap the beginning of a segment by the other speaker.

In a few instances, there are several back-channels in a row by the same speaker that have been split into different segments, that are both responding to the same previous utterance. In this case, the second back-channel is coded relative to the first back-channel, as a continue.

Revised Offsets: We now show the result of adjusting the offsets, continues and switches to remove interrupts, and correct dual-starts and delayed back-channels. In revising the offsets, we removed 100 switches, as these were interrupts. We also changed 220 switches to continues with cleaning up dual-starts, and 7 switches to continues for delayed back-channels.

The resulting offset distributions are shown in Fig. 8. The curves show the number of potential turn transitions resolved by each time and whether they resulted in a switch or continue. Table 1 gives descriptive statistics about the offset distributions for both the original simple definition of offsets and our revised definition. The rows marked with '%' show the offset length for which that percentage of the data has a smaller offset.

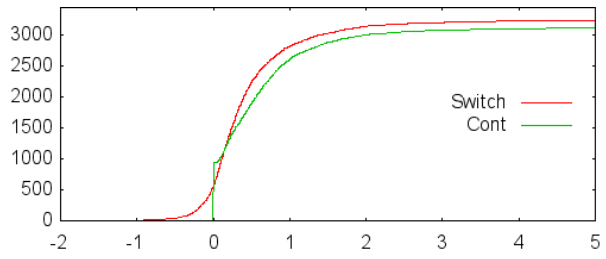

Figure 8: Revised Offsets

Comparing Fig. 8 to Fig. 3, we see that the revised switch curve has far fewer points that start before 0, and this is reflected in the positively shifted offset values in Table 1. In fact, looking at the 'Switches' in the 5% row, we see that the offset has shifted from -0.94s to -0.23s, so there are far fewer overlaps than others have found (cf., [10]), and most of the ones that do exist would not be perceptible (i.e., greater than 120ms [9]). Comparing the 'Continue' columns in Table 1, we see effectively no change due to the revised offset calculations or the increased number of 'Continue's, suggesting that the offsets that were reclassified as 'Continues' fit well within this distribution. Lastly, we see that half of our switches last at least 0.27s, and 25% are at least 0.62s; this is much longer than what other researchers have found (cf. [10]).

Comparing Corpora: Even though the corpora differ in terms of the speakers' task, all have similar distributions for switches and continues as shown in Figure 7. However, they do differ. MTD has more continues with no intervening silence. Switchboard has longer offsets. Trains has a similar number of switches and continues, MTD has more switches, and SW has more continues. Long offsets for Switchboard usually result in a continue. We will further explore these differences in future work.

	Original		Revised	
	Switch	Continue	Switch	Continue
Number	3573	2904	3246	3131
Mean	0.23	0.55	0.44	0.55
5%	-0.94	0.00	-0.23	0.00
25%	-0.11	0.00	0.07	0.00
Median	0.19	0.33	0.27	0.33
75%	0.53	0.79	0.62	0.78
95%	1.60	1.85	1.69	1.84

Table 1: Statistics about Offsets for Switches and Continues

7. Discussion

In this work, we revisited how turn-taking offsets are measured, first using a simple temporal approach to computing offsets, then refining this simple model to better account for interrupts, dual-starts, and delayed back-channels. We showed that the simple model produces an inaccurate picture, particularly in regard to the duration of overlapped speech during turn-transitions and to the number of turn-transitions vs turn-continuations. With the revised offset durations, we see that many turn-transitions take longer than previously thought, with 50% of all turn-transitions lasting at least 0.27s, and 25% at least 0.62s.

In addition, we treated turn-transitions and turn-continuations as equal alternatives. We found that the distribution of turn-taking offsets are quite similar between the two, differing primarily in the number of negative offsets – an unsurprising result given that a turn-continuation cannot overlap. This similarity in the time it takes for a turn-transition versus a turn-continue suggests that negotiation might play a role in turn-taking, rather than solely being governed by a speaker-control model, as we have argued elsewhere [8].

Given the more leisurely pace of offsets found here, it seems likely that, instead of aiming to achieve no gaps and no overlaps, speakers aim to contribute to the dialogue as quickly as they can reasonably do so. In some cases the next speech may come quickly, perhaps because the response is a simple backchannel or the speaker is continuing on. In other cases the speech may take a while, perhaps because no one clearly has the speaking floor or the intended speaker is deep in thought. Viewed this way, our findings further suggest that some turn-transitions might be negotiated, with the speaker most capable of advancing the dialogue at that moment taking the floor [15].

8. References

[1] J. F. Allen and M. G. Core, "Draft of DAMSL: Dialog Act Markup in Several Layers." Tech. Rep., Mar. 1997.

[2] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. Traum, "ISO 24617-2: A semantically-based standard for dialogue annotation," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

[3] M. G. Core and J. F. Allen, "Coding Dialogs with the DAMSL Annotation Scheme," 1997.

[4] S. J. Duncan and G. Niederehe, "On signalling that it's your turn to speak," *Journal of Experimental Social Psychology*, vol. 10, pp. 234–247, 1974.

[5] J. Geertzen, V. Petukhova, and H. Bunt, "A multidimensional approach to utterance segmentation and dialogue act classification," in *SIGDial*, 2007, pp. 140–149.

[6] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, 1992, pp. 517–520.

[7] P. A. Heeman and J. F. Allen, "The Trains spoken dialog corpus," Linguistics Data Consortium, CD-ROM, April 1995.

[8] P. A. Heeman and R. Lunsford, "Can overhearers predict who will speak next?" in *Proceedings of the AAAI Spring Symposium on Turn-Taking and Coordination in Human-Human Interaction*, Stanford, March 2015, pp. 30–35.

[9] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps." *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 508–513, Jul. 2011. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/21786916

[10] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[11] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski, "Very short utterances and timing in turn-taking," in *Interspeech*, 2011, pp. 2848–2851.

[12] J. Kane, I. Yanushevskaya, C. de Looze, B. Vaughan, and A. N. Chasaide, "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," in *Interspeech*, 2014.

[13] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language." *Frontiers in psychology*, vol. 6, 2015.

[14] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, Dec. 1974.

[15] E. O. Selfridge and P. A. Heeman, "Importance-driven turn-bidding for spoken dialogue systems," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala Sweden, Jul. 2010, pp. 177–185.

[16] M. Walker and S. Whittaker, "Mixed initiative in dialogue: An investigation into discourse segmentation," in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990, pp. 70–78.

[17] F. Yang, P. A. Heeman, K. Hollingshead, and S. E. Strayer, "DialogueView: annotating dialogues in multiple views with abstraction," *Natural Language Engineering*, vol. 14, no. 1, pp. 3–32, Jan. 2008.

[18] F. Yang, P. A. Heeman, and A. L. Kun, "An investigation of interruptions and resumptions in multi-tasking dialogues," *Computational Linguistics*, vol. 37, no. 1, pp. 75–104, Mar. 2011.