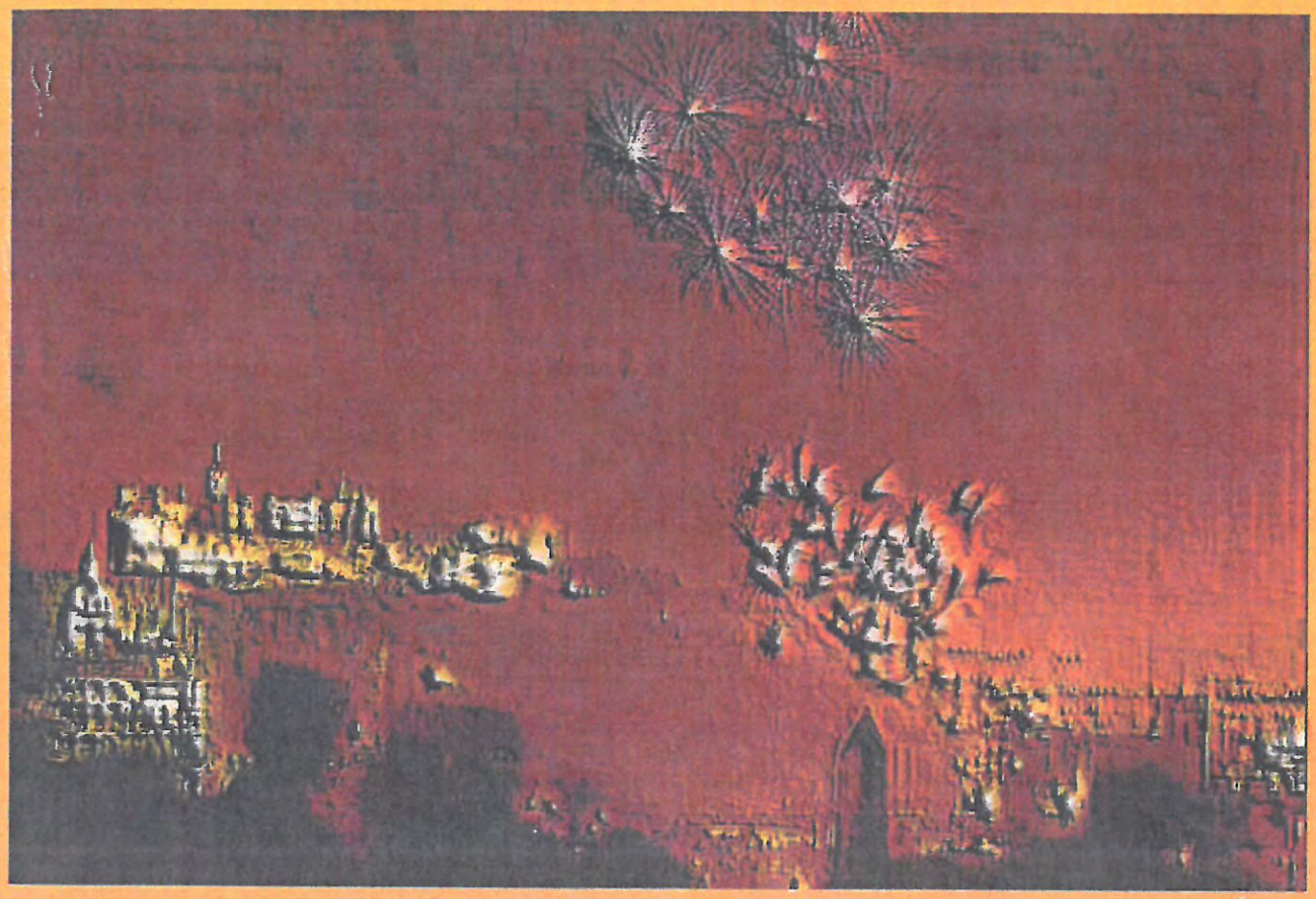
DiSS '01

Disfluency in Spontaneous Speech

ISCA

Tutorial and Research Workshop



University of Edinburgh
Scotland

August 29-31 2001

Disfluency in Spontaneous Speech, 2001

An ISCA Tutorial and Research Workshop.
University of Edinburgh, Scotland. August 29-31, 2001

Proceedings of DiSS '01

The Organisers thank the following for their support.

The International Speech Communication Association

and

The Department of Theoretical and Applied Linguistics
The Human Communication Research Centre
The Department of Psychology
(all at the University of Edinburgh)

DiSS '01 was organised by

Robin Lickley	University of Edinburgh*, UK
Liz Shriberg	SRI and ICSI, USA

LOCAL COMMITTEE

Ellen Gurman Bard	
Holly Branigan	
Mark Core	University of Edinburgh, UK
Robert Hartsuiker	
Robin Lickley	

SCIENTIFIC COMMITTEE

Ellen Gurman Bard	University of Edinburgh
Anton Batliner	University of Erlangen-Nuernberg, Germany
Holly Branigan	University of Edinburgh, UK
Susan Brennan	SUNYSB, USA
Herb Clark	Stanford University, USA
Yasuharu Den	Nara Institute of Science and Technology, Japan
Jeannie Fox Tree	UCSC, USA
Dafydd Gibbon	University of Bielefeld, Germany
Peter Heeman	OGI, USA
Herman Kolk	University of Nijmegen, Netherlands
Robin Lickley	University of Edinburgh, UK
Elmar Noeth	University of Erlangen-Nuremberg, Germany
Doug O'Shaughnessy	University of Quebec, Canada
Sharon Oviatt	OGI, USA
Albert Postma	University of Utrecht, Netherlands
Liz Shriberg	SRI & ICSI, USA

Web Page Design by Robert Eklund (http://www.ling.ed.ac.uk/DISS-01/)

Proceedings Compiled by Mark Core

Photograph on front cover from an original by Eddie Dubourg

* moving to Department of Speech and Language Sciences, Queen Margaret University College, from October, 2001.

Table of Contents

Annotation and Disfluency Types

Annotation and Analysis of Disfluencies in a Spontaneous Speech Corpus in Spanish
L. J. Rodríguez, I. Torres and A. Varona 1

Prolongations: a Dark Horse in the Disfluency Stable
Robert Eklund 5

Production

Application of EXPLAN Theory to Spontaneous Speech Control
Peter Howell and James Au-Yeung 9

Stuttering and Speech Monitoring
Nada Vasic and Frank Wijnen 13

Repeated Phoneme Effect in Japanese Speech Errors
Michiko Yoshida 17

Different Sources of Lexical Bias and Overt Self-Corrections
Sieb G. Nooteboom 21

Are Word Repetitions Really Intended by the Speaker?
Yasuharu Den 25

Gesture as an Indicator of Early Error Detection in Self-Monitoring of Speech
Mandana Seyfeddinipur and Sotaro Kita 29

Pauses in Speech by French Speakers with Down Syndrome
Laura Abou Haidar 33

Prosody and Phonetics

Prosodic Marking of Self-Repairs
Tapio Hokkanen 37

Acoustico-Phonetic Characteristics of Filled Pauses in Spontaneous French Speech: Preliminary Results
Danielle Duez 41

Interruption Glottalization in German Spontaneous Speech
Klaus J. Kohler, Benno Peters and Thomas Wesener 45

Sound and Function Regularities in Interjections
Nikolinka Nenova, Gina Joue, Ronan Reilly and Julie Carson-Berndsen 49

Filled Pauses and their Status in the Mental Lexicon
Richard Shillcock, Simon Kirby, Scott McDonald and Chris Brew 53

Human Perception and Comprehension

The Double Function of Disfluency Phenomena in Spontaneous Speech
Mária Gósy 57
Do Non-Word Disfluencies Affect Syntactic Parsing?
Karl G. D. Bailey and Fernanda Ferreira 61
Listeners' ERP Responses to False Starts and Repetitions in Spontaneous Speech
Jan McAllister, Susan Cato-Symonds and Blake Johnson 65
Grammatically Unacceptable Utterances Are Communicatively Accepted by Native Speakers, Why Are They?
Jeanne-Marie Debaisieux and José Deulofeu 69

Computational Linguistics and ASR

How to Repair Speech Repairs in an End-to-End System
Jörg Spilker, Anton Batliner and Elmar Nöth 73
Um, One Large Pizza. A Preliminary Study of Disfluency Modelling for Improving ASR
Ben Hutchinson and Cécile Pereira 77

Disfluency as a General Cognitive Phenomenon

Idiosyncratic Fillers in the Speech of Bilinguals
Caroline L. Rieger 81
Disfluencies in Writing - Are They Like in Speaking?
Åsa Wengelin 85
The Usage of Fillers at Discourse Segment Boundaries in Japanese Lecture-Style Monologues
Michiko Watanabe 89
Dialogue Moves and Disfluency Rates
Robin J. Lickley 93
Is Disfluency just Difficulty?
Ellen G. Bard, Robin J. Lickley and Matthew P. Aylett 97

Author Index

Abou Haidar, Laura 33
Au-Yeung, James 9
Aylett, Matthew P. 97
Bailey, Karl G. D. 61
Bard, Ellen G. 97
Batliner, Anton 73
Brew, Chris 53
Carson-Berndsen, Julie 49
Cato-Symonds, Susan 65
Debaisieux, Jeanne-Marie 69
Den, Yasuharu 25
Deulofeu, José 69
Duez, Danielle 41
Eklund, Robert 5
Ferreira, Fernanda 61
Gósy, Mária 57
Hokkanen, Tapio 37
Howell, Peter 9
Hutchinson, Ben 77
Johnson, Blake 65
Joue, Gina 49
Kirby, Simon 53
Kita, Sotaro 29
Kohler, Klaus J. 45
Lickley, Robin J. 93, 97
McAllister, Jan 65
McDonald, Scott 53
Nenova, Nikolinka 49
Nooteboom, Sieb G. 21
Nöth, Elmar 73
Pereira, Cécile 77
Peters, Benno 45
Reilly, Ronan 49
Rieger, Caroline L. 81
Rodríguez, L. J. 1
Seyfeddinipur, Mandana 29
Shillcock, Richard 53
Spilker, Jörg 73
Torres, I. 1
Varona, A. 1
Vasic, Nada 13
Watanabe, Michiko 89
Wengelin, Åsa 85
Wesener, Thomas 45
Wijnen, Frank 13
Yoshida, Michiko 17

Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish

L.J. Rodríguez, I. Torres, A. Varona

Departamento de Electricidad y Electrónica. Facultad de Ciencias. UPV/EHU.
Apartado 644. 48080 Bilbao. SPAIN.
{luisja,manes,amparo}@we.lc.ehu.es

Abstract

A new database consisting of 227 dialogues in Spanish was annotated with disfluencies. Then a detailed analysis of the annotations was carried out. The database had been recorded according to the well known Wizard of Oz paradigm. Seventy-five speakers were given each one three different scenarios to make queries about timetables, prices and other conditions of train travels between two spanish cities. The notion of disfluency was relaxed to include any acoustic, lexical or syntactic feature that distinguises spontaneous from read speech. A specific XML annotation scheme was developed. A simple text editor was used to insert marks, and a specific parser was implemented to find errors in annotations. The analysis of annotations revealed that disfluencies were not uniformly distributed among either user turns or speakers. Most disfluencies were grouped into certain user turns, especially the first one. On the other hand, some speakers were remarkably more prone to hesitate, repeat or correct fragments of speech than others.

1. Introduction

In the mid nineties large vocabulary continuous speech recognition technology achieved the big goal of translating read speech to text with word error rates of around 10%. This technology is now being used as a core component of broadcast news transcription systems, speech-to-speech translating systems and especially dialogue systems [1, 2, 3]. In this context the great challenge is to deal with spontaneous and somewhat unconstrained speech. This will require the acquisition and detailed annotation of generic and application specific databases for many languages. Also new modeling assumptions should be applied and more powerful algorithms should be developed.

Here we present the first milestone, which is the annotation of acoustic, lexical and syntactic disfluencies for an application specific database in Spanish language, which will serve as benchmark to study and to model this kind of phenomena. Unlike the rapidly growing number of spontaneous speech databases for English [4, 5, 6, 7], no corpus with annotation of disfluencies is available for Spanish, so this work can be considered as a pioneering effort.

The rest of the paper is organized as follows: the main features of our database are shown in Section 2; Section 3 presents the concept of disfluency applied in this work and briefly describes the inventory of speech events classified under such category; Section 4 presents the annotation format defined specifically for this work, and some details about the annotation process; statistics of disfluencies are shown and discussed in Section 5. Finally, Section 6 summarizes the conclusions of this work.

This work was partially supported by the Spanish CICYT, under project TIC98-0423-C06-03.

2. The spontaneous speech database

Our spontaneous speech database –which henceforth we will call *OZ1*– consists of the speech signals and the ortographic transcriptions of 227 Spanish dialogues, recorded at 8 kHz across telephone lines applying the well known *Wizard of Oz* mechanism: a human operator simulated the behaviour of the dialogue system, including recognition and/or understanding errors, so that users could think they were interacting with a real system [8, 9]. It must be said that the so called *users* were in fact 75 recruited volunteers, which were given three scenarios with dates, timetables and other conditions for a travel by train between two spanish cities. Actually, to adequately design the scenarios and to clarify what should be the system capabilities, a preliminary database was recorded with dialogues between real users and RENFE[1] information service operators. This preliminary database had been transcribed to plain text but not used for the adverse recording conditions [10]. Recruited users could get as much information as they wanted from the dialogue system, doing it in a natural manner, just as they would in a real call. However, some users still tended to hyperarticulate or even insert pauses between words, whereas others enlarged their turns with unnecessary explanations and often interrupted the system answers. This resulted in a great variability both in spontaneity and turn durations, these latter ranging from 0.5 to 50 seconds. The database includes 1657 non-empty user turns, lasting about 150 minutes. This gives an average of 7.3 non-empty user turns per dialogue, each one lasting an average of 5.4 seconds.

3. The inventory of disfluencies

We apply a wide definition of disfluency as any acoustic, lexical or syntactic feature that distinguishes spontaneous from read speech. In fact, we should better refer to them as spontaneous speech events. To define the inventory of disfluencies two key requirements were posed: coverage and coherence. Therefore, among all possible disfluencies, only those with enough number of samples in our database, plus some others considered significant, were included in the inventory and annotated. Before exploring the kind and frequency of the disfluencies that appear in *OZ1*, a tentative set was defined covering all the spontaneous speech events we could expect in human-machine communications, leaving aside some others which can be only expected in human-human dialogues. Starting from these considerations, a representative subset of 40 dialogues was used to validate the

[1] RENFE is the spanish public railway transportation system.

Table 1: Inventory of disfluencies, XML marks, attribute values, simplified marks and appearing counts for the database *OZ1*.

Category	XML	*source*/type	Simplified	Counts
Noises	n	*world*/generic	nw	661
		speaker/air	na	1404
		speaker/lips	nl	600
		speaker/cough	nt	9
Lengthenings of sounds	a	-	a	1019
Silence pauses	p	-	p	753
Filled pauses	f	a	fa	93
		e	fe	546
		m	fm	179
		trash	fb	210
Lexical disfluencies	l	unfinished	lu	95
		mispronounced	lm	105
Abandoned sentences	b	-	b	70
Retracings	r	repetition	rr	292
		substitution	rs	141
		insertion	ri	37
		deletion	rd	5
Discourse markers	d	open	do	150
		close	dc	189
		accept	da	78
		reject	dr	45
		explain	de	71
		request	dq	92
		fill	df	225
		exclaim	dx	15

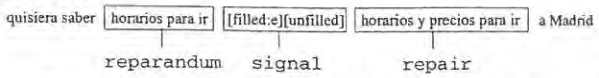
inventory of disfluencies, which evolved from the initial set to the following:

Acoustic disfluencies: this category included noises, lengthenings of sounds, silence –or unfilled– pauses and filled pauses. Noises were included because, though not disfluencies in the strict sense, they seldom appear in read speech, but are pervasive in spontaneous speech. With regard to filled pauses, various acoustic realizations were found in Spanish language, either vowels ('a', 'e') or nasalizations ('m').

Lexical disfluencies: spontaneous speech is far more relaxed than read speech, so a high number of popular or familiar expressions can be found, as well as pronunciation variants – contractions, misarticulations, non-canonical acoustic realizations of phonemes, etc.– due to dialectal or speaker specific features, high speech rates, etc. We defined lexical disfluencies as not properly –or not canonically– pronounced words; for the sake of completeness, cut or unfinished words were also included in this category.

Syntactic disfluencies: among the wide range of them that can be found in spontaneous speech (false starts, repetitions, reformulations, unfinished sentences, sentences completing a previous one, missing words, lacks of concordance, etc) we only considered two categories: *abandoned sentences* (most times false starts) and *retracings*, these latter accounting for repetitions, substitutions and reformulations with insertion or deletion of words.

We applied the structure of retracings shown in [11]: a segment to be repaired –*reparandum*–, a segment marking the correction –*signal*– which may include filled or unfilled pauses and some editing phrases like 'sorry' or 'I mean', and a third segment –*repair*– giving the replacing material, which can be a repetition, a substitution or a more complex reformulation with insertion or deletion of words, as shown in the following example, taken from *OZ1*[2]:



Discourse markers: here we consider very usual words or phrases without any specific meaning but carrying out a metalinguistic function, as opening ('hello', 'good morning'), closing ('thanks', 'good bye', 'that's all'), emphasizing ('please', 'come on'), filling ('well', 'you know'), editing ('sorry', 'I mean'), etc. Although discourse markers cannot be classified as disfluencies, but as pragmatic elements of spoken language, they were annotated to allow the definition of specific categories for them in the language model, which could improve the recognition of spontaneous speech.

4. The annotation scheme

After an exhaustive review of the formats and tools for the annotation of linguistic corpora listed by LDC [12], especially the guidelines given by the european project MATE [13], a specific XML annotation scheme was designed for disfluencies, which –as a first approach– accounted only for disfluencies happening in human-machine communications, and more particularly in *OZ1*, as explained in Section 3. The annotation scheme was accompanied by the corresponding manual [14]. Annotations could refer to instantaneous events, then they were simply inserted in the corresponding place of the ortographic transcription: <mark attribute=value/>, or could refer to a time interval, then affecting some amount of text: <mark attribute=value> TEXT </mark>. Marks were one-letter codes. Some marks needed no attributes, others required one or more attributes. For the database *OZ1* three attributes were defined: *type*, used to give a more detailed description of the disfluencies, *source*, used only for noises, and *word*, used only to supply the canonical version of a word in lexical disfluencies.

Marks were added by hand, using a simple text editor. To make easier such a tedious process, a simplified format was also defined. Each simplified annotation consisted of a short mark, usually two letters encoding both the mark and the value of the attribute *type*, enclosed between parentheses and affecting some text. The XML and the simplified annotations for the example shown above would be:

XML

```
quisiera saber
<r type="insertion">
<m> horarios para ir </m>
<s> <f type="e"/> <p/> </s>
<c> horarios y precios para ir </c>
</r>
a Madrid
```

Simplified

```
quisiera saber (ri (m horarios para ir) (s (fe)(p)) (c horarios y precios para ir)) a Madrid
```

Since machine answers were automatically generated from a predefined set of templates, only user turns were annotated –after careful listenings of the speech signals. To help the detection and correction of annotation errors, a very simple parser was implemented, which accounted not only for the parentheses and marks, but also the correctness of their contents. The parser was iteratively applied to the annotated dialogues, and

[2] translated to English as: *I would like to know timetables to go [filled:e][unfilled] timetables and prices to go to Madrid.*

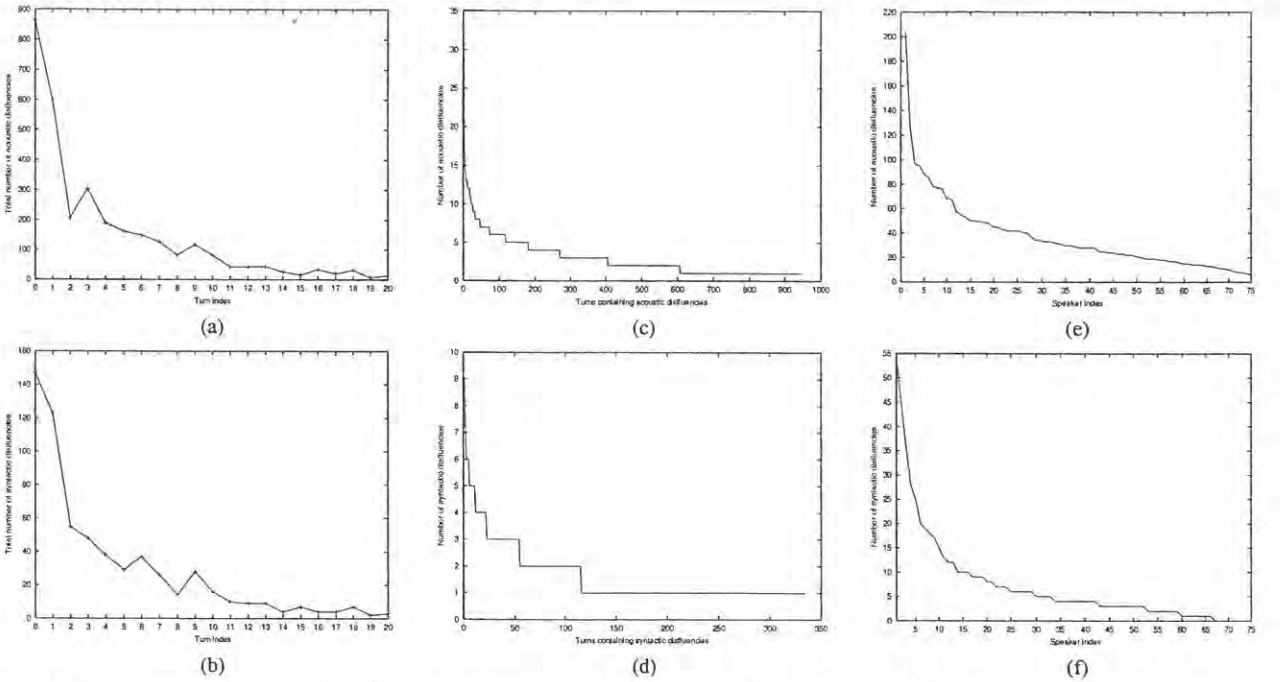


Figure 1: Graph (a) gathers acoustic disfluencies appearing in user turns with the same index, and shows the sums for the first 21 turns; graph (c) shows the counts of acoustic disfluencies for each user turn, putting them in decreasing order and leaving aside turns with no acoustic disfluencies; graph (e) shows the counts of acoustic disfluencies for each speaker, in decreasing order. Graphs (b), (d) and (f) show the same for syntactic disfluencies.

these corrected until no errors were found. By slightly modifying its source code, the parser was easily adapted to other tasks, like translating annotations from simplified format to XML, or producing various kinds of enhanced ortographic transcriptions.

Finally, to guarantee the coherence of the annotations, one single expert reviewed all of them. At the same time, speech signals corresponding to user turns were re-segmented according to the annotated phenomena, so that speech signals and their ortographic counterparts became completely coherent. The resulting database was composed of 227 text files with very reliable annotations of disfluencies and 1657 binary files containing the speech signals corresponding to coherently segmented user turns.

The categories, XML marks, attribute values and simplified marks specified in the annotation manual, along with the appearing counts for the database *OZI* are shown in Table 1. For a lack of space, the more specific marks reserved for the *reparandum* (m), the *correcting signal* (s) and the *repair* (c) in retracings are not showed.

5. Discussion

In this Section we will try to analyze the disfluencies appearing in the database *OZI*. This will help to identify those features that make spontaneous speech so difficult to recognize, and give ideas about which elements of the recognition software should be improved.

As shown in Table 1, the most common events were noises. This was due to the high degree of detail of annotations. Although most times speaker aspirations and speaker lips were hardly audible, we wanted detailed annotations to allow the recognition of various kinds of *silence*, which could improve the segmentation of speech signals, thus yielding more accurate acoustic models. After noises, acoustic disfluencies: lengthenings, silence pauses and filled pauses, were the most common events. It must be said that, although a very wide range of silence pauses was observed, we considered a difficult task to assign them a duration attribute, so the annotation of silence pauses did not include duration information. The same was applied to filled pauses and lengthenings. It was left to recognition algorithms the correct alignment of such events.

It was found a sizeable amount of retracings, especially repetitions and substitutions, which denotes the importance of modeling this kind of disfluencies, even when speakers are not real users but recruited volunteers. Significant data about the distribution of acoustic –graphs (a), (c) and (e)– and syntactic disfluencies –graphs (b), (d) and (f)– are shown in Figure 1: graphs (a) and (b) gather disfluencies appearing in user turns with the same index, and show the sums for the first 21 turns; graphs (c) and (d) show the counts of disfluencies for each user turn, putting them in decreasing order and leaving aside the turns with no disfluencies; finally, graph (c) shows the counts of disfluencies for each speaker, putting them in decreasing order.

A detailed inspection of the annotations –see graph (b) of Figure 1– revealed that most retracings, accompanied by a remarkable number of acoustic disfluencies acting as correcting signals –as shown by graph (a) of Figure 1– were grouped into certain turns, especially the first one, where users showed a hesitating behaviour. In fact, this behaviour can appear at any time in the dialogues, but it's more probable at the beginning, when the user has not defined his needs yet and does not know the system capabilities. So a high variability can be observed in the distribution of disfluencies, with many turns showing a few or no disfluencies –most times these turns consisted of a few words like *"yes, please"*, *"no, thanks"*, *"on Tuesday"*, etc– and a reduced set of turns gathering most of them, as shown in graphs

(c) and (d) of Figure 1.

A second study was made by counting disfluencies for each speaker. As shown in graph (e) and (f) of Figure 1, there was a high variability in the distribution of acoustic and syntactic disfluencies in the set of speakers. A few speakers gathered most disfluencies. This study was detailed by considering six general categories: noises (N), silence pauses (P), filled pauses and lengthenings of sounds (F), lexical disfluencies (L), syntactic disfluencies (S) –putting together abandoned sentences and retracings– and discourse markers (D). Mean and deviation values for the whole set of speakers, and counts for 10 especially selected speakers are shown in Table 2. Some speakers were remarkably more prone to hesitate, repeat or correct fragments of speech than others, yielding generally much longer dialogues (speakers 9, 11 and 30), whereas others produced very short dialogues with a few disfluencies (speakers 25 and 31). As shown in Table 2 long dialogues show a high number of disfluencies, but the amount of disfluencies was not always correlated with the length of the dialogues: speakers 4, 19, 22, 45 and 69 show very similar times but the amount of disfluencies ranges from a total number of 36 to 120. This reveals that some speakers are intrinsically more *disfluent* than others.

Table 2: Full duration of user turns and counts of disfluencies for the three dialogues carried out by each of 10 speakers selected from the database *OZ1*. The symbol N stands for noises, P for unfilled pauses, F for filled pauses, L for lexical disfluencies, S for syntactic disfluencies and D for discourse markers. Mean and standard deviation values over the whole set of speakers are shown too.

Speaker	Full duration (sec)	N	P	F	L	S	D	Total
4	91.24	29	5	3	2	1	6	46
9	378.86	124	16	72	2	12	56	282
11	394.45	135	63	140	4	54	17	413
19	70.73	29	3	3	0	0	1	36
22	89.79	19	15	35	2	12	9	92
25	42.90	14	1	6	1	1	4	27
30	478.15	160	52	75	17	45	21	370
31	38.49	10	1	9	0	0	3	23
45	125.91	14	27	41	11	19	8	120
69	72.98	20	10	18	3	9	11	71
Mean	118.63	35.65	10.04	27.29	2.67	7.27	11.53	94.45
Deviation	75.65	25.98	10.28	23.56	3.42	9.74	9.96	70.02

6. Conclusions

The main features of a spontaneous speech database consisting of 227 dialogues in Spanish were introduced. The speech events considered as disfluencies were described. Both a XML annotation format and a simplified format –to make easier the annotation process– were presented. Also a very simple parser was implemented which helped to locate and correct errors in annotations. Finally, annotation data were shown and discussed, finding that acoustic, lexical and syntactic disfluencies must be all studied and modeled for the recognition of spontaneous speech. Statistics showed that disfluencies were not uniformly distributed in the set of user turns, being more probable at the beginning of dialogues. Also a high dependence on speaker was observed. Our current work concerns two issues: first, to extend this preliminary study by recording and annotating a bigger database for the same application, but with a full dialogue system prototype and real users; and second, to model acoustic disfluencies as a first step towards a more general scheme which will include modelling approaches for lexical and syntactic disfluencies.

7. References

[1] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 100–112, January 2000.

[2] L. Lamel, "Spoken language dialog system development and evaluation at LIMSI," in *Proceedings of the International Symposium on Spoken Dialogue*, Sydney, Australia, November 1998.

[3] A.L. Gorin, G. Riccardi, and J.H. Wright, "How May I Help You," *Speech Communication*, vol. 23, no. 1-2, pp. 113–127, October 1997.

[4] Air Travel Information System (ATIS), URL: http://www.ldc.upenn.edu/Catalog/ATIS.html.

[5] Switchboard: Telephone Speech Corpus, URL: http://www.ldc.upenn.edu/Catalog/LDC97S62.html.

[6] TRAINS: Spoken Dialogue Corpus (University of Rochester), URL: http://www.ldc.upenn.edu/Catalog/LDC95S25.html.

[7] The HCRC Map Task Corpus (University of Edinburgh and University of Glasgow), URL: http://www.ldc.upenn.edu/Catalog/LDC93S12.html.

[8] J.B. Mariño and J. Hernando, "Especificación de las grabaciones mediante Mago de Oz," Technical Report BS16AV10, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, November 1999, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.

[9] A. Sesma, J.B. Mariño, I. Esquerra, and J. Padrell, "Estrategia del Mago de Oz," Technical Report BS52AV22, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, December 1999, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.

[10] A. Bonafonte and N. Mayol, "Documentación del corpus INFOTREN-PERSONA," Technical Report BS14AV20, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, June 1999, URL: http://gps-tsc.upc.es/ veu/basurde/Home.htm.

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

[12] Linguistic Data Consortium: Linguistic Annotation, URL: http://www.ldc.upenn.edu/annotation.

[13] Multilevel Annotation Tools Engineering (MATE), URL: http://mate.nis.sdu.dk/.

[14] L.J. Rodríguez, I. Torres, and A. Varona, "Manual para el etiquetado de disfluencias," Technical Report BS12BV30, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad del País Vasco, May 2000, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.

Prolongations: A dark horse in the disfluency stable

Robert Eklund†‡

† Telia Research AB, Sweden
‡ NLPLab, Dept. of Computer and Information Science, Linköping University, Sweden
Robert.H.Eklund@telia.se

Abstract

This paper studies a specific type of disfluency, *viz.* segment prolongation (PR), i.e., the "stretching out" of speech sounds as a means of hesitation. It is shown that the occurrence of PRs varies as a function of phone type, position in the word, lexical factors and word class, and that PRs are subject to phonotactic constraints in Swedish. A comparison between Swedish and Tok Pisin suggests that there are language-specific traits associated with PR production.

1. Introduction

Studies of disfluency phenomena such as filler words, repetitions, pauses, truncations, insertions, deletions and so on have become common recently.

However, one type of disfluency that has received little attention in the literature is segment prolongation, i.e., the "stretching out" of speech segments. It has been shown that PRs are more common than most other types of disfluencies, outnumbered only by filled pauses (FPs, also called "filler words") and unfilled pauses (silences) [2][4][5]. An adequate description of PRs could provide insight into human speech production, and could also help improve durational modeling for automatic speech recognizers.

Moreover, while cross-linguistic studies have shown that there are great similarities with regard to disfluencies across languages [1][2][5][10], Eklund [2] showed that differences between Swedish and Tok Pisin occur at the PR level.

The objective of this paper is to take a more detailed look at the characteristics of PRs.

2. Method

2.1. Corpora

Data from four Swedish spoken language corpora—names in table below—were analyzed. The corpora were all collected as part of the Spoken Language Translator project [8] at Telia Research AB during the period 1996 through 1998. All dialogs were task-oriented within Air Travel Information Service (ATIS) or toward the booking of general business trips [4]. Summary statistics are shown in Table 1.

Table 1: Summary statistics for the Swedish corpora. UW=Unique Word Forms. WT=Word Tokens. H=Human. "M"="Machine" (i.e., Wizard-of-Oz simulation). M=Machine.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
Type	H"M"H	H"M"	HH	HM	—
No. subjects M/F	49 26/23	23 16/7	8 6/2	16 9/7	96 57/39
No. Dialogs	84	71	24	69	248
No. Utts.	3,104	1,849	1,698	1,888	8,539
No. UW	570	792	1,157	959	3,478
No. WT	5,565	12,190	9,232	12,047	39,034

2.2. Set-up and subjects

All dialogs were made over a telephone line, and high quality recordings were made to enable acoustic analysis. The subjects were all Telia employees, and were all used to travel bookings.

2.3. Disfluency annotation

The annotation scheme used is described by Eklund [4], and is based on, and similar to, that described by Shriberg [9]. All corpora were labeled by the author.

3. Results

3.1. PR rates

Occurrence of PRs is shown in Table 2.

Table 2: Summary statistics. UW=Unique Word Forms. WT=Word Tokens.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. PRs	101	88	129	179	497
% PRs/UW	17.72%	11.11%	11.15%	18.66%	14.29%
% PRs/WT	1.81%	0.72%	1.40%	1.48%	1.27%

As can be seen, 0.7 to 1.8% of the words include prolonged segments at token level.

3.2. Durational data

The mean duration for all PRs (all data pooled) was 0.289 milliseconds (N=497). The 95% confidence interval was 0.275/0.305. Standard deviation was 0.170.

The bottom end of the durational scale is problematic from the point of view of labeling, since it is difficult to say exactly when a segment is in fact prolonged. Thus, the data were also explored with the lower quartile removed (N=373; cut-off point at 0.175). The mean duration for trimmed PR data was 0.340. The 95% confidence interval was 0.323/0.357. Standard deviation was 0.167.

3.3. PRs vs. FPs

PRs and FPs have something in common that distinguishes them from other disfluency phenomena: they both signal hesitation by means of vocalization and duration (unlike repetitions, truncations, mispronunciations etc.). There is no obvious reason to assume *a priori* that there would be any durational differences between the two types since they both serve the same purpose, i.e., signaling hesitation while still speaking (as opposed to inserted silence).

The mean duration for all FPs pooled was 0.488 (N=1,379). The 95% confidence interval was 0.474/0.501. Standard deviation was 0.255.

A *t*-test showed that FPs were significantly longer than PRs ($p<0.001$; two-tailed; mean difference 0.198). Comparing pooled PR and FP values within the same corpora creates problems with regard to whether or not the variables are to be considered dependent or not. Thus, a Wilcoxon signed ranks test and a Mann-Whitney test were also performed. Both tests showed significance at the $p<0.001$ level. Thus, it would appear safe to conclude that FPs are generally longer than PRs. However, these results need be tested at the individual level before more definite conclusions can be drawn.

3.4. Individual differences

The next thing we looked at was whether or not there were any individual differences with regard to PR production, both *per se* and relative to FP production. The observations are shown in Table 3.

Table 3: Relative frequency of PR and FP usage. Note that the sum of the four categories sometimes exceed the numbers of subjects due to the fact that the same person might appear in two cells when the lower number of comparison is zero. >="More frequent than".

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. subjects	49	23	8	16	96
FPs > PRs	35	21	7	13	76
PRs > FPs	10	2	1	3	16
FPs = PRs	4	—	—	—	4
No FPs	6	1	—	—	7
No PRs	16	3	—	—	19

For most subjects, FPs are more common than PRs. It is also more common to not employ PRs at all than it is to totally lack usage of FPs. However, the bulk of subjects who did not make use of PRs at all occur in WOZ-1. This corpus differs from the others in the way the tasks were presented [4], and contains much shorter dialogs. At least a part of the skewness could be attributed to the particulars of the task.

3.5. PR position in the word

As has been reported for Swedish [2][4], American English [5], and Tok Pisin [4], PRs are not evenly distributed within the word. A breakdown of PR position is shown in Table 4.

Table 4: Phone type and position of prolongations. For each corpus the number and percentages of phone position are given.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. PRs	101	88	129	179	497
No. Segments	24,402	52,157	32,549	96,215	205,323
% PRs/Segments	0.41%	0.17%	0.40%	0.19%	0.24%
% Initial	30/29.7	32/36.4	32/24.8	56/31.3	150/30.2
N/% V	8/26.6	8/25.0	6/18.8	10/17.8	32/21.3
N/% C +son	–/–	6/18.8	18/56.2	14/25.0	38/25.3
N/% C –son	22/73.4	18/56.2	8/25.0	32/57.2	80/53.4
% Medial	13/12.9	16/18.2	25/19.4	28/15.6	82/16.5
N/% V	2/15.4	2/12.5	5/20.0	4/14.2	13/15.8
N/% C +son	3/23.1	–/–	4/16.0	1/3.6	8/9.7
N/% C –son	8/61.5	14/87.5	16/64.0	23/82.2	61/74.5
% Final	58/57.4	40/45.4	72/55.8	95/53.1	265/53.3
N/% V	14/24.1	8/20.0	24/33.3	29/30.5	75/28.2
N/% C +son	31/53.5	28/70.0	37/51.4	57/60.0	153/57.8
N/% C –son	13/22.4	4/10.0	11/15.3	9/9.5	37/14.0

While the previously reported 30–20–50 ratio for initial, medial and final position [2][4][5], respectively, is confirmed, Table 4 shows that this tendency does not hold for all *types* of segments. While non-sonorant consonants are the most commonly prolonged segments in initial and medial position, they are the *least* frequently prolonged segments in final position. This indicates that certain types of segments are preferred in certain positions, which hints at an interaction between position in the word and segment type.

The occurrence of word-medial PRs distinguishes them from FPs, since no word-internal FPs have been encountered in our data, although FPs have been reported between lexical roots inside compounds in Swedish [5] and German [7].

3.6. Segment type

A detailed examination of segment type was then undertaken. PR frequency was normalized for overall segment frequency. Since the corpora were all transcribed according to an orthography-based, phonological scheme, making exact calculations cumbersome, all numbers must be considered approximate. The top five segments are shown in Table 5.

Table 5: Most commonly prolonged segments. For all segments, the number of prolonged segments is given divided by the total number of the same segment in the same corpus, in order to normalize for differences in general segment occurrence.

WOZ-1	WOZ-2	Nymans	Bionic
[f] 1.67%	[f] 0.90%	[f] 1.51%	[f] 2.05%
[k] 0.98%	[n] 0.56%	[s] 1.42%	[n] 1.16%
[s] 0.97%	[s] 0.48%	[n] 1.19%	[k] 1.01%
[l] 0.96%	[k] 0.43%	[o] 0.62%	[l] 0.72%
[n] 0.76%	[i] 0.32%	[l] 0.54%	[s] 0.52%

As is shown, almost all of the twenty segments are continuants. The exception is the stop [k], which often comes from the medial [k] in "klockan" (o'clock), accounting for 27% of the cases. The only vowels in the list are [o] occurring in the preposition "på" (on, for dates), which accounts for 86% of the cases, and [i], from the preposition "i" (in), which accounts for 78% of the cases.

Although segment type proper obviously is of importance, a detailed look at the words in which the prolonged segments appear implies that there is also a lexical factor to consider.

3.7. Open versus closed word classes

Given the observations in 3.6, possible differences between open and closed word classes were studied (using a traditional definition of 'open' and 'closed' word classes). Summary statistics are presented in Table 6.

Table 6: Ratio open/closed word classes. UW=Unique Word Forms WT=Word Tokens.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. Open WT	3,355	6,956	4,045	6,519	20,875
% total no. WT	60.3%	57.1%	43.8%	45.9%	53.5%
No. Open UW	497	695	993	893	3,078
% total no. UW	87.2%	87.8%	85.8%	93.1%	88.5%
No. Closed WT	2,210	5,234	5,187	5,528	18,159
% total no. WT	39.7%	42.9%	56.2%	54.1%	46.5%
No. Closed UW	73	97	164	66	400
% total no. UW	12.8%	12.2%	14.2%	6.9%	11.5%

As can be seen, the distribution is more or less 50/50 between open and closed word classes at token level, and averages one-to-eight at unique forms level.

The number of PRs in open and closed word classes are shown in Table 7.

Table 7: Rate of PRs occurring in words belonging to open/closed word classes (tokens).

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. Open	48	39	43	71	201
%	47.5%	44.3%	33.3%	39.7%	40.4%
No. Closed	53	49	86	108	296
%	52.5%	55.7%	66.7%	60.3%	59.6%

As is shown, there is a slight inclination towards prolonging words belonging to closed words classes. The difference is significant at p=0.145 (Pearson chi-square, two-tailed). Given definitional problems associated with the categories 'open' vs. 'closed', these data should be handled with some caution.

3.8. Domain dependency

We then investigated specifically which open words were prone to prolongation. Judging from the examples above, it seemed that within-domain words were more likely to be prolonged than words outside the domain. Examples of prolonged, within-domain words were "boka" (book/reserve), "hotell", "taxi" (taxi), "resa" (travel/go), "rökfritt" (non-smoking), "billigaste" (cheapest), "hemresa" (return trip), and words for dates, times and locations. The results are shown in Table 8.

Table 8: Rate of PRs occurring in words belonging to open word classes (tokens). Figures are given both for general vocabulary and domain-dependent vocabulary.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. open words w/ PRs	48	39	43	71	201
No. open words w/ PRs in domain	27	31	23	61	142
% open words w/ PRs indomain	56.3%	79.5%	53.5%	85.9%	70.6%

In all corpora, within-domain words are more often prolonged than outside-domain words. Pooling all data, the difference is significant at p<0.001 (Pearson chi-square, two-tailed). However, when pooling only WOZ-1 and Nymans, the difference is not significant (p=0.792, Pearson chi-square, two-tailed). These results might be an artefact of general corpus differences. The tasks were presented differently in WOZ-1 [4], and Nymans was human–human, while WOZ-2 and Bionic were similar both with regard to task details and setup. Consequently, no far-reaching conclusions will be drawn here with regard to the observed differences between the corpora.

3.9. Phonological length

A final issue, not to be bypassed, is that of phonological length, which is distinctive in Swedish. It is also mutually exclusive, which means that all VC syllables come either as V:C or VC: (or VCC). In recent work on dynamic segmental effects associated with focusing in Swedish, Heldner & Strangert [6] show that while focused segments in general are lengthened by an average 25%, short vowels are only marginally—not distinctively—lengthened. This observation is repeated in our data. While long vowels and both long and short consonants are subject to prolongation, there are no instances of prolonged short vowels.

4. A comparison with Tok Pisin

4.1. Tok Pisin corpus

In order to test some of the observations made above, a comparative study was made on available Tok Pisin data. The Tok Pisin corpus (TP) consists of authentic ATIS dialogs, collected on location in Kavieng, Papua New Guinea, during the period December 1999 and January 2000 [3]. TP consists of 39 authentic human–human ATIS dialogs, and was labeled by the author (who is not a native speaker of Tok Pisin). Currently, a total number of 654 utterances and 3,538 words have been transcribed, with a total number of 35 PRs [2].

4.2. Durational data

The mean duration for all PRs was 0.347 (N=35). The 95% confidence interval was 0.287/0.407. Standard deviation was 0.170. There was no significant difference between PR durations in Swedish and Tok Pisin (p=0.055, t-test, two-tailed, equal variances assumed).

4.3. PRs versus FPs

It was shown for Swedish that FPs were significantly longer than PRs. To check whether this holds true for Tok Pisin, the values for FPs in TP were explored. The mean for all FPs was 0.456 (N=80). The 95% confidence interval was 0.401/0.501. Standard deviation was 0.244.

FPs were significantly longer than PRs. A t-test resulted in p=0.018 (two-tailed, equal variances assumed), and a Mann-Whitney test resulted in p=0.008 (two-tailed).

4.4. PR position in the word

The distribution of PRs as a function of position in the word is shown in Table 9.

Table 9: Phone type and position of PRs.

	TP
No. PRs	35
No. Segments	12,840
% PRs / Segments	0.27%
% Initial phone	6/17.1%
% vowel	4/66.8%
% cons +sonorant	1/16.6%
% cons –sonorant	1/16.6%
% Medial phone	—
% Final phone	29/82.9%
% vowel	12/41.4%
% cons +sonorant	13/44.8%
% cons –sonorant	4/13.8%

As is shown, the ratio in TP for initial/medial/final position is roughly 15–0–85, which differs from the distribution reported for Swedish and American English, mentioned above.

