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

The Discourse Marker "so" in Turn-taking and Turn-releasing Behavior

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

[1]Reed College, Portland, OR, USA

erennie23@gmail.com

[2]Center for Spoken Language Understanding
Oregon Health and Science University, Portland, OR, USA

lunsforr@ohsu.edu

Abstract

Although *so* is a recognized discourse marker, little work has explored its uses in turn-taking, especially when it is not followed by additional speech. In this paper we explore the use of the discourse marker *so* as it pertains to turn-taking and turn-releasing. Specifically, we compare the duration and intensity of *so* when used to take a turn, mid-utterance, and when releasing a turn. We found that durations of turn-retaining tokens are generally shorter than turn-releases; we also found that turn-retaining tokens tend to be lower in intensity than the following speech. These trends of turn-taking behavior alongside certain lexical and prosodic features may prove useful for the development of speech-recognition software.

Index Terms: turn-taking, discourse markers, prosody

1. Introduction

In the study of turn-taking in dialogue, it is necessary to examine the pragmatic and interactional functions of the speech that speakers produce at turn boundaries. Often, turn-boundary speech consists of words or phrases that have been termed *discourse markers*; this term has been applied to a large variety of lexical items that do not fall under a neat semantic or syntactic classification. Deborah Schiffrin provides a comprehensive analysis of discourse markers, defining them as "sequentially dependent elements which bracket units of talk" [1]. She discusses their roles in creating discourse coherence, and her model includes specific reference to the exchange structure – i.e., the turn-taking structure. The "units of talk" that are bracketed by discourse markers can include turn units; it is well worth examining the functions that discourse markers may serve in constructing the exchange of turns in dialogue.

Among the terms commonly classified as discourse markers are conjunctions such as *so* and *but*, which in their most standard use are typically preceded by an introductory clause and followed by a dependent clause. However, when these terms are used at turn boundaries, one of these clauses is not present within the same turn: at the beginning of a turn, there is no introductory or causal clause, and at the end of a turn there is no conclusive or resultative clause. In some cases, often in what this study terms a Pass, one of these discourse markers may stand alone as a complete turn and thus have neither an antecedent nor a conclusion present within the same unit of speech.

Most likely, these discourse markers have evolved beyond their original semantic function, and turn-boundary instances of them demonstrate their multifunctionality and convey unique purposes – perhaps merely for turn-taking. Close analysis of the intersections of lexical and prosodic features can elucidate the functions of turn-boundary speech, and may aid in the development of systems that can approximate human turn-taking behavior. This paper focuses on the discourse marker *so* and considers prosodic features that may provide insight into its various uses; notably, its duration and intensity are interpreted as cues for turn behavior.

2. Data and Methods

2.1. Dataset

The conversational data in this study is taken from a set of three corpora of recorded dialogue: Trains [2], MTD (Multi-Threaded Dialogue) [3], and Switchboard [4]. Each session has two speakers with audio-only (non-face-to-face) speech interaction; the three corpora provide three different situations of dialogue production. In the first two corpora, the speakers are presented with a task which they must cooperatively accomplish through verbal communication, while in Switchboard the speakers are merely given a topic for conversation.

In the Trains corpus [2], one of the two speakers in each session is called the "system" and the other the "user"; they each have a different set of information about the task they must accomplish. The user must create an itinerary for transporting cargo between certain cities with a time limit, and the system has information about the available tracks and logistical limitations that must be observed. Thus, the two speakers must communicate in order to determine a solution.

The MTD [3] corpus consists of a more balanced cooperative task, in which the two speakers have equal roles and each has access to parallel information. Both "players" are using a computer program that simulates a hand of cards, and they take turns discarding and drawing cards with the goal of collaboratively completing a poker hand (e.g. a straight or a full house) between the two hands of cards. Since each player only sees half of the cards in play, they must communicate which cards are in their hands to each other at all times and cooperatively decide which cards to discard to improve their chances of winning. Meanwhile, the program intermittently interrupts the players with an additional task: colored shapes may appear on one player's screen, while the other player must ask a yes-or-no question about whether their partner has a particular colored shape.

http://dx.doi.org/10.21437/Interspeech.2016-547

The Switchboard [4] dialogues are not task-based; in these recordings, the two participants are speaking on the telephone and are given a particular topic to talk about for five or ten minutes. Since the speakers are not working to accomplish a task, these conversations tend to be more casual. In the particular subset of Switchboard dialogues used in this study, the prompt is to talk about their houses.

