Beyond On-Hold Messages: Conversational Time-Buying in Task-Oriented Dialogue

M Soledad López Gambino
CITEC, Bielefeld University
Universitätsstraße 25, 33615
Bielefeld, Germany

Sina Zarrieß
CITEC, Bielefeld University
Universitätsstraße 25, 33615
Bielefeld, Germany

David Schlangen
CITEC, Bielefeld University
Universitätsstraße 25, 33615
Bielefeld, Germany

m.lopez_gambino, sina.zarriess, david.schlangen@uni-bielefeld.de

Abstract

A common convention in graphical user interfaces is to indicate a "wait state", for example while a program is preparing a response, through a changed cursor state or a progress bar. What should the analogue be in a spoken conversational system? To address this question, we set up an experiment in which a human information provider (IP) was given their information only in a delayed and incremental manner, which systematically created situations where the IP had the turn but could not provide task-related information. Our data analysis shows that 1) IPs bridge the gap until they can provide information by "re-purposing" a whole variety of task- and grounding-related communicative actions (e.g. echoing the user's request, signaling understanding, asserting partially relevant information), rather than being silent or explicitly asking for time (e.g. "*please wait*"), and that 2) IPs combined these actions productively to ensure an ongoing conversation. These results, we argue, indicate that natural conversational interfaces should also be able to manage their time flexibly using a variety of conversational resources.

1 Introduction

How to best present information in a dialogue system is a central, and hence well-studied problem (Stent et al., 2004; Demberg and Moore, 2006; Rieser et al., 2010; Dethlefs et al., 2012b; Wen et al., 2015). What has received less attention is the question of what a system should do *until* it can present information, in the case that retrieval of this information takes time.

A simple option would be to remain silent. However, as observed in human conversation analysis, longer periods of silence appear to be marked in normal conversation and are typically avoided (Clark, 2002). As part of an effort to study online, incremental information presentation, we set up an experiment where an information provider (IP) was given their information in a delayed and piecemeal fashion, and hence was faced with the problem of having the turn before having the information to relay (Section 2). We devised a coding scheme for different types of dialogue moves used in this "time-buying phase" before task-related information is available (Section 3). Analyzing the distribution and sequencing of these moves (Section 4), we find that a variety of strategies is used, with direct requests for more time ("please wait", "one moment please") being relatively rare.

2 Data Collection

As task domain, we chose flight travel information.[1] Interactions were set up between a CALLER (C; a confederate), who had the information need, and a TRAVEL AGENT (A), who was to provide the information. The participants were assigned the role of travel agent, and assumed that they were talking to another participant. C and A were connected via audio only, through high-quality headsets. Each agent handled 10 calls (from the same caller, but treating each as separate), after two training calls. We had 10 participants (balanced for gender), all native German speakers.

To provide some control over the interaction, the task was set up so that after a greeting provided by a recording, C formulated their request in one turn (ostensibly, addressing a dialogue system that processed it) which A could hear, but not intervene

[1] A domain in which it is, to this date, realistic that a request needs significant time to be processed, as anyone who has recently used flight search engines can attest.

Figure 1: Phases of the call.

Hm, I'd like a flight from...	BEEP	A flight from Köln Bonn	to Lisbon	departure end of November	uh	one moment, please	the search for flights is in progress	There is an available flight...
CALLER'S REQUEST		ECHO: origin	ECHO: destination	ECHO: date	FILL.	WAIT REQUEST	SYSTEM STATE	ANSWER (flight offer)

Figure 2: Example interaction (gray: caller, white: travel agent)

in. C was given, as part of the experimental protocol, a schematic representation of their goal (e.g., "flight from Hannover to New York, early August, weekday, Lufthansa"), but no exact formulation. After the request was completed, the system (or so A was told) processed it and showed it in writing on a computer display placed in front of A. An audible signal was played, after which the line was assumed to be open and it was A's task to respond to the request, using information also displayed on their computer screen. This information, however, could be presented either immediately or after a certain delay (consisting of five seconds plus a random interval between 500 and 2500 ms). The information presentation itself was also varied. In 8 of the 10 calls handled by the same agent, 16 flights were presented; in the other 2, only 4. The 16 flight responses were presented either all in one go, with the 16 flights appearing individually with delays between them, or in two blocks. In some cases, flights were taken off the result list (greyed out) again after a delay. The intended effect of this presentation mode was to keep A uncertain of whether they already had the full flight list or not. Figure 1 shows a schematic illustration of the general call structure, and Figure 2 shows an example call (abbreviated, and translated from the German) with category labels explained next. Due to technical problems, some recorded calls were not useable, which left us with a total of 92 calls (1h:41min audio).