4.5. Segment type

The most commonly prolonged segments (normalized for overall segment frequency) in TP were, in descending order: [ŋ] (1.20%); [m] (0.82%); [s] (0.51%); [o] (0.41%); [u] (0.35%). That other segments are prolonged more often in Swedish than in Tok Pisin is perhaps not surprising. What is more striking is that the segments seem to be prolonged for the same reason. The phones [ŋ] and [o] mainly occur in the prepositions "long" (general preposition), pronounced [lɔŋ] or [lɔ]) and "bilong" (stronger-binding preposition, genitive marker, conjunction), pronounced [bilɔŋ] or [blɔ].

4.6. Open vs. closed word classes

Rates of words belonging to open and closed word classes and PR rates are shown in Table 10.

Table 10: Ratio open/closed word classes and PR rates in TP.

	TP
No. Open / % total no. words	1,592/45.0%
No. Closed WT / % total no. words	1,946/55.0%
No. Closed UW / % total no. Closed words	39/2.0%
No. PRs Open / % total no. PRs	6/17.1%
No. PRs Closed / % total no. PRs	29/82.9%

The tendency to prolong words belonging to closed word classes is more marked in TP than in the Swedish data. Out of 35 PRs, 29 occur either in prepositions ("long", "bilong") or in grammatical markers such as "i" (predicate marker), "bai" (future marker) or "ol" (plural marker). Moreover, three of the six prolonged words belonging to open word classes are from the domain. "fe" (fare), "ples" (place) and "tri" (three). Also, two instances of a prolonged transitive suffix "-im" are found in the words "salim" (send) and "sekim" (check). These two could arguably be analyzed as grammatical ('closed') prolongations.

5. Discussion

From a **phonological** point of view, it would seem that all segment types might be prolonged, although there is a tendency towards prolonging continuants.

Looking at **phonological length**, it is striking to find that no short vowels are prolonged in our data. This observation supports the hypothesis that phonology puts constraints on the production of PRs, which receives further support from the observations reported by Heldner & Strangert [6].

Looking at **duration** proper, our data suggest that PRs are shorter then FPs, despite their physiological, acoustic and functional similarities. The observation, if tentative, that FPs generally have longer duration could imply that FPs do have a different "status" and are viewed by the speaker as "words" in their own right. Also, that PRs, unlike FPs, are observed in word-medial position is another trait that implies that PRs and FPs do not have the same status in speech production.

From a **morphological** point of view, the favored position for segment prolongation is word-final, in both Swedish and Tok Pisin. However, the observation that the ratio initial/medial/final position differs between Swedish and Tok Pisin could suggest that PR production could be language-specific, being associated with the morphotactics of a given language.

Stepping up to **full words**, the tendency is that words belonging to closed word classes are more prone to prolongation than words belonging to open words classes.

Within the open words class group, most words with prolonged segments are within the **discourse domain**. This is not surprising, since speakers hesitate before or on items with high cognitive load, i.e. either the preposition or article before a semantically heavy item. However, words inside the domain are more likely to be uttered by many speakers, and are thus prone to over-representation, as compared to words outside the domain.

From a **cross-linguistic** perspective, the comparison with Tok Pisin shows that there are similarities between the languages. While the data in TP exhibits the same tendency to prolong segments in words belonging to closed word classes, there are significant differences at the segmental and distributional levels. This could imply that Tok Pisin speakers hesitate at the same places that Swedish speakers do, but that the hesitation affects other types of segments, given different morphological and phonotactic constraints.

6. Conclusions

In conclusion, the prototypical Swedish PR would be the final segment—preferably a continuant—of a preposition or article, or appear in a domain-dependent word which signals crucial information with regard to the task at hand.

The comparison with Tok Pisin suggests that these observations probably do not hold for all languages, and that more cross-linguistic studies of PRs need be done in order to gain deeper insights with regard to the role and function of segment prolongation in human speech production.

7. Acknowledgements

Thanks to Åsa Wengelin, Jaan Kaja and Eva Lindström for a plethora of comments. Thanks to Michael Kiefte for proofs.

8. References

[1] Den, W. & H. Clark. 2000. Word Repetitions in Japanese Spontaneous Speech. *Proc. ICSLP'00*, Beijing 16–20 October 2000, vol. 1, pp. 58–61.

[2] Eklund, R. 2000. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and Tok Pisin Human–Human ATIS Dialogues. *Proc. ICSLP'00*, Beijing, 16–20 October 2000, vol. 2, pp. 991–994.

[3] Eklund, R. 2000. Wanpela deitabeis long Tok Pisin bilong baim tiket bilong balus. (An ATIS database in Tok Pisin.) Methodological observations with regard to the collection of human–human data. *Proc. Fonetik 2000*, The Swedish Phonetics Conference, May 24–26 2000, University of Skövde, pp. 49–52.

[4] Eklund, R. 1999. A Comparative Study of Disfluencies in Four Swedish Travel Dialogue Corpora. *Proc. Disfluency in Spontaneous Speech Workshop*, Berkeley, California, 1 July 1999, pp. 3–6.

[5] Eklund, R. & E. Shriberg. 1998. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues, *Proc. ICSLP'98*, Sydney, 30 November–5 December 1998, vol. 6, pp. 2631–2634.

[6] Heldner, M. & E. Strangert. Temporal effects of focus in Swedish. Accepted for publication, *Journal of Phonetics*.

[7] Lüngen, H., M. Pampel, G. Drexel, D. Gibbon, F. Althoff & C. Schillo. 1996. Morphology and Speech Technology. *Proc. ACL–SIGPHON Conference*, 28 June 1996, Santa Cruz, pp. 25–30.

[8] Rayner, M., D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). 2000. *The Spoken Language Translator*, Cambridge University Press.

[9] Shriberg, E. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of Berkeley, California.

[10] Tseng, S.-C. 2000. Modelling Speech Repairs in German and Mandarin Chinese Spoken Dialogues. *Proc. COLING 2000*, Saarbrücken, 31 July–4 August 2000, vol. 2, pp. 864–870.

Application of EXPLAN theory to spontaneous speech control

Peter Howell and James Au-Yeung

Department of Psychology
University College London
p.howell@ucl.ac.uk james@psychol.ucl.ac.uk

Abstract

Problems for theories that explain speech errors by a monitoring process are discussed. EXPLAN theory is based on a proposal about planning and execution time, not on how errors arise. This theory is outlined and support from characteristics of fluency failure and altered feedback studies given.

1. Introduction

Spontaneous speech output differs from read output of a prepared text insofar as fluency failure is more likely. This difference may arise because a plan is supplied when a text has to be read whereas the plan has to be generated on the fly in spontaneous productions. The material that has to be read is usually syntactically correct and such a "grammatically-correct" model is taken as the benchmark against which to assess spontaneous productions that are likely to be ungrammatical (for instance Levelt's well-formedness rule for speech repairs) [1]. Most studies on read text use English, but English orthography allows little scope for prescribing the prosodic plan other than a crude indication of location of pauses, stress etc. in the punctuation. This may be why the prosodic structure differs between a text produced spontaneously and a transcript of this text read by the same speaker [2]. Two conclusions can be drawn: (1) Clarification is needed about how fluency failure arises, that looks at the dynamics of fluency failure with respect to ongoing production rather than by reference to exceptions relative to a grammatical model. (2) The read/spontaneous difference suggest that planning has an important role in shaping the characteristics of speech output. Our view is that olanning primarily restricts the availability of future speech segments. The way speakers use segments prior to a point of fluency failure reflects how they are tackling the oncoming fluency failure.

We define "fluency failure" as occurring when ".. speech control falters even though the speaker does not produce an overt error" [3]. This differs from previous definitions. Bernstein Ratner suggests that fluency failure in stuttering arises "from a disturbance in construction of the prearticulatory plan" (p.97) where the disturbance in some of the models she discusses, involves speech error [4]. Of course, the difference in definition of fluency failure depends primarily on how error is defined. We use error to refer only to incorrect phoneme selection and anticipation (not perseveration) of a preceding word or phrase. So phoneme blends (e.g. " *tab*" for *taxi* + *cab*) and transpositions (e.g. *"slow and sneet"* for *"snow and sleet"*) in Spoonerisms would be errors. The prolonged, broken and repeated parts of words, pauses of all types and perseverated whole words in *"I got on, on the seven ... fffifty ... three train t.to Mac...clesfield."* are all fluency failures but not speech errors according to our definitions. Another important point, implicit in what has already been said, is that planning limitations on future segments (mentioned under point 2 above) are not necessarily, or even usually, errors in planning future segments. More often than not, planning limitations are evident when the plan is available late rather than whether the correct one is or is not available.

The contrast between our viewpoint and the generally accepted viewpoint engenders a different perspective on the phenomena of spontaneous speech control. In particular, the difference in definitions of error and fluency failure highlight whether an explanation of the surface characteristics of spontaneous speech should be sought in the limitations during planning or error breakdown relative to a normative model. Our perspective is not without precursors but it is a perspective that has received less attention than error-approaches [2, 5]. We begin defending our minority outlook by considering a theory that illustrates the alternative position highlighting its main attraction, namely how it is able to unify several strands of evidence, and then go on to discuss some inherent problems of such a perspective.

2. A standard approach to explaining these issues

Levelt proposed a model in which a message goes through several stages before it is output, and the pattern of errors at output is regarded as providing evidence for these stages [6]. A message starts in the conceptualizer that is responsible for generating a message and monitoring to ensure that it is delivered appropriately. Monitoring devices like that in the conceptualizer take the output of a process and compare it with what the process was intended to produce (the initial input to the conceptualizer). If there is a discrepancy (an error) an adjustment is made to the output to reduce the error. The message next goes through the formulation stage. At the output of the formulation stage, the message is represented as a phonemic string. In Levelt's model, as in many others [7, 8], two sub-stages to the formulation stage are identified where the message is represented in lemma and phonological forms respectively (however, see [9] for data questioning two sub-stages). Articulation is the final step in translating the abstract phonemic representations to overt speech. The overt speech is sent via an external loop through the speech perception system to the monitor in the conceptualizer. In Levelt's model, then, one monitor suffices for checking events between intention and output action.

Repairs, as in *"Turn left at the, no, turn right at the crossroads"*, are interpreted as support for the internal monitoring process. According to a repair interpretation, the speaker gives the wrong direction, realizes this and exchanges the *reparandum* (*"left"*) with the *alteration* (*"right"*). The

substitution is not the only thing that happens: The speaker overshoots the reparandum (goes on with the words *"at the"*) and *retraces* to *"turn"* when the message recommences (the word before *"left"* that is not in error). The repair contains an interruption (here the word *"no"* though this could be a pause or a short phrase). (Most of these parsed components are optional.) Different forms of error on the reparandum support monitoring at syntactic, lexical and phonemic levels and speech can be interrupted, corrected and restarted at any point between conceptualization and articulation. Levelt also considers that errors at levels before speech output can be detected and repaired before they have been spoken overtly. These are called *covert* repairs and an example is *"Turn right at the, at the crossroads"*. Such repairs are problematic to interpret as there is no outward sign indicating the error (i.e. why the speaker hesitated after *"turn"*), nor even whether an error occurred at all. Other constructions, described as repairs, are different (D) repairs in which the topic is changed (like non sequitors) and repairs in which the speech is not pitched at an appropriate level for the addressee (appropriateness, A, repairs).

A different source of evidence, altered auditory feedback AAF, is usually cited in support of the external loop that features in models like Levelt's. Feedback is used as a term to refer to the route by which output is returned to a monitor so the message can be compared with that originally intended. Levelt's model allows the same monitor to process internal and external feedback by routing external feedback to the monitor via the speech comprehension system. The effects of AAF procedures on speech control can be illustrated with delayed auditory feedback (DAF). When this form of feedback is applied a speaker hears the voice a short time after it is spoken. DAF slows speech mainly by elongating vowels, and the speech produced has a monotone pitch and high amplitude (with the last two effects again focused on vowels) [10, 11]. The effects of DAF on fluent speakers can readily be explained by interruption to the external loop by the DAF. In this type of explanation, speakers continue to use the altered sound for voice control even though the sound is delayed. The problems observed earlier arise because the monitor acts on the misleading feedback that leads the speaker to think an error has been made. In particular, DAF leads speakers to adjust output timing even though no adjustment was needed.

The main differences between our viewpoint and internal and external monitoring explanations of these aspects of speech control are: (1) We consider that the focus on errors is limiting: (2) The patterns of speech interpreted as a result of monitoring mislead theorists and constrain them into ways of explaining events in a monitoring framework when other, more general, interpretations are possible: (3) There are problems using AAF evidence in support of the external loop.

3. Detailed critique of internal and external monitoring accounts of spontaneous speech control

Errors are infrequent events in language with some estimates as low as 0.1% [12]. Fluency failures, on the other hand, are common. For instance, Howell, Au-Yeung and Sackin estimate that fluent speakers produce around 2.57% fluency failures on function words and 0.97% on content words [13].

A monitor takes corrective action when a difference is detected between intended and actual versions (the actual version assumed to be in error). To establish whether a difference has occurred or not, like needs to be compared with like. Consequently, the monitor must have a representation of intended forms produced as actual output at any of the stages up to and including output of a message. Why should an errorless version of the intended form at lower levels be available in the conceptualizer for comparison when these lower levels generate an erroneous output? If the conceptualizer can generate a correct version for monitoring, why can this not be done (or this representation used) for output of lower stages?

Empirically, the repairs described earlier, viewed from the perspective of our definition, actually provide little support for an internal error-monitoring process. Evidence that an error has occurred is only obtained when speech output reveals this. However, this selfsame evidence for an error would suggest either that no monitoring takes place or that the monitor has not worked on this occasion as the speech goes right through to output. The only type of repair where there may be evidence for the operation of an internal monitor is where errors are intercepted and the result is a covert repair. However, in these repairs another interpretation is that no error occurred in the first place (see the later discussion of stalling fluency failures for an alternative explanation of some surface-form features considered to represent components of repairs). In the case of D repairs, the message is abandoned rather than repaired, and A repairs are a matter of style rather than anything else. Abandonment and restart of a message (as in D repairs) may be a general process (applying to all repairs) reflecting anticipated planning difficulties rather than arising from internal monitoring for errors and on-line repair to remove any detected discrepancy.

There are also some specific problems for the proposal that the external loop is monitored using AAF evidence. Borden pointed out that auditory processing takes time: A segment has to finish, then its auditory output has to be processed to establish that the sound was produced correctly before the next one can be initiated [14]. Marslen-Wilson and Tyler estimate that recognition of running speech is about 200 ms and a processing delay of this order would lead to slower speech rates than speakers achieve [15]. Second, the information provided over the external loop would have to be a veridical record of what was produced; otherwise establishing if and what error has occurred by a monitor with the intention of correcting it, would not be possible [16]. However, the representation of articulatory output provided over this loop is not veridical with respect to the intended message. The auditory representation the speaker receives while speaking is affected by internal (mainly bone-conducted) sound and external noise sources (for instance,

variation caused by each unique speaking environment) during transmission.

4. The EXPLAN model of fluency failure

We turn, now, to a brief description of our model before describing some ways it has been evaluated. Cognitive planning (PLAN) and articulatory execution (EX) of speech are independent processes in the EXPLAN model. From the outset, the contrast with auditory monitoring accounts is apparent as PLAN and EX are interdependent in monitors (EX leads to auditory output that, if error is detected, would restart or tune PLAN processes in an auditory monitor). PLAN and EX operate as a chaining process in fluent speech (when one word finishes EX, the next PLAN is picked up) and the process is intrinsically timed [17].

The situation in which speech is fluent and one form of fluency failure when PLAN cannot keep up with EX is shown in Figure 1. Time is along the abscissa and time for planning and execution are indicated by the length of the bar. So, in the fluent speech example, two segments that are quick to plan are followed by one that takes longer to plan. Execution lags production by one segment (starting after the end of the first planned segment) and execution time determines when the plan of the next segment needs to be ready. As long as there is sufficient time for the execution of one segment for the next one to be planned (or has been in the preceding sequence), speech will proceed fluently. Though not essential to the theory, we have used Selkirk's phonological words to formulate some tests on English [18]. Selkirk defines a phonological word as consisting of a content word and an optional number of function words preceding and following it. Function words are simpler than content words in English and if the simplicity is reflected in planning time, the fluent example in Figure 1 would represent a function-function-content word sequence (e.g. "in a trice"). As planning of "trice" can take longer than the time to execute "a" allows, "a" can finish execution before the plan for "trice" is ready. The speaker needs more time for planning. In fluent speakers, this can arise by pausing when the plan runs out or by repeating words that immediately precede the difficult word. Au-Yeung, Howell and Pilgrim have shown that function word repetition always occurs on those preceding content words that would buy planning time [19].

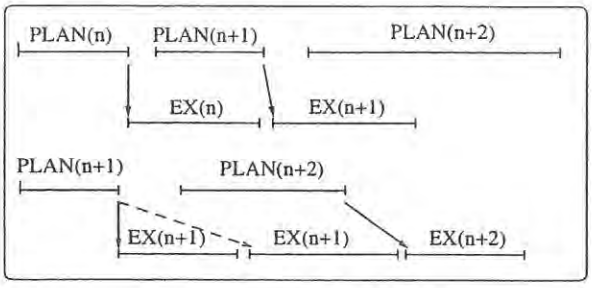


Figure 1: The EXPLAN model for fluent control and word repetition.

The property of English content words that may lead to them posing difficulty could be in the phonological structure at onset. Fluency failure occurs most often at word onset and "tip-of-the tongue" studies show that these parts are generated first. We have developed a characterization for difficulty of word onsets that incorporates whether the onset has a consonant string and whether consonants in this string are difficult to produce (indexed by the average age at which children acquire them) [20].

Another factor that can lead to variation in fluency is rate. At the outset, we noted that a text supplies a plan. According to EXPLAN, this would reduce the chance of speech execution getting ahead of planning. In contrast, if speech is negotiated at too fast a rate in local sections like "in a trice", the chance of planning getting ahead of execution is increased. Howell, Au-Yeung and Pilgrim have shown that a rapid rate in local stretches increases the likelihood of stuttering (an acute form of fluency failure) [21].

As EXPLAN does not allow interactions between execution and planning, it predicts that feedback-monitoring processes do not operate in fluency failures. This prediction appears to stand at odds with the disruption caused by alteration to auditory feedback that, as we saw, is readily explained as due to interference with monitoring processes. EXPLAN proposes that fluent speech control operates without external timing though there is an external timekeeper that regulates speech rate under specified circumstances. One of these is when DAF is presented. DAF is input direct to the timekeeper rather than transmitted from execution back to planning processes. This timekeeper is governed by number and timing relationship between inputs. Under DAF there is one extra asynchronous input that affects the timekeeper's ability to follow the response beat [22]. The lengthened period caused by the delay, adjusts an oscillator coupled to the EXPLAN process that also lowers the centre frequency of responses generated in the EXPLAN process [23]. This slows the speech of fluent speakers under DAF. For full details see [16].

The way this addresseses the two problems that were discussed in the introduction concerning an auditory monitoring account are considered with respect to whether they are problems for EXPLAN. The processes described do not require complete analysis of auditory output for speech content, they propose that any serial input to the timekeeper can affect its operation. As speech analysis is not performed on these inputs, the slowing problem that occurs in an auditory monitor does not arise. Also, the speech does not have to be veridical of the plan, only to represent rhythmic input to the timekeeper.

This work was supported by the Wellcome Trust.

5. References

[1] Levelt, W. J. M., "Monitoring and self-repair in speech", *Cognition, Vol. 14, 1983, pp 41-104.*

[2] Howell, P. and Kadi-Hanifi, K., "Comparison of prosodic properties between read and spontaneous speech", *Speech Communication, Vol. 10, 1991, pp 163-169.*

[3] Howell, P. and Au-Yeung, J., "The EXPLAN theory of fluency control and the diagnosis of stuttering", in *Current Issues in Linguistic Theory: Clinical Linguistics: Language Pathology, Speech Therapy, and Linguistic Theory*, John Benjamins, Amsterdam, in press.

[4] Bernstein Ratner, N., "Stuttering: A psycholinguistic perspective", in Curlee, R. F. and Siegel, G. M. (eds) *Nature and Treatment of Stuttering*, Allyn & Bacon, Boston, 1997.

[5] MacWhinney, B. and Osser, H., "Verbal planning function in children's speech", *Child Development, Vol. 48, 1977, pp 978-985.*

[6] Levelt, W. J. M., *Speaking: From Intention to Articulation*, MIT Press, Cambridge, MA, 1989.

[7] Dell, G. S., "A spreading-activation theory of retrieval in sentence production", *Psychological Review, Vol. 93, 1986, pp 283-321.*

[8] Dell, G. S. and O'Seaghdha, P. G., "Stages of lexical access in language production", *Cognition, Vol. 42, 1992, pp 287-314.*

[9] Caramazza, A. and Miozzo, M., "The relation between syntactic and phonological knowledge in lexical access: Evidence from the 'tip-of-the-tongue' phenomenon", *Cognition, Vol. 64, 1997, pp 309-343.*

[10] Black, J., "The effect of side-tone delay upon vocal rate and intensity", *Journal of Speech and Hearing Disorders, Vol. 16, 1951, pp 50-56.*

[11] Lee, B. S., "Effects of delayed speech feedback", *Journal of the Acoustical Society of America, Vol. 22, 1950, pp 824-826.*

[12] Shallice, T. and Butterworth, B., "Short-term memory impairment and spontaneous speech", *Neuropsychologia, Vol. 15, 1977, pp 729-735.*

[13] Howell, P., Au-Yeung, J., and Sackin, S., "Exchange of stuttering from function words to content words with age", *Journal of Speech, Language and Hearing Research, Vol. 42, 1999, pp 345-354.*

[14] Borden, G. J., "An interpretation of research on feedback interruption in speech", *Brain & Language, Vol. 7, 1979, pp 307-319.*

[15] Marslen-Wilson, W. D. and Tyler, L. K., "Central processes in speech understanding", *Philosophical Transaction of the Royal Society of London Series B, Vol. 259, 1981, pp 297-313.*

[16] Howell, P., "The EXPLAN theory of fluency control applied to the treatment of stuttering by altered feedback and operant procedures", in *Current Issues in Linguistic Theory: Clinical Linguistics: Language Pathology, Speech Therapy, and Linguistic Theory,* John Benjamins, Amsterdam, in press.

[17] Fowler, C. A., "Coarticulation and theories of extrinsic timing", *Journal of Phonetics,* Vol. 8, 1980, pp 113-133.

[18] Selkirk, E., *Phonology and syntax: The relation between sound and structure,* MIT Press, Cambridge, MA, 1984.

[19] Au-Yeung, J., Howell, P., and Pilgrim, L., "Phonological words and stuttering on function words", *Journal of Speech, Language, and Hearing Research, Vol. 41, 1998, pp 1019-1030.*

[20] Howell, P., Au-Yeung, J., and Sackin, S., "Internal structure of content words leading to lifespan differences in phonological difficulty in stuttering", *Journal of Fluency Disorders, Vol.* 25, 2000, pp 1-20.

[21] Howell, P., Au-Yeung, J., and Pilgrim, L., "Utterance rate and linguistic properties as determinants of speech dysfluency in children who stutter", *Journal of the Acoustical Society of America, Vol. 105, 1999, pp 481-490.*

[22] Howell, P., Powell, D. J., and Khan, I., "Amplitude contour of the delayed signal and interference in delayed auditory feedback tasks", *Journal of Experimental Psychology: Human Perception and Performance, Vol. 9, 1983, pp 772-784.*

[23] Large, E. W. and Jones, M. R., "The dynamics of attending: How people track time-varying events", *Psychological Review, Vol. 106, 1999, pp 119-159.*

Stuttering and Speech Monitoring

Nada Vasic and Frank Wijnen

Utrecht Institute for Linguistics OTS
Utrecht University, Netherlands
Nada.Vasic@let.uu.nl

Abstract

In ths paper we would like to argue that stuttering represents inadequate monitoring of the speech production process. The model we are proposing is *the vicious circle hypothesis.* The stuttering speakers have a malfunctioning monitor whose three parameters, namely, *focus, effort* and *theshold* are inappropriately set.

In order to test our hypothesis, we tested 20 stuttering individuals in a dual task situation. The experiment consisted of three conditions: baseline where semi-spontaneous speech was elicited and two dual task conditions. First dual task was speaking and playing a computer game at the same time where the processing resources were taken away from monitoring. The second dual task was designed to shift the monitor's focus away from habitual monitoring. Subjects were asked to monitor for a particular word in their speech. The preliminary results of our experiment show that in the dual task condition the number of disfluencies decreased in relation to the number of words, which, in turn supports our prediction that distraction has a positive effect on fluency in the case of stuttering individuals.

1. Introduction

The experiment presented here yields results that support the Vicious Circle Hypothesis which assumes that the abundance of dysfluency in stuttering speakers is due to a malfunction of the monitoring mechanism of the speech production system. We propose a psychological hypothesis that can account for both the primary and secondary characteristics of stuttering, and that is based on a well-established model of the human language-production mechanism, namely the speech production model (Levelt, 1989)[1]. Specifically, we will argue that some of the basic parameters of the monitoring process are maladapted in stuttering speakers.

2. Vicious Circle Hypothesis

The hypothesis proposed here states that in stuttering individuals three 'attention' parameters of the monitor: effort, focus and threshold are inappropriately set. A number of relatively clear and testable predictions can be derived:

(1) Effort. The vicious circle hypothesis argues that stuttering speakers invest an excessive amount of resources in monitoring. A prediction is that if you take away some of the available resources, disfluency will decrease.

(2) Focus. Persons who stutter do not just invest too much energy in monitoring, they do so for a reason, namely to 'look out' for all kinds of signs of realized or imminent disfluency. Their monitoring focus is rigid and unadaptive. The prediction we make here is that if you force them to focus on something other than temporal discontinuity, given the limited resources available, the chance of detecting discontinuities or temporal fluctuations will decrease. Hence, disfluency will decrease. Note that, in principle, focus can be altered independently of effort (it is just redirecting), so the prediction is essentially different from the foregoing.

(3) Threshold. Stutterers entertain acceptability criteria that are too strict. As a consequence, they reject speech output that other, non-stuttering people would consider completely normal. Since we assume that monitoring is based on normal perceptual processing, we predict that also in listening to speech from someone else, stutterers will entertain a more conservative standard with respect to disfluency than non-stutterers.

3. Experiment

In this experiment we concentrated on the first and the second parameter. The experiment was a dual task study, aimed at supplying further evidence for the assumption that persons who stutter do benefit from distraction of attention by a secondary, and highly demanding task, and from reorienting the monitor's focus.

3.1. Subjects

We conducted an experiment on 30 individuals, 20 stutterers and 10 non-stutterers. We defined stutterers as persons who were diagnosed as such by a speech therapist, and who considered stuttering to be a problem at the time of the experiment. The subjects ranged from mild to severe stutterers.

3.2. Tasks

The experiment comprised two conditions, namely, speech only and a dual task conditions in which speaking was accompanied/combined with one of the two tasks described below. In the primary task semi-spontaneous speech was elicited. Subjects were asked to read a newspaper article prior to each condition in order to retell the article's content. In case they stopped talking, the experimenter would call out a cue word/topic, such as 'holidays', 'family', 'hobbies' etc., and the subjects had to elaborate on the topic.

The first distraction task was supposed to take away general resources so that the monitor would be left with less energy to invest into speech monitoring. This distraction task was a computer game 'PONG'; subjects had to play table tennis against the computer. Each subject was screened for 'PONG' so that the level of the game could be tailored to the level of each individual subject.

As a contrast to the first distraction task, the second distraction task was designed to alter monitor's focus. It involved monitoring for a particular word in speech production. The subjects were asked to closely monitor their speech for the word *die* (that – indexical and relative

pronoun), which is a fairly frequent word in Dutch. They had to press a button each time they heard *die* in their speech. The computer recorded their response in order to check whether they were paying attention to the secondary task.

Each experimental condition was 10 minutes long. The first condition was a baseline condition (Speech only) in which the subjects were asked to retell the story they read in a newspaper article with no distraction. In the second condition (Dual difficult) they were instructed to retell the story and play PONG at the same time. This was the condition in which the distraction task was demanding and where the level of the game was constantly increasing in its level. The third condition (Dual simple) was also a double task condition, however, the distraction task was not as demanding as in the previous condition; subjects were invited to speak and play PONG at the same time, the level of PONG was kept constantly at its lowest (speed and acceleration). In both condition two and three the computer kept track of subjects' performance. Finally, the last condition was Monitoring *die*, subjects had to speak and react by pressing a button each time they heard themselves utter *die*.

3.3. Design

Subjects were tested individually and were taped on a digital audio recorder (DAT). After reading the instructions with a short description of the experimental tasks subjects were screened for PONG. This was used to determine the level for PONG in the Dual difficult conditions. The order of conditions was rotated across subjects such that for each block of four subjects four conditions were fully rotated.

3.4. Analysis

The first seven minutes of speech in each condition were transcribed transcribed and coded for disfluencies. Disfluencies were transcribed and classified as blocks, self-corrections, prologations, repetitions, senseless sound insertions, word breaks, unfilled pauses and filled pauses. Several occurrences of one type of disfluency on the same position were counted as one disfluency. However, two disfluencies of a different type that occurred on the same segment were coded twice.

Transcribing and coding was done in the CLAN program originally designed for the CHILDES database. In the case of our experiment we needed to count the number of words and the number of dysfluencies per condition. Disfluency is, therefore, determined in terms of the number of disfluencies per condition in relation to the number of words uttered in each condition. Filled pauses were coded but were not included in the analysis. These pauses are very common in the speech of fluent individuals and are therefore not treated as indicators of stutter-moments.

The null-hypothesis predicts no difference in the number of disfluencies between the three conditions across subjects relative to the number of words. Our model predicts a decrease in the mean number of disfluencies in the dual task conditions depending on the success with which the secondary tasks take away processing resources or manage to shift monitor's focus.

3.5. Results

It is well known that the population of stutterers is rather heterogeneous, stuttering ranges from very mild to extremely severe. It is conceivable that the effects of our experimental manipulations could interact with stuttering severity. However, the exact effect of heterogeneity in our sample was unknown beforehand, therefore we had no specific expectations on whether the results might be affected by subject variables. Additionally, the number of words produced by different subjects in different conditions varied. Therefore, we had to deal with two sources of variability in our experiment that could potentially affect the dependent variable. In order to find out how this construct affects the measurement of interest, the amount of disfluency in stuttering individuals, we performed a multilevel analysis (Goldstein, 1995)[2] on the data. This particular statistical analysis is optimally tailored to a full exploration of our data since it allows an explicit modelling of hierarchical and nested relationships between observations. The multilevel analysis allows a more complex structure of the error variances so that both differences between subjects and within subjects (between different conditions) can be removed from the residual error variance (te Riele, 1999)[3].

We wanted to check whether the number of disfluencies differed across conditions. To do so, the amount of disfluency had to be calculated in relation to the number of words produced in each condition. To this end, each individual word a particular subject produced was coded as fluent or as not fluent. A two-level regression model of the MLN computer program (Prosser, Rasbash and Goldstein, 1995)[4] was used. Its output consists of two parts: one is the fixed part of the model that gives estimates of proportions of responses, and the other one, which is the random part of the model that gives the estimates of the variances from the estimated means. Consequently, differences between estimated means were calculated and their significance was determined on the basis of chi-square.

Table 1:

Parameter estimates (logits) and proportions per type of dysfluency per total number of words

	Pong difficult		Pong simple		Speech only		Monitoring *die*	
	Logit	%	Logit	%	Logit	%	Logit	%
BLK	-2.273(.268)	.09	-2.294(.268)	.09	-2.205(.268)	.10	-2.852(.269)	.05
COR	-4.921(.193)	.007	-4.997(.195)	.006	-4.597(.186)	.009	-4.548(.185)	.01
PRL	-7.167(.576)	.0007	-5.812(.511)	.002	-6.427(.531)	.002	-5.895(.514)	.002
REP								
Cpx	-5.821(.600)	.002	-5.860(.600)	.002	-5.651(.597)	.003	-5.107(.589)	.006
Isg	-4.399(.225)	.01	-4.449(.226)	.01	-4.550(.227)	.01	-4.140(.228)	.02
Isy	-7.114(.488)	.0008	-6.718(.456)	.001	-7.158(.491)	.0007	-7.767(.565)	.0004
Wrd	-3.571(.167)	.027	-3.495(.166)	.029	-3.379(.165)	.033	-3.213(.164)	.039
Wst	-4.131(.284)	.016	-4.232(.285)	.014	-4.113(.284)	.016	-3.675(.281)	.025
SSI	-5.751(.269)	.003	-5.314(.252)	.005	-5.009(.245)	.007	-5.069(.246)	.006
UPS	-5.927(.399)	.003	-5.884(.397)	.003	-5.954(.399)	.003	-5.471(.386)	.004
WBR	-4.988(.325)	.007	-5.247(.329)	.005	-4.937(.324)	.007	-4.998(.325)	.007
TOTAL	-1.575(.124)	.17	-1.559(.124)	.17	-1.541(.124)	.19	-1.575(.124)	.17

NOTE: **BLK** = block, **COR** = self-correction, **PRL** = prolongation, **REP:cpx** = complex repetition, **REP:isg** = repetition of an initial segment, **REP:isy** = repetition of an initial syllable, **REP:wrd** = repetition of a word, **REP:wst** = repetition of a word string, **SSI** = senseless sound insertion, **UPS** = unfilled pause, **WBR** = word break.

In *Table 1*, the parameter estimates are presented per condition. More specifically, the mean number of disfluencies was calculated for each condition and is represented as totals in the table. The calculations were performed in logits, which are means expressed on a scale different than the scale in which the observations were made. The logit of a proportion P is defined as log (P/1-P). In order to convert the calculated means back to values that are interpretable, proportions calculated from the logits. If we look at the total proportions, it is obvious that the baseline condition (Speech Only) has a higher proportion of disfluencies in comparison to the other three conditions. This difference is small, nevertheless, it is significant. It is essential to note that the strength of our experiment lies in the number of observations made - words that were produced fluently or non-fluently. The total number of observations (words for all conditions for all subjects) for the whole sample that we analysed was 55177.

We found a significant difference in the mean number of disfluencies between the Pong Difficult condition and Speech Only condition ($\chi^2_1 = 15.25$, df=1 ($p<.01$)). In the baseline condition (Speech Only) subjects produced more disfluencies than in the dual task difficult condition. A significant difference was also found between Pong Simple condition and Speech Only condition ($\chi^2_1 = 11.98$ df=1 ($p<.01$)). Once again, there were more disfluencies in the baseline condition as opposed to the simple dual distraction condition. Finally, a significant difference was also found between the baseline condition and Monitoring *die* condition $\chi^2_1 = 15.51$ df=1 ($p<.01$).

There were 11 different types of disfluencies, and as can be seen in Figure 1, some of these occurred very infrequently. The most frequent ones were **bloks** and **repetitions. Blocks** were the most frequent type of all with close to 10% occurrence per total number of words across conditions. There was also a significant difference in the mean number of blocks between all dual conditions and the baseline - Pong Difficult condition vs. baseline ($\chi^2_1 = 4.27$ ($p<.01$)); Pong Simple condition vs. baseline ($\chi^2_1 = 4.80$ ($p<.01$)); Monitoring *die* condition vs. baseline ($\chi^2_1 = 189.90$ ($p<.01$)). In all instances the number of blocks increased in the baseline condition as opposed to other conditions

Repetitions were split into 5 sub-categories, which were analysed as different types. Both repetitions of **initial segments** of **initial syllables** were rather infrequent in all conditions. Only in the distraction condition where the subject had to pay attention to *die* we found an increase in the number of disfluencies for both types. **Complex** repetitions were also very infrequent (average .08%) and no difference was found between the conditions. **Word** repetitions were more numerous (average 13%) with a significant difference between Pong Difficult condition and the baseline condition ($\chi^2_1 = 8.68$ $p< .01$). It is interesting to note that there were significantly more disfluencies in the Monitoring *die* vs. all other conditions. Similar results were found for **word string** repetitions which occurred also slightly more often than word repetitions. There was an increase in the number of word string repetitions in the Monitoring *die* condition. Other types exhibited no clear pattern in relation to our predictions.

4. Discussion

Our hypothesis makes two crucial predictions with regard to the experimental study that we conducted. First one is related to the general processing resources that the monitor uses while keeping track of the speech production process. These resources can be taken away from monitoring by means of a demanding secondary task. By doing so, monitoring should become less excessive in the case of the stuttering individuals and, therefore, disfluency is expected to decrease. Second prediction is related to the focus of the monitor during speech production. We proposed that the monitor habitually over-focuses on the temporal characteristics of speech in the stuttering individuals. If we somehow push away the monitor from its habitual focus the speech should result in less disfluencies. In this case we are not taking away the resources it needs for processing but reorienting the monitor.

From the results obtained it can be concluded that it is indeed the case that when distracted, stutterers produce less disfluencies. Many previous studies failed to reproduce the same results, which many claimed to be a consequence of a secondary task that was not sufficiently engaging. In order to verify whether the effects of a dual task on speech fluency is due to the degree to which the secondary task takes away resources the first secondary task (PONG) was varied. Our study shows that even a slight distraction could be beneficial. In the case of the most simple computer game (Pong Simple condition) where the level was at its lowest and was kept constant subjects produced more fluent speech. It is also important to emphasise that the type of distraction is not crucial. Our results show that fluency increases with general distraction in the case of the computer game, but also in the experimental condition where the focus of the monitor was shifted towards attributes of the speech other than its temporal characteristics.

We found that the different types of disfluencies responded differently to the two types of secondary tasks. The frequency of blocks decreases in all three dual task conditions (Pong difficult, Pong easy, *die*-montoring). By contrast, the amount of repetitions dropped in the PONG conditions, but appears to raise in the *die*-monitoring condition. Blocks are known to be the most 'pathological' symptom of stuttering (i.e.,they are not among the disfluencies that can be considered normal). Note that in many clinical analyses of the development of stuttering in children (e.g. McDearmon, 1968; Yairi and Lewis, 1984)[5][6] all other types of disfluencies with the exception of blocks occur in both stuttering and non stuttering children. The emergence of blocks is considered to be an ominous phenomenon. Once the child starts to block, there appears to be no way back, the 'physiological' disfluency of the immature child has turned into a problem. Particularly, it has been argued (Johnson 1956)[7] that the emergence of blocks is correlated with the child's emerging awareness of his disfluency. We speculate that blocking is a learned response to the self-perception of (imminent) disfluency, which arises and remains partly under conscious (attentive) control.

Reiterating previously articulated material (repeating), on the other hand is a natural response of the language production system to trouble in planning or delivery. We may assume that part of the repetitions in stuttering are these 'normal' reactions of the language production system. Another part may be due the maladaptive monitoring process that we hypothesize. Naturally occurring disfluency is

sensitive to 'amount of work' in the production system: we have seen that disfluency rises as the speech task gets more difficult.

On the one hand, we have the perceptuo-motor secondary task (PONG) which takes away processing resources across-the-board. When performing this task stutterers are prevented from following their habit. In comparison to the baseline condition (speaking only), in the perceptuo-motor task there is no additional 'pressure' within the language production system, which has a beneficial effect on the speech of stutteres. The focus-redirecting monitoring task, on the other hand, appears to have two simultaneous effects: (1) what it's supposed to do, namely drawing the stuttering person's monitor away from what it normally focuses on, and (2) by doing so increasing the load on the production system. Consider what it means to be instructed to explicitly report every occurrence of particular word in your speech output. It means continuous controlled, attentive monitoring. It is very likely that this will interfere with normal speech planning and delivery, and that, therefore, the number of "normal" disfluencies will rise, and more so than the distracting effect will suppress them. From this perspective, it is very meaningful that the decrease in this condition is in the non-normal type of disfluency, whereas the rise is in the class of disfluencies that (at least in part) can be considered normal. This observation may provide support for our interpretation that in fact two processes co-occur, one which affects the normal "healthy" part of the language production system (→ more repetitions) and one which affects the "pathological" part, what we have named maladaptive monitoring (→ less blocks).

From our results we can conclude that even the perceptuo-motor secondary task in its easy form helps in bringing down the number of disfluencies in stuttering speakers. This seems to go with the suggestion made by us and others, (Thompson, 1985)[8] that the effectiveness of a secondary task hinges on its capacity to continuously engage attentive processing. We expected that our easy task would fail to do so, but the results suggest differently. It is likely that we have underestimated the demands made by the PONG game, particularly for relatively inexperienced players. Even when played at a low level, it may very well be that PONG is highly engaging. At the very least, the game requires a certain amount of visual attention in order for it to be played successfully.

5. Conclusions

The results of our experiment support the claim that stuttering is a consequence of maladaptive monitoring during speech production. Taking away the processing resources or shifting the focus of the monitor results in a more fluent speech.

6. References

[1] Levelt, W.J.M. *Speaking: From Intention to Articulation.* Cambridge, MA:MITPress, 1989.

[2] Goldstein, H.(1995). Multilevel Statistical Models. London:Arnold.

[3] Riele, S. te, (1999). *Early Context Effects in Spoken-word Perception.* Doctoral dissertation. Utrecht University, The Netherlands.

[4] Prosser, R., Rasbash, J. and Goldstein, H. (1995) *MLN software for multilevel analysis.* London: University of London, Institute of Education.

[5] McDearmon, J.R.(1968). Primary stuttering at the onset of stuttering: A reexamination of data. *Journal of Speech and Hearing Research, 11,* 631-637.

[6] Yairi, E. and Lewis, B. (1984). Disfluencies at the onset of stuttering. *Journal of Speech and Hearing Research, 27.* 154-159.

[7] Johnson, W. (1956). Stuttering. In W.Johnson, S.J. Brown, J.J. Curtis, C.W. Edney and J. Keaster (Eds.), *Speech in handicapped school children, revised edition.* New York: Harper & Brothers.

[8] Thompson, A.H. (1985). A test of the distraction explanation of disfluency modification in stuttering. *Journal of Fluency Disorders, 10,* 35-50.

Repeated Phoneme Effect in Japanese Speech Errors

Michiko Yoshida

The Graduate School of Languages and Linguistics
Sophia University, Tokyo
malaysia@anet.ne.jp

Abstract

Analyses of errors in the natural speech of Dutch, German, and English have shown that involuntary rearrangements of phonemes (e.g., left hemisphere → *h*eft *l*emisphere) are more likely to occur when the two words involved in the error have the same phoneme before or after the phoneme on which the error occurred (e.g., /E/ in l*e*ft h*e*misphere) [1, 2]. A study by Dell (1984) has revealed that phoneme repetition could also contribute to experimentally induced speech errors in English [3]. The present study explored the effect of repeated phonemes in Japanese speech errors by means of two error-inducing experiments. Analyses of subjects' errors showed that a sequence of syllables that share the same phoneme was more error-prone than one with a variety of phonemes, suggesting that phoneme repetition could contribute to Japanese speech errors. These results are consistent with the view that the repeated phoneme effect is common to all speakers regardless of language.

1. Introduction

Phoneme repetition has been recognized as one of factors that contribute to phonological speech errors, where one or more words are mispronounced. To take but one example, /ʎ/ and /ɳ/ in 'left hemisphere' will exchange and produce an error such as '*h*eft *l*emisphere', more often than /ρ/ and /ɳ/ in 'right hemisphere' will. In this case, the same phoneme /E/ following /l/ and /h/ induces such an error. This finding, sometimes called the "repeated phoneme effect", is important because it reveals something about the mechanisms underlying speech production.

The repeated phoneme effect operates both forwards and backwards in the serial order of speech. An analysis of German and English spontaneous spoonerisms has shown that repeated phonemes precede the phonemes on which the errors occurred as frequently as they followed them [1].

Not only repeated vowels but also repeated coda consonants lead to the difficulty in speech. Using a large collection of about 4,000 English errors, Dell [3] has shown that the rate of errors on word-initial consonants increases when the two words involved in the error have the same word-final consonant. He replicated this finding experimentally with the SLIPS technique, which was originally developed by Baars and Motley [4]. It elicits initial consonant exchanges (e.g., *b*eal *d*all instead of deal ball), anticipations (e.g., *b*eal ball instead of deal ball) and perseverations (e.g., deal *d*all instead of deal ball) using phonological priming. Subjects see several interference word pairs before a critical word pair, one at a time, with the instruction to prepare to say each pair as they see it. Eventually, they see a series of question marks that signals the subjects to speak. They must say aloud the last word pair that they saw, i.e. a critical word pair. The reason that speech errors are obtained is that critical stimuli (e.g., deal ball) are preceded by three to five interference stimuli (e.g., bet dart) that bias for a reversal of initial consonants. A final aspect of the procedure is that after saying each critical stimulus the subjects have to judge whether or not they said what they intended to say, and have to repeat slowly what they intended to say. This allows for errors of reading or memory to be separated from speech errors. The procedure demonstrated that the repeated /l/ in 'deal ball' is effective in increasing error rates of /d/ and /b/ just as the repeated /Θ/ in 'mad back' in increasing error rates of /m/ and /b/.

