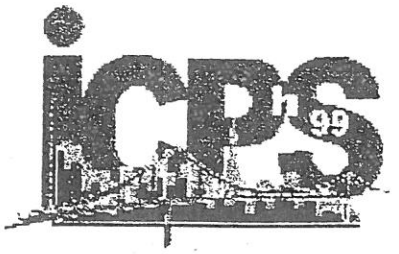

satellite meeting

Disfluency in Spontaneous Speech

U. C. Berkeley, 30 July, 1999

Organizers

Robin Lickley	University of Edinburgh
Ellen Gurman Bard	University of Edinburgh
Jean Fox Tree	UCSC
Peter Heeman	OGI
Liz Shriberg	SRI
Madelaine Plauché	UC Berkeley

ICPhS99 Satellite Meeting on

Disfluency in Spontaneous Speech

Place: 370 Dwinelle Hall,U. C. Berkeley

Date: Friday, July 30, 1999

CAFFEIN AND WELCOME

08:30-09:00 Coffee available.

09:00-09:10 Welcome - Robin Lickley

SESSION 1: PRODUCTION ISSUES. Chair, Jean Fox Tree

9:10-09:40 Robert Eklund: A comparative analysis of disfluencies in four Swedish travel dialogue corpora (p. 3)

09:40-10:10 Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, & Susan E. Brennan: Which speakers are most disfluent in conversation and when? (p. 7)

10:10-10:40 Kim Kirsner, Benjamin Roberts, & Yong-Heng Lee: Why does spontaneous speech unfold in temporal cycles sometimes? (p. 11)

10:40-11:00 Coffee

SESSION 2: PERCEPTION. Chair, Ellen Gurman Bard

11:00-11:30 Jean E. Fox Tree: Between-turn pauses and ums (p. 15)

11:30-12:00 Susan E. Brennan & Michael Schober: Uhs and interrupted words: the information available to listeners (p. 19)

12:00-12:30 Robin J. Lickley, David McKelvie, & Ellen Gurman Bard: Comparing human and automatic speech recognition of disfluent speech using word-gating (p. 23)

12:30-13:30 Lunch

SESSION 3a: ASR/CL APPROACHES. Chair, Peter Heeman

13:30-14:00 Sherri Page: Use of a post-processor to identify and correct speaker disfluencies in automated speech recognition for medical transcription (p. 27)

14:00-14:30 Sergey Pakhomov & Guergana Savova: Filled pause distribution and modeling in quasi-spontaneous speech (p. 31)

14:30-15:00 Dafydd Gibbon & Shu-Chuan Tseng: Toward a formal characterization of disfluency processing (p. 35)

15:00-15:30 Coffee

SESSION 3b: ASR/CL APPROACHES. Chair, Elizabeth Shriberg

15:30-16:00 Douglas O'Shaughnessy: Better detection of hesitations in spontaneous speech (p. 39)

16:00-16:30 Peter A. Heeman & K. H. Loken-Kim: Detecting and correcting speech repairs in Japanese (p. 43)

16:30-17:00 Mark G. Core & Lenhart K. Schubert: Speech repairs: a parsing perspective (p. 47)

17:00-17:30 General discussion

Format: Each talk 20 mins with 10 mins discussion.

Cost: $25 ($15 for students). Please pay at the registration desk by the end of the morning coffee break. Cash is preferable to checks.

Dinner arrangements being made for the speakers at a local restaurant. There may be extra places. Please sign up at registration.

A COMPARATIVE ANALYSIS OF DISFLUENCIES IN FOUR SWEDISH TRAVEL DIALOGUE CORPORA

Robert Eklund
Telia Research AB, Farsta, Sweden

ABSTRACT

This paper reports on ongoing work on disfluencies carried out at Telia Research AB. Four travel dialogue corpora are described: human–"machine"–human (Wizard-of-Oz); human–"machine" (Wizard-of-Oz); human–human and human–machine. The data collection methods are outlined and their possible influence on the collected material is discussed. An annotation scheme for disfluency labelling is described. Preliminary results on five different kinds of disfluencies are presented: filled and unfilled pauses, prolonged segments, truncations and explicit editing terms.

1. INTRODUCTION

Current automatic speech recognition (ASR) and human–computer dialogue systems have attained a technological level that allows use in every-day commercial applications, as long as the tasks are suffiently constrained. In order to allow more open-ended speech input, certain phenomena typical of spontaneous speech need to be modelled. One such phenomenon is the processing of disfluencies (pauses, truncations, prolongations, repetitions, false starts etc.), or DFs for short, where more basic knowledge is needed in order to acquire more accurate modelling of spontaneous speech. To obtain such basic knowledge, a first necessary step is thus to study DFs in application-like situations. This paper describes ongoing work at Telia Research AB, where DFs in the travel booking domain are studied.

2. METHOD

Travel booking dialogues in four different corpora are studied. The data were collected as a part of the Spoken Language Translator (SLT) project at Telia Research AB [1]. Since the SLT project work was carried out within the Air Travel Information Service (ATIS) domain [4], Swedish ATIS data were collected for language modelling and recognizer training purposes.

2.1. General Set-up and Subjects

The bookings were made over a telephone line, but high quality recordings were also made in order to facilitate acoustic analyses. The subjects were all Telia employees and were used to business travel bookings. As far as possible, the subjects were balanced over gender and age.

2.2. Corpus 1: WOZ–1 / Human–"Machine"–Human

In order to avoid too strong a colouring effect from instructions, the tasks were given in a mixture of written and pictorial form (cf. Figure 1). A more detailed description of this data collection session is found in [6].

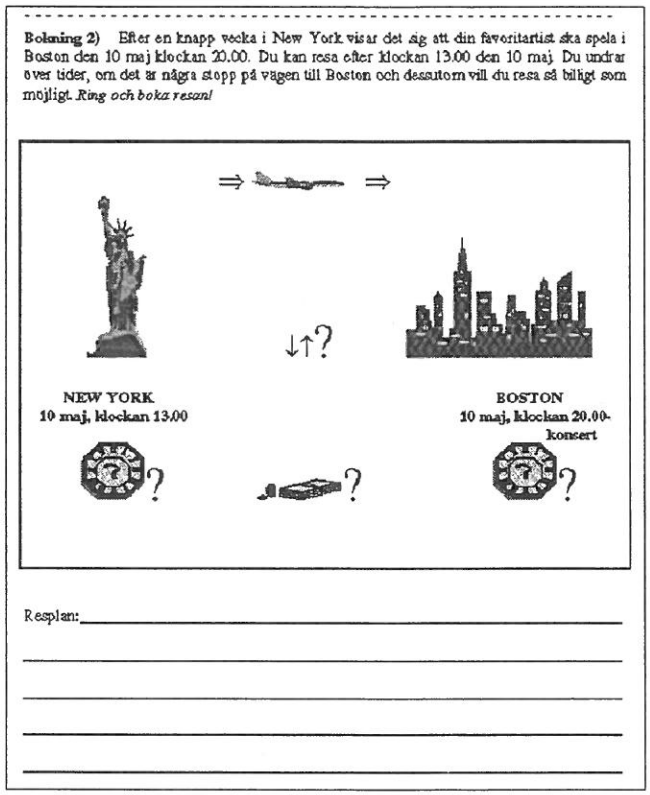

Figure 1: Task sheet number two (of ten) for the WOZ–1 corpus. The text reads: "After almost a week in New York you are told that your favourite artist is going to perform in Boston on May 10, at eight o'clock in the evening. You can leave after 1 p.m. on May 10. You want to know the times, whether there are any stopovers on the way to Boston, and moreover, you want to travel as inexpensively as possible. *Make the call and do the booking!*" The subjects were allowed to scribble down supporting notes prior to the call.

2.3. Corpus 2: WOZ–2 / Human–"Machine"

During a later phase in the SLT database (SLT-DB) project, it was decided to expand the domain from ATIS to business travel bookings. This meant that data encompassing hotel information, car rental and so on were needed. Thus, the goal was here to record business travel dialogues between users and a simulated database application. The user was given the tasks in almost exlusively pictorial form (cf. Figure 2). A more detailed description is found in [5].

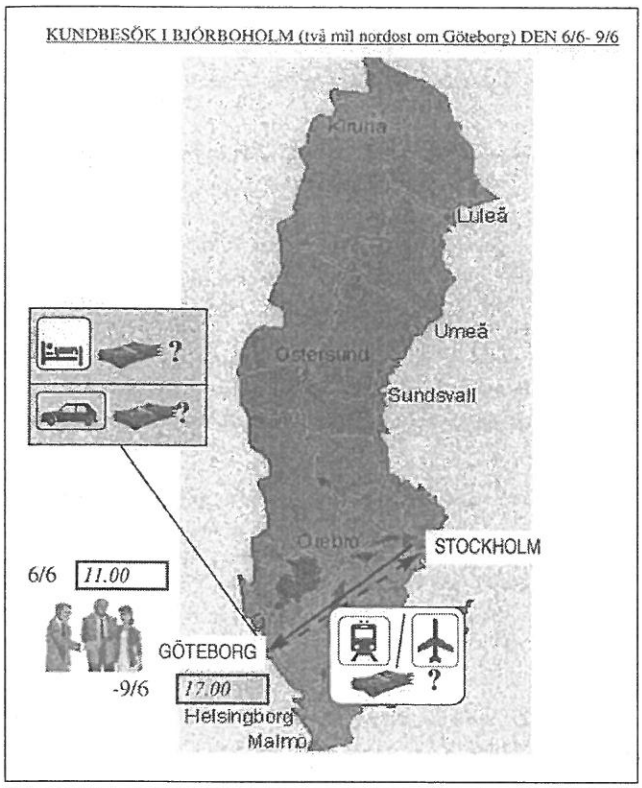

Figure 2: Task sheet number one (of three) for the WOZ–2, Nymans and Bionic corpora. The header reads: "Client visit in Björboholm (twenty kilometers north-east of Gothenburg) between June 6 and June 9." Arrows indicate departure and arrival, and icons and question marks indicate that inquiries should be made concerning the prices of trains and flights, accomodation and car rental. On this task sheet, dates and hours are given in figure format, whereas on other task sheets, icons of calendars and clocks were used.

2.4. Corpus 3: Nymans / Human–Human

In order to get a better grasp of general linguistic phenomena (grammatical, disfluencies, prosodic etc.), it was decided to collect authentic human–human dialogues. The subjects were given the same tasks as in WOZ-1 (cf. Figure 2).

2.5. Corpus 4: Bionic / Human–Machine

The goal of this data collection was to obtain authentic human–machine data. The word 'bionic' is used since authentic components were used to the extent that it was technically feasible. Since the recognizer did not cover some Swedish cities, a wizard was used to simulate recognition. Once again, the subjects were given the same instructions as were used in WOZ-2 and Nymans (cf. Figure 2).

2.6. Summary

The data thus collected are summarized in Table 1. All corpora were first transcribed orthographically by a transcription agency, but are presently being re-transcribed by the author to adhere to the annotation system described in section 3.

Table 1: Summary Statistics. The top row figures are based on the transcriptions made by the the transcription agency and are given to indicate the full size of the corpora. The bottom row (and slightly smaller) figures are based on the retranscribed and labelled material. Notes: ♠ = Only 23 subjects are labelled so far. * = Not fully transcribed. † = Note that there are eight subjects and two travel agents. ‡ = Only the subjects are transcribed and labelled, not the travel agents.

	WOZ–1	WOZ–2	Nymans	Bionic
Method	Script	Picture	Picture	Picture
No. subjects Male/Female	39 19/20	47♠ 31/16	10† 7/3	16 9/7
No. dialogues Labelled	390 79	131 55	24 20‡	* 16
No. utterances Labelled	3,722 957	3,602 1,632	2,899 1,323‡	* 517
No. words Labelled	29,645 6,181	27,277 12,142	21,611 7,159‡	* 3,117

In all corpora there are a large number of one-word utterances, e.g. confirming utterances like "ja" (yes) and so on. Since one-word utterances are less likely to be disfluent, the relative rates of these affect the number-of-DFs/number-of-utterances ratio. There is reason to believe that one-word utterances are more common in Nymans than in the other corpora, since conversation support is more likely to occur in human–human conversation. However, we have presently not developed a reliable way of excluding one-word utterances, so all figures given include one-word utterances.

3. ANNOTATION

All corpora are labelled according to an annotation scheme first presented in [2]. This system is based on the annotation scheme developed by Shriberg [7], with some extensions and minor changes. The biggest differences are the explicit labelling of prolonged segments and that durations are explicitly given for filled and unfilled pauses, as well as for prolongations. The rationale of this system is two-fold. First, the aim was to use a system that could be easily mapped between languages, in order to facilitate cross-linguistic comparisons. Preliminary results from a cross-language study of Swedish and American English are presented in [3]. Second, since the annotation scheme in Shriberg is pre-theoretical, it could be assumed that it is easily portable other languages. Although Swedish is typologically very close to English, using a method similar to that of Shriberg could serve as test case for the generality of the labelling method. If the labelling method is indeed portable between languages, we would be one step closer toward a general tool and method for disfluency labelling across languages, which is a desideratum in the quest for deeper knowledge of human speech production.

3.1. Basic Annotation Scheme

The following disfluency categories are described in this paper:

`f<` ... `>f` Filled pause (FP) Marks the beginning and end of filled pauses, most often realised as "eh" or "öh" in Swedish. The filler word is written between the **`f<` ... `>f`** markers.

u< ... >u Unfilled pause (UP) Marks the beginning and end of unfilled pauses, i.e., silence. If heavy inhalations, exhalations or other similar phenomena occur during an unfilled pause, they are labelled between the **u<** ... **>u** markers.

p< ... >p Prolongation (PR) Marks prolonged segments. Hesitations can be realized as prolonged segments. i.e., a word might be pronounced with one or more segments markedly longer than in normal, fluent speech (e.g., "I want a ffffffflight to..."). The prolonged segment is indicated within curly brackets, and it is also indicated whether it is word-initial, word-medial or word-final. For instance the label **{f-}** indicates that the prolonged segment is a word-initial [f]. Swedish makes extensive use of productive compounding, and word boundaries within compounds are indicated using a hash sign **#** .

e Explicit Editing Term (EET) Words like "Sorry", "No, wrong", "I mean" and so on. EETs can be labelled in two ways: Either each word is counted, e.g. "Sorry" = 1; "I'm sorry" = 2 (or 3), or each "EET unit" is counted "I'm sorry" = 1. Although there are arguments in favour of the latter method — DFs rarely occur within such "packages" — we have opted for the former in this paper.

/ Truncation (TR) Marks an interrupted word, either in repairs or caused by an intervening system or agent. The former, i.e., self-induced truncations, are of greater interest for our purposes, but the distinction is not made in this paper.

3.2. Repairs

Repairs are also labelled but both the labelling and the analyses of repairs are in a preliminary phase and a full discussion of these topics will have to await further work.

4. RESULTS AND ANALYSES

In this paper, we focus on FPs, UPs, PRs, EETs and TRs. The reason for choosing these types of DFs is not that they necessarily form a natural group, but rather to provide a general description of the material. Moreover, there might be more or less strong interaction between some of these types (e.g., FPs are often followed by UPs), and there is surely interaction between these types and repairs, which we have not looked at yet. The overall rates of the said DFs are shown in Table 2.

4.1. General observations

We first summarize some general observations on FPs, UPs, EETs and TRs. Since PRs is the hitherto least described DF type, these are treated in section 4.3.

4.1.1. Overall differences The first observation to be made is that the Bionic corpus exhibits the largest number of disfluencies, except for EETs. In fact, the ratio of the number of the included DFs and the number of labelled utterances is higher than 1, which indicates that there is a higher-than-100% chance that an arbitrary utterance will be in this corpus. Since the tasks were the same for WOZ–2, Nymans and the Bionic corpora, we may assume that this difference is not due to task details. It may be the case that the actual quality of the feedback—simulated synthesis in WOZ–1 and WOZ–2, human being in Nymans and real synthesis in Bionic—is the decisive factor. Even if the actors portraying the synthesizer spoke in a monotone voice and used a fixed set of standardized utterances, they did not exhibit jitter and similar phenomena associated with a real synthesizer. From this follows that one could possibly expect a larger number of disfluencies in real applications than in WOZ simulations.

Table 2: Summary of DF Rates. For each corpus, the percentages for the labelled material (as indicated in Table 1) are given, broken down by DF type. The figure is then divided by the number of labelled utterances and words in the corpus (multiplied by 100 to give percentages).

	WOZ–1	WOZ–2	Nymans	Bionic
Total no. FPs	225	400	182	145
No. FPs / no. utts	23.5	24.5	13.7	28.0
No. FPs / no. words	3.6	3.2	2.5	4.6
Total no. UPs	440	990	435	370
No. UPs / no. utts	46.0	60.7	32.9	71.6
No. UPs / no. words	7.1	8.1	6.0	11.9
Total no. PRs	94	84	107	67
No. PRs / no. utts	9.8	5.1	8.1	13.0
No. PRs / no. words	1.5	0.7	1.5	2.1
Total no. EET	20	42	10	10
No. EETs / no. utts	2.0	2.5	0.7	1.9
No. EETs / no. words	0.3	0.3	0.1	0.3
Total no. TRs	57	74	213	70
No. TRs / no. utts	6.0	4.5	16.1	13.5
No. TRs / no. words	0.9	0.6	3.0	2.2
Σ of included DFs	836	1,590	947	662
No. DFs / no. utts	87.4	97.4	71.6	128.0
No. DFs / no. words	13.5	13.1	13.2	21.2

4.1.2. Filled and Unfilled Pauses FPs and UPs are far more common than the other types of DFs, and UPs are generally twice as common as FPs.

4.1.3. Explicit Editing Terms and Truncations A somewhat surprising result is the low number of EETs in Nymans. One would assume that EETs were more common in human–human communication than to an inanimate system, but this is not the case. (Once again, the Bionic corpus is slightly different.) One possible explanation, however, could be related to the larger number of TRs; in Nymans the subjects are more often interrupted, and thus do not have either the opportunity (or reason) to put in EETs, since the experienced travel agent reacts to inconsistent information with a question. It must be pointed out that EETs are rare in all the corpora.

4.2. Durational DFs

In the set of DFs we look at in this paper, FPs, UPs and PRs stand out from EETs and TRs by being durational in nature, i.e., an FP, UP or TR can be stretched in time. Preliminary durational observations for these three types of DFs are presented in Table 3. The general tendency is that for all DFs and corpora, the pattern is PRs < FPs < UPs. Two things need be pointed out. First, whereas UPs can be stretched to great lengths, this does not occur for FPs or PRs. This explains the high standard deviation for UPs in the WOZ–2 corpus, where extreme outlier values occur, which was possible since the wizards were not instructed to take the initiative after a specified number of seconds.

Table 3: Mean Durations (in seconds) for FPs, UPs and PRs. Standard deviations are given in small figures.

	Mean FP	Mean UP	Mean PR
WOZ–1	0.53	0,62	0,25
sd	0,28	0,52	0,16
WOZ–2	0,44	1,01	0,30
sd	0,21	2,57	0,15
Nymans	0,49	0,54	0,29
sd	0,24	0,62	0,12
Bionic	0,43	0,62	0,25
sd	0,24	0,62	0,15
All	0,48	0,60	0,27
sd	0,25	0,74	0,14

Second, a distinction must be made between short UPs and very clear, computer-directed speech, where each word is uttered in isolation, but in a fluent manner. We are here dealing with two distinct speech styles. The durational values for FPs, UPs and PRs are fairly stable across the corpora, which leads us to believe that they are not that sensitive to either task details or general settings, but rather originate at lower levels in the speech production system.

4.3. Prolongations

There are basically two questions one can ask with regard to PRs: First, what kind of segments are prolonged? Second, what position in the word is favoured? Preliminary observations answering these questions are shown in Table 4.

Table 4: Phone Type and Position of Prolongations. For each corpus the percentages of phone position is given. Within each class, the percentages of phone class is given.

	WOZ–1	WOZ–2	Nymans	Bionic
% Initial phone	26.6	28.6	25.2	37.3
% vowel	1.0	0.0	14.9	8.0
% cons +sonorant	1.0	29.1	48.1	40.0
% cons –sonorant	24.6	70.9	37.0	52.0
% Medial phone	19.1	25.0	15.0	23.9
% vowel	9.6	38.1	25.0	31.2
% cons +sonorant	5.3	4.8	18.8	25.0
% cons –sonorant	4.2	57.1	56.2	43.8
% Final phone	54.2	46.4	59.8	38.8
% vowel	9.6	17.9	34.4	26.9
% cons +sonorant	30.8	71.8	53.1	61.8
% cons –sonorant	13.8	10.3	12.5	11.5

In the WOZ corpora and Nymans, initial, medial and final phone prolongations occur in the (roughly) 30–20–50 proportions mentioned in [3]. Once again, the Bionic corpus behaves slightly differently, however. Oddly enough, in no case is a vowel the preferred segment. This could depend on labelling, since there is a certain risk of judging prolonged vowels as "normal". The most important observation, however, is that all kinds of segments can be prolonged, including voiceless stops. In fact, examples like "flyge.....t" (the fligh.....t) — where the occlusion phase is prolonged — are quite common. A general problem with labelling PRs occurs at the bottom end of the durational scale. While there are clear cases of very marked prolongation, it is quite often hard to say when a segment is prolonged slightly. However, since PRs undeniably exist, one would need to develop a method of labelling them consistently.

5. DISCUSSION

Although most of the material remains to be labelled and analyzed, some tendencies are clear. First, the data collection method and set-up clearly influences the material. Thus, it seems to be the case that the use of a real synthesizer in the Bionic corpus yielded a higher rate of DFs than the two WOZ collections. An interesting preliminary result, still needing corroboration, is that WOZ simulations appear to give results that are closer to human–human interactions than to human–machine interactions, underscoring the fact that human–human control data, in this case the Nymans corpus, are important for an accurate understanding of the processes involved in human–machine interaction. Second, PRs occur in all corpora in similar proportions, and arguably serve the same function as do FPs and UPs. Third, UPs and FPs are by far the most common DFs in all corpora, while EETs, TRs amd PRs are less frequent. A final point to be made is that it seems that Shriberg's approach to DF labelling is indeed portable to Swedish.

ACKNOWLEDGMENTS

The set-up and desiderata of the data collection sessions were discussed and agreed upon between Catriona MacDermid, Camilla Eklund, Jaan Kaja, Mats Wirén and the author. Most of the work was carried out by Catriona and Camilla, with the assistance of Nils Meinhard. Thanks also to Carina Ekedahl and Lennart Svanfeldt at the travel agency Nyman & Schultz AB for participating in the human–human data collection. Also thanks to Eva Lindström and Anders Lindström for comments on draft versions of this article.

REFERENCES

[1] Becket, R., P. Boullion, H. Bratt, I. Bretan, D. Carter, V. Digalakis, R. Eklund, H. Franco, J. Kaja, M. Keegan, I. Lewin, B. Lyberg, D. Milward, L. Neumeyer, P. Price, M. Rayner, P. Sautermeister, F. Weng & M. Wirén. *Spoken Language Translator: Phase Two Report.* Telia Research AB and SRI International, 1997.

[2] Eklund, R. Interaction between prosody and discourse structure in a simulated man–machine dialogue. *Journal of the Acoustical Society of America*, Vol. 102, No. 5, Pt. 2, December 1997, 3202 [Abstract], 1997.

[3] Eklund, R. & E. Shriberg. Crosslinguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. *Proceedings of ICSLP 98*, Sydney, November 30–December 5, Paper 805, Vol. 6, 2631-2634. CD-ROM available from Causal Productions Pty Ltd, PO Box 100, info@causal.on.net, 1998.

[4] Hemphill., C.T., J.J. Godfrey & G.R. Doddington. The ATIS Spoken Language Systems Pilot Corpus. *Proceedings of DARPA Speech and Natural Language Workshop*, 96–101, 1997.
`http://www.ldc.upenn.edu/readme_files/atis/sspcrd/corpus.html`

[5] MacDermid, C. & C. Eklund. *Report on the First WOZ Simulation for the SLT-DB Project.* Technical report, Telia Research AB, 1997.

[6] MacDermid, C. & C. Eklund. *Simulering av en automatiserad översättningstjänst för resebokningar.* Technical report, Telia Research AB, 1996.

[7] Shriberg, E. *Preliminaries to a theory of speech disfluencies.* PhD thesis, University of California, Berkeley, CA, 1994.

WHICH SPEAKERS ARE MOST DISFLUENT IN CONVERSATION, AND WHEN?

Heather Bortfeld[♦], Silvia D. Leon[♣], Jonathan E. Bloom[♣], Michael F. Schober[♣], and Susan E. Brennan[♠]
[♦]*Brown University, Providence, RI;* [♣]*New School for Social Research, New York, NY;* [♠]*State University of New York at Stony Brook, Stony Brook, NY*

ABSTRACT

We examined disfluency rates in a corpus of task-oriented conversations [1] in which several factors were manipulated that could affect fluency rates. These factors included: speakers' age (young, middleaged, and older), task roles (director vs. matcher), difficulty of domain (abstract geometric figures or tangrams vs. photographs of children's faces), relationship between speakers (married vs. strangers), and gender (each pair consisted of a man and a woman). Older speakers produced only marginally higher (combined) disfluency rates than young and middleaged speakers. Overall, disfluency rates were higher both when speakers took the initiative and when they discussed tangrams, associating disfluencies with an increase in planning difficulty. However, fillers (such as *uh*) were distributed somewhat differently than repetitions and restarts, supporting the idea that fillers may be a resource for or a consequence of interpersonal coordination.

1. INTRODUCTION

Speech is notoriously disfluent [2]. Although disfluencies may not thwart speech comprehension, they are interesting for several reasons. First, they pose a problem for most theories of parsing, which are designed to handle only grammatical or "well-formed" utterances [3]. Second, by demonstrating how speech planning and articulation break down, disfluencies provide useful data about the architecture of the speech production system and the constraints upon it [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Third, in certain circumstances, disfluencies can display metalinguistic information to listeners about a speaker's confidence [15], inform listeners about a speaker's planning difficulties [16, 17], or, possibly, serve as devices for coordinating conversational interaction [18, 19, 20, 21]. Last but not least, spontaneous human speech contains disfluencies that create problems for speech recognition systems [22, 23, 24, 25, 20, 26, 27].