2.2. Segmentation and Annotation

These dialogues have been transcribed, segmented, and annotated for the purpose of turn-taking analysis. Speech-recognition software [5] was used to force-align the boundaries of each word of the transcript with the sound file, and the boundaries were cleaned up by hand as necessary using Praat [6], with careful attention paid to utterance boundaries. Speech is segmented into utterances based on the determination of **turn-interpretable points (TIPs)** using DialogueView [7]. These are defined as points of *potential* transfer of the conversational floor; in other words, the other speaker could reasonably begin speaking without it being construed as an interruption.[1] TIPs do not always result in a turn exchange; the same speaker may end up continuing or adding more speech to their contribution. TIPs divide speech into utterances based on the perception of syntactic, semantic, and/or intonational completeness – many prosodic cues affect whether a segment sounds complete or the speaker is likely to continue. TIPs were placed at points where an utterance sounded "complete" and the speaker perceptibly stopped producing sound. The criteria for determining TIPs recognize various exceptional cases, such as segments that are not syntactically complete but where the speaker cuts off abruptly with no apparent intention to continue.

The use of TIPs is intended to capture the various moments in dialogue that may be relevant to turn-taking behavior. Although these utterance boundaries do not always result in a turn exchange, they can shed light on the variety of cues that speakers use and hearers interpret to signal the interactive structure of dialogue. Of course, in the analysis, it is also useful to pay attention to the boundaries where another speaker did in fact take the turn. In the database, each utterance segment is marked either Continue or Switch depending on which speaker produced the following utterance. In addition to orderly turn exchanges and same-speaker add-ons, some TIPs result in Dual Starts (i.e. both speakers begin speaking at roughly the same time).

After segmentation, all utterances in the dialogues were annotated according to a schema designed to highlight basic dialogue acts, linked adjacency pairs, and notable characteristics of segments that relate to their roles in the exchange structure. An utterance may have several tags, each representing a different feature of its role; for example, a segment might be a Commissive, a Dual Start, and an Addon (meaning it is a syntactically dependent 'afterthought' related to the previous utterance, but with a TIP in between). Feedback and functional linkages (e.g. answers to questions) are marked with the identification of the previous utterance to which they are linked.

Among the classifications relevant to turn-taking behavior are Pass, which signifies a lack of commitment to claiming the floor, and Stall, which signals intent to hold the floor despite not knowing quite what to say yet. Segments with these tags often consist of filler words or discourse markers such as *so*; classification of these segments was based on perceptual and contextual assessments of their roles in the dialogue. Passes and Stalls are discussed in more depth in 3.5.

[1] This concept is adapted from Transition-Relevance Places as described by Sacks, Schegloff and Jefferson [8], but is based on somewhat different criteria.

2.3. Data Treatment

All data, including each utterance segment's speaker, start and end times, words, and annotation tags were stored in an XML database as a source for programmatic sorting and analysis. A script written in Tcl identified whether or not each utterance included the word *so* and in what position (utterance-initial, second, internal, and final) and whether an utterance initial *so* was also turn-initial (i.e., the last speech was produced by the other speaker).

For this study, non-discourse-marker uses of *so* (e.g. as a degree marker as in *it was so nice* or as a clause replacer as in *I hope so*) were excluded. Additionally, if the token of *so* was cut off or part of repaired speech, it was not included for analysis. Utterances that did not have at least 250 milliseconds of speech adjacent to *so* were excluded from the comparative intensity analysis.

Durations were measured for each position category using a Tcl script that output the difference between each *so* token's end time and start time. Intensities were measured using a Praat script that used the word-aligned TextGrids and each corresponding sound file to measure the maximum intensity of each *so* and of each *so*'s adjacent 250-millisecond segment (immediately following 250 ms if the *so* is utterance-initial or internal; immediately preceding 250 ms if the *so* is utterance-final).

3. Examples of *so* by utterance position

3.1. Utterance-initial

An utterance segment that begins with *so* may be a resultative or conclusive follow-up to a previous utterance by either speaker; Y's turn in (1) exemplifies this.

(1) **X: so** the quickest way to Corning is through Dansville which will take four hours
Y: so we'll get there at eleven a.m. (Trains 12.2)

However, this does not account for all utterance-initial occurrences of *so*. Some such utterances, like X's turn in (1), B's announcement in (2), and A's transition in (3), have no apparent relation to previous speech, and may in fact consist of a topic shift or an introduction of new information.