What remains a question was whether the repeated phoneme effect is common to all speakers regardless of language. Previous studies on speech errors in a few Germanic languages are not sufficient to support the hypothesis that the effect of phoneme repetition is language independent, and to suggest that this phenomenon may reflect a universal aspect of mechanisms underlying speech production. Further research on speech errors in non-Germanic languages is required to test the generality of this effect.

In the present study, two error-eliciting experiments were conducted with native Japanese speakers to test the generality of the repeated phoneme effect. In both experiments, the stimuli with the repeated phonemes induced more speech errors than those with a variety of phonemes. Thus, the cross-linguistic robustness of the repeated phoneme effect was established. A point to note is that the rate of errors on syllable coda consonants increased when Japanese subjects intended to say aloud a sequence of syllables with repeated

coda consonants. This finding throws doubt on the hypothesis that repeated phonemes induce adjacent or next-to-adjacent phonemes to slip [3].

2. Experiment 1

The aim of this experiment was to replicate the repeated phoneme effect with Japanese speakers. Dell [3] has replicated this effect in English using the pairs of real words as stimuli; this experiment, using the sequences of syllables of Japanese.

This experiment made use of error-causing agents such as fast speech rate and time pressure to induce errors. To examine the natural effect of repeated phonemes, no phonological priming to induce phonological errors was used; thus, this error-inducing technique differs from the SLIPS technique used by Baars et al. [4] and Dell [3].

2.1. Materials

Ten sequences of CV (consonant-vowel) syllables, shown in Table 1, were used as stimuli. Each sequence consisted of eight CV syllables. Every syllable in a sequence either shared a vowel (ta.sa.ha.ra.na.ma.pa.ka) or did not (to.sa.hi.re.nu.mo.pi.ke). (The mark /./ indicates the syllable boundary.) Consonants appear in the same order in all sequences (t, s, h, r, n, m, p, and k) in both conditions. None of the stimuli items constituted a meaningful word in Japanese. Materials are created in such a way that factors such as phonemic similarity or word frequency would not affect the results [1, 2, 5, 6]. To examine the influence of repeated phonemes themselves, no interference stimulus that would serve to bias subjects was used in this experiment.

Table 1: Materials used in Experiment 1

Stimuli set	
Repeated vowel condition	Different vowels condition
ta.sa.ha.ra.na.ma.pa.ka	ta.so.fu.ri.ne.ma.pu.ki
τι.si.hi.ri.ni.mi.pi.ki	ti.se.ha.ru.no.mi.pa.ku
tu.su.fu.ru.nu.mu.pu.ku	tu.σi.he.ro.na.mu.pe.ko
te.se.he.re.ne.me.pe.ke	te.su.ho.ra.ni.me.po.ka
to.so.ho.ro.no.mo.po.ko	to.sa.hi.re.nu.mo.pi.ke

2.2. Subjects and procedure

Ten students from Sophia University participated in the experiment. All were native speakers of Japanese.

Subjects were tested individually in a soundproof studio. They were seated in front of a portable PC, Toshiba Dynabook Satellite 2710 P50/4CA, which ran Frame Editor for Windows Version 1.0 that controlled the progress of the experiment. Stimuli were written in *kana*, i.e., Japanese syllabary characters, and visually presented on the screen. The subjects received each sequence of syllables one at a time. The whole stimuli set was presented twice, each time in a pseudo-random order. Four practice sequences preceded the set. They had to memorize each sequence in 8 seconds and repeat it as quickly and as many times as possible under time pressure. They were instructed to do so, because speech rate and time pressure are assumed to be error-causing factors.

Each trial had the following structure: Subjects saw a sequence on the screen following a signal sound. They had 8 seconds to read the sequence silently and to prepare to say it aloud. The sequence disappeared and then another signal sound was played. They were given 6 seconds to repeatedly recite the sequence as quickly as possible. Six seconds later, the screen presented the command with another signal: "ONCE MORE, SLOWLY" and subjects were to slowly recall and recite the sequence that they intended to say. Four seconds after this command, the screen presented the question: "DID YOU MEMORIZE CORRECTLY?" to which subjects were supposed to answer aloud yes or no, based on their judgment as to whether they memorized the sequence of syllables correctly.

The last additional recall task was included to identify errors that might have occurred during input, such as reading or memorizing. If, when presented /ta.sa.ha.ra.na.ma.pa.ka/, a subject said [ta.*ha*.*sa*.ra.na.ma.pa.ka], then recalled [ta.*ha*.*sa*.ra na.ma.pa.ka] as her intended utterance and then said yes to the question: "Did you memorize correctly?", the error was more likely a slip of the eye than a slip of the tongue. Such an error should not be counted as a speech error.

2.3. Results and discussion

In the repeated vowel condition, 56 out of 6320 segments uttered by 10 subjects were counted as errors. On the other hand, 43 out of 7745 segments were counted as errors in the different vowels condition. Errors were counted in the following way. In substitution errors, (e.g., [ne] for /te/), only one segment was counted in. Likewise, in deletion (e.g., [pka] for /pa.ka/) and addition errors (e.g., [muo] for /mo/), only one segment was counted in. However, in exchange errors (e.g., [mu.nu] for /nu.mu/), two segments were counted in. Any stuttering such as [s, s, sa] was not counted in. As stated above, the errors that might have occurred during input, such as reading or memorizing were excluded. Incomplete sequences, which were left unfinished by subjects due to the time limit, were also excluded from analysis.

Error rates in both conditions were computed for each subject. The results are shown in Figure 1. Overall, stimuli with repeated vowel elicited more errors than those with different vowels ($p<.05$, by a sign test across subjects). This

experiment provided the first clear evidence that the repeated phoneme effect is found in Japanese speech errors.

There is another thing to note. The repeated vowels in this experiment increased the error rate of syllable onset consonants. This finding seems consistent with a study by Dell [3].

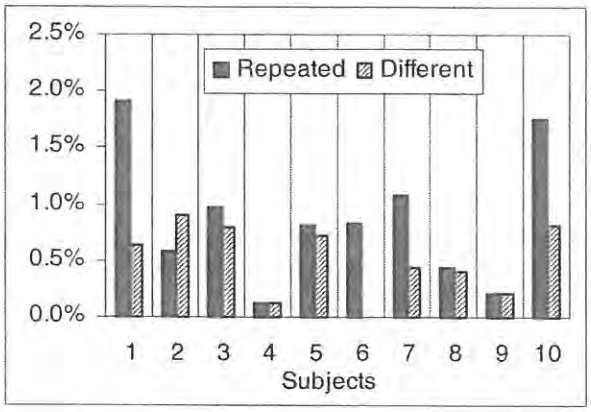


Figure 1: Error rate of each subject in repeated vowel and different vowels conditions

3. Experiment 2

The aim of this experiment was to examine the effect of repeated coda consonants. In this experiment, CVX syllables (where X stands for either a vowel or a consonant) of Japanese were used as stimuli.

3.1. Materials

Stimuli for this experiment were created with caution. There were at least two reasons. One was that the Japanese language allows only four phonemes to close the syllable. They are /N/, /Q/, /R/, and /J/ (where /N/ is a nasal, /Q/ is the first half of a geminate, /R/ is the latter half of a long vowel, and /J/ is the latter half of a diphthong such as [ai]) [7, 8]. Another reason was that the phoneme /Q/ is not allowed in the word final syllable.

Ten sequences of syllables shown in Table 2 were used as stimuli. Each sequence consisted of five CVX syllables. A sequence either contained repeated /N/ and /Q/ (e.g., taN.soQ.fuN.puQ.kiN), or did not (e.g., taR.soQ.fuN.puJ.kiR). Onset consonants appeared in the same order in all sequences (t, s, h, p and k). None of the stimuli items constituted a meaningful word in Japanese.

Table 2: Materials used in Experiment 2

Stimuli set	
Repeated coda condition	Different codas condition
taN.soQ.fuN puQ.kiN	taR.soQ.fuN.puJ.kiR
tɪN.seQ.haN paQ.kuN	tɪR.seQ.haN.paJ.kuR
tuN.sɪQ.heN.peQ.koN	tuR.σiQ.heN.peJ.koR
teN.suQ.hoN.poQ.kaN	teR.suQ.hoN.poJ.kaR
toN,saQ.hiN.piQ.keN	toR.saQ.hiN.PiJ.keR

3.2. Participants and procedure

The subjects and procedure were as in Experiment 1. A short break separated the two experiments.

3.3. Results and discussion

The procedure resulted in 210 segmental errors in total. In the repeated consonant condition, 144 out of 6753 segments uttered by 10 subjects were counted as errors. On the other hand, 66 out of 8130 segments were counted as errors in the different codas condition.

As before, error rate was computed for each subject. The results are shown in Figure 2. Overall, repeated phoneme stimuli elicited more errors than their control stimuli did ($p<.001$, by a sign test across subjects). From these results, it appears that repeated coda consonants are contributory causes of speech errors for native Japanese speakers, just as repeated vowels are.

There is another important point to note. Contrary to expectation, the procedure almost always induced coda consonants to slip (e.g., a reversal of /Q/ and /N/ in /teN.suQ.hoN.poQ.kaN/ resulted in [ten.sun.hop.pok.kan]) in both conditions, and the error-rate of coda consonants increased in the repeated coda condition. This finding is not consistent with Dell's (1984) conclusion that repeated phonemes induce adjacent or next-to-adjacent phonemes to slip [3]. The reason for the disagreement is not clear. It may be attributed to factors such as the task differences [9] and structural differences of languages. As stated above, this experiment did not use interference stimuli that would serve to bias for the reversals of onset consonants. It may be the case that repeated vowels and codas in Dell's experiments just enhanced the capacity of interference stimuli to induce errors on onset consonants: repeated phonemes themselves did not induce onsets to slip. On the other hand, repeated codas in this experiment might have enhanced the capacity of some unknown factor to induce errors on coda consonants. It might be some aspect of linguistic structure of Japanese that induced such errors. It calls for further experimental research on Japanese speech error.

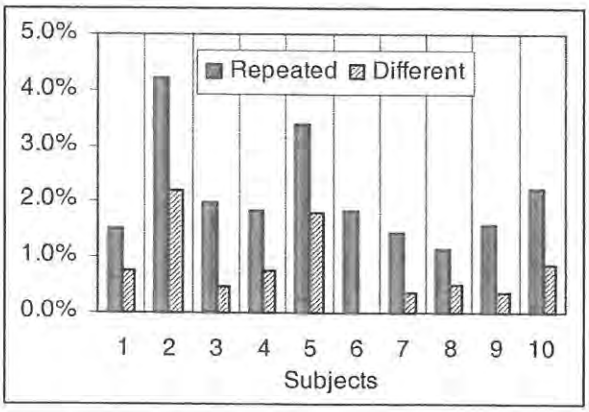


Figure 2: Error rate of each subject in repeated coda and different codas conditions

4. Conclusion

This study revealed that the repeated phoneme effect is found in Japanese speech errors. It was shown that not only repeated vowels but also repeated coda consonants lead to the difficulty in speech. This is the first demonstration that phoneme repetition plays the same role in Japanese speech errors as in Germanic languages.

It was found that repeated coda consonants could increase the error-rate of coda consonants. This finding is not consistent with a view that repeated phonemes induce nearby phonemes to slip. A satisfactory explanation for this finding was not obtained in this study. Further studies are required to identify the cause for this inconsistency.

5. References

[1] MacKay, D. G., "Spoonerisms: The structure of errors in the serial order of speech", *Neuropsychologia*, 8: 323-350, 1970.

[2] Nooteboom, S. "The tongue slips into patterns", In V. A. Fromkin (Ed.), *Speech errors as linguistic evidence*. The Hague: Mouton, 1973.

[3] Dell, G. S., "Representation of serial order in speech: evidence from the repeated phoneme effect in speech errors", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(2): 222-233, 1984.

[4] Baars, B. J., Motley, M. T., & MacKay, D. G., "Output editing for lexical status in artificially elicited slips of the tongue", *Journal of Verbal Learning and Verbal Behavior*, 14, 3822-391, 1975.

[5] Dell, G. S., "The retrieval of phonological forms in production: tests of predictions from a connectionist model", *Journal of Memory and Language*, 27: 124-142, 1988.

[6] Dell, G. S., "Effects of frequency and vocabulary type on phonological speech errors", *Language and Cognitive Processes*, 9: 313-349, 1991.

[7] Shibatani, M., *The languages of Japan*. Cambridge: Cambridge University Press, 1990.

[8] Kubozono, H., "Mora and Syllable", In N. Tsujimura (Ed.), *The Handbook of Japanese Linguistics*. Oxford: Blackwell, 1999.

[9] Stemberger, J. P., "The reliability and replicability of naturalistic speech error data", In B. J. Baars (Ed.), *Experimental slips and human error: exploring the architecture of volition*. New York: Plenum Press, 1992.

Different sources of lexical bias and overt self-corrections

Sieb G. Nooteboom

UiL OTS
Utrecht University, Utrecht
Sieb.Nooteboom@let.uu.nl

Abstract

In this paper it is argued, on the basis of a quantitative analysis of spontaneous speech errors and their corrections in Dutch, that the mechanism leading to lexical bias in speech errors cannot be same as that leading to overt self-corrections. Although spontaneous speech errors show a strong lexical bias, overt self-corrections do not. Lexical bias strongly increases with dissimilarity between target phoneme and source phoneme No such effect is found in overt self-corrections. Several possible sources of these differences are discussed.

1. Introduction

Baars, Motley and MacKay (1975) elicited spoonerisms by having subjects read aloud a target like *darn bore* preceded by bias items in which at least the first phoneme in this case was a *b*, triggering the spoonerism *barn door*. They observed that the error rate for cases such as *darn bore*, triggering lexically viable outcomes, was higher than the error rate for cases like *dart board*, triggering non-word outcomes. This lexical bias was not found when the context contained non-words only. The authors explained this result by positing an output-editing mechanism suppressing non-words that arise from speech errors in inner speech. Levelt, Roelofs and Meyer (1999) recently supported this original explanation by Baars et al. (1975) and suggested that the pre-articulatory editing leading to lexical bias is a form of covert self-correction of internal speech by the self-monitoring system that is also responsible for overt detection and correction of speech errors. A different approach has been suggested by Dell & Reich (1980), and Dell (1986), who proposed that lexical bias is caused by "phoneme-to-word" feedback during production processes, and therefore obviously not by the same mechanism that is responsible for the overt detection of speech errors. Postma (2000) recently suggested that there is production-based monitoring during speech production. If there is, such monitoring can also be responsible for lexical bias.

If overt speech errors and overt self-corrections are affected by the same perception-based monitoring system, one expects the distributions of both to reflect the same kinds of bias. If, on the other hand, overt speech errors are only affected by production-based mechanisms and overt self-corrections are affected by perception-based monitoring, one would not necessarily expect these distributions to be similar. Of course, before we can study possible similarities between these distributions it should first be assessed whether spontaneous speech errors do show an effect of lexical status. Note that this effect is compatible with both a production-based origin, as suggested by Dell & Reich (1980), Dell (1986) and Postma (2000), and with a perception-based origin (Levelt, 1989; Levelt et al. 1999). Below I will examine four predictions

1) The first prediction is that there is a lexical bias effect not only in the laboratory task used by Baars and his associates, but also in spontaneous speech errors. Garrett (1976) did not find much evidence for lexical bias in his MIT corpus of spontaneous errors. Dell and Reich (1981) report a considerable lexicality effect for another corpus. Here a new attempt will be described, on the basis of speech errors drawn from two collections of spontaneous speech errors in Dutch, and a new proposal for a null hypothesis.

2) If there is lexical bias in spontaneous speech errors and it is caused by perception-based self-monitoring, then one would expect to find a similar lexical bias in the overt corrections of spontaneous errors, meaning that spontaneous non-word errors are significantly more often corrected than real-word errors.

3) As self-monitoring is self-perception, and smaller differences are less easily perceived than larger differences, one may predict that speech errors differing minimally from the intended forms will be more often go unnoticed than speech errors differing greatly from the intended forms. One should therefore expect suppression of non-words to be less likely when two phonemes involved in a substitution differ in only one feature than when they differ in more features.

4) Obviously, if we do find an effect of phonetic similarity in lexical bias, and suppose that lexical bias is caused by the same mechanism that is responsible for overt detection of speech errors, then we expect to find the same sensitivity to phonetic similarity in the distribution of overt self-corrections. Notably we expect that errors involving a single-feature difference with the intended form are less likely to be corrected than errors involving more features. Sensitivity to phonetic similarity is compatible both with a production-based and with a perception-based mechanism. Therefore the interesting case would be if one were to find such sensitivity in the one and not in the other. This would be evidence for different mechanisms.

2. Method

2.1. Two collections of speech errors

The above predictions have been tested against errors drawn from two different collections of spontaneous speech errors in Dutch, for predictions 1 and 3, and from one of these collections for prediction 2 and 4. The reason for the difference is that one of the available collections alone provided too few relevant cases for testing predictions 1 and 3, and the oldest of the two collections could not be used for

testing prediction 2 and 4, because in that collection corrections were not reliably noted down.
- The oldest collection is basically the same as the one described by Nooteboom (1969). The errors were collected and noted down in orthography during several years of collecting by two people, the late Anthony Cohen and myself. Unfortunately, corrections were not systematically noted down. Collection of errors continued some time after 1969, and in its present form the collection contains some 1000 speech errors of various types, phonological syntagmatic errors out-numbering other types, such as lexical syntagmatic errors, blends, and intrusion errors. The collection was never put into a digital data base and is only available in typed form, each error on a separate card. Selection of particular types of errors for the present purpose was done by hand.
- The second collection stems from efforts of staff members of the Phonetics Department of Utrecht University, who, on the initiative of Anthony Cohen, from 1977 to 1982 orthographically noted down all speech errors heard in their environment, with their corrections, if any (cf. Schelvis, 1985). The collection contains some 2,500 errors of various types, of which more than 1,100 are phonological syntagmatic errors and some 185 lexical syntagmatic errors. The collection was put into a digital data base, currently accessible with Microsoft Access.

2.2. Assessing lexical bias

Lexical bias here is taken to mean that, in case of a phonological speech error, the probability that the error leads to a real word is greater, and the probability that the error leads to a nonsense word is less than chance. The problem here, of course, is to determine chance. Garrett (1976) attempted to solve this problem by sampling word pairs from published interviews and exchanging their initial sounds. He found that 33% percent of these "pseudo-errors" created words. This was not conspicuously different from real-word phonological speech errors, so he concluded that there was no lexical bias in spontaneous speech errors. One may note, however, that Garrett did not distinguish between monosyllables and polysyllables. Obviously, exchanging a phoneme in a polysyllabic word hardly ever creates a real word. This may have obscured an effect of lexical bias. Dell and Reich (1981) used a more elaborate technique to estimate chance level, involving "random" pairing of words from the error corpus in two lists of word forms, exchanging of the paired words' initial sounds, and determining how often words are thereby created, normalizing for the frequency of each initial phoneme in each list. They found a significant lexical bias in anticipations, perseverations and transpositions. In the latter, involving two errors (Yew Nork for New York) lexical bias was stronger in the first (Yew) than in the second (Nork) error.
In the current study I followed a different approach, restricting myself to single-phoneme substitutions in monosyllables, i.e. errors where a single phoneme in a monosyllable is replaced with another single phoneme, in this way optimally capitalizing on the fact that replacing a phoneme much more often creates a real word in a monosyllable than in a polysyllable. I did not, however, as Garrett (1976) and Dell and Reich (1981) did, restrict myself to initial phonemes, but took all single-phoneme substitutions in monosyllables into account. The two collections of Dutch speech errors together gave 311 such errors, 218 of which were real-word errors and 93 non-word errors. Although these numbers suggest a lexical bias, this may be an illusion, because it is unknown what chance would have given. It is reasonable to assume that a major factor in determining the lexical status of a phoneme substitution error, is provided by the phonotactic alternatives. If, for example, the *p* of *pin*, is replaced by a *b*, the phonotactically possible errors are bin, chin, din, fin, gin, kin, lin, sin, shin, tin, thin (with voiceless th), win, yin, *guin, *hin, *min, *nin, *rin, *zin, *zhin, *thin (with voiced th). In this case there are 21 phonotactic alternatives, of which 13 are real words and 8 are nonsense words.
Of course, if all phonotactic alternatives are real words (which sometimes happens), the probability that the error produces a real word is 1; and if all alternatives are nonsense words (which also happens) the probability of a real word error is zero. In the case of *pin* turning into *bin*, the chance level for a real-word error would have been 13/21=0.62. I have taken these proportions of real-word and non-word phonotactic alternatives for all individual errors included in the analysis, averaged these proportions, and I have taken the average as null hypothesis for the proportions of real-word errors and non-word errors in the subset of the collection at hand. The proportions of real- and non-word phonotactic alternatives seem to offer a workable null hypothesis for assessing any other effects.
Of course, with overt self-corrections there are fewer methodological problems in assessing lexical bias. If, among other criteria, a lexicality test is applied by self-monitoring, we may expect the correction frequency to be higher for non-word errors than for real-word errors.

3. Testing four predictions

3.1. Lexical bias in spontaneous speech errors

I have assessed the average proportions of real-word phonotactic alternatives for all 311 single-phoneme substitutions in monosyllables (not only initial phonemes), taking only into account the phonotactically possible single phonemes in that position. The average proportions of real-word and non-word alternatives are both 0.5. The expected numbers of real-word and non-word speech errors therefore are both 311/2=155.5, whereas the actual numbers are 218 and 93. There is a strong interaction between error categories and expected values based on average proportions of phonotactic real-word and non-word alternatives (chi^2=50; df=2; $p<0.0001$). Evidently there is a strong lexical bias in spontaneous speech errors, as predicted.

3.2. Lexical bias in self-corrections of overt speech errors

As we have seen, spontaneous speech errors show a strong lexical bias. If self-monitoring were responsible for lexical bias, by applying a lexicality test, then one would expect the same lexicality test to affect overt self-monitoring, as has been suggested by Levelt et al. (1999). This should lead to non-word errors being more often detected and corrected than real word errors. Indeed, if Levelt were correct in his suggestion that monitoring one's own speech for errors is very much like monitoring someone else's speech for errors, listening for deviant sound form, deviant syntax, and deviant meaning, real-word errors cannot be detected in self-monitoring on the level

of phonology. By definition, real-word errors would pass any lexicality test, and therefore could only be detected as if they were lexical errors causing deviant syntax or deviant meaning. Elsewhere (Nooteboom, submitted) I have shown that phonological real-word errors are treated by the monitor as phonological errors, not as lexical ones. The distributions of numbers of words speakers move on before stopping for correction differ significantly between phonological errors and lexical errors, but are the same for phonological real-word and phonological non-word errors. The same is true for the distributions of numbers of words included in the correction. These findings confirm evidence from Shattuck-Hufnagel and Cutler (1999), who demonstrated that lexical errors tend to be corrected with a pitch accent on the corrected word, whereas both phonological real-word errors and phonological non-word errors do not.

Clearly, phonological real-word errors are detected on the level of phonological, not of lexical processing. If, among other criteria, a lexicality test is applied by self-monitoring for phonological errors, we may expect the correction frequency to be higher for non-word errors than for real-word errors. Table 1 gives the relevant breakdown for the 315 single-phoneme substitutions, and Table 2 gives the relevant breakdown of all 1,111 phonological speech errors in the collection.

Table 1. Numbers of corrected and uncorrected single-phoneme substitutions, separately for real-word errors and non-word errors. (chi^2=1.95; df=2; p>0.3).

	Real words	Non-words
Corrected	99	69
Uncorrected	98	49

Obviously, there is no evidence of non-word errors being more frequently corrected than real-word errors. If there is any tendency in Table 1, it goes the wrong way. The data in Table 2 show that, if we consider all phonological errors instead of single-phoneme substitutions only, the probabilities for correction of real-word and non-word errors are exactly equal. It thus seems very unlikely that a lexicality test is applied in self-monitoring for overt speech errors during spontaneous speech production.

Table 2. Numbers of corrected and uncorrected phonological errors, separately for real-word errors and non-word errors (chi^2=0.117; df=2; p>0.5)

	Real words	Non-words
Corrected	218	341
Uncorrected	210	342

3.3. Lexical bias and phonetic dissimilarity

If lexical bias results from editing out of non-words by self-monitoring, one would expect errors differing from the correct form in only a single distinctive feature be missed more often than errors differing in more features. The reason is that self-monitoring is supposed to depend on self-perception (Levelt et al., 1999), and it is reasonable to expect that in perception smaller differences are more likely to go unnoticed than larger differences. As lexical bias is supposed to be the effect of suppressing non-words, one expects lexical bias to increase with dissimilarity between the two phonemes involved. To test this prediction I divided the 311 single-phoneme substitution errors into three classes, viz. errors involving 1 feature, errors involving 2 features, and errors involving 3 or more features. For consonants I used as features manner of articulation, place of articulation, and voice. For vowels features were degree of openness, degree of frontness, length, roundedness, and monophthong versus diphthong. Table 3 gives the numbers of real-word and non-word errors for the three classes.

Table 3. Numbers of real words and non-word errors, separately for errors involving 1, 2, or 3 or more features (chi^2 = 11.31; df=4; p<0.05).

	1 Feat.	2 Feat.	3 Feat.
Real words	95	96	27
Non-words	59	29	5

These results clearly suggest that lexical bias is sensitive to phonetic similarity, as predicted not only from a perception-based theory of pre-articulatory editing, but also from "phoneme-to-word" feedback (Dell & Reich, 1980; Stemberger 1985; Dell 1986).

3.4. Self-corrections and phonetic similarity

If self-corrections are sensitive to phonetic similarity, as lexical bias is, this would favour the hypothesis that both effects stem from the same mechanism. If they are not, this would suggest different mechanisms. Table 4 gives the relevant data.

Table 4. Numbers of corrected and not corrected single-phoneme substitutions, separately for errors involving 1 feature, 2 features of 3 features (chi^2=3.995; df=4; p>0.05; n.s.)

	1 Feat	2 Feats	3 Feats
Corrected	94	85	15
Not corrected	60	65	19

Obviously, there is little evidence that self-corrections are sensitive to phonetic similarity, although one would predict such an effect from perception-based monitoring.

4. A collector's bias?

Perhaps the current data suffer from a collector's bias, invalidating the otherwise plausible conclusions (Cf. Cutler, 1982). Of course, here the two possible sources of such a bias are phonetic similarity and lexical status. It seems unlikely, however, that such biases hold equally for corrected and uncorrected speech errors. The reason is that correction presents a very clear clue to the collector, easily overriding any more subtle differences due to phonetic similarity or lexical status. Thus, if there is a collector's bias due to phonetic similarity or to lexical bias, there should be an interaction between corrected versus uncorrected and lexical status combined with phonetic similarity. The data in table 5 strongly suggest that there is no such interaction. This makes it implausible that the absence of effects of lexical status and phonetic similarity in correction frequencies is due to a collector's bias.

Table 5. Numbers of Corrected and Uncorrected Single-phoneme Substitutions, Separately for Errors Involving 1, 2 or More Features, and for Real-word Errors and Non-word Errors (chi^2=3.18; df=6; p>0.7).

	1 Feat.; Real word	1 Feat.; Non-word	2/3 Feat.; Real word	2/3 Feat.; Non-word
Corr.	52	41	47	28
Not corr.	52	26	53	23

5. Discussion

What have we found? Phonological speech errors show lexical bias in the sense that in real words there are more and in non-words there are fewer such errors than expected on a chance basis. The size of the lexical bias in phonological speech errors decreases with phonetic similarity. Both effects seem compatible with the perceptual-loop theory of self-monitoring, suggested by Levelt et al. (1999). But contrary to expectation, the correction frequency of phonological speech errors is not influenced by either lexical status, or phonetic similarity. Note that these are not the only differences between lexical bias in speech errors and self-correction of speech errors. There are at least two other differences. One is a difference in speed, the other a difference in degree of consciousness.

The difference in speed is obvious: Lexical bias must be due to a mechanism operating before the error is made overt. Overt detection and correction of a speech error often, although not always, happens after the error has become overt. Another, possibly related, difference is in degree of consciousness. Speakers are often, although not always, conscious of having made a speech error, and then in many cases stop for correction. Note that a speech error that becomes sufficiently conscious to make the speaker stop for correction, has not necessarily become overt. As pointed out by Levelt (1989), stopping after an error sometimes occurs after only the first phoneme of the mispronounced word has been produced, suggesting that the stopping must have been initiated before the error had become overt. These cases can be explained by the less time-consuming inner-loop monitoring, i.e. by monitoring of inner speech via the speech comprehension system. Detecting self-produced speech errors in one's inner speech often reaches consciousness. This contrasts with the process leading to lexical bias in phonological speech errors. If lexical bias results from editing out speech errors leading to non-words, this editing process seems to be entirely subconscious. The apparent differences between lexical bias and overt self-correction strongly suggest that they are not both effects of perception-based self-monitoring. Lexical bias must have another origin. One candidate is the mechanism of "phoneme-to-word" feedback (Dell, 1986). Another is a process of production-based monitoring (Postma, 2000).

A final question is why perception-based overt detection of speech errors is not sensitive to lexical status and phonetic similarity, as one would expect from a perception-based mechanism. This is unclear. Whether speech perception really is sensitive enough to lexical status and phonetic similarity is still to be verified in perception experiments. If it turns out that perception of speech errors made by others is sufficiently sensitive to lexical status and phonetic similarity, the absence of these effects in detecting self-produced errors suggests that the perception-based self-monitoring system has immediate access to the intended form. Binary comparison of intended and produced form is an easy task, which is likely not to be sensitive to lexical status and phonetic similarity.

6. Conclusion

Lexical bias in phonological speech errors and overt detection of self-produced phonological speech errors are not products of the same mechanism.

7. References

[1] Baars, B.J. & Motley, M.T., "Spoonerisms: Experimental elicitation of human speech errors: methods, implications, and work in progress", *Journal Supplement Abstract Service, Catalog of selected documents in psychology*, (1974, fall).

[2] Baars, B.J., Motley, M.T., & McKay D., "Output editing for lexical status from artificially elicited slips of the tongue", *Journal of verbal learning and verbal behavior, 14*, 382-391, 1975.

[3] Cutler, A., "The reliability of speech error data", In: A. Cutler (ed.) *Slips of the tongue and language production*, Mouton, Amsterdam, 1982

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

[5] Dell, G.S. and Reich, P.A., "Toward a unified model of slips of the tongue", In: V.A. Fromkin (ed.) *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* New York: Academic Press, 1980, pp. 273-286.

[6] Garrett, M. F., Syntactic processes in sentence production. In: R. J. Wales and E. Walker (eds.), *New approaches to language mechanisms* ,Amsterdam: North-Holland Publishing Company, 1976, pp. 231-256.

[7] Levelt, W.J.M. *Speaking. From intention to articulation*, MIT Press, 1983.

[8] Levelt, W.J.M., Roelofs, A. and Meyer, A.S., "A theory of lexical access in speech production", *Behavioral and Brain Sciences*, 22, 1-75, 1999.

[9] Postma, A. "Detection of errors during speech production: a review of speech monitoring models", *Cognition*, 77, 97-131, 2000.

[10] Schelvis, M., "The collection, categorisation, storage and retrieval of spontaneous speech error material at the Institute of Phonetics", *PRIPU* No 10, Utrecht University, 1985, pp. 3-14.

[11] Shattuck-Hufnagel, S. and Cutler, A. (1999). "The prosody of speech error corrections revisited", [CD-ROM]. *Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, August 1-6*, San Francisco, Regents of the University of California, 1999, pp. 1483-1486.

Are Word Repetitions Really Intended by the Speaker?

Yasuharu Den

Faculty of Letters
Chiba University
den@L.chiba-u.ac.jp

Abstract

This paper compares, using our Japanese data, word repetitions with error repairs in terms of their temporal structures in order to examine whether or not the prolongation of first tokens in word repetitions, observed by Den and Clark (2000), is really an effect of the speaker's strategy. Analyses of 10 task-oriented Japanese dialogues reveal a difference between word repetitions and error repairs for the data involving cut-off in first tokens; in both types of disfluencies, the final phoneme of the first token is considerably prolonged, but the degree of the prolongation is much greater in word repetitions than in error repairs. These results support our view that prolonged first tokens in word repetitions are a product of a process under the speaker's control or intention.

1. Introduction

Spontaneous speech contains various disfluencies such as fillers, self-repairs, and repeated words. These disfluencies seem to reflect problems in speech production. When speakers cannot formulate an entire utterance at once, or when they change their minds about what to say, they may suspend their speech and produce fillers or replace words they have already produced.

There are two views of speech disfluencies. One is that they are merely accidents which are beyond speakers' control. For example, speakers may suspend a word for some unexpected reason and restart it from the beginning. The other view is that they are the results of certain strategies under speakers' control (Levelt, 1989; Clark & Wasow, 1998). Speakers may produce a filler to signal to their addressees that they are having trouble in speech production, or to inform their addressees of the kind of trouble they have. Clark and Wasow (1998), taking this 'strategic' view, proposed the *commit-and-restore* model of repeated words in English, insisting that speakers can make a *preliminary commitment* at the beginning of a major constituent, even when they are aware of its not having been well-formulated, and restart the constituent, after suspension, to restore continuity to its delivery.

Den and Clark (2000) showed that the theory of preliminary commitment also applies to Japanese, that has completely different syntax from English. They found that pauses immediately following repetitions are less frequent than those immediately preceding repetitions and those between repeated tokens. They also found that the first tokens of repetitions are prolonged whereas the second tokens are produced in the same speed as the fluent speech. These findings suggest that Japanese speakers, as well as English speakers, sometimes use word repetitions as a linguistic device to communicate to addressees their cognitive states and that they produce these tokens with the intention of making the addressees recognize as such.

These pieces of evidence, however, are still insufficient. It might be the case that the prolongation of first tokens is merely a general characteristic of the disrupted speech, having nothing to do with the speaker's strategy. In order for the prolongation to be evidence of preliminary commitment, it should be observed only when the speaker's strategy can be relevant. That is, the prolongation of tokens should not be observed when the disrupted speech cannot be viewed as a marker for intended communication.

In this study, we compare, using our Japanese data, word repetitions with error repairs in terms of their temporal structures. In error repairs, first tokens are erroneous items, which are produced by accident and cannot be viewed as a marker for intended communication. Thus, if the prolongation of first tokens is observed in word repetitions but not in error repairs, that would form strong evidence for the theory of preliminary commitment.

2. Word Repetitions in Japanese

Word repetitions in Japanese have different characteristics from word repetitions in English. First, most word repetitions in Japanese are repetitions of content words (Den, Ishizaki, & Haruki, 1997), as opposed to function words, which are frequently repeated in English. In Japanese, repetitions of function words alone are very rare. Second, as a consequence of the first characteristic, Japanese has no typical lexical items for repetitions, like *the* or *a* in English. Third, in Japanese, first tokens in repetitions are frequently cut off in the middle, which is less frequent in English. In our Japanese data, repetitions involving word cut-off amount to over 60% of the data. This is mainly due to long durations of content words, which are frequently repeated in Japanese.

A typical example of word repetitions in Japanese is the following one:

(1) ano Ya= Yamaguti-to Hirosima ari-masu-ka
uh Ya= Yamaguchi-and Hiroshima be-POLITE-Q
uh, are there Ya= Yamaguchi and Hiroshima?

In (1), the speaker suspended the word *Yamaguti* (a place name) after the initial mora *Ya*, and then restarted it from the beginning, resulting in a word repetition *Ya= Yamaguti*.[1] First tokens in repetitions are frequently cut off in the middle, as in (1), but in a few cases, speakers produce an entire word before repeating it, like *Yamaguti Yamaguti*. Both cases will be considered in this study.

This research has been partly supported by CREST of JST (Japan Science and Technology Corporation).

[1] The symbol '=' is used to indicate a word cut-off.

Den and Clark (2000) tested with Japanese data two of the three hypotheses of the commit-and-restore model, which had been proposed by Clark and Wasow (1998) to account for word repetitions in English. The two hypotheses were

The continuity hypothesis: All other things being equal, speakers prefer to produce constituents with a continuous delivery.

The commitment hypothesis: Some initial commitments to constituents are preliminary, with speakers already expecting, at some level of processing, to suspend speaking immediately afterward.

Den and Clark's findings that pauses are less frequent after restarts than before or during suspension and that the first tokens of repetitions are prolonged were consistent with the continuity hypothesis and the commitment hypothesis, respectively. The second piece of evidence, however, is not sufficient for supporting the commitment hypothesis, as mentioned in the previous section. Re-examination of the hypotheses is the topic of the present study.

3. Method

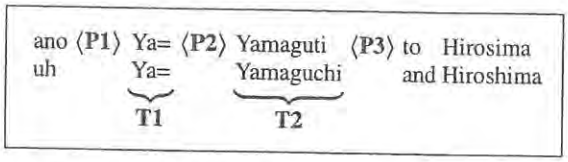
Data. The corpus used in the present study was the same as that used in Den and Clark's (2000) study.

The corpus consists of 30 task-oriented dialogues by 60 different native speakers of Japanese (total running time = apprx. 7.5 hours, number of words = 73,258). The task was collaborative route finding, in which two participants looking at slightly different railroad maps, each unseen by the other, were asked to find a connecting path from a given start to a given goal. The two maps were different in that some connections on one map were missing on the other and in that some station names were missing on either map. The participants had to find a path which was available on both maps. This setting induced the participants to exchange spontaneously and naturally utterances about connections and station names. All dialogues were digitized on a computer and transcribed with part-of-speech and disfluency annotations.

For the present study, the first 10 of the 30 dialogues were selected. These included 11 male and 9 female speakers.

Since the purpose of the analysis is comparison of word repetitions with error repairs, these two types of disfluencies were taken into account. The data for word repetitions were the same as those used in Den and Clark's (2000) study. The data for error repairs were carefully chosen from among those annotated as 'substitution repairs.' Substitution repairs are repairs in which the first tokens are replaced by the second tokens that are closely related, in syntax and/or semantics, to the first tokens (Shriberg, 1994). Substitution occurs on at least two occasions; (i) when the first token is wrong in the context and the second token corrects it, and (ii) when the second token is more appropriate in the context than the first token, which is not necessarily wrong. The first type is called *error repairs*, and the second type *appropriateness repairs* (Levelt, 1983).

For the present study, only error repairs were considered. Speakers may use a less appropriate, but not incorrect, word, as preliminary commitment, to signal their addressees that they have trouble in speech production, and replace it afterward with a more appropriate word, resulting in an appropriateness repair. Since the purpose of comparing word repetitions with other type of disfluency is to see the difference between cases where the speaker's strategy can be relevant and cases where



- Presence of a pause at Pi, longer than 60 msec.
- Duration of Ti, normalized by the mean of the durations of its fluent counterparts by the same speaker.

Figure 1: Measurement used in the analysis.

the speaker's strategy cannot be relevant, appropriateness repairs, which might be similar to word repetitions in terms of the speaker's strategy, are not suitable for the target of the comparison. Error repairs, on the other hand, are suitable for the target, for first tokens in error repairs are erroneous items and seem to irrelevant to the speaker's strategy.

The following is a typical example of error repairs in Japanese:

(2) a Na= Morioka-mo tunagatte-masu
uh Na= Morioka-as well is reachable-POLITE
uh, Na= Morioka is reachable as well.

In (2), the speaker suspended the word *Naha* (a place name) after the initial mora *Na*, and then replaced it with the word *Morioka* (a place name), resulting in an error repair *Na= Morioka*. Like word repetitions, error repairs also involve frequent cut-off in first tokens. The distinction between cases with word cut-off and cases without word cut-off will be taken into account in the analyses below.

Measurement. In the same way as Den and Clark's (2000) study, we counted the number of pauses around word repetitions and error repairs, and the durations of repeated/repaired tokens. The precise measurement is illustrated in Figure 1.

First, we counted pauses just before first tokens (P1), between first and second tokens (P2), and just after second tokens (P3). At P1 and P3, only silent pauses were considered, whereas at P2, filled pauses and editing expressions were also regarded as pauses with the corresponding durations. Silences, or fillers, of 60 msec or longer were counted as pauses.

Second, we measured the normalized durations of first tokens (T1s) and second tokens (T2s). The normalization was needed to compensate for the differences of speaking rate across speakers and of inherent phoneme length across items. For each speaker and for each item, we collected from the corpus the fluent versions of the same word by the same speaker, where a 'fluent' version means an occurrence of the word not involved in a disfluency of any type. Then, we calculated the normalized duration as the duration of a target token divided by the mean of the durations of its fluent counterparts. When a target token was cut off in the middle, we used the duration of the corresponding region in its fluent counterpart for normalization. Note that by this normalization, the value of 1.0 indicates that the token was produced at the same speaking rate as its fluent counterpart. We also calculated the normalized durations for the final phonemes of T1s and T2s by identifying these phonemes based on spectral cues.

Exceptions. The following cases were excluded from the analysis (cf. Den & Clark, 2000):

1. Repetitions/Repairs involving verbs and/or auxiliary verbs. Repetitions of (auxiliary) verbs were excluded for

they are likely to be for emphasis rather than disfluencies. Repairs of (auxiliary) verbs were excluded as well to limit the target tokens to noun phrases.

2. Repetitions/Repairs overlapping/overlapped with the other's speech. These were excluded for it was difficult to perform precise phoneme alignment for overlapping/overlapped speech (the speech signals of the two participants in a dialogue were not recorded on separate channels in our corpus).

3. Repetitions/Repairs having no fluent counterparts, or repairs in which the intended words for first tokens that are cut off in the middle are not identifiable. These were excluded simply because the normalization procedure could not be applied.

Predictions. The predictions about the difference between word repetitions and error repairs in terms of the continuity and the commitment hypotheses are as follows:

The continuity hypothesis: In both types of disfluencies, pauses at P3 are less frequent than those at P1 and P2.

The commitment hypothesis: In word repetitions, the normalized duration of T1 is longer than that of T2, whereas in error repairs, the normalized duration of T1 is as long as that of T2. Or, even if the normalized duration of T1 is longer than that of T2 in error repairs, the degree of the prolongation is smaller than that in word repetitions.

According to the continuity hypothesis, speakers prefer to produce constituents with a continuous delivery. They don't like to add a delay before every word when they have trouble in formulating an entire phrase or utterance. Rather, they are likely to suspend speaking at some point in a constituent and restore continuity to their delivery of the constituent after they have formulated it well enough. This does not depend on the source of the suspension. Whether the suspension is intended by the speaker or it is by accident, the speech after restart would be produced with a continuous delivery. Thus, we can expect, in both word repetitions and error repairs, pauses after restarts to be less frequent than pauses before or during suspensions.

According to the commitment hypotheses, on the other hand, speakers can make a preliminary commitment at the beginning of a major constituent, even when they are aware of its not having been well-formulated. They do so, being pressed by a temporal imperative; by initiating a constituent, even if prematurely, to inform their addressees that they are engaged in planning the constituent, they can escape from being heard, due to a long delay, as opting out, as confused, or as having nothing immediately to contribute. Preliminary commitments, being a signal to addressees, are marked as such by some linguistic means, e.g., prolongation (Shriberg, 1999). This is a 'strategic' use by the speaker of disfluency, and that is the case with (some) word repetitions.

The situation, however, is completely different in error repairs. In error repairs, first tokens are produced by accident, having nothing to do with the speaker's strategy. Thus, there is no reason to expect first tokens to be linguistically marked as a signal to addressees. The prediction on the normalized duration of first and second tokens has two possibilities. If the prolongation of first tokens in word repetitions, observed by Den and Clark (2000), is purely an effect of the speaker's strategy, it will not be observed in error repairs. Or, if the prolongation of first tokens observed in word repetitions is a combined effect of the speaker's strategy and of the phonological disturbance due to the disrupted speech, the prolongation will also be observed in error repairs, due to disruption, but the degree of the prolongation will be smaller than that in word repetitions.

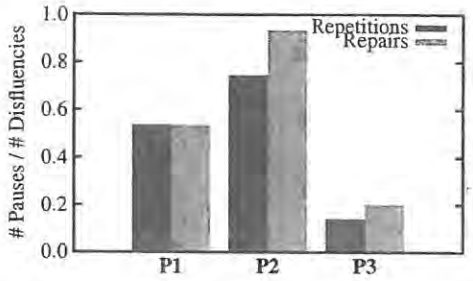


Figure 2: Frequencies of pauses at P1, P2, and P3. ($N = 43$ for repetitions and $N = 15$ for repairs)

4. Results

Pauses at P1, P2, and P3. The frequencies of pauses at P1, P2, and P3 are shown in Figure 2. For those cases where repetitions/repairs occur at the inside of an utterance, i.e., not at the beginning or the end of an utterance, a Cochran's Q test was applied separately for the repetition and the repair data.

For word repetitions, the numbers of pauses at the three locations were significantly different ($Q = 30.76$, $p < .001$). Multiple comparisons using the Ryan's procedure showed that pauses were significantly less frequent at P3 than both at P1 and at P2 ($ps < .001$). No significant difference was found between the numbers of pauses at P1 and at P2.

For error repairs, the numbers of pauses at the three locations were significantly different as well ($Q = 15.17$, $p < .001$). Multiple comparisons showed that pauses at P3 were significantly less frequent than pauses at P2 ($p < .01$), and that pauses at P1 were significantly less frequent than pauses at P2 ($p < .05$). No significant difference was found between the numbers of pauses at P1 and at P3.

Although the results for word repetitions and error repairs were slightly different, i.e., pauses were less frequent at P3 than both at P1 and at P2 in repetitions but than only at P2 in repairs, this would be due to the small data size for error repairs. The distributions of pauses for repetitions and repairs shown in Figure 2 are quite similar.

In any case, pauses after restarts are less frequent than those (before or) during suspensions, supporting the prediction from the continuity hypothesis.

Normalized Durations of T1 and T2. The mean normalized durations of T1s and T2s for word repetitions and error repairs are shown in Figure 3.

A two-factors ANOVA was applied to the data. The main effect of the token position (T1 vs. T2) was significant ($F(1, 75) = 19.36, p < .001$), but the main effect of the disfluency type (Repetitions vs. Repairs) nor the interaction between the two factors were not significant ($Fs < 1$). In both types, T2 had approximately the same duration as its fluent version (mean norm. dur. = 1.04 for repetitions and 1.02 for repairs), and T1 was considerable prolonged (mean norm. dur. = 1.60 for repetitions and 1.40 for repairs). The results for the final phonemes of T1s and T2s were similar.

These results seem to contradict with our prediction on the difference between word repetitions and error repairs in terms of the commitment hypothesis. The prolongation of first tokens

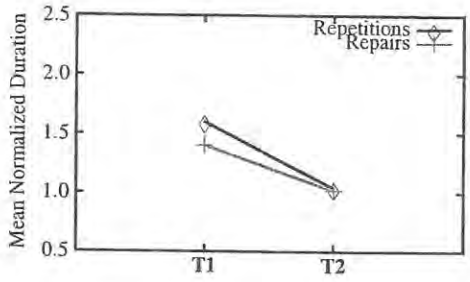


Figure 3: Mean normalized durations of T1s and T2s. ($N = 60$ for repetitions and $N = 17$ for repairs)

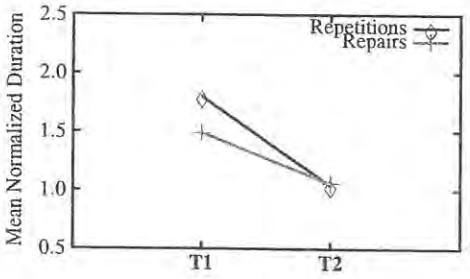


Figure 4: Mean normalized durations of T1s and T2s for the data involving cut-off. ($N = 38$ for repetitions and $N = 13$ for repairs)

was not peculiar to word repetitions, nor no additional prolongation was observed in word repetitions. However, when we focused only on the data involving cut-off in first tokens, the story changed dramatically.

Figure 4 shows the mean normalized durations of T1s and T2s for word repetitions and error repairs involving word cut-off, and Figure 5 shows the same data for the final phonemes of T1s and T2s. Although an ANOVA on the data shown in Figure 4 revealed only a significant main effect of the token position (token position: $F(1, 49) = 17.35$, $p < .001$; disfluency type: $F(1, 49) = 1.02$, $p = .32$; interaction: $F(1, 49) = 1.33$, $p = .25$), an analysis on the data shown in Figure 5 revealed a significant interaction between the token position and the disfluency type (token position: $F(1, 48) = 28.74$, $p < .001$; disfluency type: $F(1, 48) = 1.15$, $p = .29$; interaction: $F(1, 48) = 5.36$, $p < .05$). The final phoneme of T2 had approximately the same duration as its fluent version in both word repetitions (mean norm. dur. = 1.03) and error repairs (mean norm. dur. = 1.15), whereas the final phoneme of T1 was considerably prolonged and the degree of the prolongation was much greater in word repetitions (mean norm. dur. = 2.08) than in error repairs (mean norm. dur. = 1.56).

5. Discussion

In this paper, we have compared word repetitions with error repairs in terms of their temporal structures in order to examine whether or not the prolongation of first tokens in word repetitions, observed by Den and Clark (2000), is really an effect of the speaker's strategy. We have found that when we focus on the cases where first tokens are cut off in the middle, word repetitions and error repairs have different characteristics. Although in both types of disfluencies, the final phonemes of

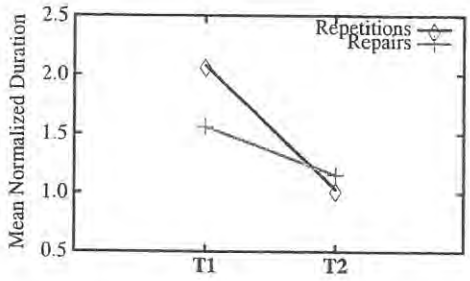


Figure 5: Mean normalized durations of the final phonemes of T1s and T2s for the data involving word cut-off. ($N = 38$ for repetitions and $N = 12$ for repairs)

the first tokens are prolonged, the degree of the prolongation is much greater in word repetitions than in error repairs. This means that, even if the prolongation can be partly attributed to the phonological disturbance due to the disrupted speech, something more is happening in word repetitions. Speakers sometimes use word repetitions as a linguistic device to signal to their addressees that they are having trouble in speech production, and produce these tokens with the intention of making the addressees recognize them as a signal.

The results of the present study seem to support our view that prolonged first tokens in word repetitions are a product of a process under the speaker's control or intention. Speakers intendedly use them as preliminary commitments in order to announce their engagement in the production of the following material, making an otherwise intolerable delay in the delivery of constituents permissible by addressees.

There are, however, still several points to be accounted for. The difference between word repetitions and error repairs was observed only in the cases involving word cut-off and only at the final phonemes of first tokens. Why the difference was not observed in other cases might be due partly to the small data size for error repairs used in the study. However, we need more detailed examination of the data and comparison with other types of disfluencies as well as disfluencies in other languages.

6. References

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37*, 201–242.

Den, Y., & Clark, H. H. (2000). Word repetitions in Japanese spontaneous speech. In *Proceedings of the 6th International Conference on Spoken Language Processing* (pp. 58–61). Beijing.

Den, Y., Ishizaki, M., & Haruki, Y. (1997). A corpus-based analysis of speech repairs in Japanese. Oral presentation at *Computational Psycholinguistics 1997*. Berkeley, CA.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.

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

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Unpublished doctoral dissertation, University of California, Berkeley.

Shriberg, E. E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the 14th International Congress on Phonetic Science* (pp. 619–622). San Francisco.

Gesture as an Indicator of Early Error Detection in Self-Monitoring of Speech

Mandana Seyfeddinipur & Sotaro Kita

Max Planck Institute for Psycholinguistics,
Nijmegen, The Netherlands
mandsey@mpi.nl

Abstract

There is a theoretical controversy regarding when the self-monitoring process interrupts the speech stream. One view holds that the speech stream is interrupted as soon as an error is detected. Another view holds that, even after an error is detected, the speaker does not interrupt immediately but continues speaking and at the same time plans the upcoming repair. We address this question by observing speech-accompanying gestures at the moment of speech disfluency. The results show that the concurrent gestural movements are typically stopped on average 240 ms before speech is stopped. In other words, the gesture suspension foreshadows the speech suspension. The gestural foreshadowing shows that the speaker must know early on that he is going to suspend speech. The gestural indication of an upcoming speech suspension suggests that the speaker does not interrupt speech at the very moment s/he detects an error. This result supports the hypothesis on speech monitoring stating that the speaker continues to talk after error detection and at the same time plans the upcoming repair.

1. Introduction

Gesture and speech are semantically and temporally tightly co-ordinated. For example, gestures are prepositioned temporally to the lexical affiliate, with which they share semantic and/or pragmatic content [1, 2, 3, 4]. The specific timing and semantic relation of speech and gesture has led to the view that gesture can serve as a window into mental processes underlying speech production [2]. In this paper, we aim to gain insight into how speakers monitor their own speech by observing accompanying gestures.

The tight coordination between speech and gesture has led to the conclusion that (at least) at the conceptual level speech and gesture production are closely interrelated, for example [5, 6, 7, 8]. Furthermore, it has been argued that self-monitoring of speech is a conceptual level process (as opposed to a formulational level process) [9]. Thus, it can be expected that gesture is sensitive to speech disfluency. By utilising the specific timing relation of speech and gesture, we test two views regarding how speakers monitor their own speech.

We will investigate this issue on the basis of a corpus of disfluencies produced during on descriptions of houses and apartments. Living-space descriptions have proven to be a useful task in order to elicit various kinds of speech disfluencies and gestures. The speaker has to transform three-dimensional space into the linear structure of speech. In addition, the speaker has to choose the appropriate words and constructions in order to convey the selected and linearised spatial information in a comprehensible way [10]. These difficulties result in a high number of disfluencies of different kinds and in a considerable use of gesture

2. Monitoring Theories

2.1. Two hypotheses about speech interruption

Speakers monitor their own delivery constantly. They control for what is going to be said is what they had intended. More specifically, they control for the appropriateness of selected words, and they check for errors (for details of foci of monitoring, see [9, 11]. If inappropriateness is detected, the speaker interrupts his speech stream and repairs the erroneous or inappropriate utterance. This whole process consists of four components: monitoring of speech, error detection, self-interruption and self-correction. Various psycho-linguistic theoretical accounts of error detection and self-interruption have been proposed (for a review, see [12]). Two of these accounts will be tested in this paper.

The Interruption-Upon-Detection Hypothesis states that the speech stream is interrupted as soon as an error is detected. This is expressed in the Main Interruption Rule: Stop the flow of speech immediately upon detecting trouble [9, 11, 13]. After the interruption of speech, the planning for reformulation takes place.

The rationale behind the Main Interruption Rule is that linguistic structures are ignored in interruption. Levelt s [11] analysis showed that speaker interrupted their speech stream at any point in the delivery. They did not attend to any linguistic boundaries like syllables, words, or phrase boundaries. One exception was that speakers tended to complete non-erroneous words, i.e. neutral or merely inappropriate ones. This led to the refinement of the model that the Main Interruption Rule only applies to cases of immediate detection of erroneous words.

The Delayed-Interruption-For-Planning Hypothesis suggests that even if an error is detected, the speaker does not interrupt his flow of speech immediately [14, 15, 16]. Upon detection of an error, the speaker will start the replanning, and interrupt, when the repair is ready to a certain degree or the speaker has run out of what can be uttered without further conceptual processing.

Blackmer & Mitton [14] based their hypothesis on the analysis of the temporal characteristics of self-repairs in spontaneous speech. They observed time intervals between the interruption point and resumption of speech that were sometimes shorter than predicted by Levelt s [11] Main Interruption Rule. According to the Main Interruption Rule, the replanning takes place only after the interruption. This implies that there has to be a time interval of some length before the resumption can take place. However, Blackmer & Mitton [14] found instances

where the suspension point was immediately followed by the correction, without any pause in between. Their results imply that the planning of the correction can take place while speaking is in progress and not only after suspension. Fox Tree & Clark [15] came to a similar conclusion but with a rather different type of evidence. They conducted a corpus study on the occurrence of the two pronunciation variants of the English article *the* (*thuh* —with the reduced vowel schwa, and *thiy* with a non-reduced vowel). They found that 81 % of the instances of *thiy* were followed by a suspension of speech. This suggests that speakers detected the problem some interval before suspending speech. By knowing in advance that he is going to suspend, the location of suspension (after *the*), and the type of suspension (pronunciation of the variant *thiy*) is planned.

2.2. Predictions

Taking the temporal and semantic interlocking of gesture and speech into account, the two theoretical approaches make different predictions concerning the gestural behavior. The **Interruption-Upon-Detection Hypothesis** predicts that any effect on gesture should be simultaneous with or following the speech suspension. There should not be any effect on gesture before the actual speech suspension. This prediction is based on two assumptions:
1. When an error is detected, a stop-signal is sent to both production modalities simultaneously (for an account of the suspension of speech and gesture production, see [5])
2. It takes longer to suspend a gesture than speech because heavier mass has to be stopped in gesture.

According to the **Delayed-Interruption-for-Planning Hypothesis**, an effect on gesture can occur even before the moment of speech suspension due to the lag between error detection and speech suspension. When speakers have detected an error or have anticipated trouble, they start to plan how to resume right away and at the same time suspend the gestural movement. In the meantime they go on speaking until the repair is ready up to a certain point or they have run out of words that can be uttered without further conceptual processing. Consequently, gesture can stop before speech stops.

3. Phases in speech disfluency and gesture

The predictions are tested by investigating the temporal relationship between different phases of self-repair and movement phases of gesture, which will be defined in the following sections.

3.1. Disfluency structure

A speech disfluency can be divided into different phases following Clark s [17] disruption schema:

	Suspension Point	Resumption Point	
on the right	⇑	**uhm** ⇑	**on the left side..**
Original Delivery		Hiatus	Resumed Delivery

The first phase is the original delivery. The speaker monitors his internal speech for appropriateness and correctness [11]. If an error is detected, the original delivery is disrupted. In the above example the original delivery is suspended at the word *right*. After the interruption a time interval (the hiatus) follows where speakers pause or utter filled pauses (*uhm, uh*) or so-called editing terms like *well, I think, I mean*. The hiatus is seen as the phase where internal reformulation processes take place [11]. The hiatus ends at the resumption point where the speaker resumes his delivery.

3.2. Gesture structure

Gestures can be segmented into qualitatively different movement phases [1, 2, 18]. The segmentation and identification of movement phases can be based purely on dynamic aspects of the hand/arm movement [18]. In the preparatory movement phase the hands move from a resting position in order to prepare for the forcefully executed part, the stroke. The preparation phase can also be followed by a static phase, where the hands are held still in the initial position. This pre-stroke hold is then released by the stroke. The stroke phase is the semiotic and dynamic nucleus of the gesture. The stroke typically displays the meaning of the gesture. In the stroke, the most force is exerted, compared to the neighbouring phases. Also after this phase a static phase might follow, which is called the post-stroke hold. A gestural unit ends when the hands retract back into resting position, e.g. on the lap.

Preparation ⇒ Hold ⇒ Stroke ⇒ Hold ⇒ Retraction

Schema: Gesture phases in a gesture unit

Of the described gestural phases, only the stroke is obligatory. Note that in natural conversation one can observe a succession of strokes without the hands going into a hold or being retracted after each stroke.

4. Method

4.1. Data

The corpus consists of six videotaped semi-natural conversations. 6 native German speakers (4 female, 2 male) were asked to describe houses and apartments they grew up in or have had lived in for a longer period to a listener. Each session lasted 30 —40 minutes. Nine minutes of the description from each speaker was transcribed. The speech data was coded for suspension points, hiatus length, and resumption points. The gestural movement phases were coded in terms of phase transitions. The temporal values were determined by a frame-by-frame microanalysis (1 frame = 40 ms). The six speakers produced an overall of 582 disfluencies, of which 267 were overt repairs. 191 overt repairs were accompanied by gestures, and 76 were not.

4.2. Coding: Stop shifts and start shifts

In the analysis of the gestural movement pattern, we focused on the transition from one phase to another. Analogous to speech suspension and speech resumption, we distinguish two different types of phase shifts: a stop shift and a start shift. In a stop shift, an ongoing gestural unit / movement phase is suspended. In a start shift, a new dynamic gestural movement phase is initiated.

Stop shift: an ongoing gestural movement is suspended or not completed:

- Shift of a dynamic phase into a static phase: an ongoing gestural movement phase (preparation / stroke) is suspended by going into a hold or by being retracted back into resting position.
- Shift of a dynamic phase into a new dynamic phase: a gesture gets suspended before being completed, e.g. a preparation phase is not followed by a stroke, for which the hands were preparing, but is followed by another preparation for the same or a different gesture.
- A dynamic phase is interrupted: a preparation or a stroke phase is prematurely truncated before a sudden abrupt halt or a sudden change in movement direction terminates the phase itself. In this case we classified the phase shift as a stop shift no matter what followed.

Start shift: a new gestural movement is started:

- Shift from a static phase into a dynamic phase: hands that are held still start a new preparation/stroke phase.
- A preparation phase is not followed by a stroke, but by a new preparation phase.
- An interrupted movement phase is followed by a new movement phase (preparation / stroke).

4.3. Analysis

We selected all utterances containing a repair that was accompanied by gesture (191). One speaker was excluded from the analysis because she did not provide sufficient data points. We analysed the occurrences of stop shifts around suspension points and the occurrences of start shifts around resumption points. In order to ensure that the observations were independent from each other, we selected all repairs (the whole disfluency unit (suspension, hiatus, resumption) that were at least two seconds apart from each other. We chose a time window of one second to each side of the suspension / resumption points and counted the number of start and stop shifts for every 160 ms slot within the window.

5. Results

5.1. Stop shifts around suspension point

Figure 1 presents the frequency of stop shifts around the speech suspension point (averaged over five speakers). The one-second window before and after the speech suspension point is divided into 160 ms intervals (0 = suspension point). Each bar shows the average frequency of stop shifts for a given time interval.

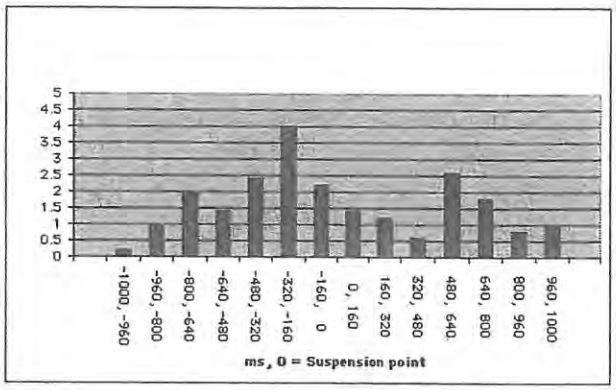


Figure 1 Average frequency of stop shifts around suspension points

As is evident from Figure 1, gesture stops before speech stops. The most common time interval in which stop shifts occur is from -320 to -160 ms (There is a secondary peak at around 400 ms to 800 ms after the suspension point The majority of these cases involve stopping of a new gesture that was generated after the suspension point. Thus, stop shifts in this peak are not directly related to the speech suspension).

5.2. Start shifts around resumption points

Figure 2 presents the frequency of start shifts around the speech resumption point (averaged over five speakers). The one-second window before and after the speech resumption point is divided into 160 ms intervals (0 = resumption point). Each bar shows the average frequency of start shifts for a given time interval.

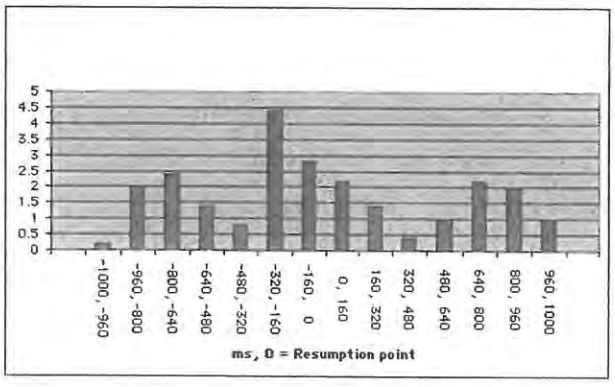


Figure. 2. Average frequency of start shifts around resumption points

As is evident from Figure 2, gesture starts before speech starts. The most common time interval in which start shifts occur is from -320 to -160 ms (There are two secondary peaks —960 to —640 and from 640 to 960. The majority of gestures that are generated around these peaks are additional to the ones around the resumption point. Thus, start shifts in these peaks are not directly related to the resumption of speech).

6. Discussion

The above results show that gesture is highly sensitive to speech disfluencies. When speech is suspended and then

resumed, gesture is also suspended and then resumed. Suspension and resumption in the two modalities are temporally coordinated in a systematic way. This suggests a highly interactive planning process that is involved in the production of both modalities.

Gesture is suspended prior to the speech suspension. This suggests that gesture can be seen as an indicator of an upcoming interruption in speech. The gestural foreshadowing of speech suspension suggests that the speaker is already aware that there is / will be trouble but he does not interrupt speech right away. This is predicted by the Delayed-Interruption-for-Planning Hypothesis, according to which speakers continue speaking after error detection. They start planning for the resumption already before the speech suspension and disrupt their delivery when the repair is ready to a certain degree or they have run out of words that can be formulated without further conceptual planning. The above result also indicate that at least some utterances are interrupted in the way not predicted by the Interruption-Upon-Detection Hypothesis, according to which gesture should be interrupted simultaneously with or even after speech suspension.

However, these two hypotheses are not mutually exclusive. A speaker may interrupt his/her speech in different ways depending on various contextual factors. For example, in order to avoid losing the floor, one might delay suspension of speech. At the same time, in order not to mislead the interlocutor, one might suspend and repair the error as soon as possible. The speaker has to always evaluate advantages and disadvantages of speech suspension at a given moment. A moment-by-moment balance among competing factors like comprehensibility and floor keeping may determine the timing, at which a speaker interrupts his/her speech.

There is an emerging view in the literature that speech interruption is not a reflex-like reaction to error detection, but a choice the speaker makes based on, for example, above mentioned factors [14, 15 16]. This study provides novel converging evidence for this idea by using speech-accompanying gesture as a window into the speaker s mind.

7. References

[1] A. Kendon, Some relationships between body motion and speech, in *Studies in dyadic communication,* A. Siegman and B. Pope, Eds. New York: Pergamon Press, 1972, pp. 177-210.

[2] D. McNeill, *Hand and mind,* Chicago: University of Chicago Press, 1992.

[3] E. Schegloff, On some gestures in relation to speech, in *Structures of social action,* J.M. Atkinson and J. Heritage, Eds., Cambridge: Cambridge University Press, 1984, pp. 266-296.

[4] P. Morrel-Samuels and R. Krauss, Word-familiarity predicts temporal asynchrony of hand gestures and speech, *Jounal of Experimental Psychology: Learning Memory and Cognition*, vol. 18, pp. 615-622, 1992.

[5] J.P. de Ruiter, The production of gesture and speech, in *Language and gesture,* D. McNeill Ed. Cambridge, Cambridge University Press, 2000, pp. 284-311.

[6] M.W. Alibali, S. Kita, and A.J. Young, Gesture and the process of speech production: we think, therefore we gesture, *Language and Cognitive Processes*, vol. 15, pp. 593-613, 2000

[7] S. Kita, How representational gestures help speaking, in *Language and gesture,* D. McNeill Ed. Cambridge, Cambridge University Press, 2000, pp. 162-185.

[8] U. Hadar and B. Butterworth, Iconic gestures, imagery and word retrieval in speech, *Semiotica*, vol. 115, pp. 147-172, 1997.

[9] W.J.M. Levelt, *Speaking: From intention to articulation*, Cambridge Mass: MIT Press, 1989.

[10] V. Ullmer-Ehrich, The structure of living space descriptions, in *Speech, place and action*, R.J. Jarvella and W. Klein Eds. Chichester: John Wiley, 1982, pp. 219-249.

[11] W.J.M. Levelt, Monitoring and self-repair in speech , *Cognition,* vol. 14, 1983, pp. 41-104.

[12] A. Postma, Detection of errors during speech production: a review of speech monitoring models, *Cognition,* vol. 77, pp. 97-131, 2000.

[13] S. Nooteboom, Speaking and unspeaking. Detection and correction of phonological and lexical errors in spontaneous speech, in *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand,* V. Fromkin, Ed. New York: Academic Press, pp.87-95, 1980.

[14] E.R. Blackmer and J.L. Mitton, Theories of monitoring and the timing of repairs in spontaneous speech, *Cognition*, 39, pp. 173-194, 1991.

[15] J.E. Fox Tree and H. Clark, Pronouncing the as thee to signal problems in speaking, *Cognition,* vol. 62, pp. 151-167, 1997.

[16] H. Clark and T. Wasow, Repeating words in spontaneous speech, *Cognitive Psychology,* vol. 37, pp. 201-242, 1998

[17] H. Clark, *Using Language,* Cambridge: Cambridge University Press, 1996.

[18] S. Kita, I. van Gijn, and H. van der Hulst, Movement phases in signs and co-speech gestures, and their transcription by human coders, in *Gesture and sign language in human-computer interaction,* I. Wachsmuth and M. Froehlich Eds. Proceedings of the International Gesture Workshop, Bielefeld, Germany, September 17-19, 1997. Berlin: Springer, 1998, pp. 23-35.

Pauses in speech by French speakers with Down Syndrome

Laura ABOU HAIDAR, Lab. « Langage & Handicap » (Tours, France)

Lauraabouhaidar@aol.com

Key words : Down syndrome – Conversation – Temporal structure – Pauses – French

Abstract

A better understanding of the control mechanisms of speech in verbal interaction is very important for the evaluation of the pragmatic competence of a mentally deficient speaker. This study focuses on pauses in the oral production of a Speaker with Down syndrome involved in a conversation : it brings to light the temporal compensation mechanisms which allow the speaker to go beyond the distortions of the segmental level. It confirms the important role of prosody in the success of a conversation, particularly with a speaker who has a handicap which disrupts language structure. Down Syndrome is a condition characterised by an overall delay in cognitive, social, linguistic and motor development. At the oral production level, it leads to deficits in segmental and supra-segmental speech patterning. The goal of this study is to bring elements of response to the following question : is the pragmatic function of language preserved in spite of significant distortions of the motor functions of the phonatory organs ? The description of the management of pauses by a speaker with Down syndrome involved in a conversation makes it possible to clarify this subject, while taking into account the various functions which are specific to them beyond the respiratory function : their role in encoding, in the delimitation of syntactic boundaries, and in the regulation of speaking turns, among others.

This study allowed us to define criteria which make it possible to characterise the oral production of a Speaker with Down syndrome. These elements relate to the variation of the frequency and the length of pauses. The results obtained are the following:

1.a high frequency of occurrence of pauses in the production of the trisomic speaker

2.a frequency of occurrence of "mixed pauses", of which the majority have very long lengths, this element revealing a lack of ease and disfluency on the production level;

3.a significant recourse to false-starts, hesitation, repetition and lengthening, to mark sound pauses;

4.a considerable number of very long pauses pauses;

5.a relatively high number of pauses located at the boundaries of or within syntagms, with rather long lengths of intra-syntagmatic uses;

We furthermore noted a rarity of long phonic sequences in the speaker with Down syndrome, these sequences seldom exceeding 2000 ms.

In spite of these results, it is important to note that we have defined parameters which show that the speaker with Down syndrome integrated rules relating to the management of pauses in verbal interaction.

1. Introduction

Down Syndrome is a condition characterised by an overall delay in cognitive, social, linguistic and motor development [1] [2]. At the oral production level it leads to deficits in segmental and supra-segmental speech patterning [3]. A better understanding of the control of vocal and gestural mechanisms in verbal interaction is essential for a comprehension of the linguistic and pragmatic competence of subjects with Down Syndrome. The goal of this study is to bring elements of response to the following question : is the pragmatic function of language preserved in spite of significant distortions of the motor functions of the phonatory organs ? The description of the management of pauses by a Speaker with Down syndrome in a verbal interaction makes it possible to clarify this subject, while taking into account the various functions which are specific to them beyond the respiratory function : their role in encoding, in the delimitation of syntactic boundaries, and in the regulation of speaking turns, among others.

Research carried out on normal speech patterns has made it possible to define several types of pauses [4] :

- silent pauses (SP) ; they can be respiratory or not ;
- non-silent pauses (NSP) or filled pauses ; they are divided into 4 sub-categories : hesitation (uh, well, etc.) ; false starts (start followed by a reformulation) ; repetitions (involuntary repetition of one unit) ; lengthened syllables ; these vocal units are produced by the speaker.

We have integrated one other sub-category in non-silent pauses : pauses of regulation, produced by the listener, which consist of indices of regulation of the type "hmh", "yeah", "mm", etc. : considering the specificity of our corpus, it appeared relevant to us to distinguish this last sub-type of non-silent pause from the others.

In addition to "silent" and "non-silent" pauses, we have introduced a third category, that of "mixed pauses". This type consists of a succession of one or several silent pauses and of one or more non-silent pauses ; the various configurations which we have grouped into this category are the following :

- a silent pause preceded and/or followed of a non-silent pause

- a silent or filled pause, produced by the speaker, preceded and/or followed of a regulatory index produced by the hearer.

Contrarily to other authors [5], considering the specificity of the sample retained within the framework of this study, namely a conversation between two speakers, it seemed necessary to us to introduce this third category of pause and to distinguish it from the preceding ones, since, on the one hand, they give an account of specific conversational strategies, and on the other hand, it is important to consider the impact which they can have on the perceptive level in terms of degree of fluidity and discursive affluence.

Research carried out on normal speech has been an attempt to explain the nature of the parameters correlated with the frequency and the length the of pauses. This reveals that the length of the pauses seems to vary according to individual [5] [6] and contextual parameters [5] [6] [7] [8] (among these parameters we can cite pauses category, type of speech, and rate).

In this study , we will thus analyse the variation of the frequency and the length of the pauses according to some selected parameters, in a conversation involving a young adult with Down Syndrome.

2. Experimental procedure

Speakers *: The speaker with Down syndrome (A) is a 22 year old male suffering from a slight mental deficiency. He was initially schooled in a normal classroom before his placement in a medico-educational institute. He can read and write. He underwent speech therapy between the ages of 5 and 14. He is severely myopic, has significant respiratory disorders due to the deterioration of the interventricular and interoricular communication. and pulmonary arterial hypertension. He has fairly good language production, no major difficulties with morphosyntactic structure, but significant problems on the articulatory level sometimes make him incomprehensible to listeners. The other speaker (B) is a 22 year old student, a co-ordinator in a centre offering various activities to young people with mental and/or motor handicaps. The two speakers know each other well since they met during the production of a play, the young Down Syndrome adult being one of the actors.*

The sample *: in this study, we are dealing with an somewhat spontaneous conversation, which actually lasted about thirty minutes and took place on the premises of the centre mentioned above. The sample retained for analysis lasts about 19 minutes. The corpus was transcribed using a Sanyo TRC-8800 transcriber : we have produced a phonetic transcription by using the IPA as well as some extensions of this alphabet [17].*

Analysis equipment *: Length measurements were taken using Winpitch II software. We took into account the oscillograph tracing as well as the amplitude curve and the fundamental frequency. In many cases, it was necessary to compare the three curves in order to obtain a reliable measurement.*

3. Results

	Speaker A	Speaker B	*TOTAL*
Phonatory time (without pauses)	17.08%	44.16%	*61.24%*
Intra-speaker pause times	17.06%	9.97%	*27.03%*
TOTAL	*31.14%*	*54.13%*	

Table I – Phonatory time and Intra-speaker pause times for Speakers A & B.

Category	*Speaker A*	*Speaker B*
SP	*34.4%*	*47.4%*
MP	*55.2%*	*40%*
NSP	*10.4%*	*12.6%*

Table 2 – Percentage of intra-speaker pause category for both speakers

	Speaker A	*Speaker B*
False-starts	*29%*	*12.79%*
Hesitations	*24%*	*40.69%*
Repetitions	*15%*	*2.32%*
Lengthening	*32%*	*44.18%*

Table 3 – Recourse to false-starts, hesitations, repetitions, lengthening in both speakers' mixed pauses

Firstly, as regards intra-utterance pauses, this study shows that speakers do not manage the difference between time allotted to pauses and to phonation in the same way: The speaker with Down syndrome allots half of his speech time to pauses, whereas in other speaker's production, the proportion is in a ratio of one to five. Pauses are more frequent and longer in the speaker with Down syndrome, and particularly mixed pauses, whereas the other speaker makes more frequent use of silent pauses. As regards "duration categories", short pauses are most frequent in both speakers, but the number of long pauses is excessive in the production of the speaker with Down syndrome.

Concerning non-silent pauses, unintentional repetitions or false starts are rarely observed in the normal speaker, who would rather uses hesitations with vocal unit lengthening. The speaker with Down syndrome uses false starts and repetitions more frequently, and when lengthening is observed, it does not exclusively concern vocalic units, but consonantal units as well.

Pauses are characterised by a syntactic function, which corresponds more frequently to a sentence frontier in the normal speaker, and the least often to a boundary between or inside phrases. In the production of the speaker with Down syndrome, we notice an impairment of the

syntactic function : pauses are quite often located between, or inside phrases.

4. Discussion

Language impairment analysis is very problematic when based on a single case study, mainly because it may be dangerous and inaccurate to distinguish elements relating to pathology from individual characteristics. It also seems hazardous to generalise from only one case. Nevertheless, beyond individual strategies, which are characteristic of normal and disabled speakers with language impairment, some of the observed phenomena are indisputably due to a language dysfunction.

Our initial aim in this study was to explore pragmatic competence in a speaker with Down syndrome involved in a conversation. The choice of temporal criteria (pauses in this contribution) seemed to be obvious : actually, temporal control is a fundamental and necessary component for successful verbal interaction. Its analysis allows us to understand the way rules of interaction are controlled (or not) by a speaker.

This study reveals that, generally speaking, the speaker with Down syndrome has obvious control over the management of some of the temporal parameters which relate in particular to the organisation of the "duration categories". The impairment appears primarily through the imbalance in the "effective" speaking time and the time devoted to pauses, which translates a lack of fluency as well as efficiency in temporal management. Time allotted to pauses is too important compared to the standard rules of conversation in the French language. But this time of pause is far from useless. To the contrary, it is firstly an "efficient time" since unquestionably it corresponds to the encoding function. Secondly, non silent pauses, especially mixed pauses, reveal how the speaker with Down syndrome is willing to take an active part in the conversation : at no time, even after very long pauses, does he give up ; in spite of his disability, he shows that he has an accurate perception of the rules of interaction, and that he is willing to participate fully as an interlocutor. From this point of view the way in which he is involved in the conversation contributes to its success. Moreover, the fact that the other speaker does not seem to be disturbed by very long pauses (and does not take advantage of them) is a guarantee that the speaker with Down syndrome's turn and speaking time belong to him completely.

Elements that contribute to the perception of disfluency in the speaker with Down syndrome are excessive pause length, as well as disproportionate use of filled or mixed pauses (especially repetition or false-starts). These phenomena make it difficult to understand him (especially for uninformed listeners). These objective, measurable elements, which correspond to a disfluency in speech production, reveal the existence of more complex problems (concerning lexical access, morpho-syntactic construction, etc). Nevertheless, it is obvious that encoding function is actually used by the speaker with Down syndrome.

It appears to us that these elements constitute a base on which specialists can rely to improve the efficiency of the speaker with Down syndrome's contribution to a conversation, so that the same elements which reveal a disfluency could sometimes be the expression of encoding problems.

Bibliography

[1] J. A. Rondal, *Langage et communication chez les handicapés mentaux : Théorie, évaluation et intervention,* Pierre Mardaga, Bruxelles, 1985.

[2], C. Chevrie-Muller and J. Narbona, *Le langage de l'enfant. Aspects normaux et pathologiques,* Masson, Paris, 1996.

[3] S.F. Warren and L. Abbeduto, The relation of communication and language development to mental retardation, *American journal on mental retardation,* 97, pp. 125-130, 1992.

[4] D. Duez, Silent and non-silent pauses in three speech styles. *Language and Speech,* 25, pp. 11-28, 1982.

[5] D. Duez, *Contribution à l'étude de la structuration temporelle de la parole en français*, Thèse de doctorat d'Etat, Aix-en-Provence, 1987.

6] F. Goldman-Eisler, Pauses, clauses, sentences, *Language and Speech,* 15, pp. 103-113, 1972.

[7] F. Grosjean and A. Deschamps, Analyse des variables temporelles du français spontané. Comparaison du français oral dans la description avec l'anglais (description) et avec le français (interview radiophonique), *Phonetica,28,* pp. 191-226, 1973.

[8] F. Grosjean and A. Deschamps, Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation, *Phonetica*, 31, pp. 144-184, 1975.

Duration Category	*SP A*	*SP B*	*MP A*	*MP B*	*NSP A*	*NS B*	**TOTAL A**	**TOTAL B**
I (0-500 ms)	*7.27%*	*8.65%*	*0.57%*	*0.96%*	*1.53%*	*1.92%*	**9.37%**	**11.53%**
II (501-1000 ms)	*2.68%*	*2.88%*	*4.40%*	*3.84%*	*1.91%*	*1.15%*	**8.99%**	**7.87%**
III (1001-1500 ms)	*1.72%*	*0.96%*	*4.40%*	*3.84%*	*0%*	*0.38%*	**6.12%**	**5.18%**
IV (1501-2000 ms)	*0.38%*	*0.57%*	*4.78%*	*1.92%*	*0%*	*0%*	**5.16%**	**2.49%**
V (2001-2500 ms)	*0%*	*0.38%*	*2.10%*	*0.96%*	*0.19%*	*0%*	**2.29%**	**1.34%**
VI (2501 – 3000 ms)	*0%*	*0%*	*1.14%*	*0%*	*0%*	*0%*	**1.14%**	**0%**
VII (> 3000 ms)	*0.19%*	*0.19%*	*1.91%*	*0%*	*0%*	*0%*	**2.1%**	**0.19%**

Table 4 – Pauses and duration category in both speakers.

Frontier type / **Duration category**	**Proposition Speaker A**	***Proposition Speaker B***	**Inter-syntagm. Speaker A**	***Inter-syntagm. Speaker B***	**Intra-syntagm. Speaker A**	***Intra-syntagm. Speaker B***
I (0-500 ms)	14.7%	*32%*	5.5%	*2.8%*	6.6%	*6.1%*
II (501-1000 ms)	13%	*20.1%*	7%	*4.1%*	6%	*3.4%*
III (1001-1500 ms)	11.4%	*12.1%*	3.3%	*2.8%*	2.7%	*2.8%*
IV (1501-2000 ms)	8.2%	*5.4%*	2.7%	*2.1%*	3.8%	*0.8%*
V (2001-2500 ms)	3.3%	*4.8%*	1%	*0%*	2.2%	*0%*
VI (2501 – 3000 ms)	2.2%	*0%*	0.5%	*0%*	0%	*0%*
VII (> 3000 ms)	5.4%	*0.8%*	0%	*0%*	0.5%	*0%*
TOTAL	**58.2%**	*75.2%*	**20%**	***11.8%***	**21.8%**	***13.1%***

Table 5 – Pauses occurrence, syntactic distribution and duration category in both speakers

Type of pause / Duration category	IBA PS	IBA PNS	IBA PM	**TOTAL IBA**	IAB PS	IAB PNS	IAB PM	**TOTAL IAB**
I (0-500 ms)	48.23%	3.53%	0%	**51.76%**	50.49%	0%	0.97%	**51.46 %**
II (501-1000 ms)	11.76%	4.71%	7.06%	**23.53%**	15.53%	0.97%	8.73%	**25.23 %**
III (1001-1500 ms)	8.23%	0%	3.53%	**11.76%**	6.80%	0%	2.91%	**9.71%**
IV (1501-2000 ms)	0%	1.18%	4.71%	**5.89%**	3.89%	0%	3.89%	**7.78%**
V (2001-2500 ms)	0%	0%	1.18%	**1.18%**	0.97%	0%	1.94%	**2.91%**
VI (2501 – 3000 ms)	0%	0%	1.18%	**1.18%**	0%	0%	0.97%	**0.97%**
VII (> 3000 ms)	0%	0%	4.70%	**4.70%**	0%	0%	1.94%	**1.94%**
TOTAL	**68.22%**	**9.42%**	**22.36%**	**100%**	**77.68%**	**0.97%**	**21.35 %**	**100%**

Table 6 – Inter-turn pauses and duration category in both speakers.

Prosodic Marking of Self-repairs

Tapio Hokkanen

Linguistics
University of Joensuu, Finland
tapio.hokkanen@joensuu.fi

Abstract

Slip studies predominantly focus on either structural or semantic properties of the errors. Since most analyses have been based on pen-and-paper collections, i.e., on-line notes, it is quite understandable that suprasegmental of errors have remained a neglected area.

The present prosodic analysis is based on acoustical measurements of 307 self-repairs. Each repair has been measured with the Praat program. In order to make the measurements psychoacoustically relevant and comparable across speakers, the changes in F_0 are expressed in terms of semitones.

In general, speakers repair slightly less than three quarters of the errors they commit whereas one quarter remains either totally undetected or at least without a repair. With respect to prosodic marking, it appears that the proportion of marked repairs in the present data is significantly larger than in previous studies: approximately two thirds of self-repairs are marked with remarkably higher pitch (>+3ST), and a total of 96.7 per cent with a somewhat heigthened pitch. It is concluded that alternations of fundamental frequency are utilized in marking self-initiated repairs.

1. Introduction

When an error occurs, speakers tend to interrupt the flow of speech and initiate a self-repair. From a communicative point of view, the abrupt interruption and the following attempts to repair the troublesome utterance are a special form of disfluency: The listener has to decide how much of the context must be rejected and where the repair will start. Levelt [1] refers to this as continuation problem. The speaker, in turn, has to signal to the listener that a repair will follow.

Here it is hypothesized that speakers need not restrict to structural matters such as editing expressions, retracings to the previous context, or fresh starts in indicating the initiation of a repair. Instead, they may also provide the listener with certain prosodic cues, such as pausing, higher pitch, loudness, and variations in speech tempo [2] that supposedly play an important role in self-repairs. The purpose of this study is to analyse the micro-prosody of self-repairs in naturally occurring slips of the tongue.

Previous studies (e.g., [3] and [4]) of prosodic marking in self-repairs have relied on auditive impressions and interpretations by the researchers themselves. Cutler [3] argues that there are two prosodic strategies for the speakers to follow. In the first one, which she refers to as **unmarked**, the speaker wants to minimize the disruptive effect of the repair to the message by keeping the pitch of the repair as close as possible to the original troublesome item. The **marked** alternative, in turn, takes advantage of noticeable up- or downward changes in the pitch.

Cutler [3] found that speakers did not use prosodic marking in repairs that follow errors at the phonetic level but that 38 per cent of lexical error corrections were marked with remarkably higher pitch. Levelt and Cutler [4], in turn, report a 45 per cent proportion in a pattern description task. Levelt [1] reminds us that there are also personal stylistic factors of the speakers when he writes, "certain speakers [- -] would, so to say, cry out the corrections".

The present study follows the argumentation of Hokkanen [5], where it was claimed that self-repairs are predominantly marked with a remarkably higher pitch.

2. The Data

The following analyses are based on a corpus of tape-recorded, naturally occurring slips of the tongue (see [5]) that have been collected from radio interviews, sportscasts, and archived samples of spoken Finnish. There are 2,202 errors in the entire corpus, approximately three fourths (N = 1,683) of which are repaired by the speaker. In this study, a sample of 307 self-initiated repairs will be analyzed prosodically. All repairs were digitized with a 22.05 kHz sample rate.

3. Method and hypotheses

In order to avoid subjective interpretations, the markedness analysis should not be based on auditive impressions but on acoustic measurements, the results of which are then interpreted with certain psychoacoustic criteria. Therefore, pitch is here acoustically correlated to musical semitones (for the advantages of expressing F_0 changes in semitones, see [6] and [7]). This method also allows us for a reliable comparison of fundamental frequency changes across various speakers and occasions.

For the purposes of this analysis, fundamental frequency has been measured at the following points in each repair: first, at the onset of the troublesome item, secondly, at the end of the troublesome item before the interruption, thirdly, at the onset of possible editing expressions, and, finally, at the onset of the repair. These points supposedly provide a rough, yet an objective idea on whether speakers mark their self-repairs prosodically, and if so, what is marked and how. The measurements were conducted with the Praat synthesis and analysis program [8]. In this program one can also convert the actual speech signal into a human-like humming, which enables the researcher to listen to the flow of F_0 pulses without the actual phones. The prosodic patterns of all repairs can thus be separated from the segmental and semantically motivated means of marking a repair.

The question now arises, what is the differential threshold expressed in terms of semitones in normal communicative situations? The criterion for markedness in this study can be found in 't Hart [9]. He argues that in order to become perceived, the changes in fundamental frequency should

exceed the limit of ±3 semitones (henceforth ST). If we adopt this criterion to the present errors, all repairs indicating an F_0 change larger than ±3 ST will be interpreted as marked with either a significanlty higher or lower pitch.