In this paper we investigate various situational and demographic factors that have been argued to affect speakers' disfluency rates. Rather than comparing disfluency rates across different corpora, where differences in rates might reflect differences in the circumstances of data collection or in coding criteria, we examine disfluency rates within one corpus of conversations [1].

WHAT MAKES SPEECH DISFLUENT?

Processing load. Speech errors and disfluencies produced by normal speakers have been studied as a window into the intermediate linguistic products and cognitive processes of speech planning since the 1950s [e.g., 4, 5, 6, 7, 8, 29, 9, 10, 11, 12, 13, 14, 30, 19]; these studies provide systematic evidence of how articulatory processes break down under increased processing load. Recently, additional evidence associating disfluencies with increased processing load has turned up in descriptive studies of speech corpora. In Oviatt's [25] study of disfluencies in six types of task-oriented conversations, long utterances had higher disfluency rates than short ones. This finding is supported by Shriberg's [20] study of disfluencies in three different task-oriented conversational corpora; the longer the sentence, the higher the rates of repeated or deleted words. The association of disfluencies with planning load is consistent with findings that disfluencies are more likely near the beginnings of turns or sentences, where planning effort is presumably higher (Boomer [31] found more fillers and silent pauses, and Shriberg [20] found more fillers, repetitions, and deletions).

The topic or domain of a conversation is another way in which the planning load of utterances may vary. In one study, social science lectures contained more disfluencies of one sort–fillers–than hard science lectures, and humanities lectures contained the most of all [32]. These findings were not due to individual differences, because rates for individual speakers did not differ when speakers all addressed the same topic. Schachter and colleagues [17] suggested that speakers use more fillers when they must choose from a larger vocabulary.

Coordination functions. A function that has been proposed for some types of disfluencies is a communicative one: Certain disfluencies may provide information that enables two people in conversation to better manage their interaction or coordinate their mental states [16, 15, 9, 20]. For instance, *time* is a resource that people manage jointly in conversations, and managing resources involves making tradeoffs. If a speaker takes a long time to produce an utterance, she risks losing her addressee's attention or her speaking turn; but if she rushes to produce one that is defective, she risks being misunderstood [32]. So she may warn her addressee of a delay in producing a word by uttering a filler such as *um*, *uh*, *ah*, and *er* [33]. There is evidence that fillers can perform this sort of function: Speakers answering general knowledge questions display accurate information about their mental search processes [15, 36]; that is, they pause longer and use more fillers before producing an answer that they lack confidence in (and that is more likely incorrect) than before one that they have a strong feeling of knowing (and that is more likely to be correct). And they pause longer and use more fillers before a *non*-answer (e.g., "I don't know") when they actually *do* know the answer but are just unable to retrieve it. This metacognitive display is a communicative one because listeners can use it to judge how likely the speaker is to know the correct answer [15].

A filler may also warn a listener that the speaker has just mispoken. Listeners were faster and more accurate in comprehending utterances like *Move to the yel- uh, orange square* when the interrupted word was followed by *uh* than when it was not [16]. Comprehension was also faster with the disfluency (e.g., *yel-uh*) than when it was replaced with an unaccounted-for pause of equal length.

There is at least one other way in which fillers may help coordinate conversation; fillers may help people to manage turn-taking. Perhaps they act as a turn-keeping cue by blocking listeners from interrupting the speaker with a new speaking turn [19]. The story of how fillers affect turn-taking may be more complicated than one where an *uh* simply helps a speaker keep an addressee from interrupting. In Wilkes-Gibbs's corpus of conversational completions (where one speaker spontaneously completed another's utterance), it appeared that fillers were sometimes interpreted as displays of trouble and requests for help, as in this example [33, 21]:

A: *and number 12 is, uh, ..*
B: *chair.*
A: *with the chair, right.*

Here B may have taken *uh* to be a request by A for help in producing the right word; if this is so, then the disfluency was used as an interactive tool. If fillers warn addressees that the speaker is still working on the utterance, then this may result in the addressee chiming in (if he can help with the speaker's problem), and otherwise waiting for the speaker to continue (if he cannot).

The idea that fillers display speakers' difficulties to addressees is not incompatible with Schachter et al's [17] finding of higher filler rates in domains with more indeterminacy. That is, when choosing words is more difficult, a speaker's need to account to her audience for any delays is presumably greater.

The idea that fillers may serve (at least in part) as a resource for interpersonal coordination is consistent with Kasl and Mahl's (unexpected) finding of a 41% increase in fillers (but not other kinds of disfluencies) in audio-only conversations between people in different rooms over conversations in the same room with visual contact [34]. Consistent with this, Oviatt [25] found that people talking on the telephone produced more disfluencies than those talking face-to-face, 8.83 to 5.50 (although she did not present filler rates separately from total disfluency rates, which also included corrections, false starts, and repetitions). Differences in disfluency rates in conversations conducted over different media may, then, be influenced by the resources these media offer for coordination.

Additional evidence that certain disfluencies are associated with coordination between speakers and listeners is gleaned from their relative distributions in speech with and without interactive partners. For instance, conversational speech is more disfluent than monologue speech; in Oviatt's [25] study there were more disfluencies in dialogues (5.50–8.83 disfluencies per 100 words) than in monologues (3.60 per 100 words).

Familiar vs. strange conversational partners. The Schober & Carstensen corpus allows us to examine whether people are more or less disfluent when talking to strangers than when talking with their spouses. The predictions to be made are unclear. On the one hand, one might expect people to be more disfluent with strangers than with intimates, because they might be more anxious with strange partners; higher disfluency rates are associated with anxiety [47]. On the other hand, to the extent that disfluencies are coordinating devices, one might expect people to be more disfluent with intimates; perhaps intimates are more likely to display their planning problems to each other and perhaps strangers plan what they say to each other more carefully.

Age. Age-related changes in cognitive, motor, and perceptual functioning may affect speech in several relevant ways. During naming tasks, speakers in their sixties have more difficulty and are slower at retrieving words than younger speakers, although the ability to define words remains intact and may even improve with age [35]. And speakers over fifty appear to use more elaborate syntactic forms than younger speakers [35]. Such age-related changes seem likely to make conversation more effortful and to generate more disfluencies. The Schober and Carstensen corpus allows us to examine how age affects disfluency rates over a fairly wide range of ages.

Gender. In Shriberg's [20] study, men produced more fillers than women did, but the sexes were equal with respect to other types of disfluency rates. Shriberg cautiously suggested that using more fillers may be a way for men to try to hold on to the conversational floor, but pointed out that in her corpora, gender was confounded with occupation and education level. In the Schober and Carstensen corpus, socio-economic status was balanced across gender, and so we can examine whether Shriberg's observation is corroborated.

Effects of these variables upon disfluencies. It is likely that the mapping of factors like cognitive load, addressee characteristics or relationship, communication medium, or speaker characteristics (such as state of arousal, age, or gender) onto disfluent speech is not a simple one. These factors may not operate independently to produce disfluent speech, but may do so in concert.

Another way in which cognitive, social, and situational factors may relate to speech production in a complex way is that different disfluencies may arise from quite different processes and situations. As we proposed earlier, perhaps some disfluencies serve an interpersonal function, such as displaying a speaker's intentional or metacognitive state to a partner, while others simply represent casualties of an overworked production system.

METHOD

Design

The corpus contained approximately 200,000 words uttered in 48 conversations, transcribed in detail, and double-checked for accuracy. In these conversations, 16 pairs of young speakers (mean age, 28.8 years), 16 pairs of middle-aged speakers (47.9 years), and 16 pairs of older speakers (67.2 years) discussed objects from a familiar domain—photographs of children--and an unfamiliar domain--black and white abstract geometric forms known as tangrams. These 48 pairs of speakers comprised 24 pairs of male and female strangers and 24 married couples, divided equally by age. They were recruited through the Stanford Alumni Association to participate for pay in a referential communication study [1]. All were married and college educated, none had significant hearing loss, and the three age groups were no different in years of postsecondary education.

There were two sets of picture cards, one for each domain (12 children and 12 tangrams). There were 4 trials; with each trial, the members of a pair alternated in the roles of director and matcher. During a trial, each member of a pair had an identical copy of a set of picture cards; the task was for the matcher to get all 12 picture cards lined up in the same order as the director's cards. Members of a pair were visually separated but could communicate freely. Half of the time, they matched children first, and half of the time, tangrams. Half of the time, females performed as the first director, and half of the time, males did. Each of the two sets of pictures were matched twice. In sum, the factors that varied systematically included the relationship between the speakers (married/strangers), age (young/middle-aged/older), topic domain (familiar/unfamiliar objects), role (matcher/director), and gender.

Coding

Speech disfluencies were categorized using a computer software program, Sequence, a HyperCard-based program for the Macintosh that enables segmenting and coding types, numbers, and sequences of behavioral events. Speech was coded as disfluent if it contained any of the following: Repeats ("just on the left left side"), restarts (e.g., "imme- just below the left side"), fillers (e.g., *uh*, *ah*, *um*, *er*), or other editing expressions (e.g, *I mean*, *rather*, *that is*, *sorry*, *oops*). Each repeat or restart was coded as one disfluency, even if the repeated or repaired phrase consisted of more than one word. When none of these kinds of disfluencies were present, a phrase was coded as fluent; when it was unclear whether a phrase contained a disfluency, it was coded as unknown.

The corpus transcript was divided into halves based on the number of words, and a different team of coders coded each half. In addition, each team coded an additional 6 trials from the other team's half; these comprised Trials 2 and 4 from one pair from each of the 6 between-subjects cells of the experimental design (Relationship X Age). So 12.5 % of the trials were double-coded. The coders were blind to which cells of the experimental design the speakers were in. Interrater reliability was excellent; there was 92.8% agreement, with a Cohen's Kappa of .91.

RESULTS

We began by examining word counts for the different types of speakers and conversations in this corpus; word counts in referential communication are assumed to be related to cognitive effort or task difficulty [33, 36]. Then we tested the effects of task role, topic, age, relationship, and gender on disfluency rates.

Word counts

Directors produced over twice as many words as did matchers, $F(1,82) = 99.23$, $p < .001$. This is as we expected, since in the director role, people typically took the initiative for establishing the identity and location of target cards and spent most of their time describing the target object, while in the matcher role, people spent much of their time giving feedback and searching for the target.

The domain of discussion mattered as well; pairs of speakers used over two and a half times as many words when discussing tangrams than pictures of children, $F(1,82) = 252.18$, $p < .001$. This domain difference was greater for matchers than for directors, interaction, $F(1,82) = 14.97$, $p < .001$. Another way to look at this is that in conversations about tangrams, the more taxing domain, matchers appeared to take on more responsibility for getting things understood, uttering 34% of the words as opposed to only 27% in conversations about faces.

There was no difference in the number of words uttered by men vs. women, $F(1,82) = .03$, n.s., nor for married couples vs. strangers, $F(1,82) = 1.26$, n.s. However, there were reliable word count differences by age: fewer words per round were uttered by younger (363) than by middle-aged (549) and middle-aged than by older (580) people, linear trend, $F(1,82) = 13.01$, $p < .001$. And there was a strong age-by-domain interaction: word counts increased more sharply with age in conversations about tangrams than conversations about children, linear trend, $F(1,82) = 11.04$, $p = .001$.

Disfluency rates

The overall disfluency rates we report are per hundred words and consist of repeated words or phrases, restarts, and fillers, unless otherwise noted. Speakers produced, on average, 5.77 of these types of disfluencies every 100 words. This is within the range found by previous studies for these types of disfluencies.

As predicted, disfluencies increased when speakers were faced with heavier planning demands. This difference emerged in two ways: first, for the task roles of director vs. matcher, and second, for unfamiliar vs. familiar domains. In the role of director, speakers produced about 6.76 disfluencies per 100 words vs. 4.78 in the role of matcher, $F(1,82) = 95.87$, $p < .001$. To break these disfluencies down further: directors produced more fillers than matchers, 3.22 vs. 1.77, $F(1,82) = 106.72$, $p < .001$, more restarts, 2.06 vs. 1.64, $F(1,82) = 21.46$, $p < .001$, and slightly but not reliably higher rates of repetitions, 1.48 vs. 1.37, $F(1,82) = 1.81$, n.s. When speakers discussed tangrams, they produced greater rates of disfluencies than when they discussed pictures of children, 6.16 vs. 5.38, $F(1,82) = 13.47$, $p < .001$. This effect was due mainly to repetitions (1.72 vs. 1.13, $F(1,82) = 8.31$, $p = .005$) and to a lesser extent, restarts (2.13 vs. 1.58, $F(1,82) = .84$, n.s.). For fillers, the difference was in the opposite direction: speakers produced slightly but reliably higher filler rates while describing faces than tangrams, 2.67 vs. 2.32, $F(1,82) = 7.85$, $p < .01$.

Older speakers produced only marginally higher disfluency rates (6.42, with repetitions, restarts, and fillers combined) than middle-aged (5.46) and younger (5.43) speakers, linear trend, ($F(1,82) = 3.61$, $p = .06$). The important factor appeared to be whether the speaker was in the older group (which ranged from 63 to 72 years of age) as opposed to in one of the other two age groups; there was no difference between the younger and middle-aged groups. This was despite that fact that the middle-aged pairs uttered more words than the young pairs and similar amounts as the older pairs.

Married couples were no more fluent in their conversations than were strangers; there were no differences by relationship in rates of restarts, repetitions, or fillers. This is contrary to what would be expected if experience or comfort with a partner were to increase fluency, or if anxiety evoked by conversing with an unfamiliar partner were to increase disfluencies. It is also contrary to what would be expected if strangers planned their speech more carefully than intimates, or if certain disfluencies are coordination devices that only intimates can use to elicit help from their partners.

Recall that Shriberg [20] found that men produced more fillers than women did. To see if our data would replicate this finding, we included speaker's gender in our comparisons. We found that while men produced no more words than women did, they had a higher rate of disfluencies overall, 6.57 to 4.97, $F(1,82) = 13.74$, $p = .001$. Why should men be more disfluent than women? When we broke this difference down further, it was due mainly to higher rates of fillers, 2.96 to 2.03, $F(1,82) = 12.86$, $p = .001$ and repetitions, 1.67 to 1.18, $F(1,82) = 8.31$, $p = .005$.

Although men produced more repetitions and more fillers than women did, these two types of disfluencies did not pattern the same with respect to other variables. Gender did not interact with role for repetition rates as it did for filler rates. And while overall, filler rates were lower when speakers discussed tangrams than when they discussed faces (2.32 to 2.67), both repetition rates and restart rates were higher (1.72 to 1.13 and 2.12 to 1.58). These distributions support the notion that fillers arise from different processes than the other types of disfluencies we coded.

General Discussion

The corpus of conversations we examined balanced task role (director vs. matcher), difficulty of domain (abstract geometric figures vs. photographs of children's faces), relationship between speakers (married vs. strangers), and gender (each pair of speakers consisted of a male and a female). This design enabled us to make direct comparisons of disfluency rates across conditions, unlike studies that have made comparisons across corpora [25].

From our data, we advance two main conclusions. First, the distributions of disfluencies in this corpus support the idea that some but not all disfluency rates increase as heavier demands are placed on the speech planning system. Second, fillers were distributed somewhat differently than repetitions and restarts, suggesting that they may also be related to processes of interpersonal coordination. If fillers help speakers coordinate with their addressees (e.g., by displaying delays in producing utterances and perhaps by soliciting help), then we should expect directors, who take most of the initiative in a matching task, to produce more fillers than matchers, and this was consistently true across both domains and for both sexes.

References

[1] Schober, M., & Carstensen, R. (1998). Do age and long-term relationship matter in conversations about unfamiliar things? Manuscript under review.

[2] Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1482-1493.

[3] 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.

[4] Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review, 93*, 283-321.

[5] Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27-52.

[6] Fromkin, V. A. (Ed.) (1973). *Speech errors as linguistic evidence*. The Hague: Mouton Publishers.

[7] Fromkin, V. A. (1980). *Errors in Linguistic Performance: Slips of the tongue, ear, pen and hand*. New York: Academic Press.

[8] Garrett, M. F. (1975). The analysis of sentence production. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9), (pp. 133-177). New York: Academic Press.

[9] Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

[10] MacKay, E. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia, 8*, 323-350.

[11] MacKay, E. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology, 3*, 210-227.

[12] MacKay, E. G. (1973). Complexity in output systems: Evidence from behavioral hybrids. *American Journal of Psychology, 86*, 785-806.

[13] Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial order mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Ed.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295-342). Hillsdale, NJ: Lawrence Erlbaum.

[14] Shattuck-Hufnagel, S., & Klatt, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18*, 41-55.

[15] Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34*, 383-398.

[16] Brennan, S. E., & Schober, M. F. (1998). When do speech disfluencies *help* comprehension? Manuscript under review.

[17] Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology, 60*, 362-367.

[18] Brennan, S. E., & Kipp, E. G. (1996). An addressee's knowledge affects a speaker's use of fillers in question-answering. *Abstracts of the Psychonomic Society, 37th Annual Meeting* (p. 24), Chicago, IL.

[19] Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15*, 19-44.

[20] Shriberg, E. (1996). Disfluencies in switchboard. *Proceedings, International Conference on Spoken Language Processing*, Vol. Addendum, 11-14. Philadelphia, PA, 3-6 October.

[21] Wilkes-Gibbs, D. (1986). Collaborative processes of language use in conversation. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

[22] Butzberger, J. W., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. In M. Marcus (Ed.), *Proceedings, DARPA Speech and Natural Language Workshop* (pp. 339-343). Morgan Kaufmann.

[23] Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting, Association for Computational Linguistics*, Cambridge, MA, pp. 123-128.

[24] Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America, 95*, 1603-1616.

[25] Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language, 9*, 19-35.

[26] Shriberg, E., Bear, J., & Dowding, J. (1992). Automatic detection and correction of repairs in human-computer dialog. In M. Marcus (Ed.), *Proceedings, DARPA Speech and Natural Language Workshop* (pp. 419-424). Morgan Kaufmann.

[27] Shriberg, E., Wade, E., & Price, P. (1992). Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In M. Marcus (Ed.), *Proceedings, DARPA Speech and Natural Language Workshop* (pp. 49-54). Morgan Kaufmann.

[28] Carletta, J., Caley, R., & Isard, S. (1993). A collection of self-repairs from the map task corpus. Technical Report.

[29] Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology, 10*, 96.

[30] Nooteboom, S. G. (1969). The tongue slips into patterns. In A. G. Sciarone, A. J. can Essen, & A. A. van Raad (Eds.), *Nomen (Society): Leyden studies in linguistics and phonetics*. The Hague: Mouton Publishers, pp. 114-132.

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

[32] Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L.B. Resnick, J. Levine, & S.D. Teasley (Eds.), *Perspectives on socially shared cognition*. Washington, DC:APA. Reprinted in R. M. Baecker (Ed.), *Groupware and computer-supported cooperative work: Assisting human-human collaboration*. San Mateo, CA: Morgan Kaufman Publishers, Inc.

[33] Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

[34] Kasl, S. V., & Mahl, G. F. (1965). The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology*, 1, 425-433 [47] Mahl, G. F. (1987). Explorations in nonverbal and vocal behavior. Hillsdale, NJ: Erlbaum.

[35] Obler, L., & Albert, M. L. (1984). Language in aging. In M. L. Albert (Ed.), *Clinical neurology of aging* (pp. 245-253). New York, NY: Oxford University Press.

[36] Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes, 23*, 119-147.

WHY DOES SPONTANEOUS SPEECH UNFOLD IN TEMPORAL CYCLES, SOMETIMES?

Kim Kirsner, Benjamin Roberts & Yong-Heng Lee
University of Western Australia

ABSTRACT

Spontaneous speech typically consists of alternating periods of continuous fluency, where fluency refers to the ratio of speech to pausing. Individual differences in fluency are substantial, with mean pause per minute ranging from less than 20 to more than 40 sec per minute in our sample of English and Mandarin speakers. While pauses have been regarded as critical clues for psycholinguistic analysis for decades, the existence of temporal cycles have been subject to extensive debate. The results of our experiments provide strong support for the presence of temporal cycles in spontaneous speech, and demonstrate in particular that fluency declines and increases prior and subsequent to topic shifts respectively. The source of temporal cycles is unclear, however. The prevailing assumption is that they reflect alternating periods of high level macro-planning, associated with low fluency, and low level micro-execution, associated with high fluency. However, a variety of alternative explanations merit consideration.

How do we talk to each other? Are the critical processes implemented automatically, or does speech depend on the allocation of scarce cognitive resources?

This paper is concerned with discourse. Spontaneous speech typically consists of alternating periods of continuous speech and pauses where fluency is used to refer to the speech : pause ratio. However, while pauses have been regarded as critical clues for cognitive and psycholinguistic analysis for decades, the further claim that spontaneous speech consists of alternating phases of high and low fluency has been the subject of extensive debate. The general interpretation, cautiously endorsed by Garrett [5] and Levelt [8] is that temporal cycles reflect alternating periods of conceptualization or macro-planning, an activity that is associated with low fluency, and low level micro-planning or execution, an activity that involves high fluency. The challenge to temporal cycles has been formidable however, including suggestions that the observed effects reflect random variation in fluency.

The first experiment was implemented to test the validity of temporal cycles. The analytic procedures were developed to test the null hypothesis, that fluctuation in fluency stems from random processes. Historically, the evidence for and against temporal cycles has been substantially based on subjective methods involving the slopes of successive functions depicting the relationship between pause duration and speech segment duration for speech units in spontaneous discourse. Fluency was calculated by reference to the proportion of speech in successive 200 msec samples, the time method, or by reference to differences between standardized durations of segments of continuous speech and the standardized durations of the pause segments which preceded the speech segments, the speech unit method. Analysis of 34 two - three minute discourse samples from eight subjects revealed that in approximately 70% of cases a cyclical pattern based on autocorrelation involving the relationship between fluency and either time or speech units explained a significant proportion of the variance, with a mean period of 19.1±7.4 seconds.

Figure 1 depicts temporal cycles in fluency for a discourse sample of approximately 50 speech units over a period of about 90 seconds. The text boxes show the speech associated with each of the seven successive speech segments that straddled the marked topic shift.

'The second experiment was implemented with six Mandarin first language speakers. The speakers provided seven minute discourse samples on each of three routine topics of conversation. Autocorrelation was used to examine the relationship between fluency and the time-based samples and a significant proportion of the variance was explained by a cyclical pattern in the majority of samples. The results with Mandarin therefore support those obtained with English, and suggest that temporal cycles in spontaneous speech may be observed in a language that is etymologically remote, a finding which suggests that the phenomenon is universal.

Topic shift has been identified as the critical determinant of fluency cycles. The assumption is that cognitive work on macro-planning peaks during topic shifts, and that macroplanning competes for scarce cognitive resources with some micro-planning processes. To explore the relationship between topic shift and fluency, we used two independent readers to identify the topic shifts in the discourse from Experiment 1 (percent agreement = 81%, kappa = 0.73, $p < 0.01$), and then explored the fluency dynamics of the speech segments adjacent to topic shifts. With

consideration restricted to cases where even micro-planning or execution is not automatic, or

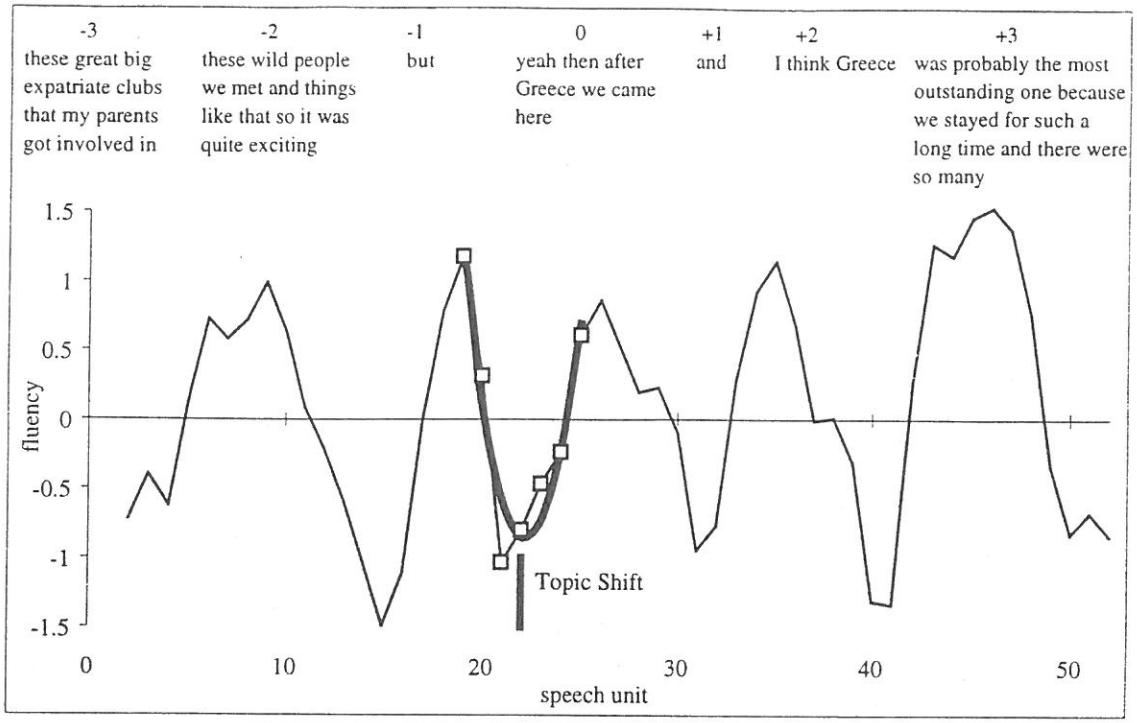

Figure 1. Topic Shift and fluency with quadratic function fitted to speech units surrounding topic shift. Numbers represents speech units relative to topic shift in the marked quadratic.

there were at least three independent speech units including a speech segment and pause on either side of the segment that included the topic shift, a quadratic function was fitted to mean fluency for the seven critical segments for each subject. The shape of the quadratic therefore reflects the presence or otherwise of fluency minima at the topic shifts. The topic shifts coincided with the minima for all subjects. The mean squared term of the quadratic functions for eight subjects was 0.063, and this value differed significantly from zero, $t(7) = 8.6$, $p<0.01$. There is therefore a systematic trend for fluency to increase and decrease before and after each topic shift, respectively. This result is also summarized in Figure 1. The figure shows the fluency analysis for a single sample of speech, including a short temporal cycle, a topic shift, and the quadratic function for the relevant discourse sample.