(2) **B: so** I just got a nine (MTD 5.2)

(3) **A:** don't worry about it just let him enjoy himself
A: so you think that you want to move away from the big city huh (Switchboard 2775)

In these cases, *so* may be classified as what Gravano et al. call a "Cue-Beginning" cue word [9]. The speakers use it as a turn-taking tool, perhaps in order to secure the attention of the listener.

3.2. Utterance-second

In some cases, a speaker begins a turn with a different cue word, such as *okay* or *alright*, immediately followed by *so*. We classified this phenomenon of utterance-second *so* as a unique category because it is not strictly initial nor does it fit in the "internal" category since it still lacks the antecedent clause. The presence of this occurrence perhaps demonstrates that turn-initial *so*

has a function beyond turn-taking, since in these cases a different word is fulfilling the "grabbing attention" function but the *so* still appears before the rest of the utterance's content.

(4) **A:** I have one heart
B: okay **so** we've got three hearts and then a eight nine ten jack right (MTD 4.2)

In (4), both *okay* and *so* may have functions beyond turn-taking: *okay* may be a signal of feedback acknowledging that B has received information from A's utterance, and *so* may be signaling the synthesis of this information with previously established information into a resultative conclusion. Second-position tokens of *so* such as this one demonstrate similar functionality to many utterance-initial tokens, but are counted as a separate category because some prosodic features may be unique to the utterance-initial position.

3.3. Utterance-internal

The utterance-internal use of *so* tends to reflect its classic resultative function as a connective between two clauses.

(5) **B:** oh now I have an eight **so** I have seven eight nine and ten (MTD 2.2)

Here, the speaker uses *so* to connect a new piece of information to a synthesized summary of information, demonstrating its causal linkage function.

3.4. Utterance-final

Occasionally, a speaker will "trail off" with a *so* tagged on at the end of an utterance, omitting the conclusion. Generally, the conclusion is mutually accessible, even obvious, to both speakers based on the context, and its ellipsis implies the most relevant assumption.

(6) **A:** I guess it's your turn to drop something **so**
B: oh (MTD 5.2)

(7) **X:** they usually get back at least two of them for the summer **so**
Y: you still need space (Switchboard 2691)

In (6), Speaker B's utterance of *oh* indicates understanding of A's implicit imperative (that B should take her turn in the game). This *oh* is a verbal indication of what Wilson & Sperber call a *positive cognitive effect* [10]. Similarly, Y's conclusion in (7) reflects comprehension of X's utterance within the context – they are discussing the use of rooms in their houses, and Y understands why it is relevant that X's kids spend the summer at home. Within Relevance Theory, *so* is functioning as an *ostensive stimulus*, a cue that draws the intended interpretation to the hearer's attention [10].

3.5. Standalones: passes and stalls

A perhaps surprisingly common phenomenon is the "standalone" *so* – an occurrence that is both preceded and followed by TIPs. Generally, these exemplify uses of *so* as a "filler word" and are annotated as either a Pass or a Stall.

A Pass is so called because it occurs when a speaker "passes up" or "passes along" the conversational floor. In other words, it is a brief turn that the speaker does not intend to keep for long. These usually occur after some period of silence or lapse in the dialogue, when neither speaker is contributing and neither is sure how to continue. Passes most commonly consist of a single word such as *so*, *but*, *um/uh*, or *well*, or a combination like *so um* or *but uh*. Speakers may employ the *so* Pass in particular when its causal implication is pragmatically relevant – that is, they may intend to prompt their partner to continue the sequential logic of the dialogue.

Stalls are much less commonly standalone, because they typically indicate that the speaker is taking the turn and wants to hold the floor. Occasionally, however, they may be followed by a TIP, but are distinguished as Stalls if they are perceived as being higher in volume and/or pitch than the quieter, lower Passes. A *so* Stall is thus similar to an utterance-initial *so* in that it may causally relate the upcoming utterance to previous speech but is also co-opted for turn-taking purposes. But unlike utterance-initials that introduce speech without a pause, Stalls are separated from the following speech and are often more emphasized or elongated to signal that the speaker wants the floor but is not quite ready to produce an utterance.

4. Results

The full database includes 23 recorded sessions – 9 from Trains, 7 from MTD, and 7 from Switchboard. The sessions vary in length: Trains ranges from 1.3 to 13.2 minutes, MTD from 13.6 to 15.9 minutes, and Switchboard from 4.6 to 10 minutes.