On the basis of previous findings by Cutler [3] and Levelt and Cutler [4], it is hypothesized, firstly, that only less than half of lexical errors are marked with a higher pitch, although errors may sometimes also be marked with a remarkably lower pitch. Secondly, the previous studies also suggest that speakers should mark only repairs following lexical errors whereas repairs of phonetic errors remain prosodically unmarked.

It is furthermore intuitively hypothesized that prosodic marking would be dependent on detection latency, *i.e.*, on the interval between the error occurrence and the moment of interruption. The further the interruption appears from the troublesome item, the greater the change in fundamental frequency. Moreover, one may also hypothesize that the longer the pause between the error and its repair, the greater the change in F_0. The latter hypothesis is based on physiological grounds: a long pause could be used for breathing, which, in turn, might result in a somewhat higher pitch.

4. The analysis

In general, it appears that 66.1% (N = 203/307), *i.e.*, approximately two thirds, of all repairs exceed the ±3 ST differential threshold level and can, thus, be regarded as prosodically marked. This figure is fairly high when contrasted to the 38 and 45 per cent proportions reported previously by Cutler [3] and Levelt and Cutler [4]. This difference can not solely be attributed to the language analyzed (*viz.* Finnish). Rather, it probably depends on the method applied: as opposed to the previous studies, the present distribution has been obtained by virtue of acoustical measurements of naturally occurring errors and their repairs.

To test the second hypothesis, namely that only repairs following lexical slips would be marked with a remarkably higher pitch [3], marking was analyzed at various linguistic levels. This comparison indicated that there arose no statistically significant differences across the rates of prosodic marking between repairs of lexical and phonological errors: the rates were 67.9 (N = 66/103) and 64.1 per cent (N = 93/137), respectively (χ^2 = 0.609, 1 d.f., not significant). This result does not give support to the first hypothesis. On the contrary, it may be concluded that prosodic marking is not restricted to the lexical level or levels above it. Instead, speakers use prosodic marking in their self-repairs regardless of the linguistic level of the error. This finding has also been interpreted [5] to indicate that it is neither structural well-formedness (phonological errors) nor communicative aspects (lexical ones) that primarily determine prosodic marking in self-repairs.

It is worth noting that none of the repairs in the present data was marked with a significantly lower pitch: the maximum downward change in F_0 was –0.66 ST which does not meet the current criterion of prosodic marking. For the sake of comparison, the maximum upward change was +19.28 ST. Since the downward changes were minimal and since there were only nine of them in the entire data, they may interpreted as falling within the limits of normal F_0 downdrift. Therefore, they become classified here as indications of prosodically unmarked repairs. The current results refer to the possibility that, as opposed to Cutler's [3] findings, speakers of the Finnish language do not mark their repairs with a remarkably lower pitch at all or do so only seldom.

With respect to the hypothesized relation between the latency of interruption and the change in F_0, no statistically significant correlation can be established here (Pearson correlation r = 0.109, n.s.). Therefore, it is evident that speakers do not use prosodic marking to signal their delayed repair initiations. An identical result is obtained in the correlation between the duration of the pause and the change in F_0 as well. It may be concluded that the purpose of the pause is something else than to provide the speaker with time to readjust the fundamental frequency. It has been suggested [5] that the duration of the pause is dependent on the linguistic level the error has arisen from.

In fact, none of the hypotheses presented so far receive direct support from the present data. Instead, it appears that the current repairs are in favor of the proposition by Levelt [1], namely, that prosodic marking of repairs is primarily semantically and communicatively motivated: speakers repair what they find necessary to repair, and prosodic marking means, first of all, marking of contrast between the troublesome item and the repair.

The suggested kind of contrast can best be seen in those instances where the prosodic marking runs counter to the normal prosodic patterns of the Finnish language. For instance, in (1) the error regards an inflectional suffix:

(1) *tulona, tules-, tulossa*
'coming'

In (1), the targeted inessive form (*tulossa*) is first substituted by the elative case (*tulona*) and then followed by a word search (*tules-*). What is essential here is that the speaker marks the repaired suffix (*i.e.., -ssa*) of the repair with an extra F_0 peak: the suffix is articulated 6.8 semitones higher (change from 101.3 Hz to 149.7 Hz) than the corresponding syllable in the troublesome word. The exceptional intonation contour of this particular repair can be best viewed in Figure 1 (see [5]):

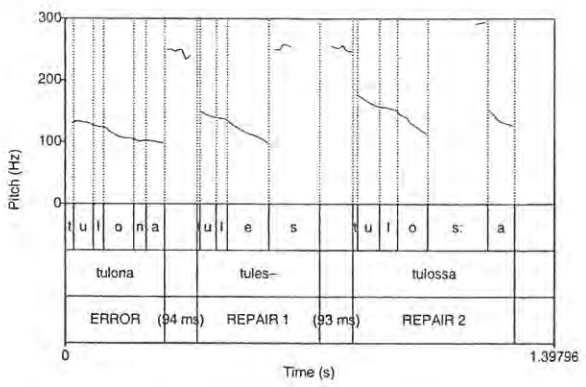


Figure 1: F_0 contour of error (1).

There are also examples regarding constituents of compounds in which an extra F_0 peak is assigned precisely to the first

syllable of the repaired constituent. These kinds of prosodic marking suggest that speakers want to make the contrast between the troublesome item and its repair more prominent by virtue of a substantially higher pitch and that this marking can be aligned exactly with the troublesome point in the utterance. It is also worth noting that the strive for marking a repair is so strong that it overrules the basic prosodic patterns characteristic of, *e.g.*, world-final syllables in polysyllabic words as well as right-hand constituents of compounds [5].

There is another strategy of prosodic marking as well, in which the F_0 level of the repair is matched with the F_0 level of the error. The technical criterion for this strategy is that the rise of the fundamental frequency contour is less than the critical +3 ST limit, yet positive. This definition distinguishes the prosodic marking from the normal F_0 downdrift. With respect to the current data, 30.6% (N = 94/307) of all repairs follow this strategy. It can be interpreted that speakers take advantage of prosodic marking not only to contrast the troublesome item and its repair but also to signal the start of the repair by readjusting the F_0 level. Alongside with structural means, such as, *e.g.*, repetitions and fresh starts, this F_0 readjustment presumably helps the listeners to easier adjust the repair to the original utterance.

The third alternative, *i.e.*, unmarked repairs, corresponds to the normal F_0 downdrift and to the speaker's choice of not marking a repair prosodically. All negative semitone values that do not exceed the –3 ST limit fall into this category. Nevertheless, these kinds of repairs are marginal, since they only account for 3.3 per cent (N = 10/307) of all repairs in the current data. They can be defended from a communicative point of view: despite the error, this strategy guarantees an undisturbed prosodic pattern and supports the continuation of the utterance. None of these repairs exceeded the –0.66 ST limit mentioned above.

Finally, there were no instances in which the repair would have been marked with a significantly (>–3ST) lower pitch. Consequently, the findings reported by Cutler [3] do not receive support from the present self-repairs.

5. Results and concluding remarks

It was argued above that speakers take advantage of changes in F_0 to mark the contrast between the troublesome item and the actual repair. A 66.1 per cent proportion of all repairs in the present data exceeded the +3 semitone limit set as a psychoacoustically motivated criterion of markedness. Instances of F_0 readjustment, in turn, account for slightly less than one third (30.6 per cent) of all repairs. Since these two strategies both deviate significantly from the normal F_0 downdrift, *i.e.*, from the third strategy, they can be treated as marked. Together the two marked strategies account for a vast 96.7 per cent majority of self-repairs, which allows for the conclusion that speakers predominantly use prosodic means in marking their repairs. In general, the proportion of prosodically marked repairs in the current data is substantially larger than those reported in previous studies.

On the basis of the present findings it is evident that speakers combine structural means with suprasegmental cues to help the listener overcome what is referred to as the continuation problem, *i.e.*, the fact that the flow of speech is abruptly interrupted and the troublesome item replaced with some other material.

It also appeared that prosodic marking is not restricted solely to the repairs following lexical errors. Instead, there arose no statistically significant differences between repairs of lower-level errors, which suggests that prosodic marking is not dependent of the linguistic level. No correlation was found between the duration of the pause after the interruption of the speech flow. This finding was interpreted here as an indication for the fact that the pause does not serve as a point for resetting the fundamental frequency.

6. References

[1] Levelt, W.J.M., *Speaking: From Intention to Articulation.* Foris, Dordrecht, 1989.

[2] Nakatani, C.H. and Hirschberg, J., A corpus-based study of repair cues in spontaneous speech. *J. Acoust. Soc. Amer., Vol. 95, 1994, pp. 1603-1616.*

[3] Cutler, A., Speakers' conceptions of the function of prosody. In: A. Cutler & D.R. Ladd (eds.), *Prosody: Models and Measurements.* Springer Verlag, Berlin, 1983. pp. 79-92.

[4] Levelt, W.J.M. and Cutler, A., Prosodic marking in speech repair. *J. Sem., Vol. 2, pp. 205-217.*

[5] Hokkanen, T., *Slips of the Tongue. Errors, Repairs, and a Model.* Finnish Literature Society, Helsinki, 2001.

[6] Hayward, K., *Experimental Phonetics.* Longman, London, 2000.

[7] 't Hart, J., Collier, R. and Cohen, A., *A Perceptual Study of Intonation. An Experimental-phonetic Approach to Speech Melody.* Cambridge University Press, Cambridge, 1990.

[8] See http://www.fon.hum.uva.nl/praat/

[9] 't Hart, J., Differential sensitivity to pitch distance, particularly in speech. *J. Acoust. Soc. Amer., Vol. 69, 1981, pp. 811-821.*

Acknowledgement

I am grateful to Professor Jussi Niemi (Joensuu, Finland) for his encouraging words and comments on an earlier version of this article.

Acoustico-phonetic Characteristics of Filled Pauses in Spontaneous French Speech: Preliminary Results

Danielle DUEZ

CNRS, UMR 6057, Laboratoire Parole et Langage
Université de Provence, 31621 Aix en Provence
Fax: (33) 04 42 59 50 96; email: duez@lpl.univ-aix.fr

Abstract

In the current analysis we examined the acoustic and phonetic characteristics of filled pauses in spontaneous French speech and their relationship to the prosody of the surrounding context. Two main results emerged : 1) There was no effect of the duration of filled pauses or their sentence location on their F0 patterns or on the differences between the highest and lowest values. 2) There was no relationship between peak-F0 values and the F0 values of filled-pause onsets, but the F0 values of filled-pause onsets and the F0-values of non-marked breath-group onsets were highly similar. The F0 values of filled-pause onsets seem to be stable within the same speaker's speech. They are speaker-dependent and strongly linked to the physiological, absolute aspects of speech production. It is assumed that filled-pause onset may be used by listeners as a reference for evaluating the speaker's pitch range.

I Introduction

This paper reports an ongoing study on the acoustic-phonetic characteristics of filled pauses in spontaneous French speech. In the last few decades, there has been a growing interest in the investigation of disfluencies. There are several important reasons for this, including the following: 1) a renewed interest in so-called spontaneous speech and the manifestations of spontaneity, 2) the investigation of speech encoding processes in various communication situations, and 3) the necessity to improve speech technologies, in particular the naturalness of speech synthesis and the reliability of speech-recognition.

A wide range of studies have concerned with filled pauses. Filled pauses are among the most common disfluencies in spontaneous English and French speech [1]. They are widely used in the speech produced by adults [2] and [3] and young people [4] and [5]. Traditionally there are two main explanations given for their occurrence. The first is based on the idea that filled pauses reflect the speaker's level of anxiety [6] and [7]. The other suggests that filled pauses are linked to the complexity of the message [8]. More recently, another explanation was proposed, based on the notion that filled pauses reflect instead speakers' attention to their speech [9]. Thus, filled pauses appear as symptoms of the attention paid to the speech production process. They occur when the speaker detects an error (or an impending error) in the encoding process, and stops to correct it.

The analysis of the distribution of filled pauses in a sentence is a source of information on the different processes involved in speech production. For example, filled pauses occurring before lexical words play a role in the lexical-selection process [10], [2], [3] and [11] while filled pauses located at the beginning of a phrase play a role in the programming of the upcoming phrase [12], [5], [13], [14], [2], [3] and [11].

The acoustic-phonetic characteristics of filled pauses and their relationship to the prosodic context have also been the subject of a certain number of studies aimed mainly at examining how filled pauses are integrated into the intonative structure of sentences and/or at defining a certain number of cues relevant to filled-pause recognition. For example, filled pauses in French were shown to be produced at a given, speaker-specific F0 value regardless of location [15]: the value was roughly equal to the average of the F_0 values of unaccented syllables and major-phrase onsets. Three main patterns have been reported for filled pauses in spontaneous French speech: (1) a slow decline, an irregular decline and a decline followed by a sudden rise [16]. Filled pauses at major syntactic boundaries as well as those within syntactic units have also been shown to have falling or flat F_0 pattern, at relatively low F_0 levels with the lowest F_0 at the end [17]. However, filled pauses at syntactic boundaries tended to start higher in F_0 and then fall, whereas filled pauses internal to a syntactic unit had lower F_0 patterns. Filled pauses at major syntactic boundaries were also found to be longer than those within syntactic units. The analysis of clause-internal filled pauses and the fundamental frequency (F_0) values of preceding peaks in dialogues and conversations in British English and American English have been shown to have higher peaks systematically associated with higher filled-pause values [18]. This was interpreted as an indication of some form of systematic relationship between the peak and corresponding filled-pause F_0 values.

In the current analysis we examined the acoustic and phonetic characteristics of filled pauses and their relationship to the prosody of the surrounding context. Our objective was to test whether filled pauses are produced at an absolute speaker-specific value or whether they are dependent on the prosodic context. Filled pauses were examined as a function of their location in sentences (within a prosodic word, and between two prosodic words or two breath groups, as defined by Vaissière [19]) in the conversations of four French speakers (two males and two females). As a measure of prosodic context, the F_0 value of the breath-group onset, the closest preceding or following peak (if any). As a measure of filled-pause F0, the F0 value of the beginning, middle (or the turning point if any), and end were used. The duration of filled pauses was also measured.

2. Methods

2.1. Subjects

The conversational speech produced by two male and two female French speakers was used for the experiment (one hour in all). The speakers were of the same sociocultural background, without strong regional accents and with normal speech and hearing. Their age ranged from 30 to 50. The subjects, who were unaware of the purpose of the recording, had to reply freely in a relaxed way to questions regarding their life, childhood, work, travels, future plans and current events. The conversations were recorded on a Sony tape-recorder at 9 cm/sec, in a quiet room at the Phonetics Laboratory of Aix en Provence.

2.2. Procedure

The conversations were transcribed orthographically by the author. The presence of filled pauses and the boundary breaks was checked perceptually. Two break levels were defined : prosodic words and breath groups [19]. Filled pauses were also checked to see whether they were accompanied by another disfluency such as a correction, an interruption or a lengthened syllable.

2.3. Filled pauses

A filled pause was any occurrence of "euh" [ø] (a filler like hm or uh in English). It could occur before or after a vowel as in "mais euh" [mɛø] or attached to a consonant as in "donc euh" [dɔ̃kø], "visite euh" [vizitø]. The cases where an optional schwa was realised and lengthened as in "que" [kø:], "le"[lø:] and "de" [dø:] were considered as lengthened syllables and not as filled pauses.

2.4. Location of filled pauses

Two main locations were considered : within a prosodic word and between two prosodic words or two breath groups phrases. The former occurred mostly before a lexical word or at the beginning of a phrase just after a conjunction or an adverb. There were different possibilities for the latter depending on the presence of a silent pause. In the absence of a silent pause or when preceded and followed by a silent pause, a filled pause was simply a between-phrase filled pause; a filled pause followed by a silent pause was a final-phrase filled pause; a filled pause preceded by a silent pause was an initial-phrase filled pause.

2.5. Reference points

There are three key points in French breath-group onset, pretonic syllable, and tonic syllable [20] and [21). Onset-F_0 values can be very high in interrogative and exclamatory sentences, and in phrases with new information, but they are generally quite stable in declarative sentences and are very close to those of pretonic syllables. Since there is no lexical stress in French, prominent syllables are mostly phrase-final syllables [19]. They may also be word-initial syllables since French also possesses an optional initial prominence on lexical words whose realisation depends on the style and speaker [22]. Peaks on a syllable mostly correspond to the realisation of initial and final prominence, but can also be the mark of an emphatic accent. Peak-F_0 values and onset-F_0 values are fundamental in defining sentence structure and in integrating the speaker's pitch range. They were taken as reference values.

2.6. Measurements

The sentences were digitised at a sampling rate of 16 kHz with a Sun computer. Measurements were carried out on spectrograms, oscillograms and F_0 curves displayed on the screen, and by listening to selected segments of the waveform in the region of interest. Filled pauses were identified both perceptually and acoustically. The duration of each filled pause was measured as the duration of the waveform between the first and the last [ø] periods. For each filled pause, five F_0 values were recorded : the first, the middle, and the last F_0 value in the filled pause, F_0 at phrase onset, F_0 at phrase-peak, and F_0 at the syllable following the filled pause (if any).

2.7. Analysis

There were three main patterns for filled pauses (flat, upward for rising F_0 and downward for falling F_0). The combination of the three basic patterns gave nine patterns in all (e.g. downward, downward-flat, downward-upward). The patterns were examined as a function of the location of the filled pauses in the phrases. The initial-F_0 values of the filled pauses were also compared with the mean values of neutral-phrase onsets (default values) and with the onset-F_0 value of the phrases in which the filled pauses were located. The same was done with peak-F_0 values. The initial-F_0 values of the filled pauses were compared with the mean peak-F_0 values and with the peak-F_0 values of the phrases in which the filled pauses were located.

3. Results

3.1. F_0 patterns of filled pauses

Table 1. Number of filled pauses as a function of sentence location (B: between two prosodic words or two breath groups ; W: within a prosodic word) and F0 pattern (D: Downward, DF: downward-flat, DU: downward-upward, F: flat, FD: flat-downward, FU: flat-upward, U:upward, UD: upward-downward and UF: upward-flat). T means total.

	D	DF	DU	F	FD	FU	U	UD	UF	T
B	35	14	9	10	1	4	2	5	1	81
W	29	22	10	13	4	2	3	5	1	89
T	64	36	19	23	5	6	5	10	2	170

The results reported in Table 1 indicate a number somewhat similar for the different patterns of the filled pauses located within a prosodic word and between-two prosodic words or two breath groups. There was no significant interaction between the distribution of filled pauses and the F_0 patterns (χ^2: 0.7, df(8), n.s.).

Decreasing F_0 patterns were the most frequent: 105 out of 170 in all. There was a continuous declining F0 for 64 of them, 36 with a descent followed by a plateau, and 5 with a plateau followed by a descent. There were 23 filled pauses with a plateau. For the remaining filled pauses, there was a combination of a rise with a plateau or a rise with a descent (and vice versa)

In general, the differences between the initial F_0 value and the final F_0 value for the filled pauses were less than 20%, which is in agreement with previous results reported in the literature [16] and [17]. However, 36 out of the 170 filled pauses were found to exhibit larger differences (ranging from 25% to 50%). The majority of them had sharp falls (16 out of

36) or downward-flat patterns (8 out of 36). For the remaining ones the distribution was as follows : downward-upward (2), flat-downward (1), flat-upward (2), upward (3), upward-downward (4). A χ^2 test performed on the results revealed a significant interaction between F_0 patterns and F_0 differences (χ^2 =19.8, p=0.01). The filled pauses with differences above 25% were located within a prosodic word (22) as well as between prosodic words or groups (3 at the beginning and 11 at the end). For 14 out of 36, there was also a silent pause. Some abrupt falls or steep rises (reaching or attaining F_0-values close to 60 Hz) for male speakers (9 cases) and female speakers (8 cases) were linked to a creaky voice.

3.2. Filled-pause duration

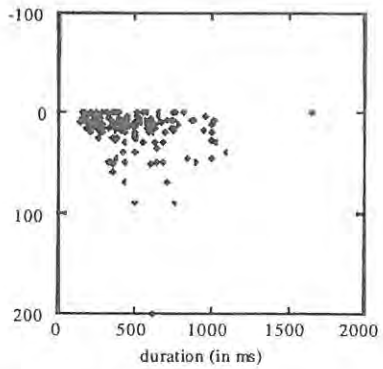
Table 2. Filled-pause duration (in ms) as a function of sentence location (B: between group; W: within a prosodic word) and F0 pattern (D:Downward, DF: downward-flat, DU: downward-upward, F: flat, FD: flat-downward, FU: flat-upward, U: upward, UD: upward-downward and UF: upward-flat). T means total.

	D	DF	DU	F	FD	FU	U	UD	UF	T
B	432	669	468	468	449	428	500	666	202	476
W	450	501	577	478	376	491	485	313	1000	519
T	441	585	522	473	412	459	492	489	601	

An ANOVA on the duration of filled pauses did not reveal a significant effect of F_0 pattern [F(152,8)=1.2, p=0.2], or location [F(152,1)=0.5, p=0.4] but a significant interaction between the two factors [F(152,8)=2.1, p=0.03].

3.3. Effect of duration on F0-differences between the lowest and highest values of filled pauses

Fig1. F0 differences (expressed as a %) as a function of duration



As can be seen in Figure 1, there was no effect of filled-pause duration on the differences (expressed as a %) between the highest value (most often the initial value) and the lowest one (most often the final value). The r^2 value (0.018) confirmed the absence of a correlation between duration and F0 differences. Thus it appears that F0 values and patterns are strongly independent on filled-pause duration. This independence is a characteristic of filled pauses.

Inital-F0 values of filled pauses and prosodic context

An ANOVA yielded significant differences between the F_0 values of filled pauses and the onsets of the phrases in which they were located [F(1, 306)=11, p=0.001], a significant effect of speakers [F(3,306)=235, p=0.0001], and a significant interaction [F(3,306)=3.2,p=0.02]. The differences were probably due to the fact that some phrase onsets were semantically or syntactically marked and exhibited high values (for example, the highest values for Spk2 and Spk3 were 220 Hz and 300 Hz, respectively).

The comparison of the F_0 values of filled pauses and those of non-marked onset phrases did not reveal significant differences [F(1,343)=3, p=0.06)], which supports the above hypothesis (Mean F_0 values can be seen in Table 3). Moreover, there was no statistically significant correlation between the initial F_0-values of filled pauses and peak F0-values (r^2=0.04).

The initial values of filled pauses clustered around the mean and were very stable. However, there were a few exceptions with low values (60 Hz, 80 Hz, and 60 Hz for Spk1, Spk2 and Spk3, respectively) due to physiological effects (creaky voice) These low values were characteristic of some filled pauses that started with a steep rise and of filled pauses located between two silent pauses.

Table 3. Mean F_0 values of filled-pause onset (FPO) and phrase-onsets. There were two different values for phrase onset: those obtained for the phrases in which the filled pauses were located (PO) and those obtained for non-marked onsets (NMFP). SD values are in parentheses.

Speaker	Male		Female	
	Spk1	Spk2	Spk3	Spk4
FPO	110	114	170	176
	(15)	(14)	(32)	(14)
PO	116	116	190	186
	(12)	(20)	(39)	(20)
NMFP	116	115	179	179
	(15)	(18)	(30)	(11)

4. Conclusions

Two main results emerge from the present analysis: 1) There was no effect of the duration of filled pauses and their location within sentences on their F0 patterns, or their highest and lowest values. 2) There was no relationship between peak-F0 values and the F0 values of filled-pause onsets, but great similarity between the F0 values of filled-pause onsets and the F0 values of non-marked breath-group onsets.

In French, four levels are usually used to represent intonation contours [23]. Level 2 is the level of non-marked onsets, level 3 is the level of final syllables of minor-continuation phrases, level 4 is the level of final syllables of major-continuation phrases, and level 1 is the finality level. A fifth level can be used for emphasis and exclamatory sentences [15].

As already mentioned [20] and [21], level 2 is a key point in the production and perception of intonation units. It is a reference value for evaluating the speaker's pitch range, and consequently, a baseline for the integration of the various key points. The fact that filled pauses are anchored in level 2 suggests that they may also constitute a reference for integrating the various levels. This interpretation is quite in line with Léon's proposal [15] to take the F_0 values of filled pauses as a reference for defining level 2. In turn, the F0 values of phrase-onsets can be used as criteria in the recognition of filled pauses.

The present results contradict the common idea that filled pauses are rest times. It is well known that vocal onsets bring into play the breath and the glottal muscles in order to produce the first vibrations of a sound. Physiologically, this consists of four different phases: 1) Tensing of the laryngeal and respiratory muscles, which makes the larynx go from its rest position to its phonatory position, 2) beginning of expiration, 3) end of vocal-fold adduction, and 4) initiation of vocal-fold vibration. The lack of adjustment between phases 2 and 3 may result in a creaky voice which is a frequent characteristic of filled pauses. There appears to be a correspondence between physiological and cognitive activity in the production of filled pauses : at the same time, the speech organs are being positioned and the phonetic program is being prepared for execution.

The F0 values of filled-pause onsets seem to be stable within the same speaker. They are speaker-dependent and strongly linked to the absolute, physiological aspects of speech production [24].

However, some filled pauses may also serve pragmatic aims and be used for specific communicative goals. In this sense, the initial F_0 values of filled pauses might also be manifestations of the relative, functional aspects of speech. Some cases observed here suggest that some F0 values may be higher than those of non-marked breath-group onsets. This assumption has to be tested with a greater number of speakers and in various spontaneous speech styles.

References

[1] Grosjean, F. & Deschamps, A., Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes :phénomènes d'hésitation, *Phonetica,* **31**, 1975, pp. 144-84.

[2] Grosjean, F. & Deschamps, A., Analyse des variables temporelles du français spontané, *Phonetica,* **26,** 1972, pp. 129-149.

[3] Grosjean, F. & Deschamps, A., Analyse des variables temporelles du français spontané, *Phonetica,* **28**, 1973, pp. 191-226.

[4] Sabin, E.J., Clemmer, E.J., O'Connell, D.C. Kowal, S., A pausological approach to speech development, in *Of speech and time : temporal speech patterns* in *interpersonal context,* A. Siegman and S. Feldstein, Eds , pp.35-55, New Jersey : Hillsdale, 1979.

[5] Candea, M., *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané,* thèse de doctorat nouveau régime, Paris III, unpublished manuscript, 2000.

[6] Jurich, A.P. and Polson, C.J. Nonverbal assessement of anxiety as a function of intimacy of sexual attitude questions, *Psychological reports*, **57,** 1985, 1247-1253.

[7] Mahl, G. F., Disturbances in the patient's speech in psychotherapy, *Journal of Abnormal and Social Psychology*, **42,** 1956, pp. 3-32.

[8] Siegman, A.W. and Pope, B. Ambiguity and verbal fluency in the TAT, *Journal of Consulting Psychology,* **30**, 1966, 239-245.

[9] Christenfeld, N., Effects of a metronome on the filled pauses of fluent speakers, *Journal of Speech and Hearing Research,* **39,** 1996, pp. 1232-1238.

[10]Goldman-Eisler, F., *Psycholinguistics. Experiments in spontaneous speech.* London and New York: The Academic Press, 1968.

[11]Maclay, H. & Osgood, C.E., Hesitation Phenomena in Spontaneous English Speech, *Word,* **1**, 1959, pp. 19-43.

[12] Boomer, D.S., Hesitation and grammatical encoding, *Language and Speech,* **8,** 1965, pp. 148-158.

[13]Duez, D., *Contribution à l'étude de la structuration temporelle de la parole en français,* Thèse de Doctorat d'état, Université de Provence, unpublished manuscript, 1987.

[14] Duez, D., *La pause dans la parole de l'homme politique,* Editions du CNRS, collection sons et parole, Paris, 1991.

[15] Léon, P., *Prolégomènes à l'étude des structures intonatives,* Studia Phonetica 2 (P. Léon and P. Martin, Eds) Montréal, Paris and Bruxelles: Didier, 1969.

[16] Guaïtella, I, *Rythme et parole : comparaison critique du rythme de la lecture oralisée et de la parole spontanée,* thèse de doctorat, Université de Provence, unpublished manuscript, 1991.

[17] O'Shaughenessy, D., Recognition of hesitations in spontaneous speech, *IEEE,* **1**, pp. 521-524, 1992.

[18] Shriberg, E. & Lickley, R., Intonation of clause-internal filled pauses, *Phonetica,* **50**, 1993, pp.172-179.

[19] Vaissière, J., Rhythm, accentuation and Final lengthening in French, in *Music, Language and Brain*, J. Sundberg, L. Nord, L. and R. Carlson, Eds. pp 108-120, Macmillan, Houndsmills, 1991.

[20]Di Cristo, A., Des traits acoustiques aux indices perceptuels. Application d'un modèle d'analyse prosodique à l'étude du vocatif en français, *Travaux de l'Institut de Phonétique,* **3**, 1976, 213-358.

[21] Di Cristo, A., *De la microprosodie a l'intonosyntaxe,* Thèse de doctorat d'etat, Université de Provence, 1978.

[22]Fonagy, I., L'accent français : accent probabilitaire, in *L'accent en Français contemporain,* I. Fonagy and P Léon, Eds; Montréal, Paris, Bruxelles : Studia Phonetica 8, 1980, 53-97.

[23] Delattre, P., (1966) Les dix intonations de base du français, *The French Review,* **40(1),** 1-14.

[24] Crystal, D. (1971). Relative and absolute in intonation analysis. *Journal of the International Phonetic Association,* **1 (1)** : 17-28.

Interruption glottalization in German spontaneous speech

Klaus J. Kohler, Benno Peters, Thomas Wesener

Institute of Phonetics and Digital Speech Processing
University of Kiel, Germany
kk@ipds.uni-kiel.de

Abstract

This paper analyzes the occurrence of phonetic interruption cues at points of syntactic irregularities (false starts and truncations) in a large annotated corpus of German dialogues and compares interruption glottalization with laryngealization in terminal low phrase-final prosodies. Glottalization (including glottal stop) predominantly marks word fragments, whereas non-verbal insertions, e.g. breathing, tend to be word-external interruption cues. Laryngealization (excluding glottal stop) predominantly signals terminal phrase boundaries in turn-final positions. Individual speakers differ a great deal as to the distribution of these phenomena.

1. Introduction

Disfluencies have been studied extensively in the past decade, to a large extent in the context of speech technology application, where the aim is to filter out syntactic irregularities for more efficient automatic speech recognition [1] [2]. In this respect, the detection of repairs plays an important role, especially when it can be related to phonetic cues, such as glottalization.

The data base for this paper originated in a similar environment [3]. In the manual orthographic transliteration of German dialogues, two categories of syntactic irregularities were distinguished and marked symbolically: false starts and truncations. False starts refer to sentences that are broken off and continued after a simple repetition or repair (of a part) of what has already been said. For future filtering, the left and right edges of the reparandum are symbolized. In the case of a truncation, a sentence is broken off and not continued, but a new sentence may be started by the same speaker. In this case only the cut-off point is symbolized. A corpus of German dialogues annotated in this way is the point of departure for the basic research into disfluencies and their phonetic exponents, especially glottalization, presented in this paper. The aim is, however, that the results from this investigation will in turn be channelled into speech technology applications.

A preliminary study of glottalization phenomena as cues to false starts and truncations in German was presented in [4]. The term 'truncation glottalization' used there is replaced here by 'interruption glottalization' to cover glottalization in both truncations and false starts unambiguously, and to follow Nakatani and Hirschberg's terminology [2].

In the following we are going to discuss the analysis of glottalization beside other phonetic exponents of false starts and truncations. 'Glottalization' refers to low frequency glottal pulsing (variable in frequency, amplitude and wave form), in alternation with, or in addition to, glottal stop. A further glottalization phenomenon is 'tight voice', which is characterized by a jump-up in F0 and by low amplitude, as well as by the auditory impression of tightness. Glottalization, glottal stop and tight voice are collectively referred to as glottalization phenomena. Other phonetic cues to these syntactic irregularities include interruptions, within lexical material, by pauses, breathing, articulatory and non-articulatory noises, hesitation particles, and hesitational lengthening[1].

It is known from previous studies on English that glottalization frequently marks the end of reparandum intervals if they end in vowels [1] [2]. Our investigation picks up this thread and inserts it into the wider context of phonetic interruption cues generally.

Irregular glottal pulsing also functions as a signal of phrase finality without disfluency. This feature is associated with terminal pitch patterns that descend to the bottom of the pitch range, and plays no role in other contours. Therefore further data are presented on phonation in phrase-final terminal prosodies (modal voice versus irregular glottal pulsing) as one phonetic cue of phrase finality. For the sake of terminological stringency, this deviation from modal voice is called 'laryngealization' as opposed to interruption 'glottalization'. Finally, the phonetic cues of interruption and of phrase-finality are compared.

2. Method

2.1. Data base

The investigation reported in this paper is based on the *Kiel Corpus of Spontaneous Speech* [5]. These data were collected in an appointment-scheduling scenario between two speakers who opened their own (and simultaneously closed their dialogue partner's) recording channel by pressing a button [6]. These recordings are, therefore, not appropriate for analyzing disfluencies triggered by speaker interaction in overlapping dialogue. For the signal files manual transliterations were produced, including words in standardized orthography, the marking of pauses, articulatory and non-articulatory noises (e.g. breathing, paper rustling), hesitation phenomena, and syntactic irregularities.

Syntactic irregularities comprise deviations from syntactic structure, morphology, and lexicon. Those phenomena that are commonly known as slips of the tongue are included if they lead to syntactic irregularities. In the marking of syntactic irregularities, four types are distinguished: word-internal or word-external false starts (=/+ or /+) and word-internal or word-external truncations (=/– or /–).

The transliteration files are automatically converted to transcription files containing canonical segmental word transcrip-

[1] Graphic signal representation and speech output of representative examples can be found at the following URL: `http://www.ipds.uni-kiel.de/publikationen/audiobsp.en.html`

tions as well as labels for the types of disfluencies mentioned above. These transcription files are the basis for the segmental labelling of the speech waves resulting in label files. The alphabet used is modified SAMPA.

Prosodic labelling is added to the label files and is done within the framework of the *Kiel Intonation Model* (KIM) [7]. The point that is relevant for this presentation is the marking of low terminal pitch contours as **&2.**, followed by a prosodic phrase boundary label.

Only the data files with complete segmental and prosodic labelling are used for this investigation. This corresponds to the signal files of all complete sessions in volumes 1 and 2 of the *Kiel Corpus of Spontaneous Speech*, with 22 speakers (13 male, 9 female) in 11 dialogue sessions; the total recording time amounts to approximately 2.5 hours (25000 words). The label files of this selected data base are entered into a structured data bank using *Kieldat* utilities [8].

2.2. Data search and data processing

The data search is carried out with reference to the marking of

- syntactic irregularities (=/+, /+, =/–, /–), data set A
- low terminal falls at prosodic phrase boundaries (**&2.**), data set B

Syntactic irregularity, represented by data set A, is the initial criterion in the search for disfluencies. It may not be coupled with phonetic exponents signalling an interruption. If there is a phonetic manifestation, it may be a glottal stop, glottalization, tight voice, pausing, breathing, a hesitation particle, or hesitational lengthening, which also includes holding a stop closure, and possibly others.

The selected data base is searched with *awk* scripts for all occurrences of syntactic irregularities (data set A), and low terminal falls (data set B) in a frame from the prosodic phrase boundary preceding to the one following any one of the respective labels. The signal portions corresponding to each of these data sets are automatically spliced together. In parallel to the signal file for each data set, two text files are generated, providing (1) the orthographic words and (2) the segmental and prosodic labels and their time points.

In the case of data set A, the next step is the automatic extraction, from the data bank, of the type of irregularity marker, the segmental contexts immediately preceding or following, and the speaker identification. For data set B, the automatically extracted information refers to the labels preceding and following the terminal contour marker as well as to the speaker identification. In both data sets, the preceding context is classified as sonorant (vowel, nasal, lateral) or non-sonorant, the following context as phonological segment or canonical glottal stop or pause/breathing or other (articulatory or non-articulatory) noises. A new label containing the automatically extracted information is introduced in the label files (2) at the time point of each syntactic irregularity or prosodic contour marker, respectively.

Canonical glottal stop refers to the automatic transcription of a glottal stop symbol before all word-initial vowels in German. This glottal stop may be realized as such or as glottalization, or not at all. If there is such a *canonical* glottal stop following a point of syntactic irregularity or a phrase-final terminal prosody, the *actual* occurrence of a glottal stop or glottalization may be ambivalent in its reference either to phonetic interruption/phrase-final laryngealization, or to the following canonical glottal stop. These ambivalent cases are excluded from further analysis.

The files for each data set are then analyzed by accessing speech wave, spectrogram, fundamental frequency, orthographic words, and labels in parallel windows with the *xassp* programme [9]. As regards data set A, the automatically generated classifiers are manually supplemented by adding information on the phonation type at the end of the reparandum, with the four-fold specification of glottal stop or glottalization or tight phonation or their absence; an additional classifier is reserved for uncertain cases. The respective label is added to the automatically inserted label string.

As regards data set B, the information on phonation type, which is added manually to the automatically generated classifiers, provides the five-fold specification of glottal stop or laryngealization or modal voice or plosive-related glottalization or uncertain. Plosive-related glottalization refers to the realization of plosives as glottal stop or glottalization in, e.g., bilateral nasal environment (*könnten* [k^{h}œn̲n̲n] 'could' [10]). If a word with such a phonotactic structure occurs in a terminal fall at the end of a prosodic phrase, the incidence of irregular glottal pulsing cannot be uniquely associated with phrase-final laryngealization; therefore these cases are excluded from further analysis.

3. Results

3.1. Phonetic cues at false starts and truncations

3.1.1. Description

There are 338 instances of marked syntactic irregularities. Of these, 41 are followed by a canonical glottal stop, 17 are labelled as uncertain phonation; these cases are excluded from further analysis. On the basis of the automatic classifications of the phonetic environments of syntactic irregularity markers, on the one hand, and the manual classifications of the phonation types at these places, on the other, a category of presence/absence of phonetic interruption within lexical material was defined in the following way:

- interruption by glottal stop (class I) or glottalization (class II) or tight voice (class III)
- interruption by pause or breathing or (non)-articulatory noises (class IV)
- interruption by hesitation particles (class V)
- interruption by hesitational lengthening (class VI)
- no interruption (class VII)

In the case of multiple cues for phonetic interruption, the following precedences determine classification: glottalization phenomena (classes I–III) over pause/breathing/(non)-articulatory noises (class IV) and hesitational lengthening (class VI); pause/breathing/(non)-articulatory noises (class IV) over hesitational lengthening (class VI); hesitation particles (class V) over glottalization phenomena (classes I–III) and hesitation lengthening (class VI). If there is a sequence of class IV and class V features, the one that comes first determines classification.

Looking at the covariance between the four defined classes of syntactic irregularities — word-internal/word-external false starts, word-internal/word-external truncations — and the seven classes of phonetic interruption across the *total speaker population* we arrive at the data distribution presented in table 1. Glottalization phenomena, summed over the classes I–III, mark 27%

Table 1: Covariance between classes of syntactic irregularities (=/+ word-internal false start, /+ word-external false start, =/– word-internal truncation, /– word-external truncation) and classes of phonetic interruption (I–VII).

	I	II	III	IV	V	VI	VII	
=/+	14	17	4	15	9	3	37	99
/+	7	12	4	43	5	6	22	99
=/–	1	1	1	4	0	1	1	9
/–	8	3	3	37	7	3	12	73
	30	33	12	99	21	13	72	280

Table 2: Classes of phonetic interruption (I–VII) for different speakers. The four categories of syntactic irregularity are merged into one.

	I	II	III	IV	V	VI	VII
TIS	4	7	3	3	1	2	7
FRS	4	3	1	2	2	1	4
JAK	2	6	0	14	0	1	4
SAR	5	0	0	11	0	0	6
OLV	0	4	0	3	3	1	9
CHD	0	0	0	13	0	0	6
ANL	0	0	0	17	4	1	9
HAH	0	0	0	12	2	4	3
	15	20	4	75	12	10	48

of all cases of syntactic irregularities; there is no difference in their distribution across the preceding sonorant or non-sonorant contexts. This relative frequency is practically identical with the 26% in class VII, which has no phonetic interruption of any kind. The highest proportion (35%) is associated with phonetic interruption by pause or breathing or (non)-articulatory noises. If classes IV and V are conflated, we get 43% non-verbal insertions.

Word-internal and word-external false starts show opposite distribution patterns. The former have a high incidence of the absence of a phonetic cue, as well as of glottalization phenomena, at the expense of non-verbal insertions; in the latter category, these distributions are reversed. Word-external false starts and truncations show similar patterns. Two lines in table 1 present extreme distributions: 1) Internal truncations have very low frequency. 2) Internal false starts have the highest frequency of glottal phenomena.

An examination of the behaviour of the *individual speakers* shows diverging trends between them. On the one hand there are speakers who have very few or no glottalizations in any of the four types of syntactic irregularities, and especially use breathing instead. On the other hand, the distribution across the four types of syntactic irregularities differs a great deal from speaker to speaker. This means that the group data may be biassed by individual speakers, particularly since the total frequency of syntactic irregularities per speaker is not very high and differs from speaker to speaker. For this reason the data presentation is broken down into the frequency distributions of individual speakers with the four categories of syntactic irregularities conflated into one. This is done for those speakers who produced more than 15 cases (maximum 26) in table 2.

Three speakers (CHD, ANL, HAH) have no glottalization phenomena, and at the same time show the highest frequency in the category of non-verbal insertions (classes IV and V). Two speakers (TIS, FRS) show the opposite trend. The remaining three speakers fall in between these two groups.

3.1.2. Interpretation

The following tentative interpretation is offered for the low frequency of word-internal truncations. In the absence of a proper dialogue situation in the recording scenario there is no overlap between speakers, so a speaker is not compelled to stop at a time when the other speaker starts speaking, which may be at any point in verbal material. Interruptions internally in a speaker's turn, on the other hand, may be predominantly of the false start type, and if they are of the truncation type they may not occur before the end of a word is reached.

The facts that the highest frequency of glottal phenomena occurs in internal false starts and that word-internal and word-external false starts show opposite distribution patterns may be seen as indicating a reinforcement of the fragment nature of the verbal material, whereas non-verbal insertions seem to be used to mark interruptions at word boundaries. This evaluation of the German data can be connected with the report by Nakatani and Hirschberg [2] that the majority of the glottalizations they found in English occur in word-fragments.

This interpretation, however, has to be taken with caution because when we look at the behaviour of different speakers we find diverging trends. There are speakers who do not seem to use glottalization phenomena as phonetic interruption cues but have a preponderance of non-verbal insertions, and there are others for whom the reverse applies. So we have to take speaker-specific preferences into consideration.

The analysis results of interruption glottalization in the German corpus differ in two important respects from the English data discussed in the literature. The relative frequency of glottalization overall and in word-fragments in particular is lower [1] [2], and the frequency of the absence of interruption cues is about as high as the frequency of glottalization. This means that although speakers may mark a syntactic irregularity by an abrupt phonetic cut-off in order to signal to the hearer that they are, e.g., going to correct themselves, they may also do the precise opposite and gloss over their false starts and truncations. So we should adopt a more differential view of the link between interruption glottalization and syntactic irregularities in that the use of different interruption cues or their absence may be related to changing intentions and/or situational constraints in one speaker, on the one hand, or characterize different speakers' behaviours, on the other.

3.2. Phrase-final laryngealization

There are 1633 instances of low terminal pitch contours at prosodic phrase boundaries. Of these, 127 can be connected with plosive-related glottalization, 205 with the phonetic realization of a following canonical glottal stop, and 180 are uncertain. This leaves 1121 cases for further analysis. Among these, 752 have modal voice, 359 laryngealization, and 10 end in a glottal stop.

The occurrence of a glottal stop in this prosodic position is negligible. Table 3 gives the distribution of the phonetic cues of laryngealization and modal voice, respectively, across the three types of phrase-finality: turn-final, turn-internal before pauses/breathing/(non)-articulatory noises, and turn-internal before verbal material. It shows a much higher incidence of laryngealization turn-final than turn-internal, and almost identical distributions of laryngealization and modal voice in the two internal types.

Table 3: Frequency distribution of laryngealization and modal voice across three types of phrase finality: TF turn-final, TINV turn-internal before non-verbal material, TIV turn-internal before verbal material.

	TF	TINV	TIV	
laryngealized	119	76	164	359
	57%	23%	29%	32%
modal voice	89	256	407	752
	43%	77%	71%	68%
	208	332	571	1111

Like interruption glottalization, phrase-final laryngealization also shows speaker-specific behaviour. There are speakers who have laryngealization quite regularly and others that have very few cases. Among the 22 speakers, 4 have relative frequencies below 10% (one 0%), 7 between 10 and 30%, 9 between 40 and 60%, and 2 have 70 and 72%, respectively. The group data, in conjunction with these individual distributions, suggest that speakers use laryngealization predominantly at the end of turns, and have diverging preferences for the use of this additional phonetic marker of terminal phrase finality.

4. Discussion

Interruption glottalization and phrase-final laryngealization differ in several respects:

a) Interruption glottalization includes the glottal stop quite frequently, laryngealization does not.

b) Interruption glottalization is associated locally with the point of interruption and sounds tense, whereas final laryngealization is realized over longer stretches, and sounds lax.

c) There are also differences of spectral characteristics between the two phenomena.

d) Laryngealization is always associated with low falling F0, glottalization occurs at the level in the F0 contour that has been reached at the utterance break, which is often high.

The impressionistic observations in b)–d) need systematic quantification as regards the extension of glottalization and laryngealization over time and number of sound elements (b), their differences in spectrum and intensity (c), and F0 context (d).

An important finding of this investigation is that both syntactic irregularities and phrase finality are signalled by multiple acoustic cues which are used in different combinations by individual speakers. Glottalization phenomena are optional markers in addition to, or instead of, other phonetic interruption features, and laryngealization is optional in addition to low terminal F0 and phrase-final lengthening. Both glottalization and laryngealization provide a strengthening of the respective signals for utterance breaks and phrase finality, and in the latter case, the turn-final position is given extra prominence. The cases for which no phonetic interruption has been recorded at syntactic irregularities require more detailed signal analysis to see whether a special pitch feature, e.g. a high F0 onset after the point of syntactic irregularity (as found in English by Bear et al. [1]), still cues a break, albeit more weakly. It must also be pointed out that classification into 'laryngealized' and 'modal voice' is very coarse, it needs further refinement into breathy beside modal voice and breathy laryngealization beside laryngealization (and possibly other phonation types).

One of the semantic functions of truncation glottalization is for speakers to indicate that they change plan and want to hold their turn before repairing or starting a new utterance. However, glottalization also occurs when a speaker is interrupted by another speaker, or attempts to take a turn without succeeding [11]. A subsequent investigation with further material from overlapping dialogues (compared with the non-overlapping ones of the Kiel Corpus) will be necessary to analyze these glottalizations in turn-holding and turn-taking strategies.

5. Acknowledgements

Part of the work reported here was funded by German Research Council Grant Ko 331/22-2 ("Sound patterns of German spontaneous speech"). We would also like to thank Michel Scheffers for writing programmes that helped us in the automatic processing of the corpus data.

6. References

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

[2] Nakatani, C. and Hirschberg, J., "A corpus-based study of repair cues in spontaneous speech", J. Acoust. Soc. Amer., Vol. 95, 1994, 1603–1616.

[3] Karger, R. and Wahlster, W., Verbmobil Handbuch – Version 3, Verbmobil Technisches Dokument Nr. 35, DFKI, Saarbrücken, 1995.

[4] Rodgers, J. E. J., "Three influences on glottalization in read and spontaneous German speech", AIPUK, Vol. 34, 1999, 177–284.

[5] IPDS, The Kiel Corpus of Spontaneous Speech Vol. 1–3, Institute of Phonetics and Digital Speech Processing, Kiel, 1995–1997.

[6] Kohler, K. J., Pätzold, M., and Simpson, A. P., From scenario to segment: the controlled elicitation, transcription, segmentation and labelling of spontaneous speech, AIPUK, Vol. 29, 1995.

[7] Kohler, K. J., "A model of German intonation", AIPUK, Vol. 25, 1992, 295–360.

[8] Pätzold, M., "*KielDat* – Data bank utilities for the *Kiel Corpus*", AIPUK, Vol. 32, 1997, 117–126.

[9] IPDS, *xassp* User's Manual (Advanced Speech Signal Processor under the X Window System), AIPUK, Vol. 32, 1997, 31–115.

[10] Kohler, K. J., "Plosive-related glottalization phenomena in read and spontaneous speech. A stød in German?", in Grønnum, N. and Rischel, J. (eds) To Honour Eli Fischer-Jørgensen, Travaux du Cercle Linguistique de Copenhague 31, C. A. Reitzel, Copenhagen, 2001, 174–211.

[11] Local, J. K. and Kelly, J., "Projection and 'silences': notes on phonetic detail and conversational structure", Human Studies, Vol. 9, 1986, 185–204.

Sound and Function Regularities in Interjections

Nikolinka Nenova, Gina Joue, Ronan Reilly, and Julie Carson-Berndsen

Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
{ nikolinka.nenova, gina.joue, ronan.reilly, julie.berndsen }@ucd.ie

Abstract

This paper investigates the relation between the sound patterns of interjections and their functional realisation in the discourse process. It considers whether certain interjection functions tend to have particular sound distributions. In order to address these questions a classification scheme for American English nonlexical interjections in terms of discourse markers is also presented.

1. Introduction

In the attempt to create a robust and relevant computational model for spontaneous speech interaction, speech system projects have only recently begun to consider dysfluencies as functional devices in the process of communication [1]. Save for the few instances in which interjections are analysed as part of the reparandum [2] or mentioned as back channelling moves [3,4], the contextual richness of interjection function has been hardly discussed [5,6]. Researchers also have casually but consistently noted that nonlexical interjections in different languages share phonetic similarities. For example, nonlexical interjections in English, Swedish [7] and Spanish [8] commonly involve infrequent or illegal phonotactic combinations. In a study involving Icelandic, English, Polish, Hungarian, Finnish, Ososo, Malagasi and Slovenian interjections, Abelin [7] noted that interjections in all these languages involve mostly labial or alveolar sounds. However, again, the phonological tendencies of nonlexical interjections have not been properly investigated.

This work contributes to filling in some of these functional and phonological gaps and demonstrates the sound regularities and the functional importance of nonlexical interjections in discourse. In this paper we contend that the sound patterns of interjections are dependent on their function, propositional meaning and position (both physical and contextual) in which they are realised in the discourse process.

Section 2.2 presents the phonological paradigm we used for the functional analysis of the constraints influencing the phonology of interjections. Then follows the analyses themselves and the discourse notions on which these analyses were based. The last section evaluates the analyses in the context of the suggested hypothesis.

2. Phonology of Nonlexical Interjections

We approached our investigation for a functional explanation of the constraints influencing nonlexical interjection based on Phonology as Human Behavior (PHB) [9,10]. PHB is a cognitive approach to phonology. Its aim is not simply to describe the systematic distributions in the sound structure of a language but also to explain these patterns. Appealing to functional and semiotic explanations, PHB purports that these patterns are directly shaped by the synergetic interactions of communicative and human physiological/behavioral constraints. That is, sounds in languages are not random because the (sometimes) conflicting goals of minimising articulatory effort and of maximising communication will tend to favour certain sounds over others. For example, most pause fillers are made up entirely by vowels (e.g. *uh*, *ah*, *oh*) as vowels require less effort to articulate than consonants. Distinctions among voiced vowels, however, become much more difficult (much subtler) with the increased number and variety of vowels which need to be distinguished in a phonological system. Therefore the speaker may have to increase efforts to enhance communication as vowels alone are limiting. Thus, although consonants are more difficult to articulate, they provide greater distinctions needed between vowels. Certain consonants and certain vowels will be more common than others. For example, consonants involving the lips and the tip of the tongue are easier to produce (and the lips being more visual so easier to perceive); therefore, these consonants occur most frequently in interjections across languages.

2.1. Interjection sound pattern hypotheses

Such a paradigm leads to a few hypotheses and explanations about the sound structure of interjections. For example, it supports Abelin's [7] observation that pause fillers tend to involve sounds produced by either the lips or the apex of the tongue, depending on their discourse function.

We hypothesise that interjections which signify **static** functions, that is those that do not change the current belief or knowledge of the participants or the intentional direction of the discourse moves (but merely indicate the speaker's attendance in the conversation, for example), will overall be much simpler and vary less phonetically than interjections indicating more **dynamic** participation. In other words, static-function interjections will most likely involve the most easily articulated sounds, which entails a more limited phonetic inventory, very simple syllable structures and most likely monosyllables. This hypothesis is motivated by the assumption that dynamic-function interjections indicate a speaker's willingness to increase articulatory effort for greater communicative holds and to produce particles with greater perceptual distinctions (or marked sounds). Likewise, static-function interjections imply more reluctance for too much articulatory effort or the avoidance of too salient sounds (or unmarked sounds).

3. The Analysis

The hypotheses outlined in the previous section, were formed to answer the following questions:

- Is there any significant difference in the sound distribution of the interjections in relation to their position in turns?
- Do certain interjection functions tend to particular sound distributions?

In order to test these hypotheses, we created a functional taxonomy for interjections that was simple enough for computational purposes but which also sufficiently captured the functions of interjections as discourse markers. We analysed the set of all interjections that were encountered in the TRAINS 91 corpus [11] based on this taxonomy. Although we did not have any phonetic transcriptions of the interjections, we assumed that the orthographic transcription of interjections are faithful to general English sound spelling rules and broadly examined them with the principles of PHB.

3.1. The choice of corpus

As was mentioned above, early research in spoken language systems filtered interjections as irrelevant to the process of communication. That is why most speech corpora transcribed for computational analyses have ignored interjections in transcription or were inconsistent in their transcription. This problem restricted our corpus choice to the TRAINS 91 dialogues. The TRAINS corpus provides orthographic transcriptions of the variations (e.g. *ohhh*) of interjection baseforms (*oh*) to approximate the actual token articulation of the given interjection. The transcription also includes overlapping speech, which, for example, was unfortunately not the case with the phonetically transcribed portion of the Switchboard corpus. The corpus is a collection of 16 task-oriented Wizard of Oz dialogues. The dialogues were approximately 80 minutes in length and included a balanced number of male and female American English speakers.

3.2. Function taxonomy

We view interjections as discourse markers, that is, the functions that they complete are based on the factors that constrain the discourse process. Three factors we identified are the information **direction** (new vs. old information), the **relation** or the hierarchical interdependency between the utterances in the dialogue (main vs. sub topic), and the participants' **intention** and expectations (what the move was intended as vs. what it was implemented as).

- The **direction** shows how the information currently presented is related to the one that has been already exchanged. When the utterance is related to a discussed topic, the direction is backward.
- The **relation** refers to the contextual position of the current utterance in the overall discourse hierarchy. It is a term that shows the focus of what *is being* said to what *has been* said. Relation realizations can be *start*, *finish* or *expansion*.
- The **participants' intentions** towards the dialogue move refer to the speaker's intention for the effect, which the current utterance would have on the other participant. When a speaker produces a move they expect this move to be responded to by a particular move or set of moves from the other participant. In our analyses this is further generalized to represent whether the utterance is intended towards the speaker themselves or the hearer. It specifies whether the utterance is a comment on current self-knowledge of the speaker or the current knowledge of the hearer. Participants' intentions may be *subjective*, where the utterance is an evaluation of self-knowledge; or they can be *objective*, which refers to evaluation of the other participant's knowledge. We also considered an additional factor: the participants' degree of evaluation of the ongoing discourse process. The degree of evaluation can be *positive*, *negative* or *neutral*. This factor is applied only to one of the functions (see Table 1)

Table 1: Function taxonomy

Function	Direction	Relation	Intention	Evaluation
Acknowledgement (Ack)	backward	finish	objective	neutral (AA), positive(AP), negative(AN)
Expansion (Exp)	forward	start, expansion	subjective (ES), objective (ER)	
Correction (Corr)	backward		subjective (CS) objective (CR)	

The interaction among these three factors establishes the three basic discourse functions (see Table1) (as opposed to syntactic or semantic). In this work we considered these functions to constrain the inference and intentional structure of discourse.

This taxonomy was used to annotate the nonlexical interjections in the TRAINS corpus.

3.3. Interjections in TRAINS

We first identified base forms by their relative high frequencies in contrastive **functional realisations**, that is, their functions (as based on the taxonomy in Table 1) and locations in turns. In correlating the variations of the interjections to their baseforms, articulatory or sound similarities are insufficient criteria because interjections with different functional realisations often have close sound structures. We found that the patterns of frequencies in functional realisations, in addition to sound proximity, provided a reliable method of identifying variants of baseforms even when frequencies were very low for some. Table 2 lists the interjections and their variants in decreasing order of frequency. Items in parentheses indicate very low frequency. Items in italics are sound synonyms.

Table 2: Non lexical interjections and their variants in TRAINS (Location: 0=constitutes turn, 1=at the beginning of turn, 2= within turn, 3 = at the end of the turn)

baseforms	variants	functions	locations
ah	(*hmm*)	Ack, Exp, Corr	2, 1
(aha)		Corr	2
(eh)		Corr / Exp	1 / 2
(err)		Corr, Exp	2,1
m-hm		Corr	0, 1
uh	uhh, uhhh, *uhm*, (*uhhm*), uhmm	Exp, Corr, Ack	2, 1, 3, 0
um	*mm*, umm, uumm, *uhm*, (*uhhm*)	Exp, Ack, Corr	2, 1, 3, 0
uh-huh		Corr	0, 2
hm	(*hmm*), m, (*mm*)	Exp, Ack	2, 1
oh	(o), (ohh), (oo), (ooh), (oooh)	Corr, Ack, Exp	1, 2, 3
(oops)	(whoops)	Corr	0, 1 / 2
[ouch]	(uch)	Corr	0
wow		Corr	1

Our analyses provide some support that interjections are context dependant and that their function is a combination of their position, their propositional meaning and the context in which they appear.

3.4. The relation between interjection position and function

Results show that the most frequent position is within the body of the utterance; however, most of these were self-expansion interjections (Exp). They show that the current speaker intends to further expand the utterance by contributing more information. The least frequent position is at the end of the utterance. Therefore, in general, interjections appear to prepare the listener in predicting the following utterances. The 2% that occur at the end are predominantly interjections, which speakers use for self-expansion (indicating an intended beginning of a turn) but were interrupted by the listener.

The second most frequent function of interjections is as indicators of change (Corr). The change includes self-repair or self-realisation (the most frequent in that type). The change of the direction of the information usually indicates that there is an update of the knowledge, or a change in the current state of the world or of the current topic in the communication. Like the general trend of interjection positions, these types of interjections tend to appear in the body of the turn; however, this case usually occurs at the beginning of a new utterance within the turn. The least frequent function of the three is that of acknowledgement (Ack).

3.5. The Phonetic Analysis

In order to test our hypotheses of the phonological properties of nonlexical interjections (Section 2.1), we classed AA, ES, and ER functions (see the taxonomy in section 3.2) as indicators of more static interaction, and the rest as more dynamic. We identified marked sounds depending on

- the complexity of syllabic structure and
- the acoustic/articulatory salience of the sounds making up the interjection.

The schwa is the most central position of the vocal tract for an American English speaker, and the closed lips the most neutral static position for no utterances; and not involve sounds of more effort such as very lengthened vowels or nonsonorant consonants. We took /m/ and the schwa to be unmarked sounds in American English. where marked of course is relative to the specific language's sound inventory. Marked sounds are rounded (e.g. *oh*), lengthened vocalisations (long vowels, *mmmmm*), noncentral or tense vowels, and nonsonorants (such as stops).

Figure 1: Relation marked/unmarked sound patterns in relation to interaction strength

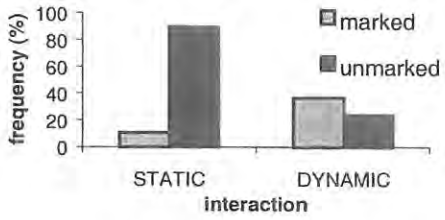
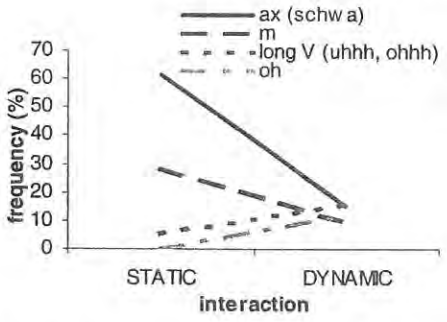
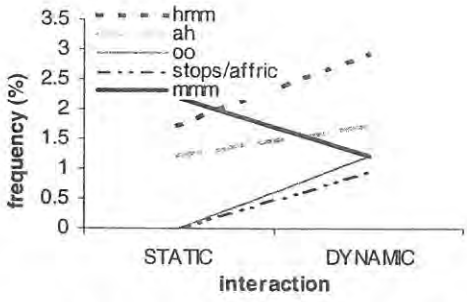


Figure 2: Relation of marked/unmarked sounds with strength of interaction





In support of our hypotheses, results confirm that the syllable structure of interjections indicating static interaction tended

away from multisyllabic forms (0.5% multisyllables) more than the dynamic interaction ones (3.4% multisyllabes). Results also show skewings indicating that the degree of markedness in the sound makeup of interjections relates directly to the degree of interaction (see Figure 1), as we also expected.

As seen in Table 2, our method of classifying marked/unmarked uncovers a direct relation between markedness and the degree of interaction in the discourse process. Figure 2 shows that sounds with less acoustic energy (such as *m*) tend to indicate less dynamic interactions (so they are mostly within turns) than those with more acoustic energy such as long vowels.

Almost all the interjections in the TRAINS corpus were monosyllabic (97.6%), as has been commonly observed in interjections across languages. The results show (Figure1) that static-function interjections (ES, ER, AA) tend to have less complex sound structures and less marked sounds than dynamic-function ones (Corr). Specifically, static-function interjections involve mostly the unmarked sounds: schwas and /m/s.

Likewise, in the group of the dynamic-function interjections include bisyllabic forms although, these are reduplication or minor variations of a very simple syllable structure. They also involve less of the perceptually weaker sounds such as schwas and more "marked" sounds. The lengthened forms *mmmmm* and *hmmmm* are neutral (such as giving the other participant a chance to interrupt) but also indicate more dynamic participation (and hence are at the beginning or at the end of turns).

Another example for the interrelation of function and sound choice are *mm-hm* and *uh-huh*. Both usually indicate more dynamic interactions. Thus, it is not surprising that they

are bisyllabic and are almost syllable reduplications. The /h/, however, also acts to increase the perceptual distinction of the second syllable from the first; without the aspiration, the speaker would have to place a pause between the *mm* syllables or a glottal stop between the *uh* syllables to ensure the perception of two syllables. Perhaps the additional syllable complexity is also balanced by the fact that both *mm-hm* and *uh-huh* involve the most neutral (least complex) sounds: /m/ and /ə/. Although /m/ is a labial nasal and thus more visual perceived and more naturally articulated, *uh-huh*, which involves a more open oral position and involving more effort function for more dynamic discourse purposes.

There were a few sounds whose frequencies were too low for us to draw any conclusions. However, as seen in Figure 2, our method of classifying marked/unmarked uncovers a direct relation between markedness and the degree of interaction in the discourse process. The only sound, which appeared not to match our predictions according to Figure 2, is the lengthened *m*, as we assumed that it is marked yet it appears more frequently in static interaction.

However, a difference does exist between lengthened /m/ and its shorter baseform. The lengthened /m/ occurs primarily at the beginning and the end of turns (thus marking the change in turns) whereas the shorter form occurs primarily within turns. This may imply that the sound structure of nonlexical interjections depend on *both* function and location and supports our hypothesis that marked sounds indicate more dynamic interactions.

4. Conclusions

In this paper, we analysed the relation between the phonetic structure and the pragmatic function that the interjections fulfil in the process of task-oriented communication. The consistencies in the sound structure of interjections in relation to their functional realisations lend support to the contention that interjections are discourse markers with functional and phonetic regularities. A stronger support of these regularities would be to conduct a cross-linguistic analysis on nonlexical interjections and investigate their dependencies on the language's sound inventory.

5. References

[1] Heeman, P.A., Byron D. and Allen, J.F. Identifying Discourse Markers in Spoken Dialog. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, Stanford, March 1998, pages 44-51. 1998

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

[3] Core, M., and Allen, J., Coding Dialogues with the DAMSL Annotation Scheme. In *Workshop Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*. 1997.

[4] Traum, D. R., *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis. University of Rochester, New York. 1994.

[5] Meteer, M. Dysfluency Annotation Stylebook for the Switchboard Corpus unpublished. 1995.

[6] Fisher, K., *From cognitive semantics to lexical pragmatics: the functional polysemy of discourse particles*. Mouton de Gruyter, 2000

[7] Abelin, A. *Studies in Sound Symbolism*. PhD Thesis. Gotehnburg Monographs in Linguistics, Göteborg University, Sweden. 1999.

[8] Montes, R.G. The development of discourse markers in Spanish Interjections. *Journal of Pragmatics*, vol. 31, pp.1289—1319. 1999.

[9] Diver, W. The Theory. In E. Contini-Morava and B. Sussman Goldberg (eds.), *Meaning as Explanation: Advances in Sign-Oriented Linguistic Theory*, pp. 43-113: Mouton de Gruyer, 1995.

[10] Tobin, Y. Phonology as Human Behavior: Theoretical Implications and Clinical Applications. Duke University Press: Durham and London, 1997.

[11] Allen, J.F. and Schubert, L.K. The TRAINS project. TRAINS Technical Note 91-1, University of Rochester, Department of Computer Science, 1991.

Filled pauses and their status in the mental lexicon

Richard Shillcock[*‡], *Simon Kirby*[†], *Scott McDonald*[‡] *& Chris Brew*[**]

[*]Department of Psychology, [‡]Division of Informatics, [†]Department of Theoretical an Applied Linguistics,
University of Edinburgh, Scotland
[**]Department of Linguistics,
Ohio State University
rcs@cogsci.ed.ac.uk, simon@ling.ed.ac.uk, scottm@cogsci.ed.ac.uk, cbrew@ling.ohio-state.edu

Abstract

We report a study of the relationship between form and meaning in the most frequent monosyllabic words in the lexicon of English. There is a small but significant correlation between the phonological distance and the semantic distance between each pair of words. To this extent, words that have similar meanings tend to sound similar. Words differ as to the size of this meaning-form correlation in their relationship with all of the other words. When the words are ranked according to the size of this correlation we find that the words which appear towards the top of the ranking are the communicatively important words. When we look at the position in the ranking of the speech editing terms, such as *er*, *oh* and *um*, we find that they are at the very top of the ranking. We argue that this position reflects the communicative importance of these items, and that it therefore makes sense to treat them as a proper part of the mental lexicon.

1. Introduction

Intuitively, speech editing terms like *um* and *er* seem to be barely worthy of inclusion in the lexicon at all. Even though they have a clear, important communicative function [1,2], it is easy to think of them as default noises of the speech production system. In this paper we explore the implications of considering such speech editing terms as lexical entries and we investigate their relationship with the rest of the lexicon.

We adopt a new perspective on the structure of the mental lexicon, which we motivate in terms of the representational strategy predominantly found in the mammalian brain. We will characterize this strategy as "structure-preserving". We see this strategy clearly in the processing of the visual world, which generates numerous topographic representations within the relevant areas of the brain. One of the characteristics of a topographic representation is that the representation of any one thing may be "triangulated" in terms of the representations of two similar things. To the extent that *C* is quite like *A* and a little like *B* in the world, the representation of *C* will be quite like *A* and a little like *B* in the brain. Language differs from, for instance, the visual world in that its processing requirements may be expected to reflect much more the representational predispositions of the human brain. The structure of the vocal apparatus and the physical medium of speech transmission are givens, and so is the scale of the problem of reference – the sheer number of different things to which we need to refer – but beyond these constraints human language is a process by which two brains communicate and we may expect to find the processing propensities reflected in the structure found in human language.

Human language is replete with structure. There are numerous possible levels of description, from the phonological to the pragmatic. Linguists have attempted to describe the structure found at each of these levels, but one crucial domain has remained outside this enterprise: the relationship between form and meaning. It has been widely accepted that the relationship between form and meaning is arbitrary. This assumption is codified in Saussure's "arbitrariness of the sign" [3]: in principle, English could carry on just as well if the referring expressions *chair* and *swan* were interchanged, *ceteris paribus*. There is nothing essentially chair-like that makes *chair* the best expression to use to refer to chairs. It is widely accepted that there are only marginal exceptions, such as onomatopoeia, to the arbitrary relationship between form and meaning. We will argue below that this belief reflects only the absence of any means of studying this relationship.

We can begin by asking what relationship between form and meaning the brain might be expected to prefer. We claim that the default relationship should be a transparent, structure-preserving one in which words sound similar to the extent that they mean similar things. If the language were built on this principle it would be maximally easy to learn and the mental lexicon would be maximally easy to organize. To the extent that a lizard resembles a snake and also a crocodile, the phonological form of the word for lizard would resemble those for snake and crocodile. The drawback to this principle lies in the fact that adult speakers need to refer to tens of thousands of different things. If it were possible to impose a structure-preserving principle comprehensively on the mental lexicon, it would mean that the resulting words would be too long to be useful. We need, therefore, to make a weaker claim, along the lines that the structure-preserving principle may be only weakly expressed in the relationship between form and meaning, or that it may only be apparent for certain words that may be taken to constitute, in some way, the "core" of the lexicon.

Below, we summarise how we have tested this hypothesis that there is a non-arbitrary, structure-preserving relationship between meaning and form. We will show that there is indeed a small but significant relationship between form and meaning in the substantial fraction of the mental lexicon which we test. We go on to explore the role of the speech-editing terms within this relationship.

2. The meaning-form relationship

We required characterize both the phonological distance and the semantic distance between any two words. Only when we had a precise number for each distance for every possible pair of words could we test for any overall correlation between the two distances. We were not concerned with relationships between words that are explicitly morphologically related, such as *walk* and *walks* or *telephone* and *telegraph*, or with words that would seem to have some historical connection, such as *circle* and *circuit*. The hypothesis we were testing involved the meaning-form relationships between superficially completely unrelated words. It is not possible, or perhaps even desirable, to exclude all traces of historical connectedness between words. Thus, *could* and *would* appear to be part of some paradigm, and *glow*, *gloom*, *glare*, ... seem to have some common proto-Indo-European ancestry. To resolve this issue we tested only the words classified in CELEX as monosyllabic and monomorphemic. .

2.1. Calculating Phonological Distance

In calculating phonological distance we were calculating the number of changes necessary to turn one word into another, to turn *chair* into swan, for instance. The phonological edit-distances were calculated using the distinctive feature representations from the Festival Speech Synthesis System[1]. Festival uses eight features: one distinguishes consonants from vowels; four classify vowels by length, height, frontness and lip-rounding; three classify consonants by type (taking the values of stop, fricative, affricate, nasal, lateral and approximant), voicing and place of articulation. Using these features we created a mismatch function assigning a penalty to each pair of two phones. Vowel features, such as length, are naturally scalar. For these we assigned a penalty of 1 for a small mismatch, and a penalty of 2 for a large mismatch. Consonant features, such as voicing, were treated as nominal variables: all mismatches attracted a uniform penalty of 1 on each dimension. Pairs involving one consonant and one vowel were assigned an additional penalty of 10. More complex, and more psychologically realistic penalty schedules could be defended on phonetic grounds, but this relatively simple one was chosen because it involves fewer assumptions. This penalty schedule was created on the basis of phonetic knowledge alone, without considering specific words and without knowledge of the semantic distances. It was never varied after its initial creation. This procedure produces a set of phone-phone distances. Distances between words were generated by applying the Wagner-Fischer edit distance algorithm [4] using the penalty schedule described above, augmented with a uniform penalty of 5 for deletions and insertions. This procedure provided us with a precise measure of the phonological distance between each word and every other word.

2.2. Calculating Semantic Distance

In calculating the semantic distance between any two words, we required a definition of meaning that could be applied equally to any word in the lexicon and also one which could be expressed quantitatively. The definition we employed was based on defining the meaning of each word in terms of the context in which it occurred. This approach echoes philosophical claims that word meaning can be defined in terms of usage [5,6]. The approach we employed has been used recently to model a variety of language processing behaviours [7,8,9]. Any use we make of the words "semantic" and "meaning" refers only to this definition, unless stated otherwise.

A semantic space was constructed from the distributional information contained in the 100 million words of contemporary English comprising the British National Corpus (BNC)[2]. Inflected words were replaced by the corresponding lemma (e.g. *walks* by *walk*). Lemmatisation reduces the sparseness of the distributional data but preserves the meaning shared by the inflectional variants. To create a context vector representation for a given word, we recorded the frequency with which each member of a list of 500 "context" words occurred within a window of five words before and five words after the critical word. Every word can be represented as a 500-dimensional context vector, but the reliability of the vector representations diminishes with word frequency [7]; hence we computed context vectors for only the 8000 most frequent words in the BNC. We defined the semantic distance between any two word vectors in the resulting high-dimensional semantic space as (1 - cosine of the angle between vectors). Word pairs that have similar distributional properties (i.e. they co-occur to a similar extent with the same set of words) receive a low semantic distance score, whereas the semantic distance is larger between words that have diverging distributional properties. This procedure provided us with the semantic distance between each word and every word in our study.

2.3. Calculating the Meaning-Form Correlation

We tested for the existence of a significant correlation between semantic distances and phonological distances across all the words in our sample. Of the 8000 most frequent words for which the BNC gave us reliable semantic distances, some 1733 were monosyllabic and monomorphemic. Although 1733 words only constitutes a small fraction of any adult native speaker's English lexicon, it still represents almost two-thirds of the number of word-types in the spoken part of the BNC. It is therefore a psychologically significant fragment of the lexicon of English.

It is important to realize that we were concerned with the distances between each word and every other word – with all the possible distances in the mental lexicon. For the 1733 words this means 1,500,778 pairs of phonological and semantic distances. This perspective gives us a comprehensive picture of the relationship between meaning and form in the mental lexicon. Previous psycholinguistic research has been almost exclusively concerned with the relationships between words for which there was some overt connection, such as semantically related words like *lion* and *tiger* or phonologically/orthographically related words like *hand* and *land*.

The correlation, *r*, between the semantic distances and the phonological distances was 0.061. There is thus only a very small overall correlation between meaning and form. We had predicted that only a very small overall correlation could be

[1] http://www.cstr.ed.ac.uk/projects/festival/

[2] http://info.ox.ac.uk/bnc/index.html

possible, due to the large number of necessarily short words in the lexicon. We tested, first, whether even this small correlation was statistically significant. We then tested whether this correlation was in some way concentrated in a particular subset of the lexicon.

Calculating the statistical significance of the correlation between 1,500,778 pairs of distances is problematic. First, the numbers within each set of measurements are not independent. Second, the number of pairs of measurements is very large and it is unclear how we should expect conventional tests of significance to behave with such numbers. The solution to these problems is to use a randomization test [10]. First, each word was randomly assigned a single, unique partner in the set. Second, the correlation was calculated using each word's own semantic context vector and its partner's phonological form, thereby randomising the relationship between the two types of distance. This procedure was repeated to generate the distribution of correlation coefficients expected by chance over 1000 randomisations. The veridical value of r was an outlier on this distribution ($z = 4.2$, $p < .001$ (one-tailed)). In summary, there is, therefore, a small but nonetheless significant relation between form and meaning across this substantial fraction of the lexicon of conversational English. The randomization test is tailored to the precise conditions and nature of the materials being tested and provides a maximally transparent test of the significance of the meaning-form correlation.

The overall correlation was highly significant, but it was also very small. Only a very small fraction of the overall variance within the distances was captured by the correlation between the two distances. We next tested the hypothesis that this correlation would not be evenly distributed across the words in our set, but would be concentrated in some particular subset of the lexicon.

For each word there were 1732 pairs of phonological and semantic distances relating it to all the other words in the set. We calculated this correlation for each word. Although the overall correlation was small, the correlations for the individual words ranged from $\tau = +0.135$ to -0.082, following an approximately normal curve positively displaced from zero. For r, the correlation ranged between +.189 and -.115. When these individual correlations were ranked, they revealed a striking ordering. The order was clearest in the ranking by r, but the same qualitative pattern obtained for τ. We discuss the ranking by r here.

The speech editing terms *oh*, *er* and *ah* were in the five most highly ranked words, ranked by r, followed by *eh* in 13th place. For a comprehensive report of the results and an extended discussion, see [11]. Overall, four psychologically important categories of words were clearly skewed towards the top of the ranking by r: speech editing terms, pronouns, proper names and swear words. We claim that these categories are all communicatively important.

3. Discussion

We have shown that, contrary to the orthodox assumption concerning the arbitrariness of the sign, there is indeed a small but significant relationship between meaning and form: words that have similar meanings tend to sound similar. Although this correlation is very small overall, it is an order of magnitude larger when the relationship of individual words with the rest of the lexicon is considered. What does this individual correlation mean? What does it mean for *oh* to have a correlation of $r = .189$ between its phonological distances and its semantic distances to the other words? It means that the rest of the lexicon conspires with the phonological form of *oh* to specify its meaning, or with the meaning of *oh* to specify its form. It means that the rest of the lexicon helps to cement the meaning-form relationship for *oh*. For a word down at the other end of the ranking by this correlation, the rest of the lexicon does not play such a role. The words *friend*, *twelve* and *frank* are the last words in the list, and they receive no such support from the rest of the lexicon for their own meaning-form relationships. When we consider what factors might determine position in the ranking by meaning-form correlation, a number of possibilities suggest themselves. Indeed psycholinguists have generated a range of dimensions of lexical variation as potential determinants of processing difficulty. Such variables include word frequency, word length, concreteness, imageability, age of acquisition and so on. A casual inspection of the ranking suggests that the words that occur towards the top of the list tend to be the shorter words, and the words which have a "less propositional" meaning, although there are very prominent exceptions – *a*, *the* and *and* all occur down near the middle of the list. More detailed analysis of the ranking by meaning-form correlation shows that the ranking reflects word length most strongly, followed by a subjective measure of familiarity. Elsewhere [11] we suggest that the correlation between meaning and form is the most basic relationship that can be tested in the lexicon, and that it is functionally significant that the communicatively important words tend to be strongly skewed towards the top of the ranking. This skewing is adaptive when considered from the perspective of how any one word is stored and accessed. Words are essentially stored superpositionally in the brain. The neural substrate over which any one word's form and meaning are stored may be quite extensive and certainly participates in the storage of many other words. In this sense we can talk about the whole of the lexicon being involved in the access of any one word. A word at the top of the ranking by the meaning-form correlation is thus stored the most securely and its representation receives indirect "assistance" from the storage of the other words.

The speech editing terms finish up at the very top of the ranked list not just because they are short and not just because they have positions in semantic hyperspace characteristic of words whose context is very varied. They are at the top of the ranked list because their phonological forms are the appropriate ones for their associated meanings.

4. Conclusions

We conclude that the speech editing terms found in the BNC can be motivated as a proper part of the mental lexicon. Their individual relationships between meaning and form are such that, in the context of the rest of the frequent monosyllabic words, they take their place as some of the most communicatively important elements of the mental lexicon.

3. References

[1] Fox-Tree, J.E. & Schrock, J.C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language, 40*, 280–295.

[2] Fox-Tree, J.E. (2001). Listeners' use of um and uh in speech comprehension. *Memory and Language, 29*, 320–326.

[3] de Saussure, F. (1916). Cours de Linguistique (eds. Bally & Sechehaye) Générale. Paris: Payot.

[4] Wagner, R.A. & Fischer, M.J. (1974). The string-to-string correction problem. J. Assoc. Computing Machinery 21, 168-173.

[5] Cruse, D.A. (1986). *Lexical semantics.* Cambridge University Press.

[6] Wittgenstein, L. (1958). *Philosophical investigations.* Oxford: Blackwell.

[7] McDonald, S. (2000). Environmental determinants of lexical processing effort. PhD thesis, University of Edinburgh.

[8] Lund, K., Burgess, C. & Atchley, R.A.. Semantic and associative priming in high-dimensional space. In (J.D. Moore & J.F. Lehman) *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, Erlbaum: Hove, UK. pp 660–665 (1995).

[9] Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods, Instrum., & Comput.*, **28**, 203-208 (1996).

[10] Cohen, P. R. (1995). Empirical methods for AI. MIT Press, Mass.

[11] Shillcock, R., Kirby, S., McDonald, S. & Brew, C. (*in preparation*). Exploring the structure-preserving nature of the mental lexicon.

The Double Function of Disfluency Phenomena in Spontaneous Speech

Mária Gósy

Kempelen Farkas Speech Research Laboratory, Linguistics Institute,
Hungarian Academy of Sciences, Budapest, Hungary
H-1068 Budapest, Benczúr u. 33., Hungary

Abstract

Disfluency in spontaneous speech is the outcome of a speaker's indecision about what to say next. The listener, however, is continuously adapted to both the language signals and the types of disfluency of the heard text. What is in the background of this adaptation process?

This paper analyses the types and characteristics of the disfluency phenomena of a 78-minute spontaneous speech sample (produced by 10 adults). The author's intention is to explain the characteristics of disharmony between speech planning and articulation within the speech production process. In order to explain the hypothesized double function of disfluency in terms of perceptual necessity from the listener's side various experiments have been carried out.

Three different samples of spontaneous speech have been selected for experimental purposes. Three groups of listeners (altogether 60 university students) participated in the experiments. One of the groups had to detect the instances of disfluency in the texts marking them on a paper sheet. The subjects of the other group listened to the same texts and then wrote down their contents. The pauses and hesitations were then eliminated from the texts. The third group of the subjects had the same comprehension task as the previous one had.

Results show that (i) instances of disfluency are consequences of the speaker's speech planning processes, (ii) their reasons and occurrences are unconsciously known by the listener as well, (iii) disfluency phenomena are relatively well predicted, (iv) the listeners need pauses and hesitations in order to comprehend the heard texts successfully.

Keywords: silent pauses, hesitations, speech planning process, comprehension strategies.

1. Introduction

Spontaneous speech is characterized by several phonetic processes like co-articulation or the variability of the phonetic form of words, and by various types of disfluency phenomena (silent and filled pauses, hesitations, prolongations of parts of an utterance, false starts, repetitions). Disfluency in spontaneous speech is the outcome of a speaker's indecision about what to say next and so shows his struggle to achieve control over planning, production, articulation.. All types of disfluency provide a useful approach to study the speech production process and so the complex phenomenon of disfluency has been widened since focusing on the types of hesitations [1]. In order to better understand the nature of disfluency we need to consider also the cognitive processes of planning and performing utterances.

The listener, however, is continuously adapted to both the language signals and the types of disfluency of the heard text. What is in the background of this adaptation process? Does the listener need disfluency in order to comprehend the text better?

Over the last 40 years, scholarly inquiry into the characteristics of various types of pauses (i.e. silent vs. filled ones) has steadily increased while many papers were devoted to other phenomena of disfluency, respectively [3]. Mahl was the first in 1956 who differentiated two types of disfluency labeling one of them *ah*-phenomenon while the other one non-*ah* phenomenon [2]. According to his definition the first category shows the speaker's uncertainty about what to say next while the second one covers other phenomena like repetitions, changes of structures, etc. Despite the widening interest, however, the results have not been enlightening, there has been no agreement on important facts, and the perceptual aspects have frequently been ignored. Among the results very little can be found about the functional aspects of disfluency in speech production and speech perception as well as in comprehension [4].

The most frequently occurring type of disfluency in fluent speech, independently of the language, is silent pause, then filled pause and the combination of these two. There are well-known factors that result in silent pauses: (i) breathing, (ii) intention of interpretation of the text, (iii) pauses determined by syntax, emotion, rhetorical and expressive emphasis, stylistic properties, and (iv) results of various kinds of disharmony between speech planning and articulation. The nature and function of juncture seems to be a matter of disagreement [5]; we understand it as a special case of "cognitive" pauses. The temporal aspects of spoken language raised the question of the duration of pauses. At the beginnings, articulatory pauses shorter than 250 ms were accepted as pauses according to the rationale given by Goldman Eisler [1]. This value was also associated with avoiding the confusion with pauses being parts of a segment (like the silent periods of voiceless stops). Silent pauses as types of disfluency are accepted generally with the durations of 130 and 150 ms [6]; however, the values of 200 and 300 ms can also be found in the literature [5, 6].

The distribution of disfluency phenomena seems to show similar tendencies across languages [7]; most of them are various kinds of pauses. In dialogues pauses appeared in 32% of all speaking time where 25% was silent while only 7% was filled pause. Another study revealed that disfluency appeared also in dialogues in every 42 seconds [8]. More research is required to learn more about disfluency being a universal or a language-specific phenomenon. There are many other questions that need answers. Should tongue tip be considered as a specific type of disfluency? Are pauses and hesitations

specific signals for the listener to perform his decoding strategy better? Or, on the contrary, do they disturb the decoding process? The present paper aims at answering these questions by carrying out several experiments with Hungarian-speaking speakers and listeners.

2. Method, material, subjects

A 78-minute spontaneous speech sample produced by 10 Hungarian speaking adults (5 females and 5 males) served as speech material. No hearing or speech defects were reported with any of the subjects. After a short period of introductory communication the subjects were asked to speak about their work. None of them were aware of the aim of the requested task, however, everybody knew that their speech was being recorded. The whole material was analyzed with respect to the occurrences and types of disfluency phenomena. Acoustic-phonetic analysis was carried out by means of the Kay Elemetrics CSL 4300B digital system.

Three different samples of spontaneous speech have been selected for experimental purposes. The data of tempo and pauses of the three texts is shown in Table 1.

Table 1: *Pauses of the three text samples (AT=articulation tempo in sounds/s, ST1=speech tempo in sounds/s, ST2= speech tempo in words/minute)*

Texts	AT	ST1	ST2	Pause time (ms/item)
1st	15.29	11.83	125.7	13.9/24
2nd	15.58	12.26	168.1	8.05/29.77
3rd	11.47	9.56	112	11.95/26

Three groups of listeners (altogether 60 university students, females and males) participated in the listening experiments. One of the groups (with 20 subjects) had to detect the instances of disfluency in the texts marking them on a paper sheet. They heard each text two times, and were asked to draw a vertical line where they thought they found the place of disfluency. The subjects of the other group (another 20 subjects) listened to the same texts and then wrote down their contents. The pauses and hesitations were then carefully eliminated from the texts (by means of the same Kay Elemetrics CSL system). The third group of the subjects (the last 20 subjects) had the same comprehension task as the previous one had. The written narratives of these two groups were compared. Analyzing the narratives the number of words used and the main ideas of the text samples were taken into consideration. To test statistical significance and statistically relevant interrelations, a two-way ANOVA was used (SPSS 8.0 for Windows statistics package).

This paper aims at investigating (i) the frequency, phonetic types and acoustic properties of disfluency phenomena of Hungarian spontaneous speech and (ii) the function of pauses from the aspects of speech production and speech comprehension. Our hypothesis is that pauses bear a double function in speech; they solve the disharmony between speech planning and articulation on the one hand and provide a better strategy for the listeners in comprehension, on the other hand.

3. Results

3.1 Disfluency phenomena. The phonetic analysis of our material showed that the following 6 types of disfluency phenomena could be detected covering 37% of the texts on average: silent pauses; hesitations; repetitions at various levels of speech, e.g. vowels, syllables, words (both function and content words), word combinations; prolongations (particularly at the end of phrases and utterances); alterations and changes (both structural and semantic); and false starts. Slip of the tongue was considered as a disfluency category. Almost all categories of disfluency could be detected with all speakers though the occurrences showed enormous individual differences ($p<0.001$). There was one (male) who had no filled pauses, three others (two females, one male) with whom neither prolongations nor false starts existed while 7 subjects did not struggle with slips of the tongue. Silent pause has a marked occurrence in all speakers' speech production. Taking disfluency phenomenon as an 'umbrella term' to be 100%, the occurrence of the 7 types shows this distribution across all speakers (on average), cf. Table 2.

Table 2: *Types and occurrences of disfluency phenomena across all speakers*

Disfluency type	Occurrence (%)	Range (%)	sd
silent pause	44.57	22.7-77.9	17.5
hesitation	12.37	0-25.7	7.84
repetition	3.25	1.3-5.33	1.51
prolongation	1.77	0-4	1.44
alteration, change	3.05	1.06-5.2	1.43
false start	1.34	0-2.85	1.06
slip of the tongue	0.36	0-1.2	0.48

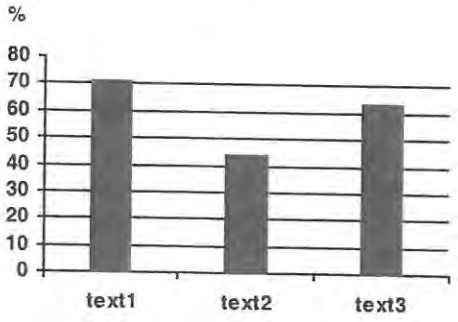
Comparing our results to those obtained with other (analyzed) languages, some interesting differences should be mentioned. Prolongations appear frequently on definite articles (*a, az*), false starts occur quite often when definite articles should be replaced (by content words), and alterations/changes concern both the words (stems) and various inflections that come from the strongly agglutinative character of Hungarian (e.g. *mentek .. mentünk* or *filmet .. filmeket*). Analyzing the correlations between paired types of disfluency, significant correlation was found only between false starts and changes ($p<0.002$).

3.2 Experiment I. For the sake of this experiment the duration of silent pauses had to be defined. Silent periods were taken to be pauses when 65% of all the listeners identified them between two lexical items. In these terms, the duration of the shortest pause was 80 ms as a perceptual agreement among the listeners.

All the hesitations – in Hungarian with the vowel [ø] and sometimes with the consonant [m] – were detected 100% correctly. Subjects were able to detect unfilled pauses in 60% of all occurrences that shows better performance than that reported in the literature (cf. 28,2%) [2]. The identification of silent pauses, however, was not independent of the actual text sample (Fig. 1), and the difference proved to be significant ($F(2, 54) = 17.238$, $p<0.0001$). The faster the speech the less

correct the perception of pauses with all listeners. However, it seems to be contradicted by the results of text1. Text1 was faster than text3 but the identification scores of pauses for text1 were slightly better than for text3. Acoustic analysis revealed that the average duration of pauses was different in the two texts: 551 ms on average in text1 while 411.7 ms in text3. The speaker of text3 produced pauses independently of structural boundaries in about 40% of all occurrences while the speaker of text1 only in 15%. This means that not only the actual speech tempo but also the average duration of pauses and their place influence the listener's correct perception.

Figure 1



Correct identification of silent and filled pauses of the three text samples

There was no significant difference in correct identification of pauses between females and males but there was a slight tendency indicating that females show more sensitivity toward correct pause perception. However, there was a significant difference concerning the false markings of pauses between females and males in text1 and text2 ($F(1, 54) = 10.205$, $p < 0.002$). The quantity of false markings was bigger with females than with males which indicates that females felt the texts to be more dissected (Fig. 2).

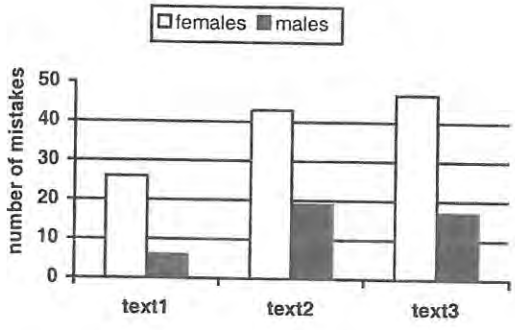


Figure 2
Number of false identification of pauses in the three texts

There was a strong correlation between the duration of pauses and their correct identification as it had been hypothesized ($r_p = 0.714$, $p < 0.0001$). The longer the pause the more correct its identification (Fig. 3). Again, factors other than actual duration also influenced the correct identification of pauses like their place within an utterance. Listeners were able to identify pauses that appeared at a structural boundary or phrase boundary significantly better (75% correct responses) as opposed to those that appeared elsewhere (36.8% correct identification).

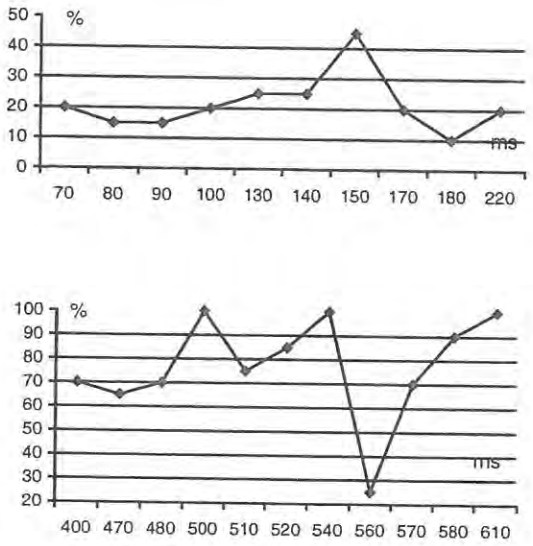


Figure 3
Interrelation of the duration of pauses and their identification (70-220 ms and 400-610 ms)

3.3 Experiment II. Recall accuracy was analyzed in the subjects' written narratives where the mean number of words and the main idea-units were taken into consideration. The number of words refers also to the quantity of narratives; the more the listener comprehended the more he could write down. Results show a significant difference depending on the presence vs. absence of pauses in the texts (Fig. 4). (Statistical analysis for words: $t(19) = 5.936$, $p < 0.0001$.)

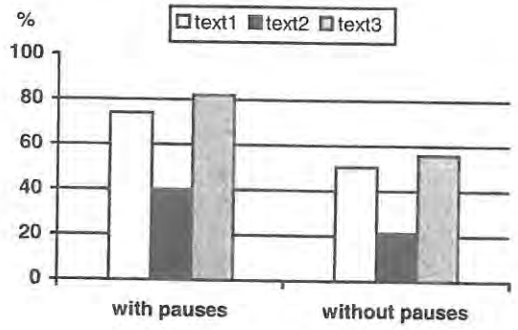


Figure 4
The differences of the number of words in the narratives

The number of words and the number of main idea-units of a narrative based on verbal comprehension predict the performance level of correct comprehension quite well [9]. A lot of extra information was found concerning the details with

those subjects who listened to the texts with their original pauses. No such information could be traced, however, in those subjects' narratives who had no pauses in their heard texts. Figure 5 shows the differences of the main ideas in percentages across the three texts (statistical analysis for the main idea-units: $t(19) = 8.320$, $p < 0.0001$).

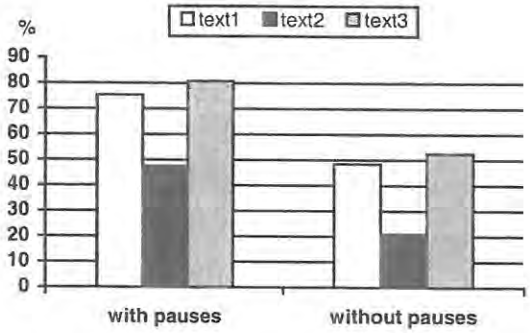


Figure 5
Differences of the main idea-units recalled in the narratives

The analysis of false statements revealed that subjects made three times more mistakes if there were no pauses in the heard texts. All these results indicate that the listener needs pauses and consequently other types of disfluency phenomena to a certain extent in order to fully comprehend the heard texts. It can be assumed, however, that too many interruptions will affect the comprehension negatively, as well.

4. Conclusion

Acoustic-phonetic analysis of spontaneously uttered texts revealed that disfluency phenomena are characteristic of the speaker and are relatively independent of the actual articulation tempo. The types and occurrences of disfluency seem to be (i) speaker-dependent and (ii) language-dependent to a certain extent. This latter means that language structure affects speech planning and speech production processes that result in specific occurrences of disfluency phenomena. In our material the most frequent types of disfluency were filled and unfilled pauses with the duration range of 70 and 1200 ms.

Listeners identified the majority of silent and filled pauses independently of their frequency, of the topic of the text, or of the speaker's fundamental frequency characteristics. There is a close correlation between the duration of pauses and their correct identification, however, a lot of other factors influence the listeners' final perceptual judgments like grammatical, syntactical predictions, semantic presuppositions, effects of suprasegmental patterns, actual articulation, etc. False pause identifications originated in (i) their unnoticed existence and (ii) their identification where no pauses existed. Sex differences were significant in this respect.

The listener, decoding verbal messages, activates all his language (and other stored) knowledge in order to follow the acoustic flow of heard speech. In order to construct meaning from a continuous acoustic stream he unconsciously controls the process of speech comprehension by his own inner speech production. The listener behaves as speaker when he/she processes speech. Objective data confirm the activation also of the Broca-area in the left hemisphere when comprehending speech. Successful comprehension requires that the listener should identify words, detect syntactic structures, and extract meaning from individual sentences, and finally build relations among the various parts of the text. Pauses are meant to provide time for such operations on the one hand, and for correction processes if they are needed, on the other hand. If the context does not enable the listeners to "fill in" the gaps created during the comprehension process, pauses or other types of disfluency can take over the role. This does not mean that speech is incomprehensible without pauses but it does mean that comprehension performance is significantly restricted in their absence. In this case comprehension is focused on key words and/or on a limited number of ideas the text contains. This would lead to an uncertain decoding process that results in a semantically restricted outcome with the possibility of false statements. It seems to be a paradox that two contradictory facts support the same explanation, i.e. pauses are not all identified, however, pauses are needed for correct decoding operations. If they are needed why cannot they be perceived absolutely correctly? Pauses can be ignored both perceptually and functionally during comprehension when they are not useful on the one hand but they are assumed to be used as parts of the decoding strategy even if they are consciously unnoticed, on the other hand.

Our results show that (i) instances of disfluency are consequences of the speaker's speech planning processes indicating their language-specific nature, (ii) they are relatively well predicted, (iii) their reasons and occurrences are unconsciously known by the listener as well, and (iv) the listeners need pauses and hesitations in order to comprehend the heard texts better. On the basis of these results the hypothesis of the double function of disfluency in spontaneous speech is confirmed.

5. References

[1] Goldman Eisler, F.: Psycholinguistics. (Experiments in Spontaneous Speech.) Academic Press. London 1968.

[2] Mahl, G.: Disturbances and silences in the patient's speech in psychotherapy. Journal of Abnormal and Social Psychology 53. 1956, 1-15.

[3] Hieke, A.E.–Kowal, S.–O'Connel, D.C.: The trouble with "articulatory" pauses. Language and Speech 26. 1983, 203-219.

[4] Duez, D.: Silent and non-silent pauses in three speech styles. Language and Speech 25. 1982, 11-25.

[5] Laver, J.: Principles of Phonetics. Cambridge University Press. Cambridge 1995.

[6] Shapley, M.: Prosodic variation and audience response. Papers in Pragmatics Vol. 1. No. 2. 1987, 66-80.

[7] Misono, Y.–Kiritani, S.: The distribution pattern of pauses in lecture-style speech. Logopedics and Phoniatrics 2. 1990, 110-113.

[8] Horga, D.: Samoispravljanje u govornoj proizvodnju. Suvremena linguistika 23/1-2. 1997, 91-105.

[9] Wingfield, A.– Tun, P.A.–Rosen, M.J.: Age differences in veridical and reconstructive recall of syntactically and randomly segmented speech. Journal of Gerontology Vol. 50. 1995, 257-266.

Do non-word disfluencies affect syntactic parsing?

Karl G. D. Bailey and Fernanda Ferreira

Department of Psychology and Cognitive Science Program
Michigan State University
karl@eyelab.msu.edu

Abstract

Although disfluencies such as *uh* are generally not treated as linguistic items, our results suggest that they can affect syntactic parsing. Using a grammaticality judgment task, we demonstrate that disfluencies are able to affect the syntactic parse of a sentence in two ways. First, disfluencies can make syntactic reanalysis more difficult by coming between an ambiguous constituent and a disambiguating item. Second, the pattern of disfluencies in spontaneous speech may be used by the listener to guide the parse of a sentence. Thus, although disfluencies have often been viewed as pragmatic phenomena, they can affect the language comprehension by influencing its parsing procedures.

1. Introduction

Comprehension of spontaneous speech is not a simple as it might initially seem. A comprehender must break down a continuous speech stream into component parts, then build a syntactic structure and determine the meaning that the speaker intended to convey. Models of language comprehension all attempt to describe this process. However, our models of speech comprehension also assume that the utterances that make up spontaneous speech are ideal deliveries [1], that is, the kind of language that we see in grammatical polished writing. In theory, disfluencies could be filtered by some component of the language comprehension system before syntactic parsing occurs; however, our data strongly suggest that utterances with disfluencies are parsed differently from those without disfluencies.

In this series of experiments, we examine whether a filled pause disfluency, *uh*, that is generally believed not to be a word or syntactic constituent [2] can have the same effect on syntactic reanalysis as linguistic material. We then examine two possible explanations to account for how disfluencies could affect parsing.

The syntactic reanalysis effect examined in this series of experiments is known as the head noun position effect. The head noun position effect [3] occurs in garden path sentences such as (1).

(1) While the boy scratched the (A) **dog** (B) *yawned* loudly.

A garden path sentence is one which is temporarily ambiguous. The ambiguity can be resolved in one of two ways, one of which is preferred. If the parser incorrectly assumes the more preferred structure, as in (1), the parser is forced to reanalyze this structure. Reanalysis is not always successful, and when it fails the subject will call the sentence ungrammatical. A baseline garden path sentence like (1) is rated as grammatical 61% of the time in reading studies [3]. In this sentence the ambiguous noun is in bold and the disambiguating verb is in italics. If modifiers such as *big and hairy* are placed at position (A), the proportion of sentences rated grammatical falls (non-significantly) to 50%. However, if a modifier such as *that was hairy* is placed at position (B), only 29% of sentences are judged grammatical. The increase in reanalysis failure that occurs when material intervenes between ambiguous head noun and disambiguating verb is the head noun position effect.

In Experiment I, we demonstrate that filled pause disfluencies (*uh*) also produce the head noun position effect. In Experiment II and III, we examine two possible reasons why the disfluency elicited the head noun position effect: one is that the disfluency delays the onset of the disambiguating verb, and the other is that the listener uses the locations of disfluencies to predict particular types of syntactic constituents.

2. Methods

2.1. Stimuli

Two garden path structures were used in this study. The first is the subordinate main structure shown in (1). The head noun position effect has been replicated several times in this structure using reading paradigms. The second structure is the coordination ambiguity structure in (2).

(2) Sandra bumped into the (C) busboy and the (A) **waiter** (B) *told* her to be careful.

The head noun position effect has never been described in the coordination ambiguity structure. Some models of parsing (e.g. [4]) suggest that the coordination ambiguity structure should not show a head noun position effect because it is reanalyzed by a different mechanism than the subordinate main structure, and therefore a comparison of (1) and (2) is of considerable theoretical interest. 30 subordinate main structures and 20 coordination ambiguity structures were created for these experiments.

Five versions of each experimental sentence were created for Experiment I: a plain ambiguous NP baseline condition, two modifier conditions with a modifier at either position (A) or (B), and two disfluency conditions with a disfluency cluster (*uh uh*) at either position (A) or (B). Material at position (A) is referred to a prenominal and material at position (B) is referred to as postnominal. For Experiment II, the disfluencies were replaced with extraneous noises that are not produced by speakers as part of spontaneous speech. The stimuli for Experiment III consisted of the baseline condition, in addition to disfluency and modifier conditions at positions (A) and (C). In the context of Experiment III position (A) will be referred to at consistent and position (C) will be referred to as inconsistent.

All of the experimental sentences were non-preferred (i.e., garden-path) structures, and were produced as a single intonational contour in order to achieve neutral prosody [5]. The sentences were normed using a prosodic acceptability task

and a sentence completion task. 50 unambiguous grammatical and 50 ungrammatical fillers were included in Experiments I and II and 20 unambiguous grammatical and 20 ungrammatical fillers were included in Experiment III.

3. Experiment I

Experiment I was an auditory analogue of the grammaticality judgment task used previously to describe the head noun position effect [3].

3.1. Procedure

Subjects were seated in front of a computer and told that they would hear sentences sampled from natural speech and that they were to judge whether the sentence that they heard was grammatical or ungrammatical. The subject indicated their judgment by pressing either a button marker "Grammatical" or a button marked "Ungrammatical". The subjects completed eight practice sentences, four grammatical and four ungrammatical, to assure that they understood the procedure. The subject pressed a button to begin each trial. The subject's judgment of grammaticality was recorded for each sentence.

The results were analyzed as a 2x2 ANOVA, with material (disfluency or modifier) and position (prenominal or postnominal) as independent variables. The subordinate-main and coordination structures were analyzed separately.

3.2. Results

Experiment I replicated the head noun position effect with spoken sentences. In the subordinate main structure (Fig. 1), the prenominal modifiers (80% judged grammatical) were more likely to be judged grammatical than the postnominal modifiers (59%). In addition, for the coordination ambiguity (Fig. 2), prenominal modifiers (93%) were more likely to be judged grammatical than the postnominal modifiers (78%), thus demonstrating the head noun position effect in a novel structure.

The disfluency conditions showed the same pattern. The subordinate main prenominal disfluencies (85%) were judged grammatical more often than the postnominal disfluencies (60%). Likewise, the coordination ambiguity prenominal disfluencies (93%) were judged grammatical more often than the postnominal disfluencies (80%).

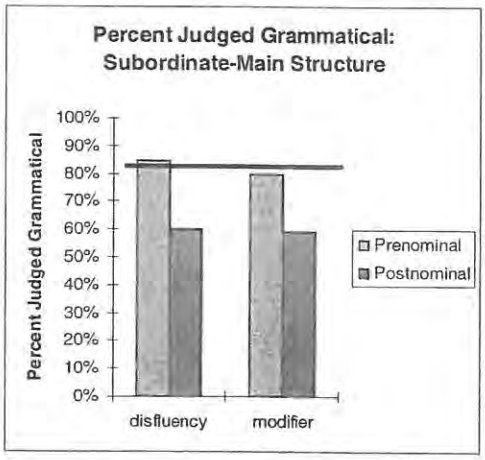


Figure 1: Percent of sentences judged grammatical for each condition of the subordinate main structure. The black line denotes the baseline condition.

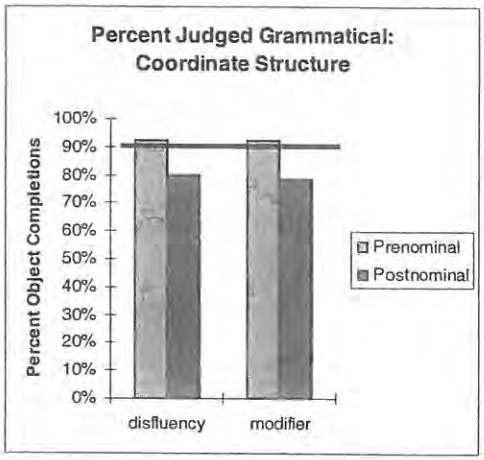


Figure 2: Percent of sentences judged grammatical for each condition of the coordination ambiguity structure. The black line denotes the baseline condition.

The main effect of position was significant (subordinate main: $F_{1,29} = 44.43$, $p<0.01$; coordination ambiguity: $F_{1,29} = 18.19$, $p<0.01$) for this experiment. The main effect of material and the interaction between material and position were non-significant for both structures.

3.3. Discussion

Experiment I shows that the head noun position effect described in reading studies also occurs in spontaneous speech. In addition, the head noun position effect has for the first time been demonstrated in the coordination structure. Thus, the head noun position effect seems to be very robust, and determining the cause of the head noun position effect could thus shed some light on the process of syntactic reanalysis.

Experiment I also indicates that disfluencies can affect the syntactic parse of a sentence. The presence of a disfluency between the ambiguous head noun and the disambiguating verb can elicit the head noun position effect. However, it is not clear from Experiment I why disfluencies might cause the head noun position effect. There are two possibilities.

The head noun position effect could be due to a delay introduced between the ambiguous head noun and the disambiguating verb. During this delay, decay in memory traces might make retrieval of the correct structure more difficult. This assumption would predict that the postnominal conditions would result in a lower proportion of sentences judged grammatical since material in the postnominal position delays the onset of the disambiguating verb. This is, of course, exactly what was found in Experiment I.

However, it is also possible that the head noun position effect elicited by disfluencies and the head noun position effect elicited by modifiers occur for different reasons. The modifier head noun position effect might be due to some process related to syntax, while the disfluency head noun position effect might be due to some special property of disfluencies. One such property might be the patterns of co-occurrence between disfluencies and syntactic constituents. Several researchers have noted that disfluencies tend to cluster around clause boundaries (e.g. [6], [7], [8]). Thus, listeners might be able to take the presence of a disfluency at a possible clause boundary as evidence of a clause boundary

and thus guide their syntactic parse accordingly. This hypothesis predicts that the postnominal disfluency condition should be judged grammatical less often than the prenominal disfluency condition, a prediction which matches the results of Experiment I.

Obviously, then, Experiment I is not sufficient to disentangle the two hypothesis. Experiments II and III were run in order to remove this confound.

4. Experiment II

Experiment II was identical to Experiment I, with one important change. In all of the experimental and filler stimuli, disfluencies were replaced with extraneous noises (cats meowing, dogs barking, telephones and doorbells ringing, people coughing and sneezing) that were not under the control of the speaker. The noises introduced a delay between the ambiguous head noun and the disambiguating verb that could not be interpreted as intentional on the part of the speaker (as an unfilled pause could be). In addition, the extraneous noises have no pattern of clustering at clause boundaries, and so the listener could not make any use of distributional information.

4.1. Procedure

The procedure was identical to the procedure in Experiment I. An additional instruction was added, indicating to the subjects that the speaker might be interrupted by an extraneous noise and that this had no bearing on the grammaticality of the sentence.

4.2. Results

Once again, the modifier conditions in both structures showed a head noun position effect. As in Experiment I, for both the subordinate main structure (Fig. 3; $F_{1,34} = 37.8$, $p<0.01$) and the coordination ambiguity structure (Fig. 4, $F_{1,34} = 4.39$, $p<0.05$) significant main effects of position were

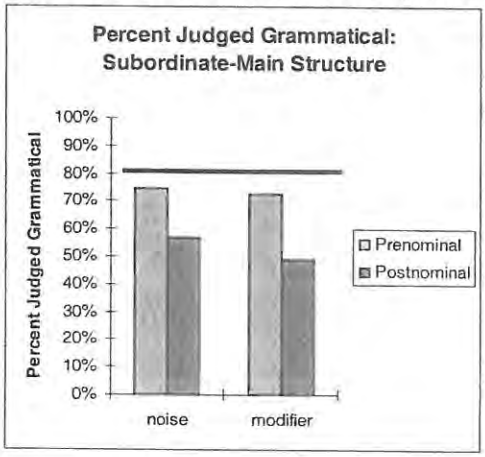


Figure 3: Percent of sentences judged grammatical for each condition of the subordinate main structure. The black line denotes the baseline condition.

observed. The main effect of material and the interaction between material and position were non-significant.

The extraneous noise conditions also showed a head noun position effect. In the subordinate main structure, the prenominal noise condition (74%) was judged grammatical more often than the postnominal noise condition (57%). In the coordination ambiguity structure, the same result was

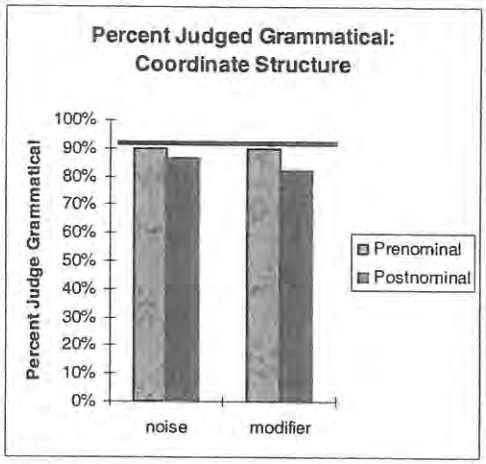


Figure 4: Percent of sentences judged grammatical for each condition of the coordination ambiguity structure. The black line denotes the baseline condition.

found, with the prenominal noise condition (90%) being judged grammatical more often than the postnominal noise condition (86%).

4.3. Discussion

The results from this experiment suggest that the head noun position effect may occur because postnominal modifiers cause the parser to be committed to the wrong analysis for a longer time period, and the longer the parser is committed to the wrong analysis, the more difficult it is to recover the correct one.

Thus, we can conclude that one way in which disfluencies can affect the syntactic parser is by delaying the appearance of a disambiguating item. Therefore, if the disfluency occurs at a place in the syntax where there is no ambiguity about the upcoming structure, the disfluency will have little effect on the parser. If, however, the disfluency occurs in a temporarily ambiguous region of a sentence, and the parser has built incorrect structure, the delay caused by the disfluency will make reanalysis more difficult.

However, we still have not ruled out the possibility that listeners could be using the distributional patterns of disfluencies found in normally produced language. The third experiment examines whether listeners make use of the tendency of disfluencies to cluster around clause boundaries in order to guide their syntactic parse.

5. Experiment III

Experiment III examined whether listeners use the presence of a disfluency to predict a upcoming clause rather than a simple noun phrase. Disfluencies or modifiers were placed at either position (A) or position (C). If disfluencies could be used to guide the parse of a sentence, then it was expected that the consistent disfluencies at position (A) would lead to a higher proportion of sentences judged grammatical, while the inconsistent disfluencies would lead to a lower proportion of sentences judged grammatical.

5.1. Procedure

The procedure for this experiment was the same is in Experiment I, with the exception that only one structure was used, and that t-tests were used to compare the two disfluency

conditions and the two modifier conditions, in addition to 2x2 ANOVA (consistency x material).

5.2. Results

Neither main effect was significant in this experiment; however there was a significant interaction between the material (disfluency or modifier) and the consistency (position (A), which was consistent with the syntactic structure, or position (C), which was not) variables (Fig. 5; $F_{1,29} = 8.53$, $p<0.01$). The t-test on the modifier conditions revealed that although the inconsistent modifier (95%) was judged grammatical more often than the consistent modifier (90%), the difference was not significant ($t_{1,29} = 1.53$, n.s.) . However, difference between the consistent disfluency (98%) and the inconsistent disfluency (90%) conditions was significant ($t_{1,29} = 6.37$, $p<0.02$).

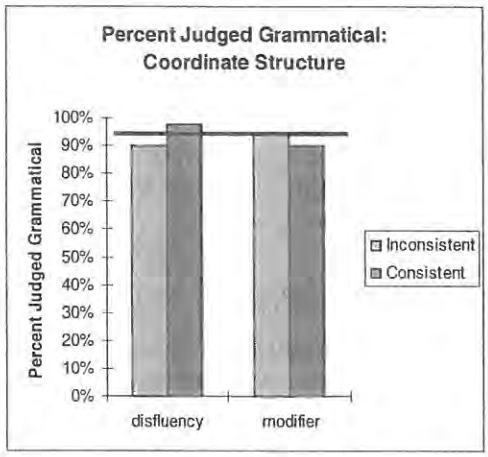


Figure 5: Percent of sentences judged grammatical for each condition. The black line denotes the baseline condition.

5.3. Discussion

The results of Experiment III indicate that the presence of a disfluency at the possible site of a clause boundary is information that can be used to guide the parse of a sentence. The presence of a disfluency at position (A) resulted in a higher proportion of grammatical judgments; thus, subjects were able to avoid the garden path more often. Thus, the head noun position effect elicited by disfluencies is probably a result of two factors working together. One is that disfluencies can cause the parser to be committed to the wrong analysis for a longer period of time, thus making recovery of the correct structure more difficult. The second is that disfluencies correlate with more complex syntactic structures, and the parser appears to be able to use the co-occurrence information to predict a more complex clausal structure and therefore to avoid being garden-pathed.

6. Conclusions

Based on these three experiments, it seems clear that filled pause disfluencies can affect the syntactic parse of a sentence. They can do so by introducing delay during the ambiguous region of a sentence, or by guiding the parse of the sentence when they co-occur with possible clausal boundaries. This suggests that models of parsing must take the effects of disfluencies into account, and that models of disfluency in spontaneous speech must account for syntactic effects in addition to pragmatic effects. The data also seem to rule out the idea that disfluencies are simply filtered from spoken utterances prior to parsing. Instead, disfluencies systematically influence the parser's basic parsing operations.

7. Acknowledgments

The authors thank Tom Beckius and Sarah Post for their help with running the experiments and preparing stimuli. This research was supported in part by a Michigan State University Distinguished Fellowship awared to Karl Bailey and by a grant from the National Science Foundation (BCS – 9976584) awarded to Fernanda Ferreira.

8. References

[1] Clark, H. H. and Clark, E. V., *Psychology and Language: An Introduction to Psycholinguistics*, Harcourt Brace Jovanovich, New York, 1977.

[2] Fox Tree, J. E., "Listeners' uses of "um" and "uh" in speech comprehension", *Memory and Cognition, Vol 29, 2001, p320-326.*

[3] Ferreira, F. and Henderson, J. M. "Recovery from misanalyses of garden-path sentences", *Journal of Memory and Language, Vol. 30, 1991, p 725-745.*

[4] Fodor, J. D. and Inoue, A., "Attach Anyway", in J. D. Fodor and F. Ferreira (Eds.) *Reanalysis in Sentence Processing* (p. 101-141), Kluwer Academic Publishers, Dordrecht, 1998.

[5] Kjelgaard, M. M. and Speer, S. R., "Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity", *Journal of Memory and Language, Vol. 40, 1999, p 153-194.*

[6] Clark, H. H. and Wasow, T., " Repeating words in spontaneous speech", *Cognitive Psychology, Vol. 37, 1998, p 201-242.*

[7] Ford, M. " Sentence planning units: Implications for the speaker's representation of meaningful relations underlying sentences.", in J. Bresnan (Ed.), *The mental representation of grammatical relations* (p. 798-827), MIT Press, Cambridge, MA, 1982.

[8] Hawkins, P. R. " The syntactic location of hesitation pauses", *Language and Speech, Vol. 14, 1971, p 277-288.*

Listeners' ERP Responses to False Starts and Repetitions in Spontaneous Speech

Jan McAllister, Susan Cato-Symonds and Blake Johnson

Department of Psychology
University of Auckland, New Zealand
j.mcallister@auckland.ac.nz

Abstract

Hindle [1] suggested that false starts and repetitions should be handled differently in a computational account of the processing of the two kinds of disfluency, and there is behavioural evidence that the human sentence processing mechanism likewise honours this distinction [2]. The same dichotomy was also evident in the electrophysiological data reported here. False starts and repetitions were identified in a corpus of spontaneous speech. Control items for the false starts were prepared by excising the reparanda to yield apparently fluent items. Continuous EEG was recorded while subjects listened to items containing the false starts, fluent false start controls, and first and second tokens of repetitions. Compared with identical words in their fluent controls, the false starts elicited a positive response similar to the P600 which is reported for syntactically anomalous words [3, 4, 5]. By contrast, second tokens of repetitions in general resulted in increased amplitude of the N400 [6]; yet, when the same repetitions were excised from context and presented list-fashion, they elicited the positive-going response which has been reported by other researchers [7].

1. Introduction

Although the speech we hear around us every day contains numerous disfluencies (interruptions to the smooth flow of speech), our conscious percept is that what we hear is, for the most part, grammatically well-formed. This suggests that our brains are organized in such a way that disfluencies are dealt with very rapidly and efficiently, and usually without conscious awareness. In this paper, we examine the brain responses that are elicited when a listener encounters a disfluency.

The following utterance contains two disfluencies:

> And he's got, he's got a little triangle hat and
> it – you can see a thin neck and a thick body.

The two disfluencies are, first, a repetition of the words "he's got", and second, a false start when the speaker abandons the original utterance at the word "it" and begins a fresh one with the words "you can see". Comprehension of these two sorts of disfluency, repetitions and false starts, was the focus of the study reported here.

When a speaker produces an utterance containing a disfluency such as "and it - you can see a thin neck and a thick body", it appears as though all or part of the original utterance "and it" is intended to be aborted and replaced with the new material. The term *reparandum* is used to refer to the material being aborted and *repair* to refer to material being substituted. Between the reparandum and repair the speaker may sometimes also insert material at the boundary between the reparandum and repair, such as a filled pause (e.g. "erm"), a silent pause or some other kind of phonological marker; such a marker is called the *editing term*.

One of the first researchers to attempt to model the comprehension of disfluent speech was Hindle [1]. A central feature of Hindle's model was the provision of distinct mechanisms for dealing with repetitions and false starts. Fox Tree [2] maintained this distinction between the two kinds of disfluency in the design of a set of monitoring experiments. She had subjects monitor for words that occurred soon after false starts or repetitions in spontaneous speech, and compared latencies with those for control materials in which the reparandum had been digitally excised. She found that monitoring times were slowed in items containing false starts, relative to their controls, suggesting that some processing cost was attached to the premature abandonment of an utterance and/or analysis of the repair. By contrast, subjects responded faster to targets that followed intact repetitions than to those in control materials without repetitions.

The goal in the present study was two-fold. First, we wished to establish whether disfluent utterances were distinguished from fluent counterparts by an electrophysiological marker. Second, we sought to determine whether Hindle's and Fox-Tree's distinction between repetitions and false starts was borne out in the data – specifically, whether different event-related potentials (ERPs) were elicited by the two kinds of disfluency.

In the last twenty years, ERP studies have provided many insights into language processing. The first report of a distinctly linguistic ERP was provided by Kutas and Hillyard [6] who found that semantically anomalous words were characterised by a negative-going wave, peaking at about 400 ms after stimulus onset, which they termed N400. The precise nature of N400 is still a matter of some debate. It is elicited in response to seemingly divergent stimuli, including line drawings, pronounceable non-words, low-frequency words and words in discourse that lacks cues to cohesion. However, the N400 is not elicited in response to syntactically anomalous material. Instead, syntactic anomaly elicits a positive-going response, termed the P600 [3, 4] or Syntactic Positive Shift [5], that may onset at 300 ms post-stimulus, or earlier in spoken language [4].

Deliberate (i.e., non-disfluent) repetitions are also associated with particular ERP effects. When repetitions occur in written or spoken lists, the response to the second token is more positive-going than that to the first token after 250 ms or 400 ms, respectively [7]. Repetitions that occur intentionally as part of connected discourse result in modulations to N400 and other components [8].

In the present study, false starts and repetitions were selected from an existing corpus of spontaneous speech. These disfluent materials, along with fluent control items that were created by digitally editing the false start materials, were played to subjects whose ERP responses to the same words in the fluent and disfluent contexts were recorded. In addition, the first and second tokens of the repetition items were excised from context and played list-fashion to the subjects, among sequences of non-repeated items, to determine whether the response to repetitions in context resembled that already reported for items repeated in lists.

2. Method

2.1. Materials

The digitized corpus of spontaneous speech from which the utterances were drawn is described by Murfitt and McAllister [9].

2.1.1. False starts

The false starts were selected first. Exclusion criteria included the presence of another disfluency in the immediate vicinity of the false start, overlapping talk by another speaker, overt editing terms at the boundary between the reparandum and repair, and word fragments in or immediately before the reparandum. One hundred and twenty eight false starts were initially identified as being potentially useable. These were excised, along with sufficient preceding and following context to permit listeners to perform the behavioural task (see below). Next, fluent control materials were made by editing copies of these experimental items so as to remove the reparandum. As previous researchers have noted, it is frequently possible to perform this editing in such a way that the resulting utterance is indistinguishable from unedited fluent speech [2]. To check the success of editing in the present materials, two volunteers checked the 128 edited items along with 20 fillers. Wherever either listener accurately identified splicing, the items were re-edited. The whole set of items was then checked by a further listener; any items where splicing was still detectable were discarded. Ninety-eight items survived this further test, and were included in the experiment. The false start materials represented the speech of nine speakers.

2.1.2. Repetitions in original context

The repetitions were next selected. The false start materials contained 17 repetitions that were not in the close vicinity of the false start itself (i.e. neither within, nor immediately adjacent to, the reparandum or the repair of a false start). From the speech of the nine speakers, a further 33 repetitions were identified, making 50 repetitions in all. The additional repetitions were taken from the digitised speech corpus along with sufficient context to make them comparable to the false start trials.

2.1.3. Tangrams

The forty tangram pictures that were used in Murfitt and McAllister's experiment [9] were digitised for use in the behavioural task.

Two experimental sequences were compiled. The same repetitions occurred in both sequences, but the false starts were divided into two sets. In one experimental sequence, half of the false starts were presented in their disfluent form while the other half were presented in their fluent form. In the second experimental sequence, items that were presented in their fluent form in the first sequence were now presented in their disfluent form, and vice versa.

Locations of the onsets of the target words were identified and these locations were used to position triggers that permitted segmentation of the continuous EEG record. The triggers were positioned at the start of the repair in the false start materials, at the start of the same word in the fluent controls, and at the start of the first and second tokens of repetitions.

Finally, utterances were paired with tangram pictures. Half the utterances were paired with a tangram picture that was consistent with the description given, and half with pictures that were inconsistent. Nine practice items and nine orientation trials were also prepared.

2.1.4. Repetitions in Lists

As noted in the introduction, well-established ERP responses are associated with the presentation of repeated items in list style, in both the auditory and visual modalities. An additional condition was therefore prepared. The 50 repetitions were excised from context and were arranged in lists along with 200 non-repeated filler items which were also drawn from the spontaneous corpus. Like the repetitions themselves, some of the non-repeated items were single words and some were multi-word sequences. Speech files were created, one for each speaker, in which the repeated and non-repeated items were presented, separated by one second of silence. One second of silence also separated the two tokens of any repeated items. As in the main experiment, the locations of the onsets of the each token of the repetitions was identified to allow placement of EEG segmentation triggers.

2.2. Subjects

Twenty members of the University of Auckland community took part in the experiment. They were paid NZ$20 for approximately 2 hours of their time. All subjects were right-handed, neurologically normal, and native speakers of English.

2.3. EEG

Electrical Geodesics 128 channel Ag/AgCl electrode nets were used. EEG was recorded continuously (250Hz sampling rate; 0.1-39.2Hz analogue band pass) during the experiment with Electrical Geodesics amplifiers (100 MΩ) and acquisition

software running on a Power Macintosh 9600/200 computer with a National Instruments PC1-1200 12 bit analogue to digital conversion card. Electrode impedance ranged from 10 to 50 kΩ. EEG was initially acquired using a common vertex (Cz) reference, and subsequently re-referenced to averaged mastoids for purposes of analysis.

2.4. Procedure

Subjects were tested individually in a quiet room. They were instructed that on each trial they would hear a spoken description which would be immediately followed by a tangram picture. They were to decide whether the picture was consistent with the description that they had just heard. The speech materials were presented via earphones and the subjects saw the tangram pictures on a computer screen. Subjects initiated trials by pressing a computer key, which triggered the presentation of the speech material. While the spoken material was being presented, a fixation cross was shown in the centre of the screen, and subjects were asked to gaze at the cross while listening to the description. At the end of the spoken description, the tangram picture replaced the cross, and subjects gave their behavioural response via the computer keyboard.

After the main part of the experiment had been completed, subjects participated in the "repetition in lists" control condition. They were told that they were now going to hear lists of speech items, some of which consisted of several words, and some of which consisted of only one word. They were instructed that they should simply listen to the speech and silently identify the words that they were hearing.

3. Results

3.1. Behavioural task

The subjects made appropriate judgments about the tangrams 91% of the time, indicating that they were conscientiously attending to the meaning of the utterances.

3.2. ERPs

Due to space limitations, only results for midline electrodes are discussed here.

3.2.1. Responses to False Starts

Waveforms associated with disfluent and fluent items at electrode Pz are shown in Figure 1. Responses elicited by

Figure 1: Averaged ERP response to false start stimuli; the heavy line shows responses to control (fluent) stimuli, the faint line, responses to experimental (disfluent) stimuli. The vertical bar indicates onset of the critical word.

disfluent stimuli, relative to their fluent controls, were characterised by a positive-going wave that was bilaterally distributed, and more pronounced at posterior sites. The positive response was similar to the P600 reported for continuous speech containing syntactic anomalies [4]. Statistical analysis of the ERP responses at the three midline sites Fz, Cz and Pz indicated a reliable early divergence of the waveforms for fluent and disfluent materials. In the 0-300 ms window, disfluent items were already eliciting a more positive-going response than fluent items ($F [1,19] = 5.87$, $p = 0.0255$); this effect was reliable only at electrodes Cz and Pz (Fluency X Electrode Site interaction: $F [2,38] = 4.24$, $p = 0.0218$). Electrode Site was not significant as a main effect ($F < 1$). In the 300-500 ms window, disfluent items were once again associated with more positive-going waveforms ($F [1,19] = 5.11$, $p = 0.0357$). The main effect of Electrode Site was also significant ($F [2,38] = 5.89$, $p = 0.0059$), with electrodes Cz and Pz more negative overall in this time window than Fz. The interaction between these variables was not statistically reliable ($F [2,38] = 2.11$, n.s.). In the 500-800 ms time window, a similar pattern emerged: fluency was once again significant ($F [1,19] = 9.45$, $p = 0.0062$), and electrode site was marginally so ($F [2,38] = 3.17$, $p = 0.0533$). There was no interaction between the variables ($F < 1$).

3.2.2. Repetitions in spontaneous speech

Figure 2: Averaged ERP response to repetition stimuli when these were heard in their original speech contexts. The heavy line shows responses to first tokens, the faint line, responses to second (i.e. disfluent) tokens. The vertical bar indicates onset of the critical word.

Figure 2 shows the responses at electrode Pz when repetitions were presented in their original (i.e. spontaneous speech) contexts. As figure 2 shows, responses to the first and second tokens elicited markedly different waveforms. The difference was apparent as an increased negativity associated with second (i.e., disfluent) tokens. This was bilaterally distributed in parietal and posterior regions.

In the 0-300 ms time window, neither the main effects nor the interaction were significant (all Fs approximately 1). In the 300-500 ms window, fluency was marginally significant ($F[1,19] = 3.24$, $p=0.0879$); neither the main effect of electrode site nor the fluency X electrode site interaction were significant (both Fs approximately 1). In the 500-800 ms window, neither main effects nor interaction were significant (all Fs approximately 1).

3.2.3. Repetitions in lists

Figure 3 shows ERP responses at electrode Pz when these same repetitions were presented in lists with 1 second of silence between first and second tokens. The lists also contained singleton filler items. Relative to the first token of the stimuli, repetitions were now characterised by greater positivity bilaterally at parietal and posterior sites in the latter part of the epoch. This positivity is broadly typical of items repeated in lists [7].

Figure 3: Averaged ERP response to first and second tokens of repetition stimuli when these were presented as items in a list. The heavy line represents the first tokens, the faint line the second tokens. The vertical bar indicates stimulus onset.

Consistent with other published studies, activity in two time windows, 0-400 ms and 400-800 ms, was analysed. The fluency effect shown in Figure 3 was significant in the earlier time window ($F[1,19 = 4.85$, $p < 0.05$), but neither the effect of electrode site nor the interaction were significant (both Fs < 1). In the second time window, there were differences between the three electrode sites overall ($F[2, 38] = 6.20$, $p = 0.0047$), and fluency ($F[1, 19] = 5.94$, $p = 0.0248$), but the two variables did not interact.

4. Discussion and Conclusion

The presence of the P600 effect associated with false starts indicates that listeners detect a syntactic anomaly when processing these items. The earliness of the divergence between the disfluent and fluent false start materials suggests that the anomaly may actually be associated with an event at the end of the reparandum, rather than the beginning of the repair. None of the experimental items contained overt editing terms such as "erm" at the boundary between the reparandum and the repair, but it is possible that other (e.g. prosodic) cues may have indicated that the item was about to become disfluent. Second tokens of repetitions, when they were heard in their original sentence contexts, gave rise to a negative response, relative to the corresponding first tokens. This finding indicates that false starts and are differently processed, as anticipated by Hindle [1] and Fox Tree [2]. It is possible that the repetition response found here may be a variant of the discourse repetition response reported by van Petten et al. [8]. Interestingly, the present result is also consistent with the ERP response to repetition blindness [10]. Repetition blindness/deafness is one mechanism that has been suggested to account for listeners' frequent failure to retain a conscious percept of the occurrence of a disfluent repetition [11].

5. References

[1] Hindle, D. ``Deterministic parsing of syntactic non-fluencies '', *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics,* 123-128, 1983.

[2] Fox Tree, J. (1995). "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech", *Journal of Memory and Language 34, 1995, p 709-738.*

[3] Osterhout, L., and Holcomb, P. "Event-related brain potentials elicited by syntactic anomaly", *Journal of Memory and Language 31, 1992, p 785-806.*

[4] Osterhout, L., and Holcomb, P. "Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech", *Language and Cognitive Processes 8, 1993, p 413-437.*

[5] Hagoort, P., Brown, C., and Groothusen, J. "The Syntactic Positive Shist (SPS) as an ERP-measure of syntactic processing", *Language and Cognitive Processes 8, 1993, p 439-484.*

[6] Kutas, M., and Hillyard, S. "Reading senseless sentences: Brain potentials reflect semantic incongruity", *Science 207, 3, 1980, p 203-205.*

[7] Rugg, M., Doyle, M.C., and Melan, C. "An event-related potential study of the effects of within- and across-modality word repetition", *Language and Cognitive Processes 8, 1993, p 357-378.*

[8] Van Petten, C., Kutas, M., Kluender, R., Mitchener, M., et al. "Fractionating the word-repetition effect with event-related potentials", *Journal of Cognitive Neuroscience 3, 2, 1991, p 131-150.*

[9] Murfitt, T., and McAllister, J. "The effect of production variables in monologue and dialogue on comprehension by novel listeners", *Language and Speech, in press.*

[10] Schendan, H, Kanwisher, N, and Kutas, M. "Early brain potentials link repetition blindness, priming and novelty detection." *Neuroreport: Neuroscience. Vol 8(8), May 1997, 1943-1948.*

[11] Bard, E.G., and Lickley, R. "Disfluency Deafness: Graceful Failure in the Recognition of Running Speech.", Proceedings of CogSci '98.

Grammatically unacceptable utterances are communicatively accepted by native speakers, why are they ?

Jeanne-Marie Debaisieux et José Deulofeu

Université de Nancy II
Université de Provence
deulofeu@up.univ-mrs.fr
debaisie@clsh.univ-nancy2.fr

Abstract

This paper aims at redefining the generally accepted notion of unfinished or elliptic sentence, which appears to be crucial in defining in turn the notion of fluency itself. It will be shown that a large part of utterances which a regularly trained linguist would consider as unacceptable and revealing some kind of disfluency of the speaker who produced them, are in fact fully accepted by the participants of a regular verbal interaction. This apparent contradiction will be explained by the fact that linguists base their judgments of well formedness of the utterances on their grammatical structure, whereas speakers interact basically by means of communicative units, which are not necessarily made up of grammatically well formed parts.

1. Introduction

Our basic assumption is that the linguistic aspects of fluency involve two distinct notions, which are generally not distinguished : traditional notion of grammatical acceptability of utterances and a new notion that we will define in this presentation both on interactional and in structural terms : communicative acceptability. Our hypothesis is that many judgments of disfluency about utterances are in fact judgments on grammatical acceptability and not on communicative acceptability.

Indeed, spontaneous speech is not ruled by grammatical acceptability but only by communicative acceptability. That is, the speakers consider as perfectly acceptable communicative acts utterances which are clearly grammatically ill formed. We will give in the first part of the paper some authentic examples of such utterances. In the following parts, we will set up a framework in which this situation will be explained not as the result of some performance disfluency of the speakers but as a natural consequence of the structural properties of language.

2. Examples of successful communication in spite of grammatical ill formedness

Example 1

cet instituteur a marqué toute ma vie/ et/ je me suis souvent posé la question/ toute ma vie de penser est-ce qu'il y a encore y en aurait ah je crois qu'y en aurait quand même mais des hommes aussi dévoués parce que il n'attendait rien /ni de mon père ni de moi et ce que je regrette c'est qu'il soit mort trop tôt [...]

[This teacher marked all my life and I asked often me this question / all my life to think are there still / would have there been / oh I think there would be anyway but as unselfish men because he didn't expected anything from my father or from me and what I regret is that he has died so soon]

Example 2

L3 des cabines téléphoniques vous en avez là sur la place ou

L5 ah non non une cabine téléphonique qui serait placée ici parce que nous avons une cabine téléphonique devant la gestion mais c'est le poste de la gestion qui s'en sert (bus E66, 16)

[L3 phone boxes, you have got some here in the square ?

L5 oh no no a phone box which would be located here because we have a phone box but it is for management staff only]

In these two examples, It is hard to see what kind of grammatical function could fulfill the NPs *des hommes aussi dévoués* and *une cabine telephonique qui serait placée ici*. The NPs are located between items like : *mais, parce que*, acting as syntactic boundaries preventing the NPs to be constructed with any governing element.

Exemple 3

A student is describing the problems she faced, when waitress in a Mac Do restaurant. She has just explained that her job conditions are hard, but she really needs this job :

.L2 en plus là cette année tu vois j'ai pris un appartement donc il va falloir que j'assume L1 ben ouaisL2 et vu que c'est le seul contrat qui me permette de payer mon loyer L1 ouaisL2 et puis c'est un CDI donc c'est à long terme (Mac.Do)

[L2 and what is more this year you see I have got an apparment so I will have to go along with that L1 yesL2 ans as it is the only contract which helps me to pat my rent L1 yeah L2 and it is a Indefinite Duration Contract so I will have it for a long time]

It is impossible to construct the subordinate clause introduced by *vu que* either with the preceding context or with the following. In both cases the result is an ungrammatical sentence :

? il va falloir que j'assume et vu que c'est le seul contrat qui me permette de payer mon loyer

? vu que c'est le seul contrat qui me permette de payer mon loyer et puis c'est un CDI

Example 4

Professors in a formal working group are discussing about how to describe a functionality of a software.

L1 mais là il faudrait préciser c'est important
L2 oh on va quand même pas être
L1 ben si justement
[L1 but here we should be more specific it is important
As far as *être* is a verb with obligatory object, the L2 first turn is an ungrammatical sentence.
We can sum up our observations in saying that in all these examples we find instances of unfinished sentences or ill formed grammatical constructions namely :
...des hommes aussi dévoués...
...ah non non une cabine téléphonique qui serait placée ici...
... et vu que c'est le seul contrat qui me permette de payer mon loyer...
... on va pas être...

3. Unfinished sentences as performance errors of disfluent speakers

One could be tempted to conclude that these unfinished sentences are due to performance errors, produced by disfluent speakers somewhat awkward in phrasing what they had in mind. One could either hypothesize that their ungrammaticality is linked to the informal status of the speech situation favoring a mere communicative use of language, in which some grammatical approximation could be accepted. But there is evidence against such an analysis. First, we can point out that most of the speakers have high social status or degree of literacy and that some situations are far from informal : in example 1, a high school teacher is telling us personal memories; in example 2, a retired senior executive is negotiating with a representative of the Town Council ; in example 4, the speakers are university staff members in a formal meeting.
From an other point of view, we must notice that these utterances are accepted as perfectly valuable parts of interactional moves by the addressees . They do not provoke any accident in the communicative flew. We cannot notice any reactions, such as clearing up requests, showing difficulties in comprehension. On the contrary, we can find pieces of evidence that the verbal interaction is going on without problems. In 3, for instance, we notice a positive feed back marker *(ouais)* and in 4, the turn *ben si justement* best shows that the speaker must have understood the previous turn as a regular statement in spite of his grammatical illformedness, as far as he expresses explicit disagreement with it. We find no evidence in conversation that such utterances elicit some negative judgments of incompleteness from the participants.
On the contrary, one can find such negative judgments in the metalinguistic comments linguist make about this kind of utterances. Look for instance at this comment of the first example by a French linguist :

"*mais des hommes aussi dévoués* ", "*mais* " est en tête **d'énoncé inachevé** et met ainsi en valeur non pas l'argument présenté mais l'élaboration de l'argumentation."(Lebre-Peytard p.126)

If, according to that comment, the argument would have only been in way of elaboration and not fully elaborated, how could this purpose oriented interaction have gone on ?
So it is the scholars and not the speakers that comment such utterances in terms of deviance or incompletion. We can hypothesize that scholars make such comments because they analyze informal speech using the tools elaborated for formal written style. In doing so they are unduly projecting properties specific to written style on spoken language structure. Namely, it is because they assume that the basic speech units necessarily obey the structural requirement of grammatical wellformedness, as it is the rule in formal written style, that the utterances appear incomplete to them. Let's now propose a different approach to these data.

4. Unfinished sentences as natural consequences of structural properties of language

4.1. macro and micro syntax : the basic heterogeneous nature of syntactic structure

If we abandon the traditional syntactic framework based on the sentence as structural unit, it is possible to explain in a natural way why an acceptable communication can be conveyed by ungrammatical sentences. Let's suppose according to the framework defined in Berrendonner [1] and Blanche Benveniste [2], that the syntactic component of language is composed of two independent sub components interacting in a modular way : micro and macro syntax. The rules of microsyntax define the wellformedness conditions of grammatical constructions, strictly understood as the projections of lexical heads, the interpretation of which is componential. The rules of macrosyntax, on the other hand, define the wellformedness conditions of other types of units from which an utterance can be built up : the communicative units. Basically, communicative units are defined as complexes of verbal and mimogestual elements interpreted by non componential semantic rules. If we shift from sentence to communicative unit as basic syntactic structural unit, and if we consider that well formed grammatical constructions are one possible but not obligatory way in which communicative units can be realized, we can solve the puzzle : structurally well formed communicative units can be either complete or incomplete grammatical constructions. In such an approach, the output "optimality" constraint stipulating that communicative units are necessarily based on grammatically well formed constructions is no longer a consequence of structural properties of language, but better belongs to a rhetorical component, where it will be described as a feature of formal written style.

4.2. Grammatical wellformedness of communicative units is not a structural property

To better understand the relationship between grammatical constructions and communicative units, let's take the case of an utterance built up from two communicative units according to the basic macrosyntactic pattern Prefix –Nucleus (Blanche-Benveniste [2, chap. 7]). This pattern can be characterized on both levels of form and content.
On the level of form, communicative units are mainly characterized by specific prosodic contours. For French, one can find a overall presentation of the relevant contours in Martin [3]. As it is widely accepted, the prefix is marked by a continuative contour and the nucleus by a range of conclusive contours. These two formal components are associated with two distinct interpretations. The nucleus conveys, according to

its contour, a specific speech act interpretation (basically, assertion as opposed to question and injunction). The prefix sets a frame of conditions of felicity for the speech act, without being itself interpretable as a speech act. A current instance of this pattern is the well known topic-comment structure :

1. (Le piano Prefix) (les doigts c'est très important Nucleus)
as for the piano the fingers are very important
2. (Le piano Prefix) (les doigts hou là là Nucleus)
the piano the fingers ... gosh !
3. (Le piano Prefix) (bof (interjection + negative mimic))

In this sub case of Prefix-Nucleus structure, the prefix is an NP interpreted as the entity about which the content of the nucleus is asserted.
We will notice first that in this type of utterances, the prefix doesn't bear any microsyntactic (grammatical) relation in reference with the categories present in the nucleus. The macrosyntactic component is totally responsible of the syntactic "togetherness" between *le piano* and the Nucleus. The macrosyntactic component should not be considered as a supplementary device adding some kind of information structure to already existing syntactic structure : it builds by itself specific syntactic structures Then, it appears that communicative units are not necessarily made up of grammatical categories. For instance, the nucleus part of the utterance in 3 above only consists of interjections or mimics. The only formal feature which is shared by all the utterances 1 – 3 is a specific prosodic pattern. We can hypothesize that speakers rely mainly on the prosodic pattern to recognize that a well formed communicative unit has been produced.
Now, as the preceding examples show that a communicative unit nucleus can be instantiated by no grammatical category at all, we can wonder whether, when some communicative unit is realized by a grammatical construction, this construction should be grammatically well formed. There are obviously instances of this pattern, but, if the two syntactic components are thought as independent means of building utterances, there is no logical reason for this parallel pattern to be the only possible one. It could equally be the case that communicative units are realized by incomplete or ill formed grammatical constructions. And it is exactly what is empirically observed in the utterances judged as showing some kind of deviance : they are perfectly well formed at the macrosyntactic level, which explains that they are accepted as valuable communicative units.
Let's apply this model to authentic "unfinished" utterances :

4.2.1. Iincomplete grammatical construction as prefix

We can mention for instance this case of grammatically ill formed PREFIX :

4. Moi je trouve les gens là qui (PREFIX-TOPIC) ben faut les éviter quoi (NUCLEUS- COMMENT)
[for me, I think , these people here who... well you must avoid them rignt)

The utterance as a whole is a well formed pattern of communicative units as we could see from its prosodic contour, which that can be shown to be basically the same as the one we would have observed if the prefix would have included a well formed relative clause. From the point of view of interpretation, recall that the way communicative units are semantically interpreted is a non componential one (for a detailed analysis of the interpretation process see Debaisieux [4]. Then the speaker gives in some way by the idiomatic use of là an instruction to the addressee to cooperate in inferring from the context the "missing" characterization of the microsyntactically incomplete NP *les gens qui*. The situation is in fact the same as in the case when the basis of inference is an interjection or a mimic (cf 2 or 3). Let's turn now to the examples of section 2.

4.2.2. Iincomplete or "missing" nucleus

Example 3 of section 2 above can be explained along the same lines : it includes a grammatically illformed communicative unit, namely the nucleus, which is made communicatively well formed by the prosodic pattern. As for semantic interpretation, the "quality" which should follow "*être*" phrased as an adjective is inferred, somewhat as "we are not no be so careful", on the basis of the sequence "quand même pas" which implies a contrast with the positive context "it is important to be very specific".
In the case of example 4 , it seems that it is the entire nucleus that the addressee has to reconstruct. So that the utterance would lack the part assuming the function of speech act. But it should be noticed that a special contour affects generally the prefix in these situations. The contour can be informally characterized as "implicative". This implicative contour acts as an instruction to reconstruct the nucleus from the context and the reconstruction task is helped by some "evidential" mimic of the speaker.

4.2.3. Complex communicative units with connectives between their subparts

What is to be noticed in the other examples (1 and 2) of section 1, is that, beside the same type of microsyntactic illformedness of some units, (in both examples the underlined part contains a "missing nucleus") they show that a standard connective morpheme (*parce que*) can link together two communicative units independently of their grammatical composition into a bigger unit. A linguist trained to analyze formal written style would declare ungrammatical such complex utterances, because there is no verb to which *parce que* can be subordinate. But in a two level syntactic component there is no a priori reason why connectives should be restricted to link grammatical units at the microsyntactic level. We can logically assume that they function on the macrosyntactic level as well as on the microsyntactic one. In that case, the apparent ungrammatical sequences like *une cabine telephonique qui serait ici parce que nous avons une cabine* in example 2 appears to be perfectly acceptable as a complex of communicative units. From the formal side, *parce que* functions as a macrosyntactic connective, and from the interpretation side, it introduces an argument (there is one phone box but it is not a public one) justifying the assertion conveyed by the preceding prefix nucleus structure in which the nucleus is semantically inferred from the context and the mimogestual attitude of the speaker : " a phone box on the square (it would be nice)". The pattern can be schematized like this :
[communicative unit 1]
[*une cabine téléphonique qui serait ici*]
(verbal part + mimogestual attitude)

link
parce que
[communicative unit 2]
(verbal)
[*nous avons une cabine mais c'est la gestion qui*
The same pattern can be applied to example 1.
We can conclude that, in all these examples a coherent interpretation can be reconstructed by the addressee from the formal cues given by the speaker. This coherence doesn't necessarily matches with grammatical cohesion : it is nevertheless related to some formal cohesion brought by macrosyntactic patterns.

4.2.4. conventionalized unfinished contructs

We can bring independent evidence for the fact that incompleteness is a structural property of utterances in noticing that some incomplete or ungrammatical structures are conventionalized in almost every language up to the point that they are no longer felt as incomplete constructions.

We can mention first instances of "headless" constructions the head of which is to be reconstructed from the context and which should in fact be considered as grammatically ill formed :

Je peux pas, je prefère pas
(I can'r, I prefer not)
Je sais pas qui
(I don't know who)

Thera are also a long list of subordinate clauses without main clauses. Unlike the unrestricted macrosyntactic non verbal patterns analyzed above, these conventionalized utterances show severe lexical restrictions. In French, we can find idiom like uses of some verbs as in :

Quand je pense qu'il voulait partir
Si tu savais ce qu'il m'a dit

These utterances can bear the prosodic contour of a nucleus, instead of one of prefix. This means that there are no longer felt as macrosyntactic complex structures where the nucleus is to be inferred from the context.

The living processes of inference we have described above are frozen here, up to a point that the result of the discourse inference :

If you knew what he told to me (you would be shocked)
has been integrated in the meaning of the idiom : he told me terrible news.

5. conclusion

As a conclusion we could say that apparently unfinished sentences or utterances used in spontaneous speech with ungrammatical sub parts are the result of the regular use of structural properties of language by the speakers. They have internalized in their linguistic competence both micro and macro system of rules what allows them to rely more or less on the cooperative participation of their addressees in verbal interaction. And consequently to produce more or less well formed grammatical utterances. The degree of microsyntactic completeness is indeed variable according to text "gender". So the same ratio of unfinished sentences will be felt natural in everyday conversation and somewhat disfluent in an explanation. This ratio is sometimes negotiated between the conversation participants. Matalinguistic statements like "you see what I mean" are used to regulate the negotiation, whereas interventions like "would you please finish your sentences" are always felt aggressive and uncooperative by conversation participants.
The case when participants systematically use well formed utterances can either be analyzed as a will to conform to the written style model or as a kind of reluctance to accept the cooperative nature of ordinary verbal interaction. We could say that the most fluent speakers, in the ordinary meaning of the word, that is, those who speak with complete and only grammatically well formed structures are also the less gifted communicators. It is interesting to notice that second language learners, who particularly need addressee cooperation, are taught to speak with "complete sentences". A certainly counterproductive strategy, in that it prevents the speakers to transfer from their mother language competence their aptitude to take advantage of verbal cooperative attitudes.

6. References

[1] Berrendonner, A., ``Pour une macrosyntaxe", in Willems, D. (ed.), *Données orales et theories linguistiques.*, p : 25-31, , Duculot, Louvain 1991.
[2] Blanche-Benveniste, C..and alii, *Le français parlé, etudes grammaticales,* Ed. Du CNRS, coll. Sciences du Langage, Paris, 1990.
[3] Martin, Ph,. "Questions de phonosyntaxe et de phonosémantique en français", *Linguisticae investigationes., Vol.2*: 93-126 , 1978
[4] Debaisieux, J-M., ``Vous avez dit inachevé : de quelques modes de construction du sens à l'oral", *Oral : variabilité et usages, Le français dans le monde, n° special* : 53-62, CLE, Paris, 2001 .

How to Repair Speech Repairs in an End-to-End System

Jörg Spilker, Anton Batliner, Elmar Nöth

University of Erlangen-Nuremberg, Germany
{spilker,batliner,noeth}@informatik.uni-erlangen.de

Abstract

If automatic speech processing wants to deal with spontaneous speech, it has to deal with disfluencies in general and speech repairs in particular as well. The paper describes the processing of speech repairs in the VERBMOBIL system and discusses the special requirements of real-time systems. With respect to this criterion, the VERBMOBIL approach and its results are compared to other work. All these results are based more or less on the evaluation of a stand alone process, not integrated in a speech system. The ultimate goal is, of course, the use and the evaluation of the impact of such a repair process in a real-time, end-to-end system. An evaluation method based on this idea is presented and some preliminary results are given.

1. Introduction

A characteristic feature of spontaneous natural human-human dialogues are disfluencies. The more speech systems are intended to deal with natural dialogues, the more important becomes the problem of handling disfluencies and in particular speech repairs. There is no exact definition of the term "speech repair", but based on the evaluation of the German VERBMOBIL corpus,[1] speech repairs in our sense comprise the following four phenomena:

- in-word repairs
- modification repairs
- pivot constructions
- fresh starts

In an **in-word repair**, the speaker interrupts within a word and corrects a part of it. A typical example is the correction in *Termei-minkalender (app-ain-ointment calendar).* **Modification repairs** correct part of the whole sentence, but do not change the syntactic construction. In contrast to other studies, we define **repetitions** as a special case of modification repairs, where the corrected part and the correction are identical. An example for a modification repair is the following sentence: *ja ist in Ordnung* Montag *<äh> Sonntag den fünften (yes it's okay* Monday *<uh> Sunday the fifth).* In a **pivot construction** (anacoluthon), the syntax of a sentence changes from the initial construction to a different one, whereby one part of the sentence belongs to both constructions. One of the few examples we found is: *ich bin vom vierzehnten bis zwanzigsten Mai <äh> <hm> bin ich ... (I am from the fourteenth to the twentieth of May <uh> <hm> I am ...).* The underlined term *vom vierzehnten bis zwanzigsten Mai (from the fourteenth to the twentieth of May)* is the Pivot, which is part of the first –unfinished– syntactic construction *ich bin vom vierzehnten bis zwanzigsten Mai (I am from the fourteenth to the twentieth of May)* and of the second – finished – syntactic construction *vom vierzehnten bis zwanzigsten Mai <äh> <hm> bin ich ... (from the fourteenth to the twentieth of May <uh> <hm> I am ...).* **Fresh starts** do not have a pivot; the construction is aborted and a completely new one is started: *also wenn wir das – das ist der Montag (so if we that – that is the Monday).* Commonly each repair is segmented in the four parts **reparandum**, **editing term**, **interruption point**, and **reparans**; an example is given in figure 1:

- **reparandum**: the "wrong" part of the utterance
- **interruption point (IP)**: boundary marker at the end of the reparandum
- **editing term**: special phrases, which indicate a repair like "well", "I mean" or filled pauses such as "uhm", "uh" (optional, most of the time missing)
- **reparans**: the correction of the reparandum

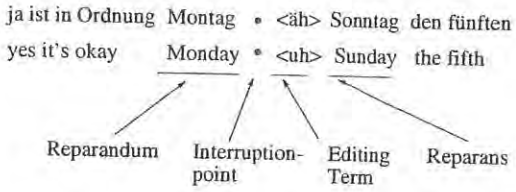


Figure 1: A repair example

2. The VERBMOBIL System

The goal of the VERBMOBIL project (1992–2000) was to build a speech-to-speech translation system supporting the tasks of appointment scheduling, travel planning and help desk. In the German part of the VERBMOBIL corpus, 21% of all turns contain at least one repair. Most of them (82%) are modification repairs. We therefore concentrate on this type of repairs. Modification repairs have a strong correspondence between reparandum and reparans. We could measure this in terms of length of reparandum and reparans (see table 1) and part-of-speech (POS) replacements. For almost all POS-categories, the speakers prefer to modify a word in the reparandum with a word which belongs to the same POS category in the reparans. Thus there is no need for a complete syntactic analysis to detect and correct most modification repairs even if repairs are characterized by violation of syntactic and semantic well-formedness [9]. We implemented a statistical approach as a filter process between the speech recognition engine and the syntactic parser. Starting with the word hypotheses graph (WHG) produced by the

[1]This work is part of the VERBMOBIL project and was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant BMBF 01 IV 701 V0. The responsibility for the contents of this study lies with the authors.

Reparandum	Type	Reparans	#
RR1	IW	RR1	580
DD1	IW		495
MM1	IR	MM1	486
RR1	IR	RR1	411
MM1 MM2	IR	MM1 MM2	111
DD1	IR		108
RR1	IW	II1 RR1	101
MM1 RR1	IR	MM1 RR1	100
MM1 RR1	IW	MM1 RR1	85
MM1	IR	II1 MM1	74
RR1 MM1	IR	RR1 MM1	53
RR1 RR2	IW	RR1 RR2	41
DD1 DD2	IR		35
MM1 MM2 RR1	IW	MM1 MM2 RR1	31
MM1	IR	II1 II2 MM1	27
RR1	IW	II1 II2 RR1	26
RR1 RR2	IR	RR1 RR2	25
MM1 RR1 MM2	IR	MM1 RR1 MM2	25
MM1 MM2 MM3	IR	MM1 MM2 MM3	24
MM1 MM2 RR1	IR	MM1 MM2 RR1	22
DD1	IW	II1	22
			2860

Table 1: Patterns for Modification Repairs (>20 tokens; 3559 patterns in the corpus, ordered by frequency); each word in the reparandum/reparans is annotated as either MM: Match, RR: Replacement, DD: Deletion, or II: Insertion; the integers (1, 2, and 3) relate the words in the reparandum/reparans with each other; IW means interruption with "w"ord fragment, IR interruption without word fragment. Example: ... *on* RR1 *Monday* MM1 IR *next* RR1 *Monday* MM1 ...

word recognizer, a prosodic module detects possible IPs. For each of these IPs, a stochastic model tries to find an appropriate repair by guessing the most probable segmentation. Repair processing is seen as a statistical machine translation problem where the reparandum is a translation of the reparans. For every repair found, a path representing the speaker's intended word sequence is inserted into the lattice. In the last step, a lattice parser selects the best path. The complete architecture is shown in figure 2.

2.1. Detection of Interruption Points

The prosodic module classifies each word boundary in the WHG as a regular or an irregular boundary. Irregular boundaries are seen as hypotheses for IPs. For each word boundary, a vector with 121 prosodic features is determined. Prosodic events like irregular boundaries are characterized by local changes in the acoustic parameters. Tests showed that a context of two words to the left and to the right of the actual word is sufficient for detection. The features are selected to give information about F0, energy, duration, pause, and POS-categories; details are given in [1]. A classification of a subsample of the VERBMOBIL database with neural networks and 559 IPs vs. 51.486 "normal" word boundaries (i.e., a relation of 1:100!) yielded the following results: recall for IPs: 90%, recall for normal word boundaries 64% which means that there are many false alarms. This is a general problem of binary statistical classifiers in cases where the proportion of the two classes is extremely unbalanced.

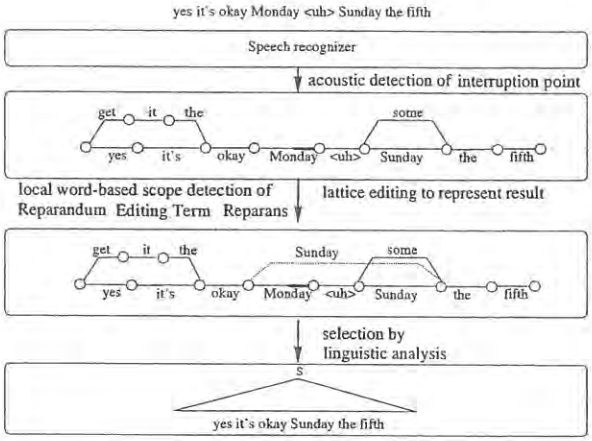


Figure 2: Architecture for repair processing

2.2. Segmentation

As mentioned before repair segmentation is mainly based on statistical machine translation (SMT) [3]. The SMT approach assumes that a speaker who produces the source sentence S originally wants to produce the target sentence T. Transferring this approach to repair processing, we assume that, if a speaker produces the reparandum (RD), he/she originally wanted to produce the reparans (RS). SMT defines a scoring function for a pair (S, T) which could be adopted for repair processing without further changes:

$$P(RD|RS) = \sum_a P(RD, a|RS)$$

a is the alignment, which describes the link between words in RD and RS; SMT is based on the hypothesis that words of the source sentence are linked to words in the target sentence.[2] If the stronger assumption is made that a word of the source sentence could only be linked to one word in the target sentence, a can be described as a vector $a_1^m = a_1 \ldots a_m$ with $a_i \in 0 \ldots l$. If the word RD_j is linked to RS_i then $a_j = i$. If it is not connected to any word in RS then $a_j = 0$. m denotes the length of RD and l the length of RS. Without any further assumptions we can infer the following:

$$\begin{aligned} P(RD, a|RS) \quad = \quad & P(m|RS) * \\ & \prod_{j=1}^{m} P(a_j|a_1^{j-1}, RD_1^{j-1}, m, RS) * \\ & P(RD_j|a_1^j, RD_1^{j-1}, m, RS) \quad (1) \end{aligned}$$

The conditional probabilities in equation (1) cannot be estimated reliably from any corpus of realistic size, because there are too many parameters. For example both P in the product depend on the complete reparans RS. Therefore we simplify the probabilities by assuming that m depends only on l, a_j only on j, m, and l, and finally RD_j on RS_{a_j}. So equation (1) becomes

$$\begin{aligned} P(RD, a|RS) = & P(m|l) * \\ & \prod_{j=1}^{m} P(a_j|j, m, l) * P(RD_j|RS_{a_j}) \quad (2) \end{aligned}$$

[2] A couple of words in the source sentence describes the same concepts as a couple of words in the target sentence.

These probabilities can directly be trained from a manually annotated corpus, where all repairs are labeled with begin, end, IP, and editing term, and for each reparandum, the words are linked to the corresponding words in the reparans. All distributions are smoothed by a simple back-off method [8] to avoid zero probabilities with the exception that the replacement probability $P(RD_j|RS_{a_j})$ is smoothed in a more sophisticated way. It is calculated by a linear interpolation of replacement probabilities for the words, the corresponding POS tags, and the semantic class

$$\begin{aligned} P(RD_j|RS_{a_j}) = \\ \alpha * P(Word(RD_j)|Word(RS_{a_j})) \\ + \beta * P(SemClass(RD_j)|SemClass(RS_{a_j})) \\ + \gamma * P(POS(RD_j)|POS(RS_{a_j})) \end{aligned} \quad (3)$$

with $\alpha + \beta + \gamma = 1$.

2.3. Processing word hypothesis graphs

The scoring function is integrated in the system on top of the prosodic annotated WHG from the recognizer. For each path through the WHG that contains an IP hypothesis, all possible segmentations, i.e., all possible (RD, RS) pairs, must be scored. In practice we reduce this set to pairs, where RD and RS are at most four words long, because we found that this restriction holds for 96% of all repairs in the VERBMOBIL corpus. Editing terms are characterized by a closed list of short phrases. Thus if after an IP such a phrase is found, it is skipped to build the (RD, RS) pair. If the score of a pair is above a heuristic threshold, the pair is accepted as a repair and an alternative path is inserted in the WHG. The resulting WHG is finally analyzed by a stochastic parser, which selects according to its model the best scored path and therefore can accept or reject the repair.

3. Constraints in real systems

Before we discuss our results and compare them to other approaches, we want to discuss the evaluation framework: the ultimate goal of our approach was a full and easy integration into a real-life system. Thus we cannot expect perfect strings as input. The state-of-the-art interface of a speech recognizer are WHGs with many co-occurring hypotheses. They represent an enormous search space, so an efficient algorithm is necessary to guarantee a time behavior in almost real time. In addition there exist no prominent repair markers, as hypothesized by some authors, to reduce the search space to the relevant points.

Another problem in real-life systems are word fragments. They play an important role in repair processing but speech recognizers are not able to mark them. With respect to these constraints, three different evaluations are carried out. The first two are stand-alone evaluations measuring the performance of the pure repair process. The third one shows the impact of the repair process on the complete VERBMOBIL system and will be described in the next section.

The first row of table 2 shows the results with the assumption that we have a perfect recognizer that produces no word errors and marks every word fragment. The test set were 441 turns, a subsample of the 559 prosodically classified turns; 118 turns were used for the training of α, β, and γ. Processing time was restricted in two ways. First there is a dynamic deadline, the real-time factor. It is set to five times the length of the turns. Second there is an absolute deadline of 10 seconds. The reference machine was a SPARC Ultra 300MHZ. The "detection" column shows the results for the repair identification task. The "correct seg." column presents the same numbers for the correct segmentation. A segmentation is defined as "correct" if reparandum and editing term are identified. In some cases within complex repairs (repairs within repairs), reparandum and editing term are not identified correctly but, if these segments are removed from the input, the resulting string is the intended word sequence. An example is:

Annotated: ... wann paßt es	*di*	*Ihn*	Ihnen denn ...
Annotated: ... when does it suit	*y*	*yo*	you ...
	RD1	RD2	RS
Recognized: ... wann paßt es	*di Ihn*		Ihnen denn ...
Recognized: ... when does it suit	*y yo*		you ...
	RD		RS

Therefore we call these results "generalized segmentation". The second row in table 2 presents the same evaluation based on an almost perfect recognizer which is not able to mark fragments.

The decrease of the recall rate from test 1 to test 2 emphasizes the importance of word fragments. Fragments are not only a prominent detection feature but they are also easier to correct. In many cases the correction is simply a deletion of the fragmented word. The decrease of the precision rate is not really a decrease in quality. It comes from the worser ratio of non-repair turns to repair-turns in test 2, where we only left out repairs with word fragments.

A direct comparison to similar work is rather difficult due to very different corpora, evaluation conditions, and goals. Not all approaches deal with the complete repair process but concentrate on either detection or correction. [10] report a recall of 83.4% and a precision of 93.9% in detecting the IP of a repair, but do not discuss the problem of finding the correct segmentation in detail. In addition their results are obtained on a corpus where every utterance contains at least one repair. [12] introduce hidden events to model the IPs of different classes of repairs in the speech recognition process. This reduces their recognition errors by about 0.9% absolute, but nothing is said about recall and precision of IP detection. Likewise they make no suggestion about getting the correct segmentation. An early and comprehensive attempt is described in [2]. They use a pattern matcher to trigger possible repairs and verify these hypotheses with a parser. The simple pattern matching algorithms achieves a recall of 76.1% and a precision of 61.8% for repair detection. 57% of the detected repairs are successfully corrected (43.6% Rec./48.1% Prec.). A second evaluation based on a different test set (26 repairs) includes a verification of the hypothesized repairs by a parser [5]. If the parser finds an unacceptable utterance, the hypothesized repairs are successively parsed until a parseable utterance is selected. In this case a detection recall of 42.3% and a precision of 84.6% is obtained. For correction the values are 30.8% recall and 61.5% precision. They comment that this procedure is not very efficient in a real-time speech system. [7] suggests a parsing approach using a deterministic parser. He assumes a perfect repair detector, so there can be no comparison as for the detection and correction algorithms. An algorithm which is inherently capable of lattice processing is proposed by [6]. They redefine the word recognition problem as identifying the best sequence of words, corresponding POS tags and special repair tags. They report a recall rate of 81% and a precision of 83% for detection, and 78%/80% for correction. The test setup was almost the same as that for test 1 (cf. Table 2). Unfortunately, nothing is said about the processing time achieved with their module. [4] build a parser on

	Detection		Correct seg.		Generalized seg.	
	Recall	Precision	Recall	Precision	Recall	Precision
Test 1	70%	86%	59%	84%	61%	84%
Test 2	48%	77%	47 %	76%	48%	76%

Table 2: Results for repair processing

top of this module in a similar way to [2]. They observe a slight improvement of about 2% in recall but a drop of about 50% in precision.

4. End-to-End Evaluation

Within real-life systems, we cannot measure the performance of the repair process in terms of recall and precision. Errors in word recognition or parsing influence the performance. If for example a word in the reparandum or reparans is misrecognized the strong correspondence between reparandum and reparans is obscured. The worst case would be that the recognition error leads to a correct sentence that the repair process should not correct. For the recall value this event is counted as a miss, but from the point of view of repair processing the behavior is totally correct. We therefore measure the impact of repair processing by the changes we found in the results after parsing. The VERBMOBIL system was tested on 276 turn with active and inactive repair processing. The turns contain 90 repairs. In 64 WHGs a repair is hypothesized, but only 12 times the parsing output is changed. A manual inspection of these changes shows that 6 repairs are correct. This means that it was a real repair or a recognition error, that could not be told apart from a repair. Two hypotheses are definitely false alarms, in two cases the hypothesis is correct, but the parser cannot analyze the corrected version. For the rest of the hypotheses, the word recognition was not good enough to decide, whether they were correct or not.

As expected there is a big difference between an idealised environment and the real-life system. But not only word fragments cause problems. Word errors[3] and parsing problems inhibit that repair processing has a greater impact on the complete system.

5. Summary and Conclusion

The term "speech repairs" denotes different phenomena, which have to be handled by different methods. In VERBMOBIL , we concentrate on the most frequent type of repairs, i.e., modification repairs. We found a strong correspondence between reparandum and reparans in syntactic and semantic features, which are utilized in a stochastic approach to repair detection and correction. The promising results on word strings could not be verified in the VERBMOBIL system. One major problem in real-life systems are word fragments, which cannot be marked by state-of-the-art word recognizers. In addition recognition errors and incomplete syntactic analyses reduce the impact of the repair process on the complete system. This first attempt of applying repair processing to a speech-to-speech system shows, that besides a necessary and possible improvement of the repair process itself, the system performance must be enhanced to benefit from such a process.

Modelling the repair segmentation as a stochastic machine translation process offers a great variety of improvements. Our approach models the replacement probability quite simple with very rough assumptions. Och et al. in [11] show and compare more sophisticated approaches, which can be applied to repair processing as well.

[3]The word error rate for the 276 turns is 24%.

6. References

[1] Anton Batliner, Jan Buckow, Heinrich Niemann, Elmar Nöth, and Volker Warnke. The Prosody Module. *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer, Berlin, 2000, pages 106-121.

[2] J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human computer dialogs. In *Proc. ACL*, pages 56–63, University of Delaware, Newark, Delaware, 1992.

[3] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June 1990.

[4] M. G. Core and K. Schubert. Speech repairs: A parsing perspective. In *Satellite meeting ICPHS 99*, pages 47–50, 1999.

[5] John Dowding, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran. Gemini: a Natural Language System for Spoken-Langugae Understanding. In *Proc. ACL*, pages 54–61, 1993.

[6] Peter A. Heeman and James F. Allen. Speech repairs, intonational phrases, and discourse markers: Modelling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571, December 1999.

[7] D. Hindle. Deterministic parsing of syntactic nonfluencies. In *Proc. ACL*, pages 123–128, MIT, Cambridge, Massachusetts, 1983.

[8] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Transaction on Acoustics, Speech and Signal Processing*, ASSP-35:400–401, March 1987.

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

[10] C. Nakatani and J. Hirschberg. A speech-first model for repair detection and correction. In *Proc. ACL*, pages 46–53, Ohio State University, Columbus, Ohio, 1993.

[11] Franz Josef Och and Hermann Ney. A comparison of alignment models for statistical machine translation. In *COLING 00: The 18th Int. Conf. on Computational Linguistics*, pages 1086–1090, Saarbrücken, Deutschland, 2000.

[12] A. Stolcke, E. Shriberg, D. Hakkani-Tur, and G. Tur. Modeling the prosody of hidden events for improved word recognition. In *EUROSPEECH '99*, volume 1, pages 307–310, Budapest, 1999.

Um, One Large Pizza. A Preliminary Study of Disfluency Modelling for Improving ASR.

Ben Hutchinson and Cécile Pereira

Syrinx Speech Systems
{ben.hutchinson,cecile.pereira}@syrinx.com.au

Abstract

A corpus of spontaneous telephone transactions between call centre operators of a pizza company and its customers is examined for disfluencies (fillers and speech repairs) with the aim of improving automatic speech recognition. From this, a subset of the customer orders is selected as a test set. An architecture is presented which allows filled pauses and repairs to be detected and corrected. A language repair module removes fillers and reparanda and transforms utterances containing them into fluent utterances. An experiment on filled pauses using this module and architecture is then described. A speech recognition grammar for recognising fluent speech is used to provide a baseline. This grammar is then enriched with filled pauses, based on their placement in relation to syntactic boundaries. Evaluation is done at the level of understanding, using a metric on feature structures. Initial results indicate that incorporating filled pauses at syntactic boundaries improves the recognition results for spontaneous continuous speech containing disfluencies.

1. Introduction

The primary application of this study is in improving automatic recognition and understanding of spontaneous speech in commercial applications. Ease of application development is one factor in a commercial setting due to constraints on time and money. The modelling of disfluencies must therefore combine ease of use and extreme robustness.

In spontaneous speech, speakers often go back and change or repeat something they have just said, or pause to think about what they want to say. Therefore, as has been reported by many authors (e.g., Heeman and Allen, 1999, 2001; O'Shaughnessy, 1999), spontaneous speech includes speech repairs and hesitations in the form, for example, of silent pauses and filled pauses (*ums* and *ahs*). Human hearers understand such speech effortlessly, but computers do not, and it presents a computational problem for automatic speech recognition and understanding (see, for example, Pakhomov and Savova, 1999). Although it has been found (Shriberg, 1996) that speakers produce fewer disfluencies when talking to computers than when talking to people, disfluencies remain in human-computer interaction. Shriberg also found that, in human-human interaction, whether or not the dialogue was goal-oriented or free conversation did not appear relevant. Moreover, as human-computer interaction becomes more natural, it is to be expected that the disfluency rate in human-computer interaction will approach that of human-human dialogues. In this paper we report on experiments that address this problem by modelling the location, syntax and semantics of speech repairs and filled pauses. A corpus of pizza orders was used to test the effects of this modelling on understanding accuracy.

The paper starts with a description of our corpus, and the disfluencies within it. We then describe the architecture of the Syrinx spoken language dialogue system, focusing on aspects relevant to the modelling and processing of disfluencies. The modelling takes advantage of explicit grammars for speech recognition and understanding. An experiment on modelling the distribution of filled pauses is then discussed. Finally, we present our conclusion and indicate directions for future work.

2. Corpus

Telephone dialogues between call centre operators of a pizza company and 162 customers (90 women, 72 men) were recorded and transcribed orthographically. The total length of the corpus was three hours and 54 minutes, and the average length of a conversation was one minute and 25 seconds.

From this corpus, 234 utterances produced by 124 customers (69 women, 55 men) were selected to form a dataset of pizza orders. A pizza order was defined as an utterance functioning as an order for pizza, irrespective of its syntactic form — most commonly, in decreasing order of occurrence, this was a noun phrase ("ah a large super supreme"), an interrogative sentence ("and could I get one chicken delight with er ah the the thick crust?"), or a declarative sentence ("I'd like a super supreme with ah extra peperoni and jalapenos"). The order could be a complete order, e.g., "Two large two large uh barbeque meat lovers thanks with a thin thin thin ah pastry", or part of an order e.g., "a small pizza", "pan fried", "super supreme", "with no olives and extra cheese". We left out replies to questions from operators about the current special, e.g., "No I just want a large ah pan pizza". The number of fluent words in an utterance ranged from one to 25, the mean number being seven.

3. Description of disfluencies

13% of all words in our dataset were either filled pauses or part of the editing term(s) or the reparandum of a speech repair.

There were 182 filled pauses in the test data. Five types were observed: *um* (84), *ah* (63), *uh* (17), *er* (13), *oh* (3) *aw* (1), and *mm* (1). 48% of filled pauses occurred at the beginning of the utterances. In these cases the utterances were syntactically mainly either a sentence (21%) or a noun phrase (24%). When the filled pauses occured within an utterance (52%), they were found most frequently before a noun phrase (18%), whether this was a complement of a verb (10%) or not (8%). Those noun phrases all referred to pizzas, and had as their heads mainly a pizza name, e.g, "two hawaian lovers", or a topping, e.g., "ground beef and onion with extra extra everything", or a crust, e.g. "the thick crust". The frequent occurrence of filled pauses before such NPs does not seem surprising in a corpus of pizza orders, since these constituents carry the essential information of the order.

For the repairs, we considered fresh starts and modification repairs, and adopted the terminology of reparandum, interruption point (ip) and alteration — following Heeman and Allen (1999; 2001). We treated filled pauses as a sub-class of pauses rather than editing terms, although we note that this may need to be revised in future work when we consider prosodic events. This means that in our data we had no abridged repairs, i.e. cases in which the repair has an editing term, but no reparandum. (1) illustrates a fresh start, where the speaker abandons the partial utterance "and can I get a um" and starts again, replacing it by the phrase "can I get another stuffed crust".

(1) Fresh start

$$\underbrace{\text{and can I get a um}}_{\text{reparandum}}\ \text{ip}\ \underbrace{\text{can I get another stuffed crust}}_{\text{alteration}}$$

Modification repairs comprise the remainder of repairs with a non-empty reparandum, and are illustrated by (2).

(2) Modification repair

$$\text{may I have}\ \underbrace{\text{a a}}_{\text{reparandum}}\ \text{ip}\ \underbrace{\text{one that has chilli}}_{\text{alteration}}$$

There were 37 repairs altogether, mainly modification repairs.

4. Architecture

For an overall description of the Syrinx spoken language system architecture, see Estival (2000). We here describe simply the utterance processing subsystem of the system, shown in Figure 1. Speech input from a microphone or telephone line is processed by the speech recogniser. The recogniser uses a grammar to constrain the sequences of words it can recognise. Thus the grammar acts as a form of language modelling. The recogniser finds the most probable utterances in the recognition grammar given the speech input. The grammar can allow disfluencies in utterances, and marks them with special tokens. These tokens are shown in Table 1. The recogniser returns the *N* most likely utterances, which may or may not contain disfluencies, along with probability scores.

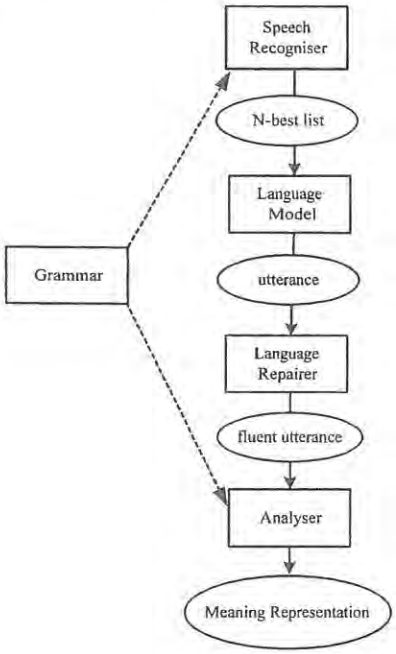


Figure 1. Architecture for processing disfluent speech.

Table 1. Mark-up tokens for disfluencies.

Mark-up token	Location
RESTART	Before a fresh start
STARTDISFