Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 23-26

The Effect of Directed and Open Disambiguation Prompts in Authentic Call Center Data on the Frequency and Distribution of Filled Pauses and Possible Implications for Filled Pause Hypotheses and Data Collection Methodology

Robert Eklund [1, 2, 3]

[1] Department of Neuroscience, Karolinska Institute/Stockholm Brain Institute, Stockholm, Sweden
[2] Department of Computer Science, Linköping University, Linköping, Sweden
[3] Voice Provider Sweden, Stockholm, Sweden
robert@roberteklund.info

Abstract

This paper studies the frequency and distribution of filled pauses (FPs) in ecologically valid data where unaware and authentic customers called in to report problems with their telephony and/or Internet services and were met by a novel Wizard-of-Oz paradigm using real call center agents as wizards. The data analyzed were caller utterances following a directed or an open disambiguation prompt. While no significant differences in FP production were observed as a function of prompt type, FP frequency was found to be considerably higher than what is usually reported in the literature. Moreover, a higher proportion of utterance-initial FPs than normally reported was also observed. The results are compared to previously reported FP frequencies. Potential implications for data collection methodology are discussed.
Index Terms: filled pauses, Wizard-of-Oz, WOZ, speech planning, speech production, many-options, data collection, open prompts, directed prompts, call center, dialog systems.

1. Introduction

It is well-known that spontaneous speech includes disfluency (to use the most common term), i.e. silent/unfilled pauses (UPs), filled pauses (FPs) like "er", segment prolongations, repetitions, truncated words and so on. Previous studies report an average general disfluency frequency of around 6% of all words uttered [1, 2, 3, 4, 5].

Disfluency needs to be considered in the design of automated human–machine dialog systems in order to make such systems both successful and attractive. Of all different kinds of disfluencies, FPs are perhaps the most important disfluencies to consider, partly given their frequency, but also given their alleged functions in spontaneous speech.

2. Filled Pauses

Except UPs – whose definition varies between sources – FPs constitute the most common type of disfluency. Already in the 1950s it was shown that FPs exhibit different distribution and behavior as compared to all other types of disfluency, as pioneered by Mahl [6] who made a distinction between FPs (*ah* disfluency) and other forms of disfluency (*non-ah* disfluency), a distinction which has subsequently been given support in a number of studies (e.g. [7]).

2.1. Filled Pause Frequency and Distribution

As is the case with general disfluency frequency, it is not completely straightforward to find exact FP frequency figures in the literature, but the sources scrutinized report FP levels around 3.5%, divided by the number of words produced. However, although these figures vary to some extent as a function of the exact nature of communicative setting and other factors that were varied and manipulated in the experiments described, an important thing to point out is that what previous reports have in common is that most of the results are based on experimental data collections, and that the subjects in most studies were aware of being recorded. Potential implications will be discussed later.

As for FP distribution, it has been repeatedly found that FPs tend to occur in utterance-initial position, or at least very early in utterances [5, 8, 9, 10, 11, 12, 13, 14]. This has most often been explained in terms of planning processes, which will be described in more detail below.

2.2. Filled Pause Hypotheses

Over the years, FPs have been explained according to a number of different hypotheses as to their possible functions in speech, ranging from the purely pragmatic-conversational to the purely physiological. For an exhaustive survey of FP hypotheses, see [5]; for a short overview, see [15].

Among the proposed FP hypotheses, the one that is of main interest in the present study is what could be labeled the *many-options hypothesis* or the *speech planning problem hypothesis*. Lounsbury [16] suggested that FPs "correspond to the points of highest statistical uncertainty in the sequencing of units in any given order", i.e. the places in the speech stream with the highest entropy, which is at the beginning of clauses, before the speaker has yet committed to anything, and where speech planning consequently is most difficult and exhibits most inherent optionality.

That FPs indeed tend to occur in utterance-initial position has been repeatedly confirmed by e.g. Beattie & Barnard [10] who observed that 55.3% of all FPs produced by customers in natural telephone conversations occurred at the beginning of utterances. Cook [9] observed that FPs tended to occur before the first, second or third word of a clause. Shriberg [11] and Eklund & Shriberg [12] reported that speakers used FPs at the beginning of utterances more often than in any other position of an utterance. Eklund [5] reported that 45.3% of FPs were utterance-initial in a large set of corpora, while Boomer [8] observed that the most frequent position for hesitations were after the first word of phonemic clauses. O'Connell & Kowal [13] reported that 45% of FPs occurred in utterance-initial position. Thus it would seem from figures alone that FP production is indeed related to planning problems and inherent optionality.

Perhaps the most striking confirmation of the hypothesis that FP production is affected by the number of options available to the speaker is found in Schachter et al. [17] who

studied hesitations in lectures within three academic disciplines with varying degrees of inherent optionality: (1) natural science, with very few options (there are very few options to describe the orbit of a planet or the outcome of a chemical reaction); (2) social science (with an intermediate degree of available options); and (3) humanities (with a high number of ways to describe, for example, what Shakespeare really meant with a certain passage). They found that lecturers within the humanities used more FPs than lecturers within social sciences, who, in turn, used more FPs than did lecturers within natural sciences.

A final comment, however, which was pointed out by Lalljee & Cook [18], is that different hypotheses with regard to FP production are not necessarily mutually exclusive, and it could well be the case that FPs might serve several different functions in conversation, and that an experiment designed to test one *particular* hypothesis might fail to produce significant results because it does not take into account other possible functions.

3. Data Collection

3.1. Call Center Data

The data were collected during a pilot project carried out at TeliaSonera between December 2004 and February 2005 at the TeliaSonera Customer Service Call Center in Sundbyberg (Sweden). The aim of the project was to prepare the ground for the launching of speech-based call routing in the TeliaSonera residential customer care, a service reached at the number 90 200, handling around 14 million calls annually. Call routing is the task of directing callers to a service agent or a self-service that can provide the required assistance, and the basic rationale is to avoid using DTMF-based decision trees that quite quickly become unwieldy when too many choices are inherent in the service.

The analysis in this paper is focused on incidence of filled pauses in customer utterances following either *directed* or *open* disambiguation system prompts, i.e. prompts that either give the callers some hints as to how to address the system, or completely open prompts that provide a maximum amount of choice on the caller.

3.2. Wizard-of-Oz Data Methodology: Traditional

Since the late 1980s data collection associated with automated system design have often been carried out within what is known as the Wizard-of-Oz (WOZ) technique (for historical descriptions of WOZ, see [5, 19, 20]), where one basically pretends having an up-and-running system, with actors playing the some or all parts of the system. Typically, subjects are given roles to play, with specified tasks to carry out while communicating with what they believe is a functioning and fully automated system, while in fact they are communicating with hidden agents (actors and/or researchers) pretending to be the system. Since basically everything in a traditional WOZ collection is "fake" – the actual setting; the subjects having "roles" assigned to them and often having instructions presented to them in spoken or written form; the agents playing roles with parts of the behavior following carefully designed instructions, etc – WOZ collections have been criticized for lack of ecological validity, (e.g. [21]).

Moreover, and importantly, subjects are also typically aware of the fact that the interaction is recorded (this was not the case in [10], where only the operators were aware of the calls being recorded), which might affect how they express themselves. It could consequently be argued that the only feature of a WOZ collection that supposedly comes close to the desired objective of the data collection is that the subjects believe they are interacting with a working system, which is quite often verified in post-experimental interviews.

3.3. Wizard-of-Oz Data Methodology: 90 200

The data in this study were collected using a novel variant of the Wizard-of-Oz (WOZ) technique, where many of the short-comings of traditional WOZ collections were avoided. First, the callers were *actual* Telia customers, with *actual* and *authentic* problems, rather than experimental subjects with tasks assigned to them. Second, all callers were *unaware* of being recorded. Third, similar to traditional WOZ collections, they believed they were interacting with a working automated system. Fourth, instead of using researchers or actors playing the part of the system, *authentic* call center agents were used as wizards, effectively playing themselves, with some restrictions on their understanding of caller utterances. The data collection is described in detail in [22].

3.4. Dialog Structure

The specific data analyzed in this paper are the caller utterances that followed a disambiguation prompt, either a *directed* prompt (where the caller is given some instructions) or an *open* prompt ("How may I help you"-style; [23]). It has been shown that users give longer utterances in response to open prompts [24], and that open prompts, when used correctly, can lead to significantly improved routing performance [25]. However, it has also been show that directed strategies are more successful for callers who lack expectations on the task structure [26]. The dialog structure is shown in *Figure 1*.

System:	*Välkommen till Telia! Här beskriver du ditt ärende med egna ord i stället för att knappa dig fram. Om du säger vad du vill ha hjälp med så kopplar jag fram dig. Vad gäller ditt ärende?* ('Welcome to Telia. Here you describe your reason for call with your own words instead of using touchpad navigation. If you tell me what kind of assistance you need I will forward you to the right agent. What is the reason for you call?')
Customer:	*Eh… Internet.*
System:	*OK, Internet.* < Directed> or <Open Prompt> played here
Directed Prompt:	Jag behöver veta lite mer om ditt ärende. Gäller det till exempel beställning, prisinformation eller support? ('I need some additional information about the reason for your call. Is it for example about an order, price information or support?')
Open Prompt:	Kan du säga lite mer om vad du vill ha hjälp med? ('Could you please tell me some more about the reason for your call?')
Customer:	<Analyzed utterances>

Figure 1: Call Center Dialog Structure.

First the (simulated) system played a welcome message and an invitation to the caller to describe their reason for call. If the utterance did not contain all information necessary to route the call to the correct destination, the system asked a disambiguation question to try to obtain the information.

3.5. Data Transcription

The dialogs between the wizards/agents and the callers were transcribed by an consulting company, STTS (**www.stts.se**), following Nuance Guidelines. Transcription did not focus on disfluency, but FPs were indicated by the item **@hes@** in the transcriptions. All instances of these hesitation labels (i.e. following the directed and open prompts) were located and listened to in order to verify that they were in fact FPs.

4. Results

4.1. General Results

In a previous paper Eklund & Wirén [15] reported that the two prompts had a dramatic effect on the syntactic-categorical form of the customers' utterances. After the directed prompt 72% of the utterances were noun-only utterances, and finite verb-sentences constituted less than 10% of the utterances. After the open prompt more than 40% of the utterances were clauses and less than 20% were noun-only utterances.

4.2. Filled Pause Frequency

FP frequency is shown in *Table 1* below.

Table 1. Summary Statistics for utterances, words and FPs, and ratios for FPs/Utts and FPs/Words.

Prompt	Utterances & Words			Filled Pauses		
	Utts	Words	Words /Utts	FPs	FPs /Utts	FPs /Words
Directed	118	216	1.8	16	0.14	0.074
Open	121	791	6.5	60	0.50	0.076
Σ	239	1007	4.2	76	0.32	0.075

As shown, there is a striking lack of difference across the two prompt settings, and FP occurrence is almost exactly the same following the two prompts. However, the figures 7.4% and 7.6% are considerably higher than what is found in previous reports. Although it is not always easy to disinter specific figures for FP frequency in the literature since they are quite often merged with general disfluency figures or are given as FPs/minute, sources where FP frequencies were specified or were easily calculated are shown in *Table 2* below.

Table 2. FP frequencies in the literature. Legend: H=Human; M=Machine/Computer; WOZ=Wizard-of-Oz

Source	FPs/Words	Comment
Maclay & Osgood (1959) [27]	3.9%	Conference data
Lallgee & Cook (1969) [28]	2.8%	High pressure
	3.4%	Low pressure
Lalljee & Cook (1973) [29]	1.9%	Females
	4.3%	Males
Lutz & Mallard (1986) [30]	3.6%	Conversation
Eklund & Shriberg (1998) [12]	3.4%	H–M WOZ
	2.6%	H–H
	0.3%	H–M push-to-talk
	3.0%	H–H
Bortfeld et al. (2001) [4]	2.6%	Conversations
Eklund (2004) [5]	3.0%	H–M WOZ
	4.1%	H–M WOZ
	2.2%	H–H
	4.4%	H–M
Moniz, Mata & Viana (2007) [31]	2.3%	Presentations

As is shown in *Table 2*, despite the fact that the sources cover almost half a century, the figures are remarkably similar. The one exception is the very low figure 0.3% in [12], but that is easily explained by the fact that those data are based on a push-to-talk service where subjects did not talk until they had prepared their speech beforehand (or at least had that opportunity). Although the list above surely is not exhaustive, it still is indicative, and it could be assumed that the figures reported in the present study are, in fact, unusually high.

4.3. Filled Pause Distribution

FP distribution is shown in *Table 3*.

Table 3. Summary Statistics for FP position

Prompt	FP position		Σ
	Initial	Other	
Directed	14	2	16
Open	36	24	60
Σ	50	26	76
Proportion	65.8%	34.2%	100%

Once again there is no statistically significant difference between the two prompt settings. However, not only do the majority of FPs occur in utterance-initial position, they do so markedly more so than was reported in the literature. e.g. in [5], where 45.3% of FPs were utterance-initial. The difference is statistically significant given a *Z*-test for two proportions (two-tailed, $p<0.01$).

5. Discussion

It could be argued that the data set studied in this paper is too small to allow any far-reaching conclusions to be drawn. However, a counterargument would be that what the data lack in *quantity* is compensated for in *quality*, thus making both observations and conclusions (if tentative) interesting, given that the data are as ecologically valid as is possible, short of wire-tapping human conversation – which basically is the case here. Compared to previous studies that to a large extent are based on either experimental settings, or in most cases have studied the speech of subjects who were aware of being recorded, any observed differences in the present data set might be of interest from both a disfluency perspective proper, and also from a strictly methodological point of view.

First, the first and obvious observation is that it that there was no difference in FP production with respect to the two different prompts. A first conclusion here would be to consider this result as providing negative support for the many-options hypothesis. However, the number of options the subjects have at hand have actually not changed as a function of the two prompts: the callers still have the same exact problem at hand. What has changed is *how* they are allowed to describe their problem, i.e. the "more options" aspect occurs only on the *form* side of the interaction, not on the *content* side. This means that to the extent that FP production is indeed indicative of planning problems, it would seem that clients are neither helped nor hampered by the use of directed (or open) prompts when interacting with a call center system, since what changes is mainly the syntactic form of their utterances, as well as utterance length. This, however could still be interesting from a design point of view.

Second, and perhaps more interesting, is the observation that significantly more FPs were produced in this study than in previous reports on FP frequency. This could be explained in different ways. The first, Occam's Razor-like, explanation could of course be that the present corpus simply is not representative in that it is comparatively small, and is based on a fairly limited set of carefully chosen utterances (following two specific disambiguation prompts). However, given the ecological validity of the data, this seems to be a little too simplistic as an explanation. As was pointed out earlier, most of the previous reports were based on data collected in experimental settings, and all of the subjects recorded were aware of the calls being recorded. As have been shown, disfluency is within speaker control (e.g. [32]) and it could be the case that the awareness of the recording devices actually have an effect on disfluency production. For example, it has been shown that speaker disfluency is decreased simply by

directing a TV camera on the speaker [33]. If so, the observed FP production in this study might, in fact, be closer to the truth than in any of the previous reports. To the extent that that is true, corroboration could be hard to carry out since it would most likely be difficult to have similar data collection studies pass ethical committees.

Third, the proportion of initial FPs in this study is among the highest reported so far, which could be considered as support for planning-problem hypotheses in general. However, as was the case with FP frequency, the different prompts did not result in statistically different distribution.

Summing up, while the two different prompts did not affect caller FP production, both FP frequency and FP distribution stand out compared to previous studies. Although there might be several possible explanations for this, *one* possible explanation for the observed difference is that analyzing real-world data might yield different results than data collected in experimental settings, and that the results presented here are indicative of the difference between experimental and ecologically valid, real-world, data.

6. Conclusions

We conclude that FP production seemingly did not vary as a function of directed or open prompts, and that freedom of expression with regard to linguistic form did not affect speech production from a planning point of view in the present study.

However, absolute FP frequency was significantly higher than in the literature, and while there are several possible reasons for this, one possible alternative is that ecologically valid data can reveal speech production processes that remain hidden in WOZ collections, no matter how well-designed.

7. Acknowledgements

The data described and analyzed in this paper are covered by a right-of-use agreement ("nyttjanderättsavtal") between TeliaSonera and Karolinska Institute (637/10, 8 Oct. 2009).

8. References

[1] Fox Tree, J.E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, 34:709–738.

[2] Oviatt, S. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Language*, 9:19–35.

[3] Brennan, S.E. & M.F. Schober. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, 44: 274–296.

[4] Bortfeld, H., S.D. Leon, J.E. Bloom, M.F. Schober & S.E. Brennan. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech,* 44(2):123–147.

[5] Eklund, R. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Dept. of Computer and Information Science, Linköping University.

[6] Mahl, G.F. (ed.). 1987. *Explorations in nonverbal and vocal Behavior*. Hillsdale, New Jersey: Lawrence Erlbaum.

[7] Christenfeld, N. & B. Creager. 1996. Anxiety, Alcohol, Aphasia, and *Ums*. *Journal of Personality and Social Psychology*, 70(3):451–460.

[8] Boomer, D.S. 1965. Hesitation and grammatical encoding. *Language and Speech*, 8(3):148–158.

[9] Cook, M. 1971. The incidence of filled pauses in relation to part of speech. *Language and Speech*, 14(2):135–139.

[10] Beattie, G.W. & P.J. Barnard. 1979. The temporal structure of natural telephone conversations (directory enquiry calls). *Linguistics*, 17:213–229.

[11] Shriberg, E.E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California, Berkeley.

[12] Eklund, R. & E. Shriberg. 1998. Cross-linguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. *Proc. ICSLP 98*, 30 Nov–5 Dec. 1998, Sydney, 6:2631–2634.

[13] O'Connell, D.C. & S. Kowal. 2005. *Uh* and *Um* Revisited: Are They Interjections for Signaling Delay? *Journal of Psycholinguistic Research*, 34(6):555–576.

[14] Pfeifer, L.M. & T. Bickmore. 2009. Should Agents Speak, Like, um, Humans? The Use of Conversational Fillers by Virtual Agents. In: Z. Ruttkay et al. (eds.): *Intelligent Virtual Agents. Proc. 9th International Conference on Intelligent Virtual Agents* (IVA), 14–16 September 2009, Amsterdam, 460–466.

[15] Eklund, R. & M. Wirén. 2010. Effects of open and directed prompts on filled pauses and utterance production. *Proc. Fonetik 2010*, 2–4 June 2010, Lund University, 23–28.

[16] Lounsbury, F.G. 1954. Transitional Probability, Linguistic Structure, and Systems of Habit-family Hierarchies. In: C.E. Osgood & T.A. Sebeok (eds.): *Psycholinguistics. A Survey of Theory and Research Problems*. Baltimore: Waverly Press, 93–101.

[17] Schachter, S., N. Christenfeld, B. Ravina & F. Bilous. 1991. Speech Disfluency and the Structure of Knowledge. *Journal of Personality and Social Psychology*, 60(3):362–367.

[18] Lalljee, M. & M. Cook. 1974. Filled Pauses and Floor-Holding: The Final Test? *Semiotica*, 12(3):219–225.

[19] Fraser, N.M. & G. N. Gilbert. Simulating speech systems. 1991. *Computer Speech and Language*, 5:81–99.

[20] Dahlbäck, N., A. Jönsson & L. Ahrenberg, Wizard of Oz Studies – Why and How. 1993. *Knowledge-Based Systems*, 6(4):258–266.

[21] Allwood, J. & B. Haglund. 1992. *Communicative Activity Analysis of a Wizard of Oz Experiment*. Internal Report, PLUS ESPRIT project P5254.

[22] Wirén, M., R. Eklund, F. Engberg & J. Westermark. Experiences of an In-Service Wizard-of-Oz Data Collection for the Deployment of a Call-Routing Application. 2007. *Proc. Bridging the Gap: Academic and Industrial Research in Dialog Technologies.* NAACL Workshop, 26 April 2007, Rochester, New York, 56–63.

[23] Gorin, A.L., G. Riccardi & J.H. Wright. 1997. How may I help you? *Speech Communication*, 23:113–127.

[24] McInnes, F.R., I.A. Nairn, D.J. Attwater & M.A. Jack. 1999. Effects of prompt style on user responses to an automated banking service using word-spotting. *BT Technology Journal*, 17:160–171.

[25] Sheeder, T. & J. Balogh. 2003. Say it Like You Mean it: Priming for Structure in Caller Responses to a Spoken Dialog System. *International Journal of Speech Technology*, 6:103–111.

[26] Williams, J.D. & S.M. Witt. 2004. A Comparison of Dialog Strategies for Call Routing. *International Journal of Speech Technology*, 7:9–24.

[27] Maclay, H. & C.E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word*, 5:19–44.

[28] Lallgee, M.G. & M. Cook. 1969. An experimental investigation of the function of filled pauses in speech. *Language and Speech*, 12(1):24–28.

[29] Lalljee, M. & M. Cook. 1973. Uncertainty in first encounters. *Journal of Personality and Social Psychology*, 26(1):137–141.

[30] Lutz, K.C. & A.R. Mallard. 1986. Disfluencies and rate of speech in young adult nonstutterers. *Journal of Disfluency Disorders*, 11:307–316.

[31] Moniz, H., A.I. Mata & M. Céu Viana. 2007. On Filled-Pauses and Prolongations in European Portuguese. *Proc. Interspeech 2007*, 27–31 August 2007, Antwerp, 2645–2648.

[32] Siegel, G.M., J. Lenske & P. Broen. 1969. Suppression of normal speech disfluencies through response costs. *Journal of Applied Behavior Analysis*, 2:265–276.

[33] Broen, P.A. & G.M. Siegel. 1972. Variations in normal speech disfluencies. *Language and Speech*, 15(3):219–231.