4.1. Duration

The average duration of the word *so*'s production will be compared across the *so*-position categories of utterances. The first hypothesis to be tested predicts that utterance-initial productions of *so* are generally shorter in duration than utterance-final or standalone occurrences, in order to achieve the speaker's goal of "grabbing" the turn and quickly transitioning to the substantial content of the turn. Additionally, Standalone Stalls are predicted to include the longest durations of *so*, since the speaker lengthens it to convey an intention to hold the floor while buying time to collect their words.

In the dataset, there are a total of 671 occurrences of *so* available for duration analyses. Table 1 shows the distribution across categories. Note that these categories are exclusive; that is, the standalones do not count as utterance-initial or -final. Also, standalones and utterance-finals that are tagged as cut-off (abandoned) speech have been excluded from analyses.

Table 1 also summarizes the average durations, in milliseconds, of the word *so* across each category of utterance. Here we see that, in general, turn-retaining *so* tokens are shorter in duration than turn-releasing tokens. Comparing the groups pairwise, we found no statistically significant differences between utterance-initial and utterance-2nd, and utterance-2nd and utterance-internal (independent two-tailed t-test, all p's$>$0.05, NS[2]). However, the utterance-internal productions were marginally shorter than utterance-initials (independent two-tailed t-test, df=495, p$<$0.09).

Similarly, we found no statistically significant difference between utterance-final and standalone-passes, or utterance-final and standalone-stalls (independent two-tailed t-test, all p's$>$0.05, NS), but did find a significant difference between standalone-stalls and standalone-passes (*a-priori* independent one-tailed t-test, p$<$0.05).

For further comparison, we collapsed the data into two groups: Group 1 (turn-retaining) includes utterance-initial, utterance-2nd, and utterance-internal, and Group 2 (turn-releasing) includes utterance-final, Pass, and Stall. In essence,

[2]not statistically significant

Group 1 tokens are immediately followed by additional speech, and Group 2 tokens are not. The mean duration in Group 1 was 203 ms and 480 ms in Group 2, a significant difference by independent two-tailed t-test, (df=670), $p<0.001$. There were also significant differences found between Group 1 and Passes (independent two-tailed t-test, df=650, $p<0.001$) and between Group 1 and Stalls (independent two-tailed t-test, df=598, $p<0.001$).

Further exploring how the duration of *so* might play a role in turn-taking, we also compare the durations of utterance-initial *so* productions that initiate a speaker change to those that initiate a new utterance by the same speaker. The durations are shown in Table 2. Here we see two interesting phenomena. First, the productions occurring as part of a speaker change are significantly shorter (independent two-tailed t-test, (df=418), $p<0.01$), but the difference is small enough (33ms) that it might fall below the threshold of human perception. Second, less than half as many utterances with an initial *so* occur as part a speaker change as compared to the same speaker continuing.

Table 1: Durations of *so* productions, both by position, and grouped by whether the *so* was followed by additional speech.

so Position	n	Mean Duration (ms)	Group Mean (ms)
Utterance-init	424	208	
Utterance-2nd	93	192	203
Utterance-internal	72	183	
Utterance-final	10	497	
Standalone-pass	62	467	480
Standalone-stall	10	547	
	671	237	

Table 2: Durations of utterance-initial *so* productions that initiate a speaker change versus those that initiate a new utterance by the same speaker.

Turn	n	Mean Duration (ms)
Speaker change	130	231
Same speaker	289	198

4.2. Relative intensity

To illustrate how the *so*'s relate to the localized acoustic environment, we measured the maximum intensity of within-utterance speech immediately adjacent to *so*. Specifically, we measure speech preceding utterance-final tokens and speech following initials and internals. For the adjacent speech segment, we chose to measure only 250 ms as this was near the mean *so* duration of 237 ms. For these analyses 18 utterances were excluded in which the adjacent speech was not at least 250 ms long.

For this comparison, we chose to use maximum intensity rather than mean. This is because, due to the generally lower-intensity production of the unvoiced /s/, mean intensity is not a particularly helpful measure for the segment *so*; maximum intensity highlights the more salient vowel production, which the listener is more likely to perceive as the volume of *so* when compared with adjacent speech.

Table 3 displays the averages of these measures for each category, and the difference in intensity between the *so*'s and their adjacent speech. Here we see that utterance initial, 2nd, and internal *so*'s are significantly quieter than the following adjacent speech, by paired two-tailed t-test, all p's<0.01. In contrast, although the utterance-final *so*'s appear to be louder than the adjacent speech, no significant difference was found (Wilcoxon Signed-rank test, N=9, W=8, NS).

Table 3: Comparing average maximum intensity (in dB) of *so* and adjacent speech (250 ms).

Intensity	Utt-initial	Utt-2nd	Utt-internal	Utt-final
(n)	(410)	(90)	(72)	(9)
so	57.77	59.40	57.98	65.72
adjacent	60.37	62.37	59.69	59.84
difference	-2.60*	-2.97*	-1.70*	5.88

5. Discussion & Conclusions

The duration data support the hypothesis that turn-retaining productions of *so* tend to be shortest while standalone Stalls tend to be longest. Passes are significantly longer than the turn-retaining group, but shorter than Stalls; perhaps a medium length is optimal for conveying a turn-release. Stalls appear to be insufficient attempts to claim the floor; the speaker lengthens *so*, perhaps with other intonational cues, to differentiate it from a turn-releasing Pass and attempt to indicate intent to hold the floor despite not immediately contributing content. In contrast, in utterance-initial cases where the *so* often leads right into the following speech without a perceptible pause, speakers may be eager to contribute the substantive clause, and thus produce *so* for only long enough to "grab" the turn and ensure that the hearer is paying attention. In addition, more than half of utterance-initial uses of *so* are produced by the same speaker as the previous turn, perhaps demonstrating a causal linkage to their preceding content and thus behaving similarly to utterance-internal *so* (which also tends to be short in duration).

The intensity data, while yielding no significant data regarding utterance-final tokens, reveal that volume tends to increase following an utterance-initial, second, or internal *so*. Perhaps these tokens start off quieter and are followed by an increase in intensity in order to emphasize the remaining speech (the perhaps more contextually important "body" of the utterance).

These patterns may shed light on prosodic turn-taking cues in dialogue. The duration data reveal that utterance-final occurrences of *so* are longer than utterance-initial "turn grabs" but shorter than turn-holding stalls, suggesting that a longer but not too drawn-out production of *so* can signal a turn release. The relative intensity data reveal that turn-retaining tokens of *so* tend to be quieter than the following segment of speech, suggesting a tendency to emphasize utterance content.

These findings are potentially applicable to the fields of speech recognition and natural language processing. The duration data provide a model for how a system can differentiate a token of *so* that will be followed by more speech from one that releases the floor. Although more research into this pattern is merited, these results may extend to other discourse markers and provide a generally applicable model of lexical and prosodic cues for turn-taking in conversation.

6. Acknowledgements

This work was funded by the National Science Foundation under grant IIS-1321146 and an associated REU.

7. References

[1] D. Schiffrin, *Discourse markers*. Cambridge University Press, 1988, no. 5.

[2] P. A. Heeman and J. F. Allen, "The TRAINS 93 Dialogues (No. TRAINS-TN-94-2)," Rochester University, NY, Dept of Computer Science, Tech. Rep., 1995.

[3] P. A. Heeman, F. Yang, A. L. Kun, and A. Shyrokov, "Conventions in human-human multi-threaded dialogues: a preliminary study," in *Proceedings of the 10th international conference on Intelligent user interfaces*. ACM, 2005, pp. 293–295.

[4] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[5] S. Sutton, R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, R. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, M. Massaro, and M. Cohen, "Universal speech tools: the CSLU toolkit," in *International Conference on Spoken Language Processing*, Sydney Australia, Nov. 1998, pp. 3221–3224.

[6] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[7] P. A. Heeman, F. Yang, and S. E. Strayer, "DialogueView: An Annotation Tool for Dialogue," in *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue - Volume 2*, ser. SIGDIAL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 50–59. [Online]. Available: http://dx.doi.org/10.3115/1118121.1118129

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

[9] A. Gravano, J. Hirschberg, and v. Beňuš, "Affirmative Cue Words in Task-Oriented Dialogue," *Computational Linguistics*, vol. 38, no. 1, pp. 1–39, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1162/coli_a_00083

[10] D. Wilson and D. Sperber, "Relevance theory," in *Handbook of pragmatics (eds L. Horn & G. Ward)*. Blackwell, 2002, pp. 607–632.