This mode of analysis is also being implemented with Mandarin, and, given evidence that the structure of Mandarin is dominated by the distinction between topic declaration and topic elaboration, analyses are also being implemented with a classification based on this variable.

According to the modal explanation of temporal cycles, the reduction in fluency around topic shifts reflects competition for scarce cognitive resources, an account which involves the additional assumption that else it would not compete with macro-planning. However, for this explanation to be valid, it must also be assumed that macro-planning and micro-planning compete for the same pool of cognitive resources, or else the reduction in fluency associated with macro-planning would not be evident.

Perhaps the clearest evidence that spontaneous speech involves cognitive work comes from a study implemented by Greene and Lindsey [7]. The critical manipulation in this study involved the contrast between single and dual goal conditions. In the former, subjects simply explained to a job applicant that their application was unsuccessful, whereas, in the dual task condition, they were required to convey concern and encouragement to the applicant at the same time as they were informing them that their application had been unsuccessful. Greene and Lindsey reported a significant reduction in fluency under dual task conditions. Perhaps the most vexing problem with the analysis of fluency variation in normal speakers involves the possible impact of style. Thus, although it is possible that the dual goal manipulation used by Greene and Lindsey, it is also possible that the effect reported by them reflects manipulation of style rather than workload. Duez [2], for example, in a study involving French politicians, reported that fluency was significantly reduced when the politicians were making a prepared speech in public as distinct from

answering questions.

A third experiment was implemented to test the proposition that the generation of spontaneous speech is insensitive to cognitive workload. Sixty subjects were shown a picture of the face of a person unknown to them, for 20 seconds, and then expected to create a story about the life of that person. Following the dual goal manipulation used by Greene and Lindsey [7], half of the subjects were warned that they would later be asked to recall their story, and half were not. This manipulation was designed to mimic the single and dual goal manipulation used by Greene and Lindsey, while avoiding the stylistic changes that might have accompanied their manipulation. There is in addition evidence from systematic dual task studies that memorization is a cognitively expensive task [1]. The results were clear. The additional memorization required produced a small and non-significant reduction in fluency, from a mean speech-to-pause ratio of 1.91 to 1.85. The assumption that spontaneous speech production involves cognitively expensive workload, while intuitively appealing cannot be taken for granted, and requires systematic analysis.

One additional finding of interest involves individual differences in fluency. Although the distinction between fluent and dysfluent aphasics was developed more than 30 years ago, and early studies included systematic analyses of fluency, there is virtually no work on individual differences in fluency in spontaneous speech, a remarkable comment given the prominence of individual differences in this variable, and its role in classification. While the experiments reported here used different procedures for eliciting spontaneous speech, the individual differences within each study are substantial. Thus, on average, each minute of discourse in Experiment 1 included a mean of 32.7±6.4 sec of pause, and similar values were obtained in Experiment 2, at 29.9±6.2 sec. Across the two experiments, the mean values for this parameter ranged from 19 to 40 sec. In future work we plan to measure individual differences in fluency in order to explore the relationship if any between this variable and working memory, and to provide normative data for the analysis of dysfluency in aphasia.

The research introduced in this paper has been influenced by several broad issues. The first of these involves the validity and generality of temporal cycles. It is now evident that spontaneous speech unfolds in alternating phases of high and low fluency. However it is not yet clear whether the underlying pattern is periodic or aperiodic, or whether non-detection under certain conditions reflects variation in a component process, involving working memory for example, or a detection problem associated with the signal to noise ratio of temporal cycles. It is possible for example that the underlying phenomenon is aperiodic except when the system approaches its boundary condition, when periodic effects may be observed.

The second issue involves the basis of temporal cycles. The source is not clear. According to Levelt [8], for example, temporal cycles reflect alternation between (macro-)planning, a process that is under executive control, and micro-planning, implementation of formulation and articulation (Levelt, 1989, p 126). However this claim is difficult to reconcile with the assumption that formulation and articulation are automatic processes, a claim made elsewhere by Levelt [8, p21]. There may be a paradox here. If micro-planning involves automatic processes, it should not compete with macro-planning, particularly if, as seems likely, the latter involve qualitatively different processes and, therefore, the processing conditions that should minimize competition for scarce resources. The assumption that temporal cycles in spontaneous speech involve a scheduling problem -- between macro- and micro-processes -- raises questions that stand outside the mainstream on cognitive models of scheduling. For one thing, such models have usually been applied to problems that are resolved in less between 1/10th to 1/100th of the time associated with the low fluency phase in spontaneous speech. For another, recent evidence suggests that scheduling is provoked by competition between peripheral information processing systems rather than central systems, a key element in Levelt's account.

The assumption that language production depends on voluntary control processes may also be inferred from Garrett [5]. According to Garrett "non-fluent speech is ever-present and breaks in fluency may be regarded as indication that rate of speech output has overrun the rate of *decision making* either about what we will say or how we will say it. Hence idea that distribution of hesitation types will reveal basic organizational features of language" (italics supplied). According to this view then, language production involves continuous competition between scarce resources of various types, where fluency is a direct measure of the efficiency of one or more decision processes.

Temporal cycles can be explained without reference to competition between macroplanning and microplanning, however. According to Greene for example [7], pauses in the low fluency phase of spontaneous speech reflect the operation of assembly processes, processes. which determine the order, content and timing of information extracted from a mixture of more and less accessible records. According to this account, then, declining fluency before topic shifts reflect the fact that the next packet is not ready for transmission, a state that might lead to a lower speech to pause ratio if speaker s tried to maintain control while they were waiting for completion of assembly operations. According to this point of view, then, low fluency phases and therefore temporal cycles are more likely to occur when less

accessible or more complex records must be used in production, an assumption for which there is considerable support at the lexical level at least.

Each of the first two accounts is consistent with the assumption that the capacity of working memory is a critical determinant of fluency, although the level of the assumed relationship is unclear. One possibility involves the phonological loop [6] But the duration of the units involved in this system correspond more closely to the duration of speech segments than cycles. A second possibility involves a model of working memory more akin to that described by Eriksson and Kintsch [3], where working memory includes cues to information in more permanent storage systems.

A closely related question concerns the role of consciousness in language production. But is consciousness part of the explanation or part of the data? According to one point of view, working memory and by extension conscious planning and decision-making play a critical role in language production. For Garrett [5, p37] this is explicit, and "voluntary intellectual effort coincides with subjective experience of 'decision making micro-crises'". However this is not the only possible account. Another possibility is that consciousness is relevant to the selection of communication targets, and monitoring, but that it plays no other direct role in the planning and production of language. According to this point of view then, conceptualization apart, we become aware of what we have said after rather than we have said it. This is, intriguingly, a point of view that has attracted interest in research into artificial problem solving systems. According to Michie [10], for example, problem-solving is primarily the work of visualization supported by automatized skills, and consciousness operates at the level of goal-setting and monitoring, and of the construction and communication of after-the-event commentaries, not the critical problem solver. Essentially the same idea can be extracted from Fauconnier [4, p1], where he states that, "visible language is only the tip of the iceberg of invisible meaning construction that goes on when we think and talk".

We cannot answer the question posed in the title of this paper. Temporal cycles could reflect scheduling owing to competition between cognitively expensive processes where cycles only emerge when load exceeds some threshold, delay in the time required to assemble complex production processes, or discontinuity in more basic processes concerned with the management and control of effector systems generally. And, finally, the role of consciousness is unclear though of great interest. Are conscious contributions restricted to target selection and monitoring, or are they an integral element in a broad range of micro- and macro-planning processes?

REFERENCES

[1] Craik, F.I.M., Naveh-Benjamin, M., Govoni, R. & Anderson, N.D. (1996). The effects of Divided Attention on Encoding and Retrieval Processes in Human Memory. Journal of Experimental Psychology: General, 125 (2), 159-180.

[2] Duez, D. (1982). Silent and non-silent pauses in three speech styles. Language and Speech, 25(1), 11-28.

[3] Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. Psychological Review, 102(2), 211-245.

[4] Fauconnier, G. (1997). Mappings in thought and language. Cambridge, Mass.: Cambridge University Press.

[5] Garrett, M. F. (1982). Production of speech: Observations from normal and pathological langauge use. In A. Ellis (Ed.), Normality and pathology in cognitive functions, . London: Academic Press.

[6] Gathercole and Baddeley, A.D. (1993). Working memory and language. Lawrence Erlbaum.: Hove.

[7] Greene, J. O., & Lindsey, A. E. (1989). Encoding processes in the production of multiple-goal messages. Human Communication Research, 16(1), 120-140.

[8] Levelt, W. J. M. (1989). Speaking: From intention to articulation. Cambridge, Mass.: MIT Press.

[10] Michie, D. (1995). Consciousness as an engineering issue, Part 2. Journal of Consciousness Studies, 2(1), 52-66.

Between-Turn Pauses and Ums

Jean E. Fox Tree
University of California, Santa Cruz

ABSTRACT

Pauses and *um*s are often treated as two versions of the same thing, with the traditional label for *um*s, filled pauses, emphasizing this seeming interchangeability. To explore this hypothesis, I compared how overhearers interpreted a speaker's contribution to a conversation depending on whether the speaker responded immediately, paused and responded, or said *um* and responded. Overhearers answered a series of questions about the turn exchanges they had heard. The questions measured their interpretations of the second speakers' speech production difficulty, honesty, comfort with the topic discussed, familiarity with the interlocutor, and desire to have further contact with the interlocutor. In two experiments, the type of turn exchange was found to influence overhearers' interpretations. Results supply information about both the signalling properties of *um*s and the relationship between *um*s and pauses of varying lengths in the environment of a turn exchange.

1. INTRODUCTION

Researchers have identified a number of potential signals *um*s could provide when classed together as a group with *uh*s. They might aid in utterance processing by displaying information about the upcoming utterance, such as its syntactic structure [11, 12] or its discourse structure [20], or by displaying speakers' production difficulty [3, 5, 6, 8, 12, 14, 16, 18, 21]. They might aid in turn-taking by displaying a desire to hold or gain a turn [11, 15, 17, 18]. Finally, they might provide interpersonal information, such as indicating a speaker's anxiety level [9, 10, for extended review, see 2].

Usually, *um*s and *uh*s are considered to be different pronunciations of the same thing. But there is reason to suspect that *um*s and *uh*s are functionally different. For one thing, *um*s are more common before a long pause than *uh*s are [19]. *Um*s are also more likely than *uh*s to occur at phrase beginnings [20]. *Um*s and *uh*s also have different effects on word monitoring [4]. To investigate the use of *um*s further, in the current research I tested off-line interpretations of *um*s used in turn exchanges.

I compared two kinds of turn exchanges, precisely timed turns and turns separated by 3 s pauses. Long pauses between turns are often viewed as a result of second speaker error. But they can also be described as a product of both participants. The person who left the floor may have done so prematurely or inappropriately. The one who did not take it up may have followed turn-taking cues or not. I focus in the current research on how inter-turn pauses and *um*s affect the perception of the second speaker.

In the experiments presented here, people heard turn exchanges and then answered five questions probing (1) how well overhearers thought the two interlocutors on the tape knew each other, (2) whether overhearers thought the second speaker would be likely to seek further contact with the other person, (3) how much speech production difficulty overhearers thought the second speakers had, (4) how deceptive overhearers thought the second speakers were being, and (5) how comfortable overhearers thought the second speakers were with the topics discussed.

Long inter-turn pauses should result in more negative interpretations than short pauses [13]. But it is not clear how *um* may influence these negative interpretations. More specifically, there is reason to expect that adding *um*s will cause no change in interpretations above and beyond the pauses that follow them, so that an *um* plus short pause will act like a short pause and an *um* plus long pause will act like a long pause. There is also reason to expect that adding *um*s will counteract negative effects of long pauses, so that both an *um* plus short pause will act like a short pause, and an *um* plus long pause will act like a short pause. Finally, there is reason to expect that adding *um*s will make effects more negative.

There are at least two reasons to expect *um*s not to influence interpretations above and beyond the influences of subsequent pauses. One is that *um*s may be filtered out and ignored. The second is that *um*s may be equivalent to pauses, and to the extent that *um*s are closer in length to brief inter-turn pauses of under 1 s as opposed to long inter-turn pauses of 2 s or more, they may not engender negative interpretations, just like brief pauses between turns do not engender negative interpretations.

There are also at least two ways *um*s may counteract the negative effects of long pauses. One is that they may provide an explanation for upcoming pauses [1, 19]. For example, by indicating that speakers are thinking about what to say, a positive quality, *um*s may provide a more favorable interpretation of upcoming pauses than listeners would have come up with in the absence of *um*s. A second is that they made add politeness. *Um*s enable speakers to maintain smooth exchanges even when they are not ready to take the floor, demonstrating speakers' willingness to prevent awkward silences and to do their part in keeping the conversational ball rolling. This positive attribute may offset any negative contributions of a subsequent long pause.

In contrast to predictions of no effect, there are at least two reasons to expect that adding *um*s will make effects more negative. One is that adding an *um* before a long pause rests all the responsibility for negative effects squarely on the second speaker's shoulders. That is, instead of sharing negative effects between speakers in the pause-alone condition, adding an *um* before the pause may shunt all the negativity to the second speaker. With this hypothesis, the negative influence of *um*s may only appear when the *um*s are

followed by long pauses, as short pauses do not have extra negativity to split between speakers or load on one speaker. Another reason *um*s may make effects more negative is that they may in themselves convey negative interpretations.

As a first test, I compared the three turn intervals of a 1 s pause, a 3 s pause, and a .5 s pause plus *um* plus 1 s pause. The 1 s pause is the amount of silence comfortably sustained before a conversational participant tries to fill the gap, either by the current speaker continuing or by the next speaker beginning [7]. The 3 s pause is an uncomfortably long silence [13]. A comparison of these two pause-alone conditions was anticipated to replicate earlier research that uncomfortably long inter-turn pauses yield negative attributions. The third condition, the .5 s pause plus *um* plus 1 s pause, was chosen to test whether adding an *um* made any difference to the interpretations of a pause of acceptable length. More specifically, would the *um* have no effect or a negative effect? If a negative effect, would this effect be as great, or greater, than that of a 3 s inter-turn pause? The *um* was preceded by a .5 s pause to make the turn-exchange sound natural, as an *um* coincident with the offset of the last speaker's turn sounded premature.

As a second test, I compared the three turn intervals of a .5 s pause, a 3 s pause, and a .5 s pause plus *um* plus 3 s pause. This test once again provides end-points for measuring the effects of an *um* on the interpretations of pauses. But this time, the *um* precedes an usually long pause. The 1 s inter-turn pause was reduced to .5 s to more closely match the pause preceding the *um* in the *um* condition and to assure the perception of a smooth turn exchange in the .5 s condition. This experiment explicitly contrasts a long pause by itself, where part of the responsibility for it could be shared by first speaker, to a long pause that follows an *um*, where responsibility for the *um* may be unambiguously attached to the *um*-producer.

The two experiments together will allow teasing apart of the different predictions outlined above. For example, if *um* plus 1 s patterns similarly to 1 s, it may be because *um* is ignored. But if *um* plus 3 s also patterns similarly to a smooth turn exchange, then *um* can be taken to counteract the negative effects of a 3 s pause. Alternatively, if *um* plus 1 s patterns like 3 s, this would suggest that both *um*s and 3 s pauses contribute negatively towards interpretations. But if *um* plus 3 s also patterns similarly to 3 s, then an additional conclusion can be reached: the negative contributions of *um*s and pauses are not additive. Another possible outcome is that *um* plus 3 s is worse than 3 s, suggesting that effects of *um*s and pauses are additive.

2. EXPERIMENT 1

Thirty spontaneously produced turn exchanges were edited to create three versions. For example, the exchange "(Speaker A) Are you here because of affirmative action? (Speaker B) It helped me out a little bit." was digitally manipulated so that the turn interval consisted of either a 1 s pause, a .5 s pause plus *um* plus 1 s pause, or a 3 s pause. Materials were presented to overhearers in three counterbalanced lists such that a particular person heard only one version of a turn exchange. After hearing each trial, overhearers responded to the five questions using a 7-point Likert scale.

Results were that the type of turn exchange affected overhearers' interpretations for every question. As expected, a 1 s interval between turns always led to more positive interpretations than a 3 s interval. With a shorter interval, overhearers rated interlocutors as more familiar with each other and more likely to seek out future contact, and rated the second speakers as having less production difficulty, being more honest, and being more comfortable with the topic discussed.

The *um* plus 1 s interval always fell within the boundaries of the other two intervals. It was never more positive than the 1 s interval, nor more negative than the 3 s interval. With respect to judgements of likelihood of future contact, *um* plus 1 s was more similar to a 1 s interval, but with respect to judgements of familiarity, honesty, or comfort with the topic, it was more similar to a 3 s interval. With respect to judgements of speech production difficulty, *um* plus 1 s fell between 1 s and 3 s, similar to neither. So, when making judgements of likelihood of future contact, *um*s appear to be overlooked, but when making other judgements, *um*s appear to be taken into account, with three of the remaining four questions having *um* 1 s patterning with 3 s.

3. EXPERIMENT 2

The same thirty turn exchanges were edited to create three versions. This time the turn interval consisted of either a .5 s pause, a .5 s pause plus *um* plus 3 s pause, or a 3 s pause. Other aspects of the design and procedure were the same as before.

As with Experiment 1, the type of turn exchange affected overhearers' interpretations for every question asked. Also as with Experiment 1, the .5 s interval between turns always yielded the most positive effects, this time more positive than either other condition for every question.

Like Experiment 1, the *um* interval was at the 3 s boundary for the familiarity and honesty questions. Having an *um* or a 3 s pause made interlocutors appear less familiar or less honest, but having both did not boost the unfamiliarity or dishonesty. For these questions, *um*s will have the same effect whether they are followed by 1 s or 3 s. This would not occur if respondents were seen as taking more responsibility for a pause after an *um* than a pause alone. If this were happening, interpretations should be more negative in the *um* 3 s condition than the 3 s condition.

Unlike Experiment 1, the *um* interval was at the 3 s boundary instead of the 1 s boundary for the likelihood of future contact question. This effect can be seen as a result of the 3 s pause alone; the same interpretations would be predicted were *um* ignored.

Also unlike Experiment 1, the *um* interval did not fall within the boundaries of the other two intervals for the speech production and comfort with topic questions. In fact, for these questions, interpretations appear to be additive. With either an *um* or a 3 s pause, interlocutors were though to have more speech production difficulty or to be more uncomfortable with the topic, and even more so with both.

4. GENERAL DISCUSSION

When people begin speaking after someone else has stopped, they can choose to wait and then start speaking, or to say *um* and then wait, among other things. What implications does each choice have? In two experiments, I compared a 1 s pause, a .5 s pause plus *um* plus 1 s pause, and a 3 s pause (Experiment 1), and a .5 s pause, a .5 s pause plus *um* plus 3 s pause, and a 3 s pause (Experiment 2). In both experiments, precisely timed turns were interpreted more favorably than turn exchanges with 3 s gaps. Smooth turn exchanges were taken to indicate greater familiarity between participants and increased likelihood of future contact, and also to indicate second speakers' relative ease of speech production, greater honesty, and greater comfort with the topics discussed.

*Um*s before turns either had no effect relative to smooth exchanges or made interpretations more negative. In no case did *um*s make interpretations more positive than smooth turn exchanges, nor did they ever counteract the negative effects of 3 s pauses. In fact, in some cases, preceding a 3 s pause by an *um* made the interpretations more negative.

Comparing across experiments, effects of *um*s and pauses appeared additive for judgements of speech production difficulty and comfort with the topic discussed, but nonadditive for judgements of familiarity and honesty. That is, for the additive questions, both an *um* and a long pause contributed negatively to interpretations, with both being worse than either alone. For the nonadditive questions, adding an *um* or a pause contributed negatively, but having both did not push the effects further into a negative direction. For the remaining question on likelihood of future contact, having an *um* seemed to do nothing once the pause effects were taken into account.

Another way to summarize the results is that *um*s with a short pause are interpreted as negatively as a 3 s pause when it comes to judgements of respondents' current motivations and feelings (familiarity, honesty, comfort with topic). They do not have this effect on predictions of future behavior (likelihood of future contact) or on judgements of current cognitive processes (speech production difficulty). *Um*s with 3 s pauses keep judgements of motivations and feelings on the negative side, sometimes as additive effects with pauses, sometimes not. They still do not have an effect on prediction of future behavior, although they now do exacerbate judgements of production difficulty when added to a 3 s pause.

The experiments presented here and the kinds of interpretations probed provide clear evidence that *um*s and pauses are not the same thing, and that *um*s are not ignorable. The experiments also provide evidence that *um*s can effect interpretations at multiple levels and in different ways. For example, *um*s may cause speakers to appear to have production problems, a cognitive attribution, while at the same time causing them to appear deceptive, an interpersonal attribution. Said another way, ascribing *um*s to production problems doesn't appear to override the parallel association of *um*s to deception.

ACKNOWLEDGEMENTS

This research was supported by faculty research funds granted by the University of California, Santa Cruz. Correspondence may be addressed to Jean E. Fox Tree, Psychology Department, University of California, Santa Cruz, Santa Cruz, CA, 95064.

REFERENCES

[1] Christenfeld, N. (1995). Does it hurt to say um? *Journal of Nonverbal Behavior, 19*(3), 171-186.

[2] Christenfeld, N., & Creager, B. (1996). Anxiety, alcohol, aphasia, and *um*s. *Journal of Personality and Social Psychology, 70*(3), 451-460.

[3] Christenfeld, N., Schacter, S., & Bilous, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research, 20*(1), 1-10.

[4] Fox Tree, J. E. (1997). *Listeners' uses of ums and uhs in on-line speech processing.* Paper presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia, PA.

[5] Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition, 62*(2), 151-167.

[6] Jefferson, G. (1974). Error correction as an interactional resource. *Language in Society, 3*(2), 181-199.

[7] Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: An Interdisciplinary Perspective* (pp. 167-195). Philadelphia: Multilingual Matters, Ltd.

[8] Kasl, S. V., & Mahl, G. F. (1987). Speech disturbances and experimentally induced anxiety. In G. F. Mahl (Ed.), *Explorations in Nonverbal and Vocal Behavior* (pp. 203-213). Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

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

[11] Maclay, H., & Osgood, C. E. (1959). Hesitation in spontaneous English speech. *Word, 75*, 19-44.

[12] Martin, J. G. (1967). Hesitations in the speaker's production and listener's reproduction of utterances. *Journal of Verbal Learning and Verbal Behavior, 6*(6), 903-909.

[13] McLaughlin, M. L., & Cody, M. J. (1982). Awkward silences: Behavioral antecedents and consequences of the conversational lapse. *Human Communication Research, 8*(4), 299-316.

[14] Reynolds, A., & Paivio, A. (1968). Cognitive and emotional determinants of speech. *Canadian Journal of Psychology, 22*(3), 164-175.

[15] Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research, 2*(1), 51-81.

[16] Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology, 60*(3), 362-367.

[17] Schegloff, E. A. (1981). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Georgetown University Round Table on Languages and Linguistics 1981* (pp. 71-93). Washington, DC: Georgetown University Press.

[18] Siegman, A. W. (1979). Cognition and hesitation in speech. In A. W. Siegman & S. Feldstein (Eds.), *Of Speech and Time* (pp. 151-178). Hillsdale, NJ: Lawrence Erlbaum.

[19] Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32*, 25-38.

[20] Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*, 485-496.

[21] Tannenbaum, P. H., Williams, F., & Hillier, C. S. (1965). Word predictability in the environments of hesitations. *Journal of Verbal Learning and Verbal Behavior, 4*, 134-140.

UHS AND INTERRUPTED WORDS: THE INFORMATION AVAILABLE TO LISTENERS

Susan E. Brennan* and Michael F. Schober‡

*State University of New York at Stony Brook; ‡New School for Social Research

ABSTRACT

Speech disfluencies are generally assumed to harm comprehension. Our studies investigated whether this is true, or whether certain disfluencies might actually *help* comprehension by marking for listeners which information the speaker intends to repair. We tested two hypotheses: (1) whether an interrupted word signals that the word was produced in error, and (2) whether a filler such as *uh* after an interrupted word signals an error. Listeners heard fluent instructions and disfluent ones whose reparanda contained completed words, interrupted words, or interrupted words with fillers, and then responded to these instructions. Responses to mid-word interruptions were no faster than to between-word interruptions, although there were fewer errors when less of the unintended word was heard. Responses to mid-word interruptions with *uh* were faster and more accurate than controls without disfluencies. With more complex displays, the response time advantage (but not the error rate advantage) diminished, suggesting that an interrupted word followed by *uh* tells listeners what the speaker does *not* mean. A fourth experiment showed that it is not the presence of the *uh* per se, but the additional time after the interrupted word that is the source of this "disfluency advantage."

1. INTRODUCTION

Spontaneous speech is notoriously disfluent--speakers pause, restart, repeat words, and produce non-words such as *er*, *um*, and *uh*--and yet most psycholinguistic experiments and theories of parsing are designed as if speech were entirely fluent [1]. On most views of parsing, disfluencies should make utterances more difficult to process. Speech repairs are assumed to pose a *continuation problem* for listeners [2], who need to filter out or ignore the disfluencies in a reparandum in order to determine which part of an utterance was intended by the speaker. Indeed, in studies using a gating task, disfluencies disrupted word recognition; that a disfluency was present was detected before a disfluent word was recognized [3].

Another possibility is that certain kinds of disfluencies may not affect comprehension at all. For one thing, disfluencies are difficult to detect and transcribe accurately; it has been suggested that listeners are relatively deaf to the words in a reparandum, or the part of a disfluent utterance that needs to be repaired [4]. Using a word monitoring task and spontaneously produced utterances, Fox Tree [1] found that repetitions did not seem to harm processing, while some false starts did. It is possible that disfluencies such as fillers are produced with intonation that is discontinuous from the rest of an utterance; this may help listeners distinguish them [5]. So certain disfluencies might be ignored during processing, along with other surface idiosyncracies of the speech signal.

Some have suggested that the form of disfluencies may actually *help* listeners solve the continuation problem (e.g., [6, 7, 8, 2]). On a pragmatic level, disfluencies such as *uh* and *um* have been demonstrated to be meaningful; they provide information that listeners can use to reliably discern the speaker's commitment to an utterance [9].

In our experiments, we used a comprehension task to investigate whether the form of certain disfluencies enables listeners to solve the continuation problem, therefore helping (or at least not harming) comprehension. Levelt [2] suggested that "by interrupting a word, a speaker signals to the addressee that that word is an error. If a word is completed, the speaker intends the listener to interpret it as correctly delivered" (p. 481). This we take as our first hypothesis: If a speaker intends a listener to "move to the orange square" but instead says, "move to the yell- orange square," the interrupted word may let the listener know sooner that the speaker is either having difficulty with or intends to replace the color name, than if the speaker had said "move to the yellow- orange square. A second hypothesis is that fillers, such as *uh*, *er*, or *um*, serve as editing expressions that "warn the addressee that the current message is to be replaced" ([2], p. 481).

To test these hypotheses, we had listeners select target objects on a simple display in response to fluent and disfluent versions of the same and similar spoken instructions. We measured comprehension in an indirect yet realistic way, by seeing how quickly (relative to the onset of the target color word) as well as how accurately listeners could select the target object.

2. METHOD

2.1. Subjects

Students at the State University of New York at Stony Brook volunteered to participate either for research credit to fulfill a course requirement or for a small honorarium. All were native speakers of English and naive to the purpose of our experiments. Disfluencies were elicited from a total of 32 speakers; 38 additional students made norming judgments, and a total of 180 others participated in one of the four comprehension experiments.

2.2. Design and stimuli

Fluent and disfluent utterances were elicited by having naive speakers perform an interference-intensive task, modified from van Wijk and Kempen's [10]: watching a computer display and giving instructions about a highlighted target that occasionally changed unexpectedly. We found that it was necessary to have an addressee in the room, actually carrying out the instructions and responding with backchannels, for the speaker's' instructions to sound at all natural. The critical stimuli we collected were based on three types of disfluent utterances (reparanda are in boldface): (1) between-word replacements, e.g. *Move to the* ***purple-*** *yellow square*, (2) interrupted words with replacements, e.g. *Move to the* ***pur-*** *yellow square*, and (3) interrupted words with the filler *uh*,

e.g. *Move to the* ***pur- uh****, yellow square*). For each spontaneous disfluency we created two types of edited controls: *sanitized* (the reparandum was edited out) and *pause-edited* (the reparandum was replaced with a pause of equal length). Experiments 1-4 used the same set of stimulus utterances: 68 fluent instructions, 34 spontaneously disfluent ones (comprised of 14 between-word replacements, 14 interrupted words with replacements, and 6 interrupted words marked with *uh* and followed by replacements), 34 sanitized versions, and 34 pause-edited versions. So 1/5 of the utterances contained an overt lexical disfluency and 4/5 did not; 3/5 were spontaneously produced and 2/5 were digitally edited. Experiment 4 used 12 additional items based on the mid-word interruptions marked with *uh*: half of these had the interrupted word removed from the reparandum, and half had the *uh* removed. For both these types of items, the removed material was replaced with a silent pause of exactly the same length.

The main set of 170 utterances was normed to make sure that listeners could not hear the electronic edit points. A group of listeners who did not participate in any of the other experiments made ratings by listening once to each utterance and deciding whether it had been played in its original form as the speaker produced it, or whether it had been electronically edited. Listeners could not tell the pause-edited utterances from the naturally disfluent ones. In the analyses, we used the pause-edited version of each naturally disfluent instruction as its own within-item control.

Experiment 1 examined the speed and accuracy of responses to these instructions, in the context of a 2-object display. Experiment 2 replicated the findings from Experiment 1, counterbalancing for the side of the display that the target object appeared on (in Experiment 1 there was a slight advantages for objects on the left side of the display). Experiment 3 increased the complexity of the display from 2 to 3 objects. Experiment 4 examined the nature of the cue behind the disfluency effect, to see whether it was due to (1) the *uh* itself, (2) the combination of the mid-word interruption with *uh*, or 3) the time after the mid-word interruption that elapses while *uh* is being produced.

2.3. Procedure

For the comprehension experiments, each listener was seated before a graphics display that, for each trial, presented horizontally arranged circles or squares followed by a spoken instruction of the form *Move to the X*. They were told to press a response key corresponding to the target item the speaker intended them to move to. The keys corresponded spatially to the layout of the shapes on the screen. Listeners were told to respond both quickly and accurately. After completing 15 practice trials that contained both fluent and disfluent utterances, they did the experimental trials. Each utterance was presented only once.

3. RESULTS

We compared error rates and response times relative to the onset of the target color words for the naturally disfluent utterances and the fluent, sanitized, and pause-edited versions. Of particular interest were the comparisons of between-word to mid-word interruptions, of disfluencies marked by *uh* to those not so marked, and of naturally disfluent to pause-edited utterances. With this last comparison, for each disfluent utterance that led to a correct response, we subtracted reaction times to its pause-edited version from those to its unedited version. A positive score constitutes a *disfluency advantage*.

The pattern of results was consistent across the experiments (see Figure 1). By themselves, reaction times failed to support Hypothesis 1, that a mid-word interruption alone should act as a better cue that a speaker intends to replace the interrupted word than a between-word interruption. However, the error rates *did* support this hypothesis; listeners made fewer errors after mid-word than between-word interruptions. There was stronger support for Hypothesis 2: Mid-word interruptions marked by *uh* speeded responses relative to mid-word interruptions without *uh*. This did not happen at the expense of accuracy; responses to interruptions with *uh* were just as accurate as responses to fluent utterances. Disfluencies without *uh* had higher error rates than any other kind of item (fluent, pause-edited, and sanitized). It is not at all surprising that error rates to disfluent utterances were higher, since listeners first heard all or part of a misleading color word before hearing the correct target color word, and once they committed themselves to a response they could not undo it. What is worth noting is that this cost to accuracy disappeared in the mid-word interruptions with *uh*.

In Experiment 3, when the display included three rather than two objects, listeners were slower overall in choosing targets. There was still a response time advantage for mid-word interruptions with *uh* over their pause-edited versions, compared to the disfluencies without *uh*, but this advantage diminished to 28 ms (down from 71 and 78 ms in Experiments 1 and 2; see Figure 1). *Basically, when the disfluency was less informative about what the speaker did not mean, interruptions with *uh* were less disadvantaged than other disfluencies, suggesting that in this situation *uh* tells listeners what the speaker does *not* mean (rather than achieving its advantage by acting as a more general alerting signal). The error rates supported both hypotheses: mid-word interruptions were again more accurate than between-word interruptions, and mid-word interruptions with *uh* were more accurate than those without (once again, no worse than fluent utterances).

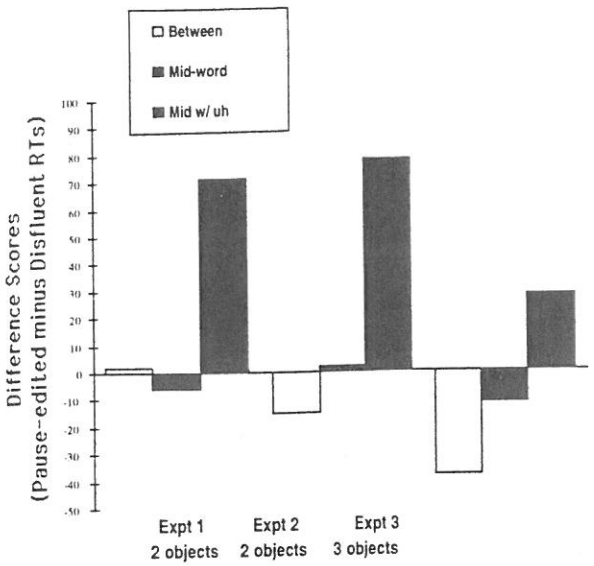

Figure 1: Disfluency advantage (pause-edited minus disfluency difference scores) in ms.

Experiment 4 examined the disfluency cue more closely to determine whether the disfluency advantage is due to *uh*, to the combination of the interrupted word plus *uh*, or merely to the extra time that elapses during *uh* after the interruption and before the target word. This experiment included six additional controls that edit out the *uh* and six that edited out the interrupted color word preceding it (replacing the removed material with a pause of equal length). As before, there was no disfluency *dis*advantage in the response times. The comparisons of interest are the response times and error rates for the unedited disfluencies with *uh* vs. those with interrupted words or *uh*s edited out. When only the *uh* was edited out and replaced with a pause of equal length (leaving the interrupted color word with the same interval to the target word), response times and error rates were the same as those for unedited disfluencies with *uh*. When the interrupted word was replaced with a pause, leaving the *uh* before the repair, response times slowed by about 80 ms.

4. DISCUSSION

Consistent with Fox Tree's findings [1], our data show that disfluencies need not harm processing; in fact, under some circumstances they may be informative. In our experiments, the disfluency advantage of mid-word interruptions marked with *uh* appears to be due mainly to the additional time that elapses during *uh* after the interrupted word, rather than to the presence of *uh* itself. Apparently the interrupted word acts as a cue that supports an inference by the listener about what the speaker did *not* mean, and when this cue is gone, the response time advantage goes too (Experiment 4). When the informativeness of the disfluency cue is diminished, so is the disfluency advantage (Experiment 3). Together, these data favor an inferential process over an attentional one in this particular task: If the disfluency advantage were due simply to alerting the listener to an upcoming repair or to heightening attention to the upcoming target word, then it should not have diminished with a 3-item display or when the word fragment before an *uh* was removed.

When a speaker interrupts an unintended word rather than completing it before a repair, this confers a distinct advantage on the listener: The listener is less likely to make a premature commitment to the wrong interpretation when she hears less misleading information. And in a logically constrained context, when there is sufficient time between the interrupted word and the target word, response time is speeded. The pattern of data supports our first hypothesis about interrupted words as potential cues, as well as a modified version of our second hypothesis: *uh* helps because it buys the listener time.

Our studies have focused on listening to disfluent utterances in a logically constrained context. We constrained the context in order to control the informativeness of the disfluencies, for this first demonstration that speakers *can* use the cues available in disfluencies to improve their performance in a behavioral task requiring them to follow simple instructions. The elicited utterances we used were semantically repetitive, which may have led listeners to attend to the disfluencies in ways they would not in the real world. This was probably not an issue for the filler *uh*, since there were only 6 of these items in the set of 170 instructions (3.5%). But it may well have affected people's responses to the between- and mid-word interruptions, as the speaker changed the color word in 20% of the utterances. It remains to be seen whether an interrupted word in a more diverse set would still cue listeners that the speaker intended to replace the word (as Levelt [2] originally suggested). That is, interrupted words might lose their value as cues if the instructions sometimes included utterances like "Move to the yel- yellow square."

While restarts such as this are not uncommon in naturalistic speech corpora, they were surprisingly rare in our elicited corpora. In the course of eliciting disfluencies, we recorded spoken instructions from two sets of speakers who referred to highlighted objects on a computer display. One of the speakers from the first set of 12 (with a confederate acting as his addressee) produced the fluent and disfluent utterances that we used in the current experiments. The second set of 15 speakers was recorded doing the same task, but with naive addressees who actually carried out the task of moving a cursor to the geometric objects in the instructions. In both of these corpora, there were surprisingly few instances of disfluencies involving the same color word repeated. This happened even though we programmed the displays in a way we thought would elicit such disfluencies (about 1/5 of the time the highlight on the object the speaker was supposed to refer to flickered but reappeared in the same place; about 2/5 of the time it jumped to another object, and about 2/5 of the time it stayed put).

The possibility that listeners might be using a strategy tailored to the discourse goals or context does not diminish the relevance of our findings; if strategic processing is taking place, then people have remarkable ability to discern the contextual informativeness of paralinguistic cues in spontaneous speech and to act on these cues.

ACKNOWLEDGMENTS

This brief report for the Disfluency in Spontaneous Speech Satellite Meeting of the 14th International Congress of Phonetic Sciences summarizes some work reported in detail in [11]. We are grateful to Jonathan Bloom, Ricardo Carrion, Julia Kung, Angela Lawrence, Maria Malzone, Kimiko Ryokai, Leora Schefres, and Darron Vanaria for their help with the experiments, and to Richard Gerrig and Arty Samuel for

helpful discussions. This material is based upon work supported by the National Science Foundation under Grant No. IRI9402167.

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] Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

[3] Lickley, R. and Bard, E. G. (1998). When can listeners detect disfluency in spontaneous speech? *Language and Speech, 41*, 203-226.

[4] Bard, E. G. and Lickley, R. (1998). Disfluency deafness: Graceful failure in the recognition of running speech. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*. University of Wisconsin- Madison.

[5] Shriberg, E.E. and Lickley, R.J. (1992). The relationship of filled-pause F0 to prosodic context. In *Proceedings of the IRCS Workshop on Prosody in Natural Speech*, Technical Report IRCS-92-37, 201-209, University of Pennsylvania, Institute for Research in Cognitive Science. Philadelphia, PA.

[6] Clark, H. H. (1994). Managing problems in speaking. *Speech Communication, 15*, 243-250.

[7] Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

[8] Fox Tree, J. E. (1993). Comprehension after speech disfluencies. Unpublished doctoral dissertation. Stanford University, Stanford, CA.

[9] Brennan, S. E. and Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34(3), 383-398.

[10] van Wijk, C. and Kempen, G. (1987). A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology, 19*, 403-440.

[11] Brennan, S. E. and Schober, M. F. (1999). When do disfluencies *help* comprehension? (Under revision)

COMPARING HUMAN AND AUTOMATIC SPEECH RECOGNITION USING WORD GATING

Robin Lickley, David McKelvie & Ellen Gurman Bard
Human Communication Research Centre, University of Edinburgh, Scotland

ABSTRACT

This paper describes a study in which we compare human and automatic recognition of words in fluent and disfluent spontaneous speech. In a word-level gating study with confidence judgements, we examine how the recognition and confidence of recognition of words by humans develops over utterances and show how disfluency disrupts the process. We give an automatic recogniser the same task and compare its performance with the humans'. With both systems, subsequent context supports word recognition: confidence in word recognition peaks after subsequent words have been heard. With both systems, disfluency adversely affects recognition of words in the immediate vicinity of the disfluent interruption (for repeats and repairs): disrupted subsequent context disrupts the recognition process.

1. INTRODUCTION

Spontaneous speech is disfluent: speakers need time to formulate utterances and often make changes on the fly, so pauses, repetitions and restarts abound. Until recently, models of speech recognition, both psychological and computational, have focussed on corpora of read or rehearsed speech and ignored the problems posed by spontaneous speech and disfluency. The most obvious way to characterise the problems posed by disfluency is in terms of detection and then resolution in order to produce a fluent representation of what the speaker intended to say. Specifically, following this view, we need to detect the *interruption point* and remove the *reparandum,* replacing it with the *repair* [5]. Under certain experimental conditions, we can detect disfluency reliably as soon as we hear the onset of the repair [7]. Yet, as listeners, we seem to filter out many disfluencies with great ease. Perception experiments have demonstrated that people regularly mistranscribe disfluencies or even miss them altogether [2, 6, 8].

We have evidence that people miss disfluencies because of the way they naturally process spontaneous speech. People do not recognise words on the basis of their acoustic shape alone, for words excerpted from their running speech contexts can be extremely difficulty to recognise. Instead, people depend on the context in which the word appears. As a consequence, people do not always recognise words one by one and in the order in which they are produced. Word-level gating studies have shown that many words - mostly short, unstressed function words - can only be correctly recognised when subsequent stress-bearing content words have been heard [3]. How far a word's recognition is delayed beyond its offset is a function of the inherent intelligibility of the word token, and of both prior and subsequent material. Words at the beginning of utterances are more difficult to recognize by their own offsets, presumably because the listener has less prior context to recruit to the task of disambiguating the speech sounds. Words before minor prosodic boundaries tend to be more delayed in recognition than those before major boundaries [10]. Major boundaries usually follow stress-bearing content words and end longer phrases or clauses, so that semantic, syntactic, and prosodic considerations will all tend to make such items consolidation points for recognition of the earlier, weaker words.

Disfluencies disrupt human word recognition because they disrupt this process. The interruption points in disfluent speech leave words in reparanda with truncated subsequent contexts. Any eventual continuation may be non-contiguous, so that the usual consolidation points may be lacking or so delayed that the undeciphered material may be lost before it can be resolved [2]. Delay is made more telling because the repair portion of a disfluent utterance is like the beginning of a new utterance: the words after the interruption have a disrupted prior context and are more prone to late recognition than words at the same serial position in fluent utterances [2, 7].

The critical factor here is the dependence on subsequent context, which we can measure in a tendency toward delayed recognition. We report a study in which we try to determine whether such a tendency can be found in a standard phoneme-based HMM automatic recognizer with a bigram language model. We conjecture that if it can be, and if it resembles the similar tendency in human listeners, we may be able to adapt the model to edit out disfluencies by failing to recognize them, just as people do.

In this study we compare human and HMM performance on a word-level gating task with matched fluent and disfluent utterances. Word-level gating allows us to look at the effects of prior and subsequent context on word recognition. Stimuli are presented in chunks which increase in length by one word on each presentation, beginning with just one word. Subjects respond by typing what they think the latest word is and making any desired alterations to previous word-judgements. In this experiment, subjects were also required to give a confidence score on a scale of 1 to 5 to show how sure they were of each word: 1 represented very low confidence and 5, certainty. In

previous gating studies [2, 3], we have looked at immediate, late and failed word recognition. In this study, we focus on the point at which confidence judgements for correctly guessed words reach their maximum level: the consolidation point. One feature of word-level gating which makes the recognition task easier than it is in normal listening is that the word-segmentation problem is solved: subjects know that each new presentation of any stimulus increases its length by one word and they can assume that the number of spaces on the current line of their answer grid represents the number of words that they ought to have transcribed. While this makes the task less natural than continuous listening, it also makes the demonstration of late and failed recognition all the more striking.

2. EXPERIMENTS

2.1 Materials.

Materials were the same for both the human and automatic recogniser tests: All stimuli were sampled from the HCRC Maptask Corpus, a set of 128 task-based spontaneous dialogues [1]. The corpus has time-aligned word-level transcriptions and is fully annotated for disfluencies (see [4] for some analysis of disfluencies in the corpus).

A total of 360 utterances were sampled and stored on a Sun Sparcstation:

- 120 *disfluent utterances*. These contained repetitions, deletions, insertions, substitutions or complex disfluencies were first selected (e.g. *along to the left towards the left of the banana tree*);
- 120 *length controls*. These were fluent utterances matched for speaker and length in words with the disfluent items formed one set of controls, so that comparisons of word recognition effects for serial position could be made (e.g. *keep going up until you're horizontally level with the rope bridge*);
- 120 *word controls*. These were fluent utterances matched for speaker with the disfluent set and containing the same sequences of words as occurred in the reparanda of the disfluent items formed the second control set (e.g. *move to the left about an inch*)

None of the stimuli contained filled pauses or other non-words.

All stimuli were segmented at word level by hand using signal annotation software: this provided the end points of the words for use in the gating experiment. Interruption points in the disfluent stimuli were also marked, as were the equivalent points in the fluent controls.

Word-level gating experiments take a long time to run. For this reason, the stimuli were divided into 4 groups of 90 stimuli which were distributed by Latin Square into 3 subgroups such that no subject would hear both the disfluent stimulus and either of its fluent controls. Thus, the materials were split into 12 groups of 30 items.

2.2 Procedure and Participants.

Human Subjects. Stimuli were presented over semi-closed headphones in a quiet computer lab. The experiment was run using a computer program and responses were typed into a graphical interface on Sun Sparcstations. Each response was saved to a computer file when the return key was hit. Null responses were not allowed. Subjects were able to do the experiment at their own pace and all completed it in one session of about an hour.

Subjects were 72 students at the University of Edinburgh, who participated either as a course requirement or for a small fee. Six subjects were assigned to each group. All subjects were native speakers of English and none reported having any hearing difficulty.

ASR. A phoneme-based HMM recogniser with a bigram language model trained on a subset of the HCRC Maptask Corpus was adapted to mimic the human perception experiment. Word-string hypotheses were generated for each point in the gating experiment and the recogniser was constrained to produce hypotheses for the correct number of words at each point: At end of the *n*th word, the best sequence of *n* words found by the recogniser was determined and its confidence measured by the probability assigned to the path.

3. RESULTS

Results from the automatic recogniser were pretty poor. So for the purposes of this study, we report results for one quarter of the experiment: 90 stimuli, consisting of 30 disfluent utterances and their two sets of 30 fluent controls. The 30 disfluent utterances were those which were best recognised by the automatic recogniser. We compare the consolidation points for correct word recognition by humans with those of the automatic recogniser. For humans, we define the consolidation point for a correct word as the point at which total confidence scores across subjects peaked for that word. For the automatic recogniser, we define the consolidation point as the point at which the HMM's hypothesis for that word reached its maximum divergence over other word hypotheses.

3.1 Points of Recognition.

Previous studies [2, 3] would predict that between 15 and 20% of words would be recognised late. The confidence judgements allowed a more refined test of when listeners were sure that they had recognised a word. Whereas previous work suggests that the large majority of words can be recognised on first presentation in a gating experiment, this version of the experiment shows that listeners' *confidence* in their guess does not usually peak until later. Only 168 of 788 non-final words (21.3%) formed their own consolidation point for humans, showing full confidence on first presentation. Interestingly, the rate of immediate consolidation for the automatic recogniser was very similar (166 (21.1%)) (Figure 1).

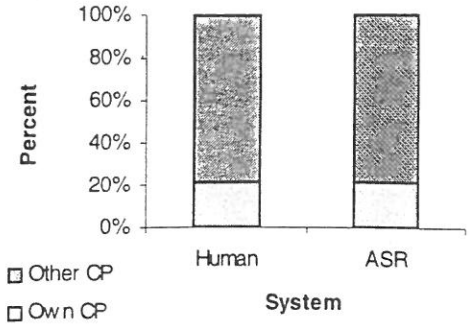

Figure 1: Consolidation points are on later words in most cases, for both human and automatic recognition

3.2 Consolidation Points

Consolidation points did not occur randomly, but clustered, both for humans and for the automatic recogniser: of 788 possible sites for consolidation (i.e. the total number of non-final words) humans consolidated at 351 sites and the automatic recogniser at 347. Figure 2 illustrates the consolidation points for the two systems for two fluent stimuli.

The examples in Figure 2 suggest that the relationship between word types and consolidation points and between consolidation points for people and for the automatic recogniser is not straightforward. There is some evidence of consolidation points occurring at heads of phrases and on content words, but closer analysis of where points of consolidation occur is needed and planned.

3.3 Disfluency Effects

In the analyses that follow, we focus on the words that matter most for the analysis of the effects of disfluency on word recognition: the words closest to the interruption point in the disfluent test items and words in matched positions in the two sets of controls. We discuss outcomes for the word which ended at the interruption point and for the two words before and the two words after that word.

People and the automatic recogniser are affected by disfluency and in similar ways. Both take longer to consolidate their word-hypotheses in disfluent utterances than in length-matched fluent controls around the interruption point ($F(1,143)=6.71$, $p<.015$). Figure 3 illustrates, first of all, the tendency to greater delays at early parts of the utterance, mimicking curves for recognition [3, 9]. Mean delay to consolidation from first presentation of a word

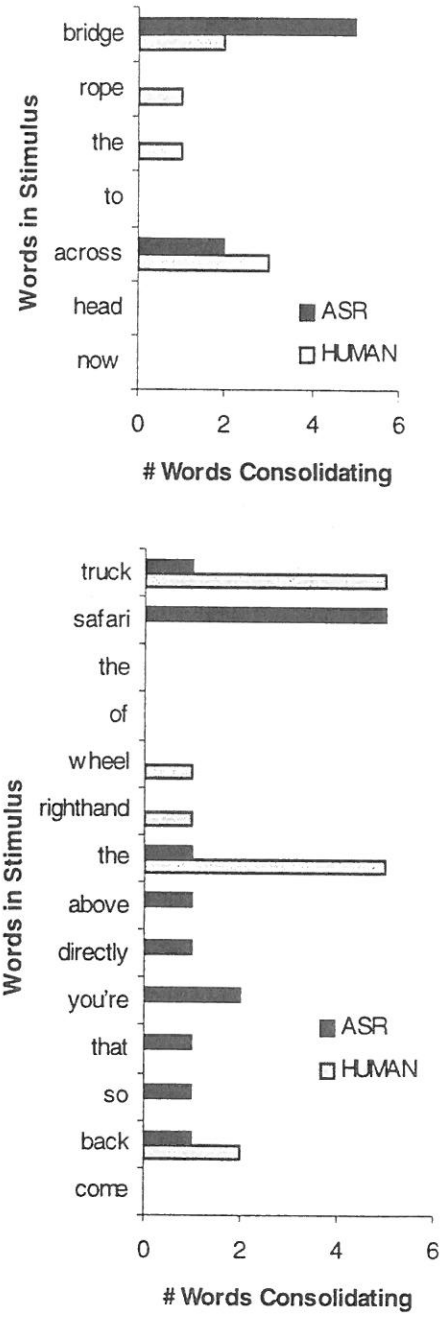

Figure 2: Consolidation Points in two fluent stimuli (*"now head across to the rope bridge"; "come back so that you're directly above the right-hand wheel of the safari truck"*) for Human subjects and for the Automatic Recogniser (read from bottom up).

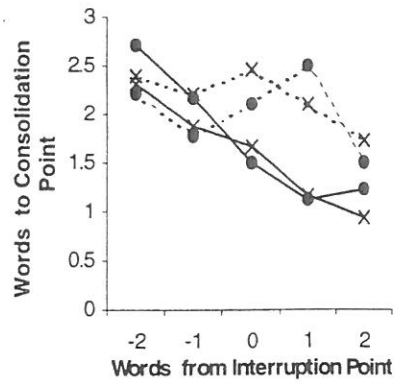

Figure 3: Disfluency delays Consolidation Points for Word Recognition for both human and automatic recognition. Solid lines are fluent stimuli, broken are disfluent; circles are for Human listeners, crosses for ASR.

decreases with serial position in the fluent controls, but hiccups across the interruption point in the disfluent cases, for both recognition systems.

There is some difference between the human and automatic recognisers. The point where consolidation is most delayed for the automatic system is on the word at the interruption point, where for humans the delay to consolidation has its peak at the word immediately after the interruption.

Similar results relating word recognition success to serial position in the utterance were found for mean human confidence scores on first presentation of each word: words later in an utterance were given higher confidence scores except around the interruption point in disfluencies, where confidence dropped.

4. DISCUSSION

The work described here represents the first steps in the analysis of a large scale recognition experiment and this paper begs many more questions than it addresses. However, a few things are clear from these early results:

- Adding confidence judgements to the word-level gating method gives a more refined view of when words are recognised in running speech.
- Humans and our ASR reach confidence peaks similarly late.
- Disfluency leads to lower confidence for the words around the interruption point, for both humans and ASR.
- Disfluency delays both human and ASR consolidation relative to delays in fluent utterances.

More detailed analyses will examine the roles of parts of speech and syntactic and prosodic phrasing in the locations of consolidation points and will take greater account of actual recognition confidence levels.

ACKNOWLEDGEMENTS

We thank Henry Thompson for writing the software on which the human experiment was run. The authors gratefully acknowledge the support of EPSRC SALT grant number GR/L50280.

REFERENCES

[1] Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. & Weinert, R. 1991 The HCRC Map Task Corpus. *Language and Speech*, 34, 351-366.

[2] Bard, E.G. & Lickley, R.J. 1998, Graceful Failure in the Recognition of Running Speech, *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, University of Wisconsin, 108-113.

[3] Bard, E.G., Shillcock, R.C. & Altmann, G.T.M. 1988, The Recognition of Words after their Acoustic Offset: Effects of Subsequent Context, *Perception and Psychophysics 44(5), 395-408.*

[4] Branigan, H., Lickley, R.J. & McKelvie, D. 1999, Non-Linguistic Influences on Rates of Disfluency in Spontaneous Speech, *Proceedings of the ICPhS*, San Francisco.

[5] Levelt, W.J.M. 1983, Monitoring and Self-Repair in Speech, *Cognition, 14, 41-104.*

[6] Lickley, R.J. 1995, Missing Disfluencies, *Proceedings of the ICPhS, vol 4, 192-195,* Stockholm.

[7] Lickley, R..J. & Bard, E.G. 1998, When can listeners detect disfluency in spontaneous speech? *Language and Speech 41(2), 203-226.*

[8] Martin, J.G. & Strange, W. 1968, The Perception of Hesitations in Spontaneous Speech, *Perception and Psychophysics 3(6) 427-438.*

[9] Pollack, I. & Pickett, J.M. 1964, Intelligibility of Excerpts from Fluent Speech: Auditory vs. Structural Context, *Journal of Verbal Learning & Verbal Behavior 3, 79-84.*

[10] Shillcock, R.C., Bard, E.G. & Spensley, F. 1988, Some Prosodic Effects on Human Word Recognition in Continuous Speech., *Proceedings of SPEECH '88, 7th FASE Symposium*, 819-826.

Use of a Postprocessor to Identify and Correct Speaker Disfluencies in Automated Speech Recognition for Medical Transcription.

Sherri Page[1]
University of Minnesota, Minneapolis, USA

ABSTRACT

Medical practitioners speak in a quasi-spontaneous monologue when they dictate a chart note, letter, or patient history. Prior research has largely ignored the issue of disfluency in dictation, arguing that speakers can control recording and start over if necessary. In 550,000 words of hand transcribed medical dictation, however, we find numerous filled pauses, repetitions, and other self-repairs. This paper describes: a pre-theoretical classification of disfluencies, developed to identify patterns useful in automatic text processing; the patterns of disfluency found in a corpus hand tagged with this classification, which include repetitions in combination with substitutions, insertions, and deletions; and, preliminary results of implementation of a disfluency pattern matcher and filter in a postprocessor developed for commercial use.

1. INTRODUCTION

While it is true that speaker disfluencies can reduce the quality of automated speech recognition, that is not the focus of this paper. Even when recognized with perfect accuracy, disfluencies introduce extraneous text which must be removed when the desired end product is not a verbatim transcript. The aim of my research is to identify the patterns of repair disfluencies in such a way that they can be identified automatically and the extraneous text removed. This paper represents a first step towards developing an analysis that can be automated in a speech recognition system.

2. CORPORA

This study uses two corpora consisting of hand transcribed medical dictation, containing an estimated total of 550,000 words, including filled pauses. Many of the speakers are represented in both corpora. All speakers are M.D.s or Nurse Practitioners. The first corpus is used primarily for identifying patterns of disfluency, and the second is for testing those patterns.

2.1. Corpus 1 (MED_TAG). Size: approximately 37,000 words. Number of speakers: 21; 6 female, 15 male. Corpus consists of 338 files of digitized speech and hand transcribed, aligned text. Each file represents a full or partial dictation containing up to one minute of speech. Every file was hand tagged for repairs by three different people while listening to the recorded audio. Tagging methods are described further in section 3 below. Word fragments were added to these transcripts at time of tagging.

2.2. Corpus 2 (BIG_MED). Size: approximately 495,000 words, after exclusion of approximately 20,000 words of text shared with MED_TAG. Number of speakers: approximately 20, mixed gender. Corpus consists of 18 files of hand transcribed speech, containing approximately 16,000 dictations, text only. Word fragments are not transcribed in these files.

3. DISFLUENCY CLASSIFICATION

The classification system introduced below was developed specifically for this research project, in an attempt to create categories of repair disfluencies fitting the following criteria: that they be mutually exclusive, for ease of both human and machine classification of individual repair sites, and that they allow for cross-comparison of repair types and further sub-classifications. In this paper I use the term repair site to mean something closely corresponding to Shriberg's definition of disfluency region [6]: it consists of an ill-formed first part, the *reparandum*, followed by an *interregnum* (which Clark terms the *hiatus* [1]) -- which may be empty or contain filled pauses or editing expressions -- followed by the well-formed part, the *repair*. My analysis differs somewhat in that a site is considered to include any filled pauses bordering it (i.e. directly adjacent to either the reparandum or the repair).

3.1. Classification outline.

3.1.1. Exact repetition (category 1). This category includes single or multiple word repetitions, separated optionally by silence, filled pauses, editing expressions, or any combination of these. E.g. "the um the", and "with a with a". The smallest possible site is two words in length.

3.1.2. Exact substitution (category 2). This category includes single or multiple word substitutions, separated optionally by silence, filled pauses, editing expressions, or any combination of these. E.g. "five correction seven". The minimal length for this type of repair site is two words.

3.1.3. Repetition and substitution (category 3). This category includes substitutions with repeated material to the left or the right. E.g. "does not did not", and "um for two for one". The smallest possible site for this type is four words.

3.1.4. Repetition and insertion (category 4). This category includes repetitions with a new word inserted before or medially. E.g. "to clean to try to clean". Minimum size for this type is three words.

3.1.5. Repetition and deletion (category 5). This category includes repetitions with a word omitted either at the start of the repeat or medially. E.g. "no spotting dysuria or abnormal correction no spotting or dysuria". The smallest possible site is five words.

3.1.6. Editing expressions (categories A and B). Editing expressions such as "correction", "I'm sorry", and sometimes "or", are not considered words in this classification. They really are words, of course, but they are excluded because they are not intended to be part of the finished utterance. Rather, they are signals to the listener to replace something that was just said with something else. Repair sites containing editing expressions are classified as category B, and sites without them are category A. All repair sites are classified with both a letter and a number.

3.1.7. Filled pauses. Filled pauses (FPs) are not considered words in this classification, so two FPs in a row, for instance, would not count as an exact repetition. Any FP occurring at the edge of a site is analyzed as belonging to that site.

3.1.8. Word fragments. Word fragments are considered words in this classification, although they are not considered identical to the words they may be partial utterances of. For instance, something like "dislo- dislocated" might be classified as a *fragment repeat* in some research [2], but under this system it is analyzed as an exact substitution.

3.2. Hand tagging of MED_TAG corpus.

3.2.1. Insertion of textual markers. Three undergraduate students were trained in the system of classification outlined above. Listening to the digitized audio and reading the verbatim transcripts (both of which had been sanitized to remove doctor or patient identifying information), the taggers annotated the transcripts of the MED_TAG corpus by inserting textual markers:

- at the beginning of a repair site (i.e. the left edge of the reparandum), using a number and a letter to denote the repair type, followed by '>';
- at the point of resumption of fluent speech (i.e. the right edge of the hiatus, or left edge of the repair proper), using the character '|'; and finally,
- at the end of the repair site, or the right edge of the repair proper, using the same number and letter preceded by '<'.

3.2.2. Insertion of word fragments. Because word fragments may signal that speakers are changing their minds about what they have just said [1], the taggers were instructed to insert orthographic approximations of any fragments they perceived, ending each with a dash '-'. As Shriberg notes, fragments are difficult even for trained transcriptionists to notice, particularly when adjacent to another disfluency [6], perhaps because of some sort of pre-conscious filtering that applies to normal speech perception. Thus it is possible that a few word fragments were missed by the taggers. Because their spellings varied wildly, it was difficult to assess the reliability of fragment transcription; however in terms of location it appeared to be quite good. E.g. one tagger's "hydras-" was transcribed by another as "hydrasa-" and the third as "hydros-", but all of these preceded "hydrocortisone".

3.2.3. Inter-tagger reliability. Reliability in general appeared quite good. In the first pass of analysis, I selected only the repair sites identified by two or more taggers. Later I verified each site by hand and found that almost every one of the sites identified by two or more taggers was in fact a repair site, whereas almost none of the sites marked by only a single tagger were.

4. ANALYSIS

The following notation is used in each of the analyses below:

Symbol	Definition
w1 w2 .. w*n*	individual words
FP	filled pause
EE	editing expression
?	zero or one
*	a sequence of zero to three
|	point of resumption

Table 1. Pattern notation.

4.1. Pattern and distribution of repair types.

4.1.1. Exact repetitions. As shown in the table below, the vast majority of repetitions found in MED_TAG involved a single word, specifically: 75 sites out of 87, or 86%; and repetitions of longer sequences diminish in frequency. Only 1 instance of a repetition containing an editing expression was found, and it involved a single word repeated once. In only 4 instances was a (single) word repeated twice.

Words	Sites	%tot	Pattern
1	75	86%	FP* w1? w1 EE? FP* | w1 FP*
2	5	6%	FP? w1 w2 | w1 w2
3	5	6%	FP? w1 w2 w3 FP* | w1 w2 w3
4	2	2%	w1 w2 w3 w4 | w1 w2 w3 w4 FP?
5+	0	0%	--
Total	87	100%	--

Table 2. Exact repetition.

As shown in the table above, filled pauses occurred optionally at the beginnings and ends of disfluency regions, and at the resumption point. It is interesting that FPs did not occur intra-sequentially, i.e. between w1 and w2, or between w2 and w3, in any of the multi-word repetition sequences. However only 12 of such longer sequences occurred, total. The next step is to look for these sequences in BIG_MED, and see whether FPs exhibit the same behavior in the larger corpus.

4.1.2. Exact substitutions. In Table 3, below, the 1st column contains the number of distinct words in the reparandum portion (RM) of the disfluency region, while the 2nd column contains the number of distinct words on the repair side (RR). (In the table above, "Words" denoted both RM and RR, because in exact repetitions those numbers are always identical.) This category unexpectedly accounts for half of all the repair sites identified in MED_TAG: 192 out of 383.

RM	RR	Sites	%tot
1	1	140	73%
1	2	12	6%
1	3+	8	4%
2	1	7	4%
2	2+	11	6%
	Other	14	7%
	Total	192	100%

Table 3. Exact substitution, counts.

Patterns	Sites	%tot
FP* w1 FP? EE* FP? | w2 FP*	140	73%
FP* w1 FP* | w2 FP? w3 FP?	12	6%
FP? w1 FP* EE? FP* | w2 w3 w4 w*n**	8	4%
FP? w1 w2 FP* | w3 FP?	7	4%
FP? w1 FP? w2 FP? EE* FP? | w3 FP? w4 wN*	11	6%
--	14	7%
--	192	100%

Table 4. Exact substitution, patterns.

Nearly three-quarters of all the exact substitutions found in MED_TAG were single word substitution (140 out of 192), as shown in Table 3, above. Not shown is that over half of the single word substitutions (75 out of 140) were simply "w1 \| w2", containing no filled pauses or editing expressions. Also, nearly two-thirds of all substitutions involved word fragments in the RM (122 out of 192). Most surprisingly, less than 10% (19 out of 192) of the total substitutions contained any editing expressions.

4.1.3. Repetition with substitution. This category can also be called anchored substitution. It differs from category 2 in that it has extra material that helps "anchor" the new words in their intended positions in some sense. As noted by Tannen, repetitions in narrative allow a speaker to create a frame to slot new information into [7]. Repetition can serve a similar function in repairs.

Patterns	Sites	%tot
FP? w1 w2 FP? EE* | w1 FP? w3 FP?	23	27%
FP? w1 w2 FP? EE* | w3 w2 FP?	8	10%
FP? w1 w2 FP? w3 FP? EE? | w1 w2 FP? w4	8	10%
--	45	54%
Total	84	100%

Table 5. Repetition with substitution.

As you can see in the table above, patterns in this category (category 3) were much less homogenous than those in categories 1 and 2 (shown in 4.1.1. and 4.1.2., above). The three most frequent patterns shown account for about half of the data; the rest of the patterns (54%) consist of wildly variant singletons.

4.1.4. Repetition with insertion.

Pattern	Sites	%tot
FP w1 FP? EE? | w2 w1 FP?	6	32%
FP? w1 FP? | w2 w3 w1 w*n**	4	21%
--	9	47%
--	19	100%

Table 6. Repetition with insertion.

Very much like 4.1.3., above, the two most common patterns account for only half of the category 4 sites identified. The rest of the patterns were likewise irreconcilable singletons, completely dissimilar from the first two patterns and from each other.

4.1.5. Repetition with deletion. Only a single instance of category 5, deletion, was found! The exact repair was: "malignancies ah you should have a ah correction period ah \| malignancies period ah". This appears to be a complex repair, in fact, containing two separate attempts at deleting the same elements. The speaker first tried to indicate a complete restart (i.e. deletion of "you should have a") with the editing expression "correction" followed by the punctuation word "period", then apparently felt a need to make the deletion more explicit by repeating the words "malignancy" and "period" with nothing in between. This could actually be reanalyzed as a category 2 substitution embedded within a category 1 repetition, as follows: "1A> malignancies 2A> ah you should have a ah correction \| period ah <2A \| malignancies period ah <1A". The separate category of anchored deletion might thus turn out to be unnecessary. In any case, one category 5 site (out of 383) seems to indicate a certain degree of rarity.

4.2. Summary of repair types found in MED_TAG.

Category	Sites	%type	%tot
1A	86	99%	22.5%
1B	1	1%	0.3%
Total 1	87	100%	22.7%
2A	173	90%	45.2%
2B	19	10%	5.0%
Total 2	192	100%	50.1%
3A	56	67%	14.6%
3B	28	33%	7.3%
Total 3	84	100%	21.9%
4A	14	74%	3.7%
4B	5	26%	1.3%
Total 4	19	100%	5.0%
5A	0	0%	0.0%
5B	1	100%	0.3%
Total 5	1	100%	0.3%
Total A	329	--	85.9%
Total B	54	--	14.1%
Grand Total	383	--	100.0%

Table 7. Summary of repairs in MED_TAG.

This summary gives the ratios of absence (A) vs. presence (B) of editing expressions in each of the 5 repair types, information not included in the pattern tables of the previous section. See 4.4., below, for the actual words used in these EEs.

4.3. Pattern and distribution of filled pauses (FPs).

As can be inferred from all of the patterns listed in the previous tables, filled pauses can and do occur: preceding the reparandum, within the hiatus, and after the repair. They can also occur between sequences of words either in the reparandum or the repair, though this placement appears much less often (and not at all in exact repetitions). A more detailed analysis of the distribution of FPs within repair sites is in progress.

4.4. Pattern and distribution of editing expressions (EEs).
As shown in Table 7, above, only 14% of the repair sites identified in MED_TAG contained EEs. This is a bit lower than Levelt reported [3], which is likely due to the exclusion of filled pauses from consideration here as editing expressions. Certainly a great many FPs do occur in the hiatus, with or without EEs. Interestingly, two EEs sometimes occurred in a row (which is implicitly stated in patterns containing "EE*"). The EE "or" for instance, was found preceding "actually", "correction", and "sorry". Total occurrences of each EE are listed in the table below. (Note: these do not correspond to the number of sites containing EEs, because some sites contain multiple EEs.)

EE	Instances	%EEs
"or"	26	42.6%
"correction"	18	29.5%
"I'm sorry"	7	11.5%
"sorry"	2	3.3%
"pardon me"	3	4.9%
"excuse me"	2	3.3%
"actually"	1	1.6%
"make that"	1	1.6%
"lets see here"	1	1.6%
Total	61	100.0%

Table 8. Editing expressions.

The most common EE was unfortunately "or", which of course is not exclusively used in repairs. The word "correction" is a better EE for repair identification purposes, although it can occur in non-repair phrases such as "surgical correction". The best indicators of repair are actually "sorry" and "I'm sorry": all of their occurrences in BIG_MED were found to be EEs, every one!

4.5. Testing repair patterns in BIG_MED.
The testing phase has only recently begun. Preliminary indications are that repair patterns in MED_TAG also exist in the much larger corpus, BIG_MED, but results are not ready at time of writing.

5. DISCUSSION

As shown in the previous analyses, the majority of repairs found in our corpus involved single words. This is consistent with the findings of Levelt [3],[4] and Nooteboom [5], that repairs usually occur immediately after an error. In addition, relatively few of the repairs contained editing expressions, substantiating Shriberg's proposal that editing phrases may be less frequent in disfluencies than previously thought [6].

6. POSTPROCESSOR IMPLEMENTATION

6.1. Exact repetitions.
A postprocessing program to identify single word repetitions has already been implemented. If successful, this alone could filter up to a fifth of the extraneous text introduced by speaker disfluency. The program already suppressed identified filled pauses; now it suppresses the first instance of a repetition as well. Unfortunately, there are a number of cases where repetition is actually desired. The obvious cases are "that that" and "very very"; slightly less obvious cases have been found to include number series, such as "nine nine", and some pronouns, e.g. "her her". The latter case occurs with any word that can both end and begin a sentence, if sentence boundaries are not explicitly indicated (as they generally aren't, in automated recognition, unless they are dictated by the speaker). Further research is needed to determine which words or types of words the exclusion list must include in order to avoid mis-suppression of legitimate repetitions.

6.2. Exact substitutions.
It is very difficult to identify exact single word substitutions automatically, much less multiple ones. However, part-of-speech (POS) might be used to filter some types of single word substitution. For example, series of determiners such as possessives and articles can be filtered automatically, because they normally cannot occur in a row: e.g. "the his leg". Certain prepositions also have the property of not sequencing with other prepositions, e.g. "at"; but identifying these would require further subcategories of prepositions: POS alone is clearly insufficient to identify allowable sequences.

6.3. Repetitions combined with substitutions, insertions, or deletions.
Many of these patterns can be identified automatically, such as "w1 w2 w1 w3", with the output to be filtered as "w1 w3", but constraints must be put on the kind of word that can be considered the repair "anchor" (i.e. the repeated item). For instance, "every hour every day" should not be assumed to represent an intended utterance of "every day". More investigation is needed to determine what types of words can serve as reliable anchors.

AKNOWLEDGEMENTS

Sincere thanks to Joan Bachenko, Ph.D., and Professor Amy Sheldon, for encouraging this project both at LTI and UMN, and for facilitating undergraduate assistance (via UROP). I wish also to thank the taggers themselves, Kristin Bergquist, JoElle Kangas, and Hannele Nicholson, for their comments as well as painstaking attention to detail. This research was partly supported by NIH grant HD07151.

NOTES

The author can be reached via email at pagex016@tc.umn.edu.
1. Also at Linguistic Technologies, Inc., 105 South 3rd St., St. Peter, MN 56082.

REFERENCES

[1] Clark, H. H. 1996. *Using language*. Cambridge: Cambridge Univ. Press.
[2] Clark, H. H., & Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology*, **37**, 201-242.
[3] Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition*, **14**, 41-104.
[4] Levelt, W. J. M. 1989. *Speaking*. Cambridge, MA: MIT Press.
[5] Nooteboom, S. G. 1980. Speaking and unspeaking: detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance* (pp. 87-95). New York: Academic Press.
[6] Shriberg, E. E. 1994. *Preliminaries to a theory of speech disfluencies*. Unpublished Ph.D. dissertation, University of California, Berkeley.
[7] Tannen, D. 1990. *Talking voices: repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge Univ. Press.

Filled Pause Distribution and Modeling in Quasi-Spontaneous Speech

Sergey Pakhomov and Guergana Savova[1]
University of Minnesota, Minneapolis, USA

ABSTRACT

Filled pauses (FP's) are characteristic of spontaneous speech and present considerable problems for speech recognition by often being recognized as short words. Recognition of quasi-spontaneous speech (medical dictation) is subject to this problem as well. An *um* can be recognized as *thumb* or *arm* if the recognizer's language model does not adequately represent FP's. Representing FP's in the training corpus substantially improves recognition. Several techniques of conditioning a training corpus with FP's were evaluated to show that a bigram probability method, as well as a uniform distribution probability method, centered around the average sentence length, yield better recognition results. The best method of conditioning a training corpus with FP's may have to target clause boundaries despite the fact that inserting FP's at clause boundaries by using a limited set of clause boundary anchors failed.

1. INTRODUCTION

Filled Pauses are not used at random but have a systematic distribution and well-defined functions in discourse. [1,2,3,5,8,9,16]. Cook and Lalljee [4] make an interesting proposal that FP's may have something to do with the listener's perception of disfluent speech. They suggest that speech may be more comprehensible when it contains filler material during hesitation: a FP may serve to preserve continuity and may serve as a signal for drawing the listener's attention in order for the listener not to lose the onset of the following utterance. Perhaps, from the point of view of perception, FP's are not disfluent events at all. This proposal bears directly on the domain of medical dictations. Some physicians who use old voice operated equipment train themselves to use FP's instead of silent pauses so that the recorder won't cut off the beginning of the following utterance.

Filled pauses, false starts, repetitions, fragments, are characteristic of spontaneous speech and present considerable problems for speech recognition. FP's are often recognized as short words of similar phonetic quality. Recognition of quasi-spontaneous speech (medical dictation) is subject to this problem as well. For example, in our system, an *um* is recognized as *thumb* or *arm* if the language model does not adequately represent FP's. The FP problem becomes especially pertinent where the corpora used to build language models are compiled from text with no FP's. Shriberg [12] has shown that representing FP's in a language model helps decrease the model's perplexity. She finds that when a FP occurs at a major phrase or discourse boundary, the FP itself is the best predictor of the following lexical material; conversely, in a non-boundary context, FP's are predictable from the preceding words. Shriberg [10] shows that the rate of disfluencies grows exponentially with the length of the sentence, and that FP's occur more often in the initial position (see also Swerts [16]).

In a previous study, Pakhomov [9] shows that a language model based on a training corpus populated with FP's using bigram probabilities significantly improves recognition over a language model that contains no FP's. He also finds an improvement with the bigram approach over the uniform distribution model which represents FP's inserted into the training corpus at random with insertion points centered around every 15th word (empirically established average frequency of FP's).

In this paper we present a method of conditioning training corpora with FP's at clause boundaries in addition to the bigram approach and four uniform distribution models. We suggest that, although the method of inserting FP's at clause boundaries does not yield satisfactory recognition results, it may prove to be more fruitful with an improved clause boundary identification mechanism. We also show that using recognition accuracy as the only gauge to determine the goodness of a FP model is not sufficient due to FP overrepresentation effects. We also show advantages and disadvantages of populating training corpora with FP's at random.

2. QUASI-SPONTANEOUS SPEECH

The term quasi-spontaneous speech reflects the fact that medical dictations used for analysis in this study are very different from unprepared monologues as well as read text and tend to retain features of both. Family practice dictations are pre-planned and follow an established SOAP format: Subjective (informal observations), Objective (examination), Assessment (diagnosis) and Plan (treatment plan). The Subjective part tends to resemble unrehearsed monologues where the rest of the dictation is more like read speech. Physicians are aware of their audience and often address the medical transcriptionists directly by thanking them and telling jokes.

3. TRAINING CORPORA AND FP MODELS

This study used three base and five derived corpora. Base corpora are collections of medical dictations used for two purposes: analyzing FP distribution and FP conditioning. Derived corpora are collections of the same dictations after FP conditioning. Brief descriptions of each follow in sections 3.1 and 3.2.

3.1. Base

- Balanced hand transcribed training corpus (BHT_CORPUS) that has 75, 887 words of word-by-word transcription data evenly distributed among 16 talkers. This corpus was used

to build a bigram model, that controls the process of populating a no-FP corpus with artificial FP's (BIGRAM_FP_MODEL). A more detailed description of the model follows in section 3.3.1.

- Unbalanced hand transcribed training corpus (UHT_CORPUS) of approximately 500,000 words of all available word-by-word transcription data from approximately 20 talkers. This corpus was used only to calculate the average frequency of FP use among all available talkers and the average frequency of pronounced punctuation.

- Finished transcription corpus (NOFP_CORPUS) of 13,537,262 words contains all available dictations and no FP's. It represents over 200 talkers of mixed gender and professional status. The corpus contains no FP's or any other types of disfluencies such as repetitions, repairs and false starts. The language in this corpus is also edited for grammar.

3.2. Derived

- BI_FP_CORPUS is a version of the finished transcriptions corpus populated with FP's based on the BIGRAM_FP_MODEL. (FP count: 2, 294, 909)

- CBFP_CORPUS is derived from the NOFP_CORPUS conditioned with FP's via a method that favors clause boundaries as discussed in section 3.3.3 (FP count: 1, 068, 938)

- RND_FP Corpora
  Four corpora were derived from the NOFP_CORPUS by populating it with FP's uniformly distributed in four ranges. The FP distributions correspond to our perception of average syntactic phrase length and empirically determined average sentence length and FP frequency. (see Table 1)

RND_FP Corpora	Motivation	Range	FP count
RND_FP_CORPUS_3	Theoretical (short syntactic phrase length)	0-6	3,867,789
RND_FP_CORPUS_5	Theoretical (long syntactic phrase length)	0-9	2,707, 842
RND_FP_CORPUS_10	Empirical (Avg. Sentence length)	0-19	1,289,796
RND_FP_CORPUS_15	Empirical (Avg. FP frequency)	0-29	873, 538

Table 1: Uniform distribution based corpora.

3.3. FP conditioning methods

Three distinct methods of corpus conditioning were used in this study: Bigram method, Random method and Clause Boundary method. Description of each follows in sections 3.3.1, 3.3.2 and 3.3.3 respectively.

3.3.1. Bigram method. Here a bigram model is constructed prior to conditioning the NOFP_CORPUS with FP's. This model contains the distribution of FP's obtained from BHT_CORPUS by using the following formula:

$$P(FP|C_{w-1}) = \frac{C_{w-1FP}}{C_{w-1}}$$

$$P(FP|C_{w+1}) = \frac{C_{w+1FP}}{C_{w+1}}$$

Thus, each word in a corpus to be populated with FP's becomes a potential landing site for a FP and does or does not receive one based on the probability found in the BIGRAM_FP_MODEL.

3.3.2. Random method. This method determines FP locations using uniformly distributed random spacings. It allows to control the overall frequency of FP's by controlling the distribution range.

3.3.3. Clause Boundary method. This is a pseudo-random method that makes limited use of linguistic knowledge while populating the NOFP_CORPUS with FP's. Similar to the Random method, Clause Boundary method involves distributing FP's uniformly in the range between 0-29; however, the insertion points are shifted towards clause boundaries.

We used the following lexical items as clause boundary anchors: "period", "that", "which", "if", "whether", "who", "when", "what", "where", "why", "how", "because", "so", "however", "although", "though." This is a very limited set of complementizers which can be expanded with a parser.

Using the word "period" as a clause boundary marker presents a problem because the talker pronounce punctuation on average only 20% of the time. Calculating this number is not straightforward in the absence of corresponding literal and finished transcription corpora. At our disposal we have 500,000 words of literal transcriptions (UHT_CORPUS) which contains spoken "periods" but no punctuation. We also have a 13.5 mil (NOFP_CORPUS) word corpus that is punctuated but does not necessarily represent what was said. The average length of a sentence in the NOFP_CORPUS turns out to be around 10 words. This average length, when applied to the UHT_CORPUS, gives us an estimate of about 50,524 sentences. Given that the word "period" is found in the UHT_CORPUS 10,097 times, we can roughly estimate that about 20% of sentences are terminated with the word "period" actually spoken. This means that to insert a FP at sentence boundaries, the program has to perform the insertion around the "." in NOFP_CORPUS prior to eliminating 80% of punctuation for training the language model.

4. TRIGRAM LANGUAGE MODELS

The following trigram models were built using ECRL's Transcriber language modeling tools [6]. Bigram cutoffs were set at 0 and trigram cutoffs were set at 1.

- NOFP_LM was built with the NOFP_CORPUS with no FP's.
- BIFP_LM was built with the BI_FP_CORPUS.
- RNDFP_LM_3 was built with the RND_FP_CORPUS_3
- RNDFP_LM_5 was built with the RND_FP_CORPUS_5
- RNDFP_LM_10 was built with the RND_FP_CORPUS_10
- RNDFP_LM_15 was built with the RND_FP_CORPUS_15
- CBFP_LM was built with the CBFP_CORPUS.

5. EVALUATION METHODS

Three different evaluation methods were used in this study: Recognition accuracy, FP correctness, False FP measure. Description of each follows.

5.1. Recognition accuracy

Recognition accuracy was obtained with ECRL's HResults tool and is summarized in Chart 1.

5.2. FP correctness

This metric produces a percentage ratio of correctly recognized FP's to the total number of FP's in a given dictation. The results are summarized in Chart 1.

5.3. False FP measure

This measure is computed by averaging the number of times a real word or part of a word is recognized as a FP in any given dictation by a talker across all dictations. For example, the word "umbilical" may be recognized as "um build". Chart 2 displays the results.

6. RESULTS AND DISCUSSION

Speech data comes from 23 talkers selected at random represents 3 to 5 (1-3 min) dictations for each talker. The talkers are a random mix of male and female medical doctors and practitioners.

Modeling FPs consistently increases total recognition results, even for talkers who use no FP's. All six FP models we built recorded an increase in recognition accuracy 6.083% to 9.15%. This is in accord with Pakhomov's [9] report that LMs incorporating FPs systematically decrease the models' perplexity and increase recognition accuracy.

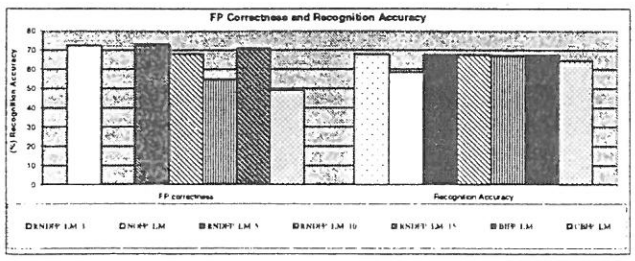

Chart 1: Filled Pause correctness and Recognition accuracy by Model

Total recognition results for the six FP models do not show significant statistical difference (see Chart 1). To be able to gauge the models' performance in detail, we introduced an additional measure – FP correctness (see Chart 1). Measuring correctness of FP recognition separately from overall recognition accuracy allows us to evaluate the FP component of the language model. An increased FP representation in a language model may lead to an increased recognition of previously misrecognized words together with a substitution with FP's of previously correctly recognized words. The overall recognition accuracy may not change, which makes it impossible to detect the direction of recognition performance. It is important to single out the FP component in order to evaluate a FP model's recognition performance more accurately.

Measuring FP correctness on the six models shows a considerable difference among the models' performance. Naturally, NOFP_LM had 0% FP correctness. FP correctness appears to grow proportionately with the frequency of FP's in the training corpus. The top four FP models - RNDMFP_LM_3, RNDMFP_LM_5, RNDMFP_LM_10 and BIFP_LM - do not exhibit a significant variation, but each appears to perform better than RNDFP_LM_15 and CBFP_LM. False FP measure, however, shows that RNDFP_LM_3 and RNDFP_LM_5 produce significantly more false FP's that any other model, which leaves RNDFP_LM_10 and BIFP_LM as the two optimal models in terms of the trade-off between overall recognition accuracy, FP correctness and false FP rate.

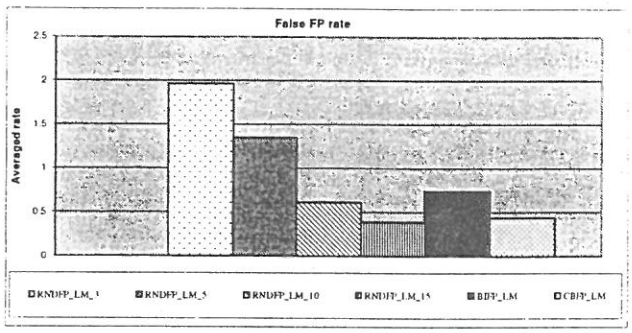

Chart 2: False FP rate by Model

The frequency of FP's represented in RNDFP_LM_10 corresponds roughly to FP's inserted around sentence boundaries. Our other model, CBFP_LM, intended to reflect linguistic knowledge about clause boundaries showed only satisfactory results. Nevertheless, the fact that RNDFP_LM_10 (average sentence length) is among the best models suggests that the clause boundary approach could be promising.

We correlated the results for the top two models, BIFP_LM and RNDFP_LM_10, in terms of FP correctness to actual FP rate in Chart 3. BIFP_LM(with 2,294,909 FP's) performs slightly better, although not significantly, on low FP users (FP rate less than 10%). For high FP users, RNDFP_LM_10 seemed to yield slightly better results. That model has 1,289,796 FP's. For these two models, the raw number of FP's in them is not a good predictor for group performance (low vs high FP users). Thus, we are concluding that it is not the raw FP frequency but rather a combination of frequency and pattern of distribution of the FP's in the training corpus that correlates with FP recognition correctness. An intuitive suggestion to populate a corpus with

fewer FP's when tailoring a model to low FP users would not necessarily yield better results. On the contrary, it might even hurt the FP correctness rate.

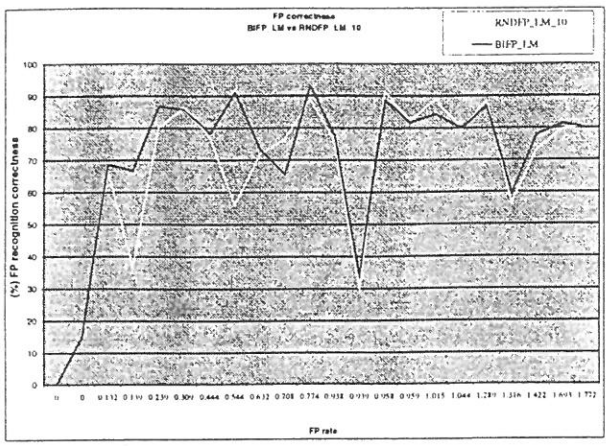

Chart 3: FP correctness – BIFP_LM vs RNDFP_LM_10 Sorted Left to Right by FP Rate

BIFP_LM implicitly incorporates linguistic knowledge since it is based on a bigram model built from a hand-transcribed corpus that includes FP's. However, the linguistic knowledge is not generalized in a scheme that we could explicitly used. Our 'true' linguistic model, CBFP_LM, does account for a very limited set of lexical anchors for clause boundaries. Despite its mediocre performance, we do think that a balanced patterned representation of FP's in the training corpus may eventually yield better results than the purely random FP insertions.

Interstingly, talkers who do not use FP's at all or use them very sparsely have improved recognition accuracy when a FP LM rather than NOFP LM is used. This side effect may be linked to fact that introducing FP's into the training corpus decreases the model's perplexity [9] and results in better recognition overall. At this point, there is not enough data to make any conclusion. We are planning to investigate this issue further.

7. CONCLUSION

The results of our study indicate that, for speech recognition purposes, FP distribution can be modeled on bigram probabilities or with a uniform distribution centered around average sentence length. We were unable to obtain satisfactory results by conditioning the training corpus with FP's based on limited linguistic knowledge, which we attribute to under representation of clause boundary rules. To improve the representation one would have to make use of a parser. We have also shown that using recognition accuracy as the only measurement to determine the goodness of a FP model is not sufficient – other methods such as FP correctness and false FP detection must be used for accurate evaluation.

AKNOWLEDGEMENTS

Our thanks go to Joan Bachenko, Ph.D., and Michael Moon, Ph.D., Linguistic Technologies, Inc., Edina, Minnesota for their invaluable advice, warm encouragement and endless patience.

NOTES

The authors can be reached at pakh0002@tc.umn.edu and savo0014@tc.umn.edu

1. The authors are also affiliated with Linguistic Technologies, Inc., 105 South 3rd St., St. Peter, MN 56082

REFERENCES

[1] Chen, S., Beeferman, Rosenfeld, R. 1998. "Evaluation metrics for language models," In DARPA Broadcast News Transcription and Understanding Workshop.

[2] Christenfeld, N, Schachter, S and Bilous, F. 1991. "Filled Pauses and Gestures: It's not coincidence," Journal of Psycholinguistic Research, Vol. 20(1).

[3] Cook, M. 1977. "The incidence of filled pauses in relation to part of speech," Language and Speech, Vol. 14, pp.135-139.

[4] Cook, M. and Lalljee, M. 1970. "The interpretation of pauses by the listener," Brit. J. Soc. Clin. Psy. Vol. 9, pp. 375-376.

[5] Cook, M., Smith, J. and Lalljee, M 1977. "Filled pauses and syntactic complexity," Language and Speech, Vol. 17, pp.11-16.

[6] Valtchev, V. Kershaw, D. and Odell, J. 1998. The truetalk transcriber book. Entropic Cambridge Research Laboratory, Cambridge, England.

[7] Lalljee, M and Cook, M. 1974. "Filled pauses and floor holding: The final test?" Semiotica, Vol. 12, pp.219-225.

[8] Maclay, H, and Osgood, C.1959. "Hesitation phenomena in spontaneous speech," Word, Vol.15, pp.19-44.

[9] Pakhomov, S. 1999. "Modeling Filled Pauses in Medical Dictations." Proc. ACL'99.

[10] Shriberg, E. E. 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis, University of California at Berkely.

[11] Shriberg, E.E. and Stolcke, A. 1996. "Word predictability after hesitations: A corpus-based study," In Proc. ICSLP.

[12] Shriberg, E.E. 1996. "Disfluencies in Switchboard," In Proc. ICSLP.

[13] Shriberg, E.E. and Bates, R. and Stolcke, A. 1997. "A prosody-only decision-tree model for disfluency detection" In Proc. EUROSPEECH.

[14] Siu, M. and Ostendorf, M. "Modeling disfluencies in conversational speech," Proc. ICSLP, 1996.

[15] Stolcke, A and Shriberg, E.1996. "Statistical language modeling for speech disfluencies," In Proc. ICASSP.

[16] Swerts, M, Wichmann, A and Beun, R. 1996. "Filled pauses as markers of discourse structure," Proc. ICSLP.

TOWARD A FORMAL CHARACTERISATION OF DISFLUENCY PROCESSING

Dafydd Gibbon and Shu-Chuan Tseng
Universität Bielefeld, Germany

ABSTRACT

Inherent structural characteristics of speech disfluencies are the prerequisite for the fulfilment of detecting and correcting speech disfluencies in spontaneous speech. However, a considerable number of recent research works on speech disfluencies focus on the surface patterns of speech disfluency editing structure, instead of looking into the relations between editing structure, the syntactic structure and the prosodic structure of speech disfluencies. In this paper we present first results of a new line of research, using feature structures modelled by finite state transducers, on the formal modelling of speech disfluencies in unplanned speech, in relation to all three levels of description.

1 INTRODUCTION

Recent studies of speech disfluencies have focussed mainly on exploring the structural characteristics of speech disfluencies, with the goal of developing psycholinguistic models. However, another line of attack is developing: an engineering need to provide robust human language technology systems, with the ability to cope with disfluencies in speech recognition, either as 'noise' or as functional components of speech, or even perhaps to introduce elements of disfluency into speech synthesis in an effort to simulate more natural and intelligible speech.

Empirical studies have shown that disfluencies are not arbitrary but can be characterised systematically. To describe the internal structure of speech disfluencies in spontaneous speech, most approaches have adopted an 'autonomous template model', without considering the relation of disfluency patterns to the syntactic contexts or the prosody of the elements concerned, for example Levelt [9], Shriberg [13], Heeman & Allen [5] and Bear & al. [2]. Heeman & Allen and Bear & al. were concerned with annotation systems for speech data. They used pattern-based detection of one-word and two-word speech repetitions, insertions and adjacent replacements, also without explicitly using syntactic context.

In this paper we present initial results of research into developing a more explicit declarative dimension to disfluency processing, in the expectation that this will make disfluency processing easier to integrate into a modern processing model. After a discussion of the functionality of disfluency we discuss a family of FST (finite state transducer) models for disfluency and illustrate the application of an FST model to data taken from a German instruction dialogue corpus [3]. The results are documented in detail in [14].

2 DISFLUENCY DETECTION

Psycholinguistic interest in disfluency is partly concerned with disfluency production and perception processes *per se*, and partly with the light that dysfunctionalities can cast here, as in other areas of language performance, on representations and processes of perception and production in general.

Current models of disfluency perception are essentially experimental. Lickley & Bard [11], for example, carried out gating experiments with the aim of finding out what kinds of linguistic cue can help human listeners to detect disfluency. Their results indicate that prosodic cues play a more decisive role in the detection of disfluency than explicit lexical cues.

Linguistic methodology is essentially based on the distributional analysis of corpora, whether small and model-directed, or large. This is the methodology of the present approach, in which computational models of patterns in a corpus of unplanned speech [3] are developed. Distributional linguistic or computational linguistic methods yield results which are in principle neutral with regard to of perception (parsing) and production (generation), though perhaps closer to production. Levelt [9], Tseng [14] and others have shown using large corpora that the majority of speech repairs, especially of complex forms, have a regular internal form, for which a *three event model* can be formulated: Levelt [9] used the categories *reparandum* (stretch of speech to be repaired), *editing term* and *alteration* for the three events, and Tseng [14] used the categories *problem item*, *editing phase* and *corrected item* in a related three event model. Levelt's three event model represents the classical template approach to disfluency structure:

Template: <OrigUtt,EdPhase,Repair>
OrigUtt=<X,reparandum,delay>
EdPhase=editterm
Repair=<retrace,alteration,Y>

The original utterance contains the reparandum, the editing phase consists of editing terms and the repair

contains the alteration, which is the correction of the reparandum.

However, in Tseng's data, the majority of complex speech disfluencies turned out to involve *items* which were *phrases*, and which are thus best characterised as *problem phrase*, *editing phase* and *corrected phrase*, where *phrase* is a unit dominated by a syntactic category (such as NP, VP, PP). This distributional result demonstrated the importance of the linguistic unit *phrase* in the production of speech disfluencies, and the need for explicit phrasal models, in contrast to the 'autonomous template' models used in earlier work which did not take phrasal syntactic structure explicitly into account,

Approaches to disfluency modelling within the engineering context of human language processing, Heeman & Allen [5], Bear & al. [2] and Nakatani & Hirschberg [12], have all used template–based annotation systems to label their data. However, more complex processing models have been used. Hindle [6] built a procedural parser to automatically detect and correct syntactic non-fluencies. Langer [8] set up normalisation rules on the basis of finite state automata to detect and correct syntactic speech repairs. Althoff et al. [1] used a finite state transducer as a word lattice parser in a speech recognition system to correct disfluencies in compound words.

In addition to syntactic disfluency modelling, other linguistic categories have been dealt with. The results on prosody by Lickley & Bard [11], have already been noted; Levelt & Cutler [10] also reported that prosodic marking was present in speech repairs, These results were confirmed, with different methodology, by Tseng [14].

From studies such as these, the conclusion can be drawn that regular patterns for the detection of disfluencies are available, and that these regular patterns may be suitable for use in disfluency detection models for cognitive processing, and in disfluency detection components of human language technology systems.

3 DISFLUENCY PROCESSING

The phase of disfluency detection is logically (not necessarily temporally) followed by the phase of disfluency processing (it is conceivable that disfluency signals may trigger hypotheses about possible repairs before the disfluency has completed its editing and alteration phases, either in the speaker or in the hearer).

3.1 Template models. As already noted, in general, disfluency models have been template-based, i.e. finite structures with slot-filler characteristics, as with the Levelt three event model. A recent integrative template-based approach is developed in Tseng [14], in which complex disfluencies in noun and prepositional phrases are formally described.

But while templates express a form of declarative 'observational adequacy', it is necessary to understand their formal properties in order to be able to suggest plausible processing models. As a first approximation, it may be suggested that disfluency templates are finite structures, and therefore by definition trivially describable by regular grammars (equivalently, finite state automata), and that correction mechanisms may be implemented as finite state transducers (FSTs).

3.2 FST models. Empirical evidence shows that although disfluency sequences can be rather short, they are in principle of arbitrary length, so that a finite template model is not helpful, and more general finite state automata with cyclic structures must be considered. General cyclic models are clearly over–powerful; there are narrow performance constraints on length and consequently additional (perhaps statistical) length constraints must be considered.

A number of empirically validated FST models have indeed been proposed, such as Langer's Disfluency Filter model Langer [8], Tseng's Disfluency Repair model [14], and the Broken Compound model of Althoff & al. (1996) [1]. The latter has been operationally validated by in the form of an implementation as a component of a a speech recognition system.

It can be shown, however, that while standard FST models are adequate for many disfluency types, more complex models are also required, which take prosody and linguistic structure into account (cf. also Lickley & Bard [11] for the detection of disfluency, relying on prosodic cues rather than explicit lexical cues), and which go beyond the classical structures of FSTs. This means that syntactic and prosodic contexts of speech disfluencies influence the production form as well as the production length of speech disfluencies. The results of modelling the distributional data as an FST are shown in Figure 1; the relation between straightforward lexico–syntactic information and 'metalocutionary' editing is coded in style of the transition graphs; for the statistical properties of the FST, see [14]; length heuristics are not considered here.

3.3 Multitape FSTs. Formally, an FST (finite state transducer) can be seen as a finite state automaton (FSA) whose transitions are labelled with elements of a vocabulary of pairs or longer tuples, rather than the atomic elements of garden variety FSAs. A standard FSA is said to accept a *regular language*, while a standard FST is said to accept a *regular relation*. The

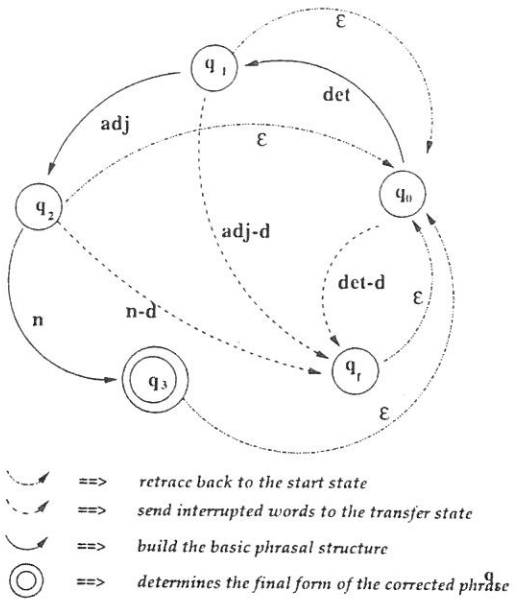

Figure 1: Relation between 'lexico-syntactic' and 'metalocutionary' information.

minimally structured (standard) FST accepts a binary relation, with the left-hand element of each pair in the relation being regarded as an *input* symbol and the right hand pair as an *output* symbol; binarity is not an essential condition, however. FSTs need not be thought of only as input/output devices; they can also be interpreted as processors for parallel streams of information. Kaplan & Kay [7] discuss the application of such FSTs to phonology.

We propose multi-tape FSTs as devices for formalising the relations between the different structural levels involved in the detection and processing of disfluencies, and that the parallel streams of information which are being processed are essentially the following:
lexico-syntactic information stream: reparandum & alteration;
prosodic information stream: pitch & duration;
metalocutionary information stream: editing term & the phonetic (and semantic) *change operations* over reparandum and alteration.

It has been shown by Carson-Berndsen [4] that FSTs can be used to interpret (i.e. represent a model for) an *event logic* model of the sequential and parallel events which make up utterances: constraints between parallel streams are mapped to (underspecified) feature structures, and sequential constraints are mapped to transitions between states of the automaton. This approach has been operationally validated in an experimental spoken language recognition system. The Levelt autonomous template model is a useful abstraction away from details of FST processing, once these details have been established, and the representation of the model as a feature structure can be regarded as a step towards a plausible underlying representation for the third, metalocutionary information stream.

The sequential constraints which map to phonetic sequences in the metalocutionary information stream can be described in terms of concatenation; the treatment of parallel constraints between information streams will be discussed briefly below.

Using a conventional feature structure notation, with co-indexing of structure, to represent the Levelt template as an abstraction from the FST, both phonetic interpretation and semantic interpretation can be related to each other systematically:

$$\left[\begin{array}{ll} \textit{disfluency} & \\ \text{PHON} & = [a] \frown [1] \frown [b] \frown [2] \frown [c] \frown [3] \\ \text{SEM} & = [2]\,([3] \sqcup_{default} [1]) \\ \text{OrigUtt} & = \left[\begin{array}{ll} a & \text{X} \\ 1 & \text{reparandum} \\ b & \text{delay} \end{array}\right] \\ \text{EdPhase} & = \left[\begin{array}{ll} 2 & \text{edit term} \end{array}\right] \\ \text{Repair} & = \left[\begin{array}{ll} c & \text{retrace} \\ 3 & \text{alteration} \end{array}\right] \end{array}\right]$$

The main semantically relevant constituents are marked with numbers, and other elements are marked with letters.

The *syntactic and prosodic control* constraints between parallel information streams can be represented by *association lines*, as shown in Figure 2.

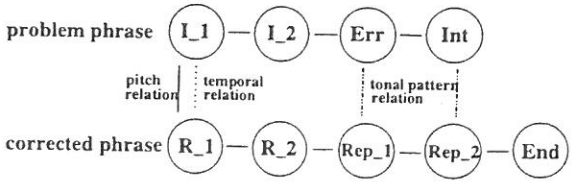

Figure 2: Prosody-metalocutionary synchronisation (association).

But why is semantic intepretation also relevant, and what might be the interpretation of the operator $\sqcup_{default}$ and the functional structure? Just as the phonetic realisation of the reparandum is a fact which remains, and is not in fact — despite current terminology — 'altered', 'repaired', or 'corrected', but simply *supplemented* by the repair, so the semantic

interpretation of the reparandum is also a fact which, particularly in the case of a contradiction (sometimes a 'Freudian slip') or in the case of a co–hyponymous meaning, remains and can be integrated into a complex semantic interpretation of the whole disfluent expression. Examples from the corpus, where semantic interpretation of the reparandum is relevant, are:

`ist bei mir auf der rech--, ich kann es auch umdrehen`

'is on my side on the ri—, I can also turn it round'

The likely semantic interpretation of the reparandum 'rech', 'right hand side' is available to the hearer if needed.

`Uhmm jetzt fügst Du mit dem mit dem an mit den drei mit den fünf Löchern UHMHM mit dem langen Stück`

'er now you start with the with the with the three with the five holes um with the long piece'

The series of abortive repairs yields a set of semantic interpretation hypotheses: is 'the long piece' the same object as the 'five hole' object?

So disfluencies are not just *noise*. The operator $\sqcup_{default}$ is *default unification*, essentially overriding of the meaning of the reparandum by the meaning of the repair (often, but not always, leading to identity) within the lexico–syntactic information stream. The metalocutionary function is qualification of the result of default unification by the the operator from the metalocutionary information stream, for instance by indicating a focus shift.

4 CONCLUSION

On the basis of distributional data collated and formalised in [14], it has been shown that disfluency structures can be represented with formal means which are already in use in computational phonology its applications to the human language technologies.

We suggest that in work in this area, well–understood representation systems and procedures for manipulating them should be used in order to facilitate the integration of 'exotic' facts about speech such as disfluencies into representations of the more familiar parts of the linguistic universe, in particular to syntax and prosody. With a strategy such as this, previous usefully illustrative, but *ad hoc* notations and diagramme styles can be superseded by formalisms rather than notations, which promise both greater generality and precision, and the hope of explanatory power within the context of language representation and processing as a whole.

We believe we have shown the feasibility of such a programme in the present paper; present work marks only a beginning, however, and leaves many gaps, such as symbolic and numerical length constraints, exact mappings to phonetic correlates, or fine details of the semantic interpretation operations, for future research.

REFERENCES

[1] Althoff, F., Drexel, G., Lüngen, H., Pampel, M. and Schillo, C. 1996. The Treatment of Compounds in a Morphological Component for Speech Recognition. In Gibbon, D. (ed.), *Natural Language Processing and Speech Technology.*

[2] Bear, J., Dowding, J. and Shriberg, E. 1992. Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog, In *ACL*, 56-63.

[3] Brindöpke, C., Häger, J., Johanntokrax, M., Pahde, A., Schwalbe, M. and Wrede, B. 1995. Darf ich dich Marvin nennen? Instruktionsdialoge in einem Wizard–of–Oz–Szenario: Szenario–Design und Auswertung. SFB 360 Report 95/16, Universität Bielefeld.

[4] Carson–Berndsen, Julie 1998. *Time Map Phonology: Finite State Methods and Event Logics in Speech Recognition.* Kluwer Academic Press: Dordrecht.

[5] Heeman, P. and Allen, J. 1997. Detecting and Correcting Speech Repairs. In *ACL*, 295-302.

[6] Hindle, D. 1983. Deterministic Parsing of Syntactic Non-fluencies. In *ACL*, 123-128.

[7] Kaplan, Ron M. & Martin Kay 1994. Regular models of phonological rule systems. *Computational Linguistics* 20(3), 331–78.

[8] Langer, H. 1990. Syntactic Normalization of Spontaneous Speech. In *COLING-90*, 180-183.

[9] Levelt, W. 1983. Monitoring and Self-Repair in Speech. *Cognition*, 14: 41-104.

[10] Levelt, W. and Cutler, A. 1983. Prosodic Marking in Speech Repair. *Journal of Semantics*, 2(2): 205-217.

[11] Lickley, R.J. and Bard, E.G. 1992. Processing Disfluent Speech: Recognising Disfluency Before Lexical Access. In *ICSLP*, 1499-1502.

[12] Nakatani, C. and Hirschberg, J. 1993. A Speech-First Model for Repair Detection and Correction. In *ARPA Workshop on Human Language Technology*, 329-334.

[13] Shriberg, E. 1994. Preliminaries to a Theory of Speech Disfluencies. PhD Thesis. University of California at Berkeley.

[14] Tseng, S.-C. 1999. Grammar, Prosody and Speech Disfluencies in Spoken Dialogues. PhD thesis. University of Bielefeld.

BETTER DETECTION OF HESITATIONS IN SPONTANEOUS SPEECH

Douglas O'Shaughnessy
INRS-Télécommunications
16 Place du Commerce, Verdun, Quebec, Canada H3E 1H6

ABSTRACT

Practical speech recognizers must accept normal conversational voice input (including hesitations). However, most automatic speech recognition work has concentrated on read speech, whose acoustic aspects differ significantly from speech found in actual dialogues. Hesitations, of which the most frequent are filled pauses, are common in natural speech, yet few recognition systems handle such disfluencies with any degree of success. Filled pauses (e.g., "uhh," "umm"), unlike most silent pauses, resemble phones which form words in continuous speech. The work reported here further develops techniques to allow automatic identification of filled pauses. Such identification, if reliable, would reduce potential confusion in determining an estimated textual output for an utterance. The Switchboard database (of natural telephone conversations) provided data for the study. While most automatic recognition methods rely entirely on spectral envelope (e.g., low-order cepstral coefficients), identifying filled pauses requires using a combination of spectra, fundamental frequency and duration. High precision and a low false alarm rate for filled pauses are feasible without excessive computation.

1. INTRODUCTION

We study here the acoustical phenomena of disfluencies in spontaneous speech. In particular, we model filled pauses, how they manifest themselves in spontaneous speech, and how they may be automatically located, for the purposes of assisting automatic speech recognition (ASR). Filled pauses are sounds usually resembling individual vowels, which are inserted into speech typically when speakers think about what to say next. The alternative of silent pauses also occurs in spontaneous conversations, but filled pauses serve the additional purpose of 'holding the floor' (i.e., hindering interruption by listeners), while silent pauses invite interruption. Filled pauses are usually centrally-articulated vowels (i.e., those requiring minimal articulatory effort, which frees the speaker to think of other things). Thus, in English, the most common form is 'uh', which greatly resembles the schwa vowel. Alternative forms, such as 'er' and 'um' (the latter being a sequence of two sounds), occur but much less often.

In fluently read speech, speakers tend to pause regularly at locations which are logical from a syntactic viewpoint (e.g., at sentence and major phrase boundaries). Furthermore, speakers rarely use filled pauses in read speech, except if they lose their place in the text. In spontaneous speech, on the other hand, hesitation pauses and restarts are widespread, and these phenomena have large effects on many aspects of the speech signal. While there are other differences between read and spontaneous speech, disfluencies are a major cause of the reduced success of ASR on spontaneous speech. Other factors rendering spontaneous speech harder to recognize include: a more variable (and often faster) speaking rate, greater use of short words and function words, and a tendency towards more coarticulation and generally less precise articulation.

A better model of filled pauses should find application in ASR. Most recognizers ignore the effects of disfluencies. Despite the appropriate use of stochastic models in ASR, many acoustical variations are poorly accounted for in the current frame-based hidden Markov model (HMM) approaches to ASR. For example, timing patterns, and duration generally, are not handled well in current systems. Some low-level phonetic duration aspects are directly encoded as geometric probability distributions (pdfs) in phonemic HMMs. In word-based HMMs, the durational effects of word-level stress can also be so encoded. However, larger durational variations, e.g., due to speaking rate changes and sentence-level stress, are mostly ignored in HMM systems, which can hinder performance. Continuous-speech recognizers are largely based on phonemic HMMs, whose state transition pdfs are biased toward average phone durations. When disfluencies cause much slower speech, the frame independence assumption of HMMs leads to spurious phoneme insertions. Thus, a better knowledge of how disfluencies affect spontaneous speech should be of assistance in designing future ASR that may handle conversational speech. While current recognizers accept different types of user speech, general spontaneous speech has not been recognized well so far.

A large database of spontaneous speech (the Switchboard database) was analyzed in terms of spectral, durational, and fundamental frequency measurements. For recognition purposes, a simple spectral analyzer was developed to distinguish filled pause sounds from other speech sounds. A primary application of this study lies in improving the performance of ASR, for applications that must accept an input

of spontaneous speech (e.g., verbal conversations with computer databases). For such purposes, we wish to eliminate the filled pause from the input sound sequence, so that the recognizer will operate on only a sound sequence of the desired words. In addition, information about where a speaker has inserted a filled pause could also be useful in deciding among alternative textual interpretations (e.g., among an N-best list of hypotheses).

Filled pauses can increase difficulties for ASR, which usually makes no provision for unpredictable sounds that resemble speech. In virtually all current recognition systems, filled pauses are interpreted as vowels and thus as parts of word hypotheses for the textual component of the recognizer. They can also cause difficulties in having a proper interpretation in the language-model component (since the language model is often trained only on fluent text, of which there is much more available). Reliable transcriptions of spontaneous speech (i.e., ones faithfully describing all acoustic events) are rare and/or small in size (e.g., such transcriptions often describe a cleaned-up version of the spontaneous speech, omitting the disfluencies). Even the transcribed Switchboard database is not fully accurate in its transcriptions.

Our previous work reported on hesitation phenomena in spontaneous speech in general, and focussed on pauses (both filled and unfilled) [7]. The current paper discusses the task of recognizing filled pauses. We present here a more comprehensive analysis of filled pauses than has usually been found in the literature, including examination of the duration and pitch of the words surrounding the pauses in spontaneous speech. In addition, we give intuitive explanations for the phenomena, based on a theory of using prosodics to cue semantic information to a listener.

2. PREVIOUS STUDIES ON DISFLUENCIES

Acoustical analyses of disfluencies with a view toward speech recognizers have only been done in the last few years. Earlier work dwelled often on the length of the word-repeat sequences (and occasionally on pause durations). Most of the work on disfluencies that has been reported in the literature has treated the phenomena in a general qualitative or overly simple quantitative fashion. For example, the linguistics literature describes where such disfluencies are likely to be found in broad terms of syntax and semantics, but gives little quantitative detail. The cognitive psychology literature gives simple statistics regarding disfluencies, in terms of frequency of occurrence. As far as we know, few reports have linked the intonational cues of both F0 (fundamental frequency) and duration to disfluencies in a way that could be useful to automatic speech recognition. Indeed, very few recognition systems use intonational cues, especially F0, at all. In this paper, we examine how these latter parameters could be exploited directly.

In examining a corpus of speech produced by people spontaneously describing colored images, Levelt [1] found in 51% of disfluency restarts that the speaker halted immediately after the word to be corrected, while 31% of the time the speaker stopped one or more words after the incorrect word (e.g., "...from green left to pink - er, from blue left to pink"). How far speakers 'back up' has been the subject of recent study [3].

Levelt found that the filled pause "uh" occurred in 30% of restarts. He noted that uttering such a neutral sound (i.e., filling the pause) may help the speaker prevent an interruption by another speaker. The implication is that listeners often interpret unfilled pauses (i.e., silence) as a cue to start speaking, but they tend not to interrupt a filled pause. Levelt noted that restarts can be either marked prosodically by changes in intonation (in the speech before and after the pause) or unmarked prosodically (i.e., no change in intonation). Cases of simple mispronunciation tended to be unmarked, whereas lexical changes (replacement of a word with a different sense) were marked. While Levelt's work is of interest here, it gave few quantitative details other than simple statistics of occurrence; in particular, F0 and durational distributions were rarely mentioned.

One major study of disfluencies in 1994 examined their detection in the ATIS corpus [6]. The authors tried to discriminate different types of word boundaries: fluent from disfluent, using a wide range of features (e.g., pause durations, energy, Fo, accent, and parts-of-speech). Using 350 utterances, each having at least one disfluency, their decision-tree algorithm found 192 of 233 repair sites, while having 19 false alarms.

Also recently, a research group at SRI tried to automatically locate filled pauses. It appears that their work has not examined direct filled-pause detection from speech, but rather by augmenting more general ASR methods. For example, in 1997 using Switchboard conversations, they reported a 92% recall rate (percentage of occurrences of filled pauses detected from among all the actual pauses) and a 13% false alarm rate (percentage of words incorrectly labeled as filled pauses) [2]. As they noted, their experiments were not a fair test of filled pause detection in the sense that, in addition to the input speech signal, they assumed knowledge of the correct word boundaries (which an ASR system would not have a priori). Proper detection of filled pauses in excess of 90% would certainly be useful in a practical recognizer, although incorrectly signaling every seventh word as a filled pause limits the usefulness of their prosody-only approach (the 1997 work did not exploit spectral detail). Their 1998 work [4] examined detection of filled pauses (among other disfluencies) through the use of a more general ASR system, including relevant language models.

Thus previous attempts at filled-pause detection seem to have significant constraints on their functioning (e.g., requiring a priori word boundaries or large-scale ASR systems). It is our intent in the current work to examine how a simple filled-pause detector might work, without a full-scale ASR system, i.e., without complex acoustic mod-

els (e.g., thousands of triphone Gaussian-mixture hidden Markov models) and without large-scale language models (e.g., trigram statistics with vocabularies of thousands of words). In the case of practical ASR, such calculation may be needed, but we wished here to see how accurate a simple processor could be in finding and delimiting filled pauses. The SRI work [2] [3] seemed to require additional calculations beyond most ASR systems (e.g., F0) and/or assumed a priori linguistic knowledge (other than directly furnished by the speech signal). We indeed use F0 as an important parameter in filled-pause detection, but exploit no database-dependent language models nor require complex acoustic modeling. Our analysis is based strictly on the speech input, and does not assume access to the correct text (as would be the case in an actual recognition situation), nor even on language models (which can vary widely, depending on the application, and on the subject of current conversation).

The importance of identifying filled pauses has been noted in recent ASR work. For example, to create an efficient language model for human-machine dialogues, Reichel et al [5] manually remove all "words representing noise in the input, such as 'uh' and 'um.'" Another very recent work noted that filled pauses tend to occur very often in user requests over the telephone (e.g., 7189 such pauses in 10600 responses to "How may I help you?") [8]. Direct accounting for such disfluencies assisted in performance improvements in this practical ASR system.

3. SPEECH DATABASE

We examined disfluencies in a standard speech database (used by several speech recognition research groups in North America), called Switchboard. It contains about thirty hours of actual conversations (each about five minutes in duration), over normal switched telephone lines, each one between two strangers conversing on an assigned topic. As such, it is fairly natural spontaneous speech (although less so than in the CallHome database of conversations among friends and family), and thus has many disfluencies. It differs significantly from other databases: unlike the Wall Street Journal database, it is telephone-based and spontaneous; unlike the ATIS database, it has real conversations; unlike both, it has yielded relatively poor recognition rates so far (e.g., less than 70% word accuracy). While CallHome may present even more recognition difficulties, Switchboard is more realistic for applications since it involves strangers talking to each other, as would be the case in many commercial dialogue requests. While one could argue that ATIS may be more appropriate in that it simulated travel agent requests, Switchboard used the telephone network (which most applications would do) and its dialog interaction was much more real and rapid than the Wizard-of-Oz simulation for ATIS.

4. ANALYSIS METHOD

The filled pause detection algorithm was developed by analysis of many utterances from different speakers, in the training set of utterances. A separate set of utterances (not examined during training) were used to test the recognition algorithm, with speakers who were different from those in the training set (and chosen at random from the database). All recognition phases of the task used automatic spectral, F0 and duration estimation, in conjunction with a simple expert-system recognition. Spectral analysis involved examining each 10-msec frame of speech data, using a Hamming window of 25.6 msec, to obtain a DFT amplitude spectrum. This spectrum was then automatically scanned to extract the highest-energy harmonics and broad spectral peaks (roughly, but not exactly, the major formants). F0 estimates were provided by a simple analysis of these narrowband spectra, searching for harmonics (and assuming a continuity of F0 across successive frames).

Silent pauses were easily located using a silence energy threshold, relative to an estimated level of background noise (which, in practice, was not high in most Switchboard conversations). However, silent periods of short duration (e.g., those typically less than 120 ms) can be actually part of intended speech (i.e., stop closures). Since we were not attempting full ASR in these experiments, we did not concern ourselves with such otherwise important issues, but simply viewed silences of greater than 120 ms as effective pauses (and hence ignored shorter silences, even though some of these actually were brief pauses, rather than stop closures – to discriminate these, one would have to examine phonetic context and the ensuing spectra in more detail, for possible stop burst releases and formant transitions into the next vowel).

Instead, the main focus of this research concerned filled pauses. Their presence was estimated and located as long, steady vowels (exceeding 120 msec) with low F0 (relative to the calculated average F0 for each speaker during a conversation) and a neutral spectral pattern (again, compared against an average vowel spectral pattern computed over the conversation). Specifically, strong, periodic (i.e., vowel-like) sounds whose spectra closely resembled central vowels were the only candidates considered, although those with an ensuing long nasal were candidates for the rarer 'umm' filled pause. In comparing with the ATIS-database filled pauses [7], we relaxed the F0 restriction to be very low, since Switchboard filled pauses include many examples with rising F0 (especially at turn-taking points, where one conversant paused and the other was not ready to take over the conversation). The duration of such phonetic events had to be 120 ms or more to be considered as a filled pause. Examples of shorter cases were mostly simple vowels as parts of words. A significant minority of filled pauses were preceded and/or followed by silent pauses. The presence then of a steady central vowel, preceded and/or followed by a silence of more than 120 ms was a very reliable indicator of a filled pause.

Since full ASR was not necessary here, we did not use more complex spectral analysis, as found in other systems.

In particular, the mel-based cepstrum and determination of mixtures of Gaussian probabilities were not needed. To distinguish filled pauses from the rest of speech, a simpler spectral analysis can suffice. We need only enough spectral detail to see whether the sound is fairly steady over 100 ms or so and resembles a neutral vowel. More precise specification is of course needed for other ASR tasks, but given the variability of speakers and recording conditions, we decided that simple analysis would be best for this preliminary filled-pause detector.

Such simple spectral analysis does not require formant tracking (which can be very difficult in general). It can nonetheless allow more precision in specific spectral matching than HMMs, without requiring the more complex methods used in many recognizers (e.g., with cepstral coefficients).

5. ACOUSTICAL ANALYSIS RESULTS

In the many utterances examined, each conversation (average of 5 minutes) averaged about a dozen or more filled pauses. Filled pauses tended to occur: 1) often at the start of a conversation turn, 2) often within the first three words of a clause (e.g., right after 'I, you, and, but, so....'), 3) at major syntactic boundaries (e.g., after a clause), 4) right after 'the' or 'a' or 'is', 4) less often at the end of a turn (in this last case, the filled pause acted as a cue for the other person to speak, unlike the hold-the-floor role the filled pause had elsewhere).

Average durations for filled pauses were: 1) at the start of a speaker turn: a median and mean of about 220 ms (typical range: 90-500 ms), 2) later in a turn: a median and mean of about 170 ms (range: 40-400 ms), and 3) with no adjacent silence: a median and mean of 150 ms. Thus the most difficult filled pauses to recognize are the third case, where no silence is adjacent to the filled pause, and the only cue is a long schwa-like sound (albeit shorter than silent-adjacent filled pauses).

Possible confusions between a filled pause and an intended speech sound include the following: 1) the word 'a' (at usually 40-80 ms, this article is much shorter than virtually all filled pauses), 2) word-initial schwa (e.g., 'about,' 'announce' – here too, non-filled-pause cases are typically much shorter), 3) word-final schwa (much rarer in English words), 4) 'uh-huh' (typical background agreement response of a listener). To distinguish between a filled pause and 'uh-huh: the latter never has a pause between syllables, it almost always occurs right after another talker's speech, and has a total duration of 140-360 ms (much longer than filled pauses).

Filled pause frequency was highly speaker-dependent (i.e., some speakers are much more disfluent than others). In terms of filled pauses per 100 words of speech in Switchboard, the most fluent speakers used only two, while some used fifteen. The mean and median were 6-7 filled pauses per 100 words. In comparison with the less-spontaneous ATIS data, the filled pauses were shorter and had more variable F0 patterns.

6. CONCLUSION

This paper has detailed acoustic phenomena of filled pauses in a multi-speaker database of spontaneous, continuous speech, and has given intuitive explanations for them, based on a theory of using prosodics to cue semantic information to a listener. It has also given an approach for automatic location of filled pauses, capable of high precision and low false alarm rates. Based on an analysis of the acoustic data alone (no language modeling), such location is feasible, and can be of assistance in the context of an automatic speech recognizer. The filled pauses can be distinguished acoustically, via an analysis of duration, F0 and spectral detail in the neighborhood of a pause.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] Levelt, W. *Speaking: From Intention to articulation.* 1989. Cambridge, MA: MIT Press.

[2] Shriberg, E., Bates, R. and Stolcke, A. A Prosody-only decision-tree model for disfluency detection. 1997. In *Eurospeech-97*, Rhodes, 2383-2386.

[3] Shriberg, E. and Stolcke, A. How far do speakers back up in repairs? A quantitative model. 1998. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, 2183-2186.

[4] Stolcke, A. et al. Automatic detection of sentence boundaries and disfluencies based on recognized words. 1998. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, 2247-2250.

[5] Reichel, W., Carpenter, B., Chu-Carroll, J. and Chou, W. 1998. Langage modeling for content extraction in human-computer dialogues. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, 2315-2318.

[6] Nakatani, C. and Hirschberg, J. 1994. A Corpus-based study of repair cues in spontaneous speech. Journal of the Acoustical Society of America, vol. 95, 1603–1616.

[7] O'Shaughnessy, D. 1992. Recognition of hesitations in spontaneous speech. In *Proc. of the Intern. Conference on Acoustics, Speech and Signal Processing*, 593–596.

[8] Rose, R. and Riccardi, G. Modeling disfluency and background events in ASR for a natural language understanding task. 1999. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, 341–344.

DETECTING AND CORRECTING SPEECH REPAIRS IN JAPANESE

Peter A. Heeman* and K. H. Loken-Kim†
*Oregon Graduate Institute, Portland OR
†fonix Corporation, Woburn MA

ABSTRACT

One of the characteristics of spontaneous speech is the abundance of speech repairs, in which speakers go back and repeat or change something they have just said. In other work [7], we proposed a language model for speech recognition that can detect and correct speech repairs in English. In this paper, we show that this model works equally as well on a Japanese corpus of spontaneous speech. The structure of the model captures the language independent aspect of speech repairs, while machine training techniques on an annotated corpus learn the language dependent aspects.

1. INTRODUCTION

One of the biggest challenges in recognizing and understanding spontaneous speech is dealing with speech repairs, where speakers go back and change or repeat something they have just said. The following illustrates an English speech repair from the Trains corpus [6], a corpus of human-human task-oriented spoken dialogs.

Example 1

we'll pick up a tank of uh the tanker of oranges
reparandum *ip* *et* *alteration*

In this example, the speaker replaced "a tank of" by "the tanker of". The speech that was replaced is referred to as the *reparandum* of the repair, and the speech that replaces it is referred to as the *alteration*. The end of the reparandum is the *interruption point* [10] and is sometimes followed by an *editing term*, such as "uh", "let's see", "okay", and "well". In order to understand a speaker's turn, speech repairs need to be detected and the extent of their reparandum and editing terms determined, which we refer to as *correcting* the repair.

Sometimes, editing terms occur by themselves in the middle of an utterance, as illustrated in the following.

Example 2

we need to um manage to get the bananas to Dansville
ip *et*

Following Levelt [10], we view such phenomena as speech repairs, and refer to them as *abridged repairs*. Such repairs also need to be detected and corrected, which might not be trivial (cf. [2]). Words, such as "well" and "okay", can be ambiguous as to whether they are part of an editing term or are part of the sentential content. Furthermore, there might be word correspondences across the editing term that are just spurious. In the above example, the correspondences between "need to" and "manage to" should not be taken as evidence that the reparandum is "need to".

Speech repairs are not limited to English. The following is a Japanese speech repair, with its English translation written underneath [13].

Example 3

hotelu kara shitsureishmashita kaigijyo kara
hotel from I am sorry conference center from
reparandum *et* *alteration*

Examples of editing terms in Japanese are "eh", "ah" and "shitsureishmashita". They also exhibit the same phenomena, such as word correspondences between the reparandum and alteration.

Speech repairs are a natural part of spontaneous speech and the extra words that they add in constitutes 10% of the words in the Trains corpus of English human-human task-oriented dialogues, with a repair occurring in approximately 54% of all speaker turns with at least ten words [7]. Listeners are able to process speech repairs without conscious effort [14], and very quickly after they occur [11]. This suggests that hearers process speech repairs early on in processing, possibly using local information.

Automatic detection and correction of speech repairs is difficult since it requires combining multiple sources of evidence [2]. This resolution must be done early on, most probably during speech recognition, due to the interactions that exist between speech repairs and predicting the next word [7]. In fact, we feel that the local context contains sufficient evidence to detect and correct most repairs.

In other work [8, 7], we proposed a model for detecting and correcting English speech repairs. Our model for resolving speech repairs results from redefining the speech recognition problem so that it not only hypothesizes the word sequence, but also hypothesizes the existence of speech repairs and their corrections, intonational phrasing, discourse markers and shallow syntactic analysis (part-of-speech tags). The repair processing model itself is not language specific; rather, it provides an architecture for accounting for lan-

There tends to be word correspondences between the reparandum and alteration [10, 2]. The word correspondence might be a word repetition or a word replacement, but with the same part-of-speech (POS) tag. Word correspondences tend to be cross-serial [7].

The alteration will be a fluent continuation of the words before the reparandum [9].

The reparandum onset can depend on the syntactic context [10].

Speech repair (and editing term) occurrence can depend on the syntactic context.

Pauses might co-occur with the interruption point of speech repairs as well as at the end of the editing term.

Not all repairs have an editing term, and editing terms can consist of more than one word or phrase.

Words that can be used as editing terms can be ambiguous as to whether they are being used in a repair or not.

Table 1: Speech Repair Assumptions

guage usage. Table 1 gives some of the aspects of speech repairs that can be captured in the model.

Within the above framework, the model is parameterized by deriving probability estimates from training data. These probability estimates capture the language and domain specific information needed to instantiate the model. These parameters include how likely certain word sequences are to be used as editing terms, what syntactic contexts repairs are likely to occur in (both the interruption point and the reparandum onset), and what constituents syntactic well-formedness. The lack of syntactic well-formedness can discourage a fluent interpretation for a stretch of speech while the presence of syntactic well-formedness is used in finding fluent continuations between the words before the reparandum onset and the alteration. The parameters also indicate what constituents likely correspondences between the reparandum and alteration.

In this paper, we demonstrate that our repair processing model is not language specific by training and testing it on a corpus of Japanese task-oriented dialogues. The training data is used to tune the parameters of the model, such as which Japanese words tend to used as editing terms (such as "ah"), and the syntactic contexts in which they tend to occur. In the rest of the paper, we first discuss the Japanese corpus. We then briefly describe our statistical language model, report the results of detecting and correcting the speech repairs, and compare these results with others that have been reported in the literature. We then conclude with a discussion of the implication of the results.

2. CORPUS

The Japanese corpus is a collection of human-human task-oriented dialogs [12]. The ATR Japanese Morphological analysis manual [16] was used for tokenization and for assigning part-of-speech tags. The tagset consisted of 30 POS tags. There was 304,000 words of data in the corpus and 659 speech repairs (excluding abridged repairs). Speech repairs were annotated using a scheme that captures the extent of the reparandum and editing terms [13]. However, only non-abridged repairs were considered in this experiment. Intonational phrases were not annotated, nor was the amount of silence between words available.

3. MODEL

Our statistical language model associates seven variables for each word that is postulated.[1] Five of these variables are for detecting and correcting speech repairs.

- E_i indicates if word i starts, continues or ends an editing term.
- R_i indicates if word i is the interruption point of a speech repair. For repairs with editing terms, the repair is marked on the end of the editing term.
- O_i indicates which previous word is the reparandum onset (only applicable if R_i indicates a repair).
- L_i indicates which word in the reparandum corresponds (or licenses) the current word (only applicable if processing a repair).
- C_i indicates the type of correspondence between word i and its licensor. This can be a word match, a replacement by a word of the same POS tag, or other.
- P_i indicates the POS tag for the current word.
- W_i indicates the word identity for the current word.

Using the above seven variables, the speech recognition problem is defined as follows, where the recognizer searches for the best sequence of words, POS tags, and speech repairs given the acoustic signal.

$$
\begin{aligned}
&\hat{W}\hat{P}\hat{C}\hat{L}\hat{O}\hat{R}\hat{E} \\
&= \arg\max_{WPCLORE} \Pr(WPCLORE|A) \\
&= \arg\max_{WPCLORE} \frac{\Pr(A|WPCLORE)\Pr(WPCLORE)}{\Pr(A)} \\
&= \arg\max_{WPCLORE} \Pr(A|WPCLORE)\Pr(WPCLORE)
\end{aligned}
$$

[1] Our full model [7] also includes a variable for intonational phrase boundaries and a variable to account for silences between words.

The second term in the last equation above is the new language model, and assigns a probability to how likely the sequence of words, POS tags, and speech repair markings are. We rewrite this term as follows.

$$\begin{aligned}
&\Pr(WPCLORE) \\
&= \prod_{i=1}^{N} \Pr(E_i|W_{1,i-1}P_{1,i-1}C_{1,i-1}L_{1,i-1}O_{1,i-1}R_{1,i-1}E_{1,i-1}) \\
&\quad \Pr(R_i|W_{1,i-1}P_{1,i-1}C_{1,i-1}L_{1,i-1}O_{1,i-1}R_{1,i-1}E_{1,i}) \\
&\quad \Pr(O_i|W_{1,i-1}P_{1,i-1}C_{1,i-1}L_{1,i-1}O_{1,i-1}R_{1,i}E_{1,i}) \\
&\quad \Pr(L_i|W_{1,i-1}P_{1,i-1}C_{1,i-1}L_{1,i-1}O_{1,i}R_{1,i}E_{1,i}) \\
&\quad \Pr(C_i|W_{1,i-1}P_{1,i-1}C_{1,i-1}L_{1,i}O_{1,i}R_{1,i}E_{1,i}) \\
&\quad \Pr(P_i|W_{1,i-1}P_{1,i-1}C_{1,i}L_{1,i}O_{1,i}R_{1,i}E_{1,i}) \\
&\quad \Pr(W_i|W_{1,i-1}P_{1,i}C_{1,i}L_{1,i}O_{1,i}R_{1,i}E_{1,i})
\end{aligned}$$

From the above, we can see that the language model is made up of seven probability distributions, one corresponding to each variable that needs to be hypothesized for each word. Each probability distribution depends on the hypothesized values of the preceding variables. For instance, after the editing term variable is hypothesized for the current word, we use its value as part of the context for hypothesizing the value of the repair variable. In searching for the best interpretation of the variables, we follow the technique proposed by Chow and Schwartz [3] and only keep a small number of alternative paths by pruning the low probability paths after processing each word.

We use a decision tree learning algorithm to estimate the probabilities. The decision tree algorithm learns a hierarchical set of equivalence classes of the context and uses interpolated estimation to compute the probabilities [1]. The set of questions that we allow the decision tree algorithm to ask takes into account our assumptions about speech repairs that we give in Table 1. For instance, the context for hypothesizing the reparandum onset, O_i, allows the decision tree to make generalizations about the syntactic contexts in which the reparandum might start. By encoding these assumptions into the questions, the decision tree should be able to make better generalizations about the nature of speech repairs, thus giving better probability estimates, especially for limited amounts of training data.

We also do hierarchical clustering of the words and POS tags [5] that groups together similarly behaving words and tags. This results in a binary encoding for each word and POS tag. The decision tree can then ask questions about the words or POS tags in the context by asking questions about the bit encodings.

4. RESULTS

To make the best use of our limited data, we used a six-fold cross-validation procedure: each sixth of the data was tested using a model built from the remaining data. We also changed all word fragments into the token <**fragment**> with POS tag **FRAGMENT**. Changes in speaker are marked in the word transcription with the special token <**turn**>. Since current speech recognition rates for spontaneous speech are quite low, we restricted our algorithm to only consider the hand-annotated word transcriptions.

		Japanese Results	Trains: Comparable	Trains: Full Model
Detection	recall	78.6	64.9	68.2
	precision	74.9	80.0	81.0
	error rate	47.8	51.3	47.8
Correction	recall	71.8	58.8	62.3
	precision	68.4	72.5	74.0
	error rate	61.5	63.6	59.6

Table 2: Results on Japanese and English Corpora

Table 2 gives the results of running the model on the Japanese corpus and on the English Trains corpus. The results for the Japanese corpus are given in the second column. We achieved a detection recall rate of 78.6% with a precision of 74.9%, and a correction recall rate of 71.8% with a precision of 68.4%.[2] To provide a fair comparison with the Trains corpus, we run a version of our model on this corpus that models the detection and correction of speech repairs but does not model intonational phrasing nor use silence information. Unfortunately, we were unable to remove the modeling of abridged repairs, which were not used in the Japanese corpus. The results of this model are given in column three, with the results for the modification repairs and fresh starts combined. As can be seen from the error rates, the Japanese results are comparable to the results on the Trains corpus. The fourth column adds silence information and intonational modeling to the Trains results. The results in this column give an indication of how much the Japanese results can improve. Furthermore, as other acoustic cues are added into the model, the results should further improve.

5. COMPARISON

This work expands on previous work on detecting and correcting speech repairs in Japanese. Sagawa *et al.*[15] proposed a parser-first approach for detecting and correcting repairs (c.f. [4]). If the parser failed on an utterance, the

[2] A repair is counted as correctly detected if the interruption point is correctly identified along with the type of repair. A repair is counted as properly corrected if the interruption point is correctly identified, and its editing terms and reparanda are correctly found. In the case of overlapping repairs, as long as the total extent of the reparanda are identified, all repairs are counted as correct.

utterance would be passed to a *translator* that would search for a repair and correct it, and pass the resolved utterance back to the parser. To deal with utterances that contain more than one repair, this process was repeated until the utterance could be parsed. The translator has a set of rules that it tries to apply for detecting and correcting speech repairs. As we argued elsewhere [7], not all speech repairs are syntactically ill-formed (at least in English); rather, there are a number of sources of evidence that need to be combined in order to decide if a repair has occurred. Furthermore, modeling repairs is strongly intertwined with speech recognition. Hence, resolving them after speech recognition will make it more difficult to correctly recognize the speech.

Kikui and Morimoto [9] proposed a method for correcting Japanese speech repairs. They start with the word transcription, its POS analysis, the utterance boundaries, and the speech repair interruption points. Their method combines both looking for strong similarities between the reparandum and alteration, as well as finding a syntactically well-formed continuation from the speech before the reparandum to the alteration (in terms of allowed POS adjacency). With this method they are able to correct 94% of all of the speech repairs. However, information about the proposed corrections is strong evidence about whether a repair actually occurred [8]. Hence, speech repair detection and correction should not be separated.

6. CONCLUSION

In this paper, we have shown that the model that we proposed in previous work, also works on a Japanese corpus. Only general characteristics of speech repairs are hard-coded into the model, such as being able to have correspondences between their reparanda and alterations, having reparandum onsets that have certain syntactic regularities, having a lack of syntactic well-formedness across the interruption point, and having editing terms. All other parameters are learned from the corpus. We find that this algorithm is indeed able to model Japanese speech repairs, and is able to detect and correct 71.8% of the repairs with a precision of 68.4%. These results are comparable to results obtained on the English Trains corpus.

7. ACKNOWLEDGMENTS

Preliminary research work was completed while both authors were at ATR Interpreting Telecommunications Laboratory in Japan. The authors wish to thank ATR for the use of the Japanese corpus, Aki Yokoo, Tsuyoshi Morimoto and Yoshihiro Kitagawa. The first author is partially funded by a grant from Intel.

8. REFERENCES

[1] L. Bahl, P. Brown, P. deSouza, and R. Mercer. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1001–1008, 1989.

[2] J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 56–63, 1992.

[3] Y. Chow and R. Schwartz. The n-best algorithm: An efficient procedure for finding top n sentence hypotheses. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 199–202, 1989.

[4] J. Dowding, J. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran. Gemini: A natural language system for spoken-language understanding. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 54–61, 1993.

[5] P. Heeman. POS tags and decision trees for language modeling. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Maryland, June 1999.

[6] P. Heeman and J. Allen. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium, 1995.

[7] P. Heeman and J. Allen. Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialog. *Computational Linguistics*, 25(4), 1999.

[8] P. Heeman, K.H. Loken-Kim, and J. Allen. Combining the detection and correction of speech repairs. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-96)*, pages 358–361, 1996.

[9] G. Kikui and T. Morimoto. Similarity-based identification of repairs in Japanese spoken language. In *Proceedings of the 3rd International Conference on Spoken Language Processing*, pages 915–918, 1994.

[10] W. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[11] R. Lickley and E. Bard. Processing disfluent speech: Recognizing disfluency before lexical access. In *Proceedings of the 2nd International Conference on Spoken Language Processing*, pages 935–938, October 1992.

[12] K.H. Loken-Kim, F. Yato, L. Fais, T. Kurihara, M. Yoshigawa, and Y. Kitagawa. Transcription of spontaneous speech collected using a multimodal simulator–EMMI in a direction finding task. TR-IT-0029, ATR-ITL, 1993.

[13] K.H. Loken-Kim, Y. Kitagawa, and Y. Ohta. Linguistic analysis of speech disfluency in the atr language database. TR-IT-0107, ATR-ITL, 1995.

[14] J. Martin and W. Strange. The perception of hesitation in spontaneous speech. *Perception and Psychophysics*, 53:1–15, 1968.

[15] Y. Sagawa, N. Ohnishi, and N. Sugie. A parser coping with self-repaired Japanese utterances and large corpu-based evaluation. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 593–597, 1994.

[16] N. Uratani, T. Tashiro, H. Yamada, and K. Matsumoto. Users manual for japanese morphological analysis in the ATR spoken language database. TR-IT-0009, ATR-ITL, 1993.

SPEECH REPAIRS: A PARSING PERSPECTIVE

Mark G. Core and Lenhart K. Schubert
mcore,schubert@cs.rochester.edu
http://www.cs.rochester.edu/u/mcore
Computer Science Department
University of Rochester
Rochester NY 14627

ABSTRACT

This paper presents a grammatical and processing framework for handling speech repairs. The proposed framework has proved adequate for a collection of human-human task-oriented dialogs. both in a full manual examination of the corpus, and in tests with a parser capable of parsing some of that corpus. This parser can also correct a pre-parser speech repair identifier producing increases in recall varying from 2% to 4.8%.

1. MOTIVATION

In the discussion below, we adopt the convention of using the term speech repair to include hesitations. Many speech repairs have associated editing terms (*I mean, um*), and abridged repairs [6]) consist solely of editing terms (i.e. they have no corrections).

Speech-based dialog systems often attempt to identify speech repairs in the speech recognition phase (prior to parsing) so that speech repairs will not disrupt the speech recognizer's language model ([6],[7],[8]). In such a system, it is then tempting to remove conjectured reparanda (corrected material) and editing terms from the input prior to further processing. There are two issues that need to be addressed in such an approach, one pertaining to dialog interpretation and the other to parsing. First, how can the dialog manager of the system access and interpret these editing terms and reparanda, if the need arises? Such a situation could occur in an example such as *take the oranges to Elmira, um, I mean, take them to Corning*; here reference resolution requires processing of the reparandum. Also, the system might want to access the reparanda and editing terms to see the speaker's original thoughts and any hesitations, for instance as indicators of uncertainty. For more details see [3]. Second, if speech repair identification occurs before parsing, should the parser be made aware of reparanda?

We believe that the answer to the second question should be yes. The parser has more information about the possible grammatical structures in the input than a pre-parser repair identifier and can possibly correct errors made by it. This point applies not only to each speaker's contributions in isolation but also to the interactions between contributions. An example is provided by utterances 11-15 of TRAINS dialog [5] d91-6.1 (Figure 1) where the interleaving of speaker contributions can help identify repairs.

(To fit the example on one line, we have abbreviated the initial part of *u*'s contribution, *move the engine at Avon engine E to*, to *move engine E to*.) Repair detection and correction typically act on only one speaker's stream of words at a time. If for some reason, the corrections *E one*, *en-*, and *engine E one* were not recognized by a pre-parser repair detector, the parser's knowledge of *s*'s correction might help find these repairs. If the dialog parser treats the words of the two speakers as a single stream of data (as ours in fact does), *s*'s correction appears right after the phrase it corrects.

This paper presents a framework that addresses both sorts of issues above. The framework allows for complete phrase structure representations of utterances containing repairs without removing reparanda from the input. Thus structural analyses of repairs are made available to the dialog manager. The idea is to create two or more interpretations for each repair; one interpretation for the corrected utterance and one possibly partial interpretation for what the speaker started to say. Editing terms are considered separate utterances embedded in the main utterance.

The focus of this paper is on the second issue, i.e., testing the ability of a parser to improve pre-parser speech repair identification. We show that by applying the parser's knowledge of grammar and of the syntactic structure of the input to the hypotheses made by the pre-parser repair identifier, we can improve upon those hypotheses.

2. HOW THE PARSER ACCOMODATES REPAIRS

The parser deals with reparanda and editing terms via metarules. The term metarule is used because these rules act not on words but on grammatical structures. Consider the *editing term metarule*. When an editing term is seen[1], the metarule extends copies of all phrase hypotheses ending at the editing term over that term to allow utterances to be formed around it. This metarule (and our other metarules) can be viewed declaratively as specifying allowable patterns of phrase breakage and interleaving [2]. This notion is different from the traditional linguistic conception of metarules as rules for generating new PSRs from given PSRs.[2] Procedurally, we can think of metarules as creating new (discontinuous) pathways for the parser's traversal of the input, and this view is readily implementable.

The repair metarule, when given the hypothetical start and end of a reparandum (say from a language model such as [6]),

```
U: move engine E to            E one    en- engine E one to Bath
S:                  engine E one     okay
```

Figure 1: Utterances 11-15 of TRAINS dialog d91-6.1

extends copies of phrase hypotheses over the reparandum allowing the corrected utterance to be formed. In case the source of the reparandum information gave a false alarm, the alternative of not skipping the reparandum is still available.

For each utterance in the input, the parser needs to find an interpretation that starts at the first word of the input and ends at the last word. This interpretation may have been produced by one or more applications of the repair metarule allowing the interpretation to exclude one or more reparanda. For each reparandum skipped, the parser needs to find an interpretation of what the user started to say. In some cases, what the user started to say is a complete constituent: *take the oranges I mean take the bananas*. Otherwise, the parser needs to look for an incomplete interpretation ending at the reparandum end. Typically, there will be many such interpretations; the parser searches for the longest interpretations and then ranks them based on their category: UTT > S > VP > PP, and so on. The incomplete interpretation may not extend all the way to the start of the utterance in which case the process of searching for incomplete interpretations is repeated. Of course the search process is restricted by the first incomplete constituent. If, for example, an incomplete PP were found then any additional incomplete constituent would have to expect a PP.

Figure 2 shows an example of this process on utterance 62 from TRAINS dialog d92a-1.2. Assuming perfect speech repair identification, the repair metarule will be fired from position 0 to position 5 meaning the parser needs to find an interpretation starting at position 5 and ending at the last position in the input. This interpretation (the corrected utterance) is shown under the words in figure 2 The parser then needs to find an interpretation of what the speaker started to say. There are no complete constituents ending at position 5. The parser instead finds the incomplete constituent `ADVBL -> adv • ADVBL`. Our implementation is a chart parser and accordingly incomplete constituents are represented as arcs. This arc only covers the word *through* so another arc needs to be found. The arc `S -> S • ADVBL` expects an ADVBL and covers the rest of the input, completing the interpretation of what the user started to say (as shown on the top of figure 2). The editing terms are treated as separate utterances via the editing term metarule. For more details including a discussion of second speaker interruptions see [2],[4]

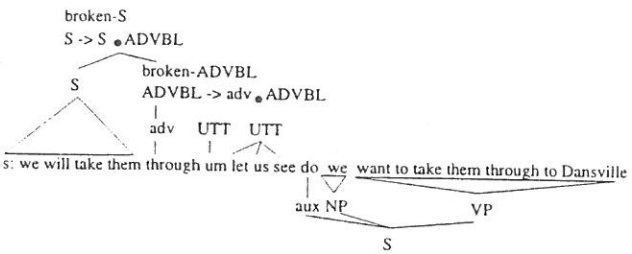

Figure 2: Utterance 62 of d92a-1.2

3. RESCORING A PRE-PARSER SPEECH REPAIR IDENTIFIER

Given that the parser can accept input containing reparanda and editing terms, a pre-parser repair identifier does not have to "clean up" input by removing hypothesized reparanda and editing terms. It can instead give the parser its n-best hypotheses about possible reparanda and editing terms. For this paper, we put aside the question of how the parser determines when utterances end. In the experiments below, the parser will always be given utterance endpoints. For each utterance, the parser can try various hypotheses from the repair identifier. Based on the grammaticality of these hypotheses and any scores previously assigned to them, the parser decides which one is correct.

To test whether such post-correction would improve recall, the parser described in section 2 was connected to Heeman's speech repair identifier [6]. The latter produced up to 100 hypotheses about the speech repairs, boundary tones, and parts of speech associated with the words of each turn in the test corpus. Each hypothesis was given an estimated probability.

Both the parser and Heeman's speech repair identifier were developed and tested on the TRAINS corpus [5]. However, Heeman's testing data was broken into two streams for the two speakers while the test data for the parser merged the two speakers' words into one data stream. The differences in segmentation resulted in different speech repair annotations.

3.1 Experiment One

The first experiment used the parser's speech repair annotations. The version of Heeman's module used is prior to the one reported in [6]. Correspondingly, the recall and precision of this module are lower than current versions. The recall and precision of the model on the test corpus is shown in table 1. The test corpus consisted of 541 repairs, 3797 utterances, and 20,069 words.[3]

To correct Heeman's output, the parser starts by trying his module's first choice. If this results in an interpretation covering the input, that choice is selected as the correct answer. Otherwise the process is repeated with the module's next choice. If all the choices are exhausted and no interpretations are found, then the first choice is selected as correct. This approach is similar to an experiment in [1] except that Bear et al. were more interested in reducing false alarms. Thus, if a sentence parsed without the repair then it was ruled a false alarm. Here the goal is to increase recall by trying lower probability alternatives when no parse can be found.

Repairs correctly guessed	271
False alarms	215
Missed	270
Recall	50.09%
Precision	55.76%

Table 1: Heeman's Speech Repair Results from Exp 1

Repairs correctly guessed	284
False alarms	371
Missed	257
Recall	52.50%
Precision	43.36%

Table 2: Augmented Speech Repair Results from Exp 1

The results of such an approach on the test corpus are listed in table 2. Recall increased by 4.8% (13 cases out of 541 repairs), showing promise in the technique of rescoring the output of a pre-parser speech repair identifier.

One factor relating directly to the effectiveness of the parser at correcting speech repair identification is the percent of fluent or corrected utterances that the parser's grammar covers. In a random sample of 100 utterances from the corpus, 65 received

some interpretation. However, 37 of these utterances are one word long (*okay*, *yeah*, etc.) and 5 utterances were question answers (*two hours*, *in Elmira*); thus on interesting utterances, likely to have repairs, accuracy is 39.7%. What happens when a fluent or corrected utterance cannot be parsed is that the parser may pick a low scoring repair hypothesis that eliminates the unparsable material (this may be most of the utterance). This situation results in a false alarm and actual repairs in the input may be missed.

A question raised by this experiment was the effect of knowing utterance boundaries in choosing a repair hypothesis. All of Heeman's repair hypotheses were truncated to fit within utterance boundaries. However, this may have resulted in obviously incorrect hypotheses that the parser could easily eliminate. If Heeman's module had known of utterance boundaries at the outset it could have eliminated these possibilities itself. The baseline measures of the second experiment were adjusted to control for this advantage.

3.2 Experiment Two

In the second experiment, the most recent version of Heeman's repair identifier was used; a baseline measure considering the effect of utterance boundaries was calculated; and Heeman's segmentation of the TRAINS corpus was used. Heeman's segmentation broke the input into two parts, one for each speaker, and further divided those into turns. The author broke turns into utterances as defined by the parser's grammar. Heeman's scoring module worked on a per-turn basis, meaning if a turn had several utterances the parser was not allowed to pick one hypothesis for the first utterance and a different one for the second. The parser scored the different hypotheses based on the number of words that parsed for each hypothesis. So if one hypothesis allowed two utterances to parse, one containing 5 words and another containing 7 words, its score would be 12. The hypothesis with the highest score was picked. In the case of ties, the hypothesis with the higher probability (as assigned by Heeman) was chosen.

To construct a baseline measurement taking into account the effect of utterance boundaries, hypotheses output by Heeman's module that crossed utterance boundaries were eliminated. The top scoring hypothesis out of those remaining was selected for each turn. The resulting recall and precision are shown in table 3. The test corpus for this experiment includes one additional dialog (d93-10.5) giving a total of 20,213 words. The additional dialog and different segmentation and repair annotations result in a corpus of 2295 turns, 3953 utterances and 695 speech repairs. Involving the parser as described above produces the results shown in table 4. Recall increases by 2% (9 repairs out of 695). Actually, there are 30 cases where the parser corrected the output of Heeman's module, but there are also 21 cases where the parser incorrectly rejected Heeman's first choice creating a false alarm and causing a repair to be missed. These instances occurred when the parser's grammar did not recognize the corrected utterance.

Because three aspects of the experiment were changed between experiments one and two, it is difficult to say whether 2% is a more valid measure of increase in recall than the 4.8% measured in experiment one. As a preliminary test, we measured the parsability of 60 turns randomly drawn from this corpus and containing 100 utterances. 63.3% of the turns parsed but if we do not consider turns consisting of one-word utterances and phrasal question answers then only 31.3% of these non-trivial turns parsed. Since experiment one was utterance-based and had a parsing rate of 39.7% on non-trivial utterances, the change in segmentation could have affected the recall rate. Clearly more experiments need to be run to get the correct figure.

Repairs correctly guessed	445
False alarms	125
Missed	250
Recall	64.03%
Precision	78.07%

Table 3 Heeman's Speech Repair Results from Exp 2

Repairs correctly guessed	454
False alarms	749
Missed	241
Recall	65.32%
Precision	37.74%

Table 4: Augmented Speech Repair Results from Exp 2

3.3 Discussion

The first question to be answered about these results is how to address the drop in precision. Up to this point the probabilities assigned by Heeman's module were only used to break ties. Combining these probabilities with the percentage of words parsed and using this score to rank hypotheses could offset the effect of lower probability hypotheses that remove unparsable but fluent material from the input.

A wider coverage grammar would also help, but the parser would still be judging repairs solely on whether they occur in the interpretation constructed by the parser. In addition to grammatical disruption, the parser could also measure syntactic parallelism between a potential reparandum and its correction. This ability needs to be investigated in further detail. Phrase-level parallelism will not likely be enough. An informal search of the test corpus revealed that only 11% of repairs were corrections of complete phrases or clauses. One could modify a statistical parser to return the most likely incomplete and complete constituents at every position in the input. Having incomplete constituents for comparison might allow a useful syntactic parallelism score to be constructed. Or perhaps the role of the parser should merely be to decide whether a particular repair hypothesis fits in the most highly probable parse of the input.

The results of these experiments are promising. Even with low grammatical coverage the parser was able to increase the recall. The remaining missing examples were not recovered either because the parser's grammar did not cover the corrected utterance or Heeman's repair module did not include the repair. Post-hoc analysis is needed to determine whether the majority of errors were the result of the parser or whether we also need to consider how to find repairs not posited by a module such as Heeman's.

In the case of grammar failure, the parser cannot interpret the utterance even if the correct repair hypothesis was chosen. An experiment described in [4] measured utterance parsing accuracy on a corpus of 495 repairs from the TRAINS dialogs. Even though the parser was given perfect speech repair information, only 144 of the 495 repairs appeared in utterances having a complete parse. Thus, the 9 additional repairs (out of 695) found in experiment 2 and the 13 additional repairs (out of 541) in experiment 1 should be considered in light of the fact that

these repairs are in utterances that parse whereas even if the other repairs in these corpora were corrected they might not parse. So the effect of the 9 and 13 repairs on the comprehensibility of the corpora is somewhat greater than the 2% and 4.8% increases in repair recall measured above.

4. CONCLUSION

The dialog parsing framework and implementation presented in this paper show how to extend standard parsers to handle speech repairs. Such an approach allows the parser's knowledge of the possible grammatical structures of the input to impact speech repair identification, resulting in increases in recall varying from 2% to 4.8%. This approach also provides a structural representation of reparanda enabling a dialog system to track the speaker's "train of thought" (or as mentioned, to support reference resolution).

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grants IRI-9503312 and 5-28789. Thanks to James Allen and Amon Seagull for their help and comments on this work. Thanks to Peter Heeman for providing data and guidance for the paper.

NOTES

1. The parser's lexicon has a list of 35 editing terms that activate the editing term metarule.
2. For instance, a traditional way to accommodate editing terms might be via a metarule,
X -> Y Z ==>$ X -> Y editing-term Z, where X varies over categories and Y and Z vary over sequences of categories. However, this would produce phrases containing editing terms as constituents, whereas in our approach editing terms are separate utterances.
3. Specifically the dialogs used were d92-1 through d92a-5.2; d93-10.1 through d93-10.4; and d93-11.1 through d93-14.2. The language model was never simultaneously trained and tested on the same data.

REFERENCES

[1] Bear, J., Dowding, J., and Shriberg, E. 1992. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL 92)*, 56-63.

[2] Core, M. and Schubert L. 1998. Implementing parser metarules that handle speech repairs and other disruptions. In Cook. D. (ed.), *Proceedings of the 11th International FLAIRS Conference*. Sanibel Island.

[3] Core, M. and Schubert L. 1999. A model of speech repairs and other disruptions. Working notes of the *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*. Cape Cod.

[4] Core, M. and Schubert L. 1999. A syntactic framework for speech repairs and other disruptions. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 99)*. College Park.

[5] Heeman, P.A. and Allen, J. F. 1995. the TRAINS 93 dialogues. TRAINS Technical Note 94-2, Department of Computer Science, University of Rochester, Rochester NY 14627-0226.

[6] Heeman, P. A. and Allen, J. F. 1997. Intonational boundaries, speech repairs, and discourse markers: modeling spoken dialog. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 97)*. Madrid, 254-261.

[7] Siu, M.-h. and Ostendorf, M. 1996. Modeling disfluencies in conversational speech. In *Proceedings of the 4rd International Conference on Spoken Language Processing (ICSLP-96)*, 386-389.

[8] Stolcke, A. and Shriberg, E. 1996. Statistical language modeling for speech disfluencies. In *Proceedings of the International Conference on Audio, Speech, and Signal Processing (ICASSP)*.