3 Annotation

Time-Buying Stretch In this paper, we focus on what we call the "time-buying stretch", that is, the time from after the beep (when A gets the turn) until the moment at which A offers information about a specific flight, or declares definitely that no flight matches the request. One of the authors identified these stretches in the calls. There is one such stretch in each call, the length of which depends on the information delay mode (see previous section) and the individual selection speed of A. These stretches vary in duration from 4 seconds to 50 seconds, with the majority being shorter than 20 seconds.

Time Buyer Categories To enable a fine-grained analysis of the strategies for bridging the time until information presentation, we annotated dialogue moves that do not directly move the task at hand forward (as per the definition of time-buying stretch). We started out from the general DAMSL scheme (Core and Allen, 1997) but, somewhat contrary to our expectations, found that the dialogue moves in our data correspond to various backward and forward-looking actions coded in different parts of the DAMSL hierarchy. Thus, we opted for a flat scheme, allowing us to label conversational actions specific to our domain. The categories are shown together with examples in Table 1. It is important to note here that we allow for multi-functionality of the dialogue moves. Moves in the "echo" category, for example, clearly also have a conversational grounding function (Clark, 1996; Bunt, 2011); however, our focus is on their function to avoid giving task information or being silent[2].

The TB stretches were segmented and annotated by one of the authors. An independent second annotator also labelled a randomly selected set of 20% of the time buyers, using the information from Table 1 as a guideline. For these segments, we calculated Cohen's $\kappa = 0.93$, indicating that the categories are well-recognisable.

4 Analysis

The first observation to make is that there is a similar amount of speech (629 seconds) and of silence (771 seconds) in the time-buying stretches. It seems clear, hence, that our agents do something else than just wait until they have task-related in-

[2]Interestingly, given our task setup, confirmation of the search parameters was not really necessary for A, as these were displayed on A's screen.)

Category	Description	DAMSL	Examples
acknowledgment	signaling understanding of the request/ acceptance of task	Signal Understanding → Acknowledge	C: I want to fly to Bristol. A: *Okay*
echoing	repeating the request or part of it	Signal Understanding → Repeat / Statement → Reassert	C: I'm looking for a flight to Izmir at the beginning of August. A: *A flight to Izmir . beginning of August*
conf./exp./rep. request	A asks C to clarify, repeat or expand on request	Influencing addressee future action → Directive → Info-Request	*Did you say Lufthansa?*
filler	conventional hesitation sound	?	*Uh, uhm, mm, etc.*
wait request	A asks C to wait	Influencing addressee future action → Directive → Action-Directive / Information Level → Task Management	*One moment, please*
agent/system state	providing information about factors which prevent A from offering information	Information Level → Task Management	*The search for flights is still in progress. I'm not sure if Emirates flies this route.*
commitment	expressing that A is (still) engaged in performing the task	Committing Speaker Future Action → Commit	*Let's have a look...*
availability	announcing information without presenting it	Statement → Assert / Committing Speaker Future Action → Commit	*I could offer you a number of flights...* Hmm, you said Quito, is that correct?
partial match	presenting information which only matches the request partially	Statement → Assert / Signal-Understanding → Repeat	*There's a flight to Sidney on 2.8 at 07:15*, but you would prefer to fly after lunchtime, so let's keep looking...
temporary non-availability	announcing lack of information at the current moment	Statement → Assert	*Until now I haven't found any flights for your request*, let's keep looking...
incomplete	partial utterance	Communicative Status → Abandoned	*Maybe I can find...*

Table 1: Time buyer categories (C: Customer, A: Agent)

Category	%
echoing	21
filler	19
agent/system state	10.4
acknowledgment	9.4
commitment	8.8
incomplete	6.7
wait request	6.3
conf./exp./rep. request	5.9
availability	5.1
other	3.5
partial match	2.2
temporary non-availability	1.6

Table 2: Distribution of time buyer categories

formation to provide. Table 2 shows the overall distribution of time buyer categories. As can be seen, echoing occurs frequently, as does production of fillers. Direct requests to wait are comparatively rare. As Figure 3 shows, there is considerable variation between speakers in their distribution of time buyer categories, in particular for echoing and filler, which can occur very frequently or rarely depending on the speaker. Finally, Figure 4 illustrates the temporal sequencing of the TB categories. The plot shows percentages of TB type for the first seven time-buyers uttered in each episode (where available). As this indicates, there seems to be a certain structure to the sequencing of these acts: First, taking over the floor (and accepting the task) is acknowledged, then some time is filled with echoing parts of the request; when information becomes available, the parameters are made present again through clarification / expansion requests, or announcements of partial or full availability. Task- and grounding-independent acts such as fillers, announcements of system state, or direct wait requests, are available at any time, but are most relevant after the initial grounding has been done and before partial information is available for presentation.

5 Related Work

To the best of our knowledge, delayed information presentation has so far not been systematically studied. Various systems, however, addressed the problem in an ad-hoc manner. The

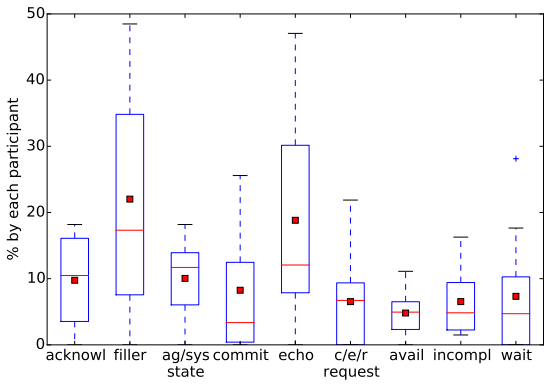

Figure 3: Distribution of TB categories per speaker (only categories with an overall frequency higher than 5%)

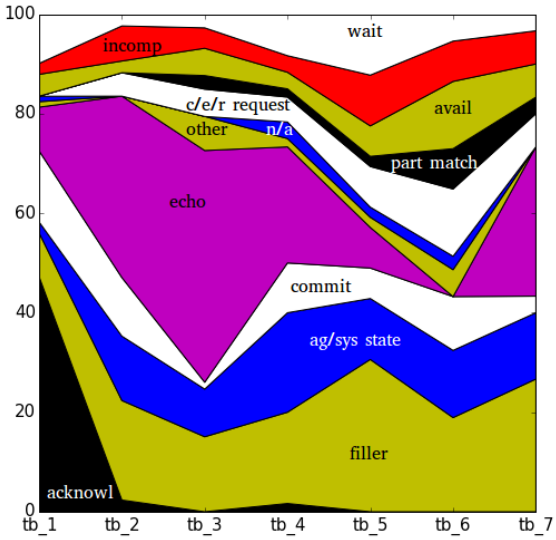

Figure 4: Distribution of time buyer categories for the first seven time-buyers in each episode (where available)

TRIPS system (Stent, 1999), for example, deals with pauses during language generation by inserting "turn-keeping" utterances, such as *um* and *wait a minute*. Funakoshi et al. (2008) conducted a Wizard-of-Oz experiment with a robot which blinked a light on its chest during long pauses, and participants successfully understood this signal as meaning that the robot was processing the incoming utterance. Wigdor et al. (2016) carried out an experiment in which a robot using "pensive fillers" (utterances such as *good question* and *let me think*) was viewed as more alive by participants than one which only postponed information by producing pauses. Although the motivation behind this experiment is not related to a real need to buy time, its results suggest that explicitly addressing collateral aspects of the task before conveying primary task information is not only not detrimental to the interaction, but might in fact be beneficial.

From a broader perspective, we see this study on time buying as contributing to research on incremental generation and information presentation for dialogue systems, cf. (Skantze and Hjalmarsson, 2010), and incremental processing in general (Schlangen and Skantze, 2009). In this line of research, it is typically acknowledged that dialogue systems should be set up in a way such that they are able to start speaking before a complete plan of what to say has been built. Skantze and Hjalmarsson (2010) present a model for incremental generation that includes the ability to insert small speech segments for hesitations and fillers, in case the system has not fully planned the current utterance. It is unclear how such a system would be able to deal with scenarios similar to the ones we have investigated in this work. Similarly, other work has looked at appropriate timings of feedback and barge-in in spoken dialogue systems (Dethlefs et al., 2012a; Meena et al., 2013), dealing with situations where the system does not need to buy time pro-actively.

6 Conclusions and Further Work

It is often difficult to systematically elicit conversational phenomena in human-human dialogue (Gustafson and Merkes, 2009), at least to an extent that would support robust data-driven systems for conversational dialogue. We have presented an experiment designed to investigate conversational strategies used to bridge time until a task can be fulfilled, or to say something before fully knowing what to say. We found that such phenomena can be successfully and systematically triggered by manipulating and delaying the information that an agent has to communicate in a typical travel information setup. Our analysis focused on the time-buying stretch, i.e. the phase of the interaction where the information provider cannot offer factual information. Even in this stretch, task- or interaction-management related acts are clearly preferable over explicit requests for more time.

In future work, we plan to analyze the remaining phases of the recorded interactions where agents actually provided information. This will allow us to compare conversational strategies in this initial time-buying stretch to grounding-related strategies used in the information presentation phase. It would also be interesting to analyze in-

formation postponing in actual telephone interactions from customer/passenger service lines, and see whether time-buying in the real world exhibits similar characteristics to those in our recordings. Clearly, this is subject to the possibility of obtaining access to such data.

On the other hand, it is necessary to devote more efforts to understanding the variation between the use of time-buyers by different individuals (Figure 3), as well as along time (Figure 4). In addition, while we think that the setup we devised is representative for the travel information domain specifically, it remains to be seen how information-postponing occurs in other conversational contexts. A similar remark could be made in connection to other languages: Since tolerance to silence has been shown to differ significantly across cultures (Lundholm Fors, 2015), observation of the phenomenon in non-German interactions might also prove revealing.

Finally, we still need to establish how to incorporate these insights in a human-agent interaction scenario. While our taxonomy was useful for annotation and analysis, it could be necessary to adjust it in order to implement time-buying in an actual system.

Acknowledgments

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Special thanks to our student assistants for their help with transcription and annotation, and to Julian Hough and Ting Han for enlightening discussions on the topic of time-buying.

References

Harry Bunt. 2011. The semantics of dialogue acts. In *Proceedings of the Ninth International Conference on Computational Semantics*. Association for Computational Linguistics, Stroudsburg, PA, USA, IWCS '11, pages 1–13.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Herbert H. Clark. 2002. Speaking in time. In *Speech Communication*, Elsevier Science, volume 36, pages 5–13.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*. Boston, MA, volume 56.

Vera Demberg and Johanna Moore. 2006. Information presentation in spoken dialogue systems. In *Proceedings of EACL*.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012a. Optimising incremental dialogue decisions using information density for interactive systems. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 82–93.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012b. Optimising incremental generation for spoken dialogue systems: Reducing the need for fillers. In *Proceedings of the Seventh International Natural Language Generation Conference*. Association for Computational Linguistics, pages 49–58.

Kotaro Funakoshi, Kazuki Kobayashi, Mikio Nakano, Seiji Yamada, Yasuhiko Kitamura, and Hiroshi Tsujino. 2008. Smoothing human-robot speech interactions by using a blinking-light as subtle expression. In *Proceedings of the 10th International Conference on Multimodal Interfaces*. ACM, New York, NY, USA, ICMI '08, pages 293–296. https://doi.org/10.1145/1452392.1452452.

Joakim Gustafson and Miray Merkes. 2009. Eliciting interactional phenomena in human-human dialogues. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 298–301.

Kristina Lundholm Fors. 2015. *Production and Perception of Pauses in Speech*. Ph.D. thesis, University of Gothenburg.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2013. A data-driven model for timing feedback in a map task dialogue system. In *14th Annual Meeting of the Special Interest Group on Discourse and Dialogue-SIGdial*. pages 375–383.

Verena Rieser, Oliver Lemon, and Xingkun Liu. 2010. Optimising information presentation for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1009–1018.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*. Athens, Greece, pages 710–718.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th Annual Meeting*

of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, Stroudsburg, PA, USA, SIGDIAL '10, pages 1–8.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics, page 79.

Amanda J. Stent. 1999. Content planning and generation in continuous-speech spoken dialog systems. In *Proceedings of the KI'99 workshop "May I Speak Freely?"*.

Tsung-Hsien Wen, Milica Gašić, Dongho Kim, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics.

Noel Wigdor, Joachim de Greeff, Rosemarijn Looije, and Mark A. Neerincx. 2016. How to improve human-robot interaction with conversational fillers. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pages 219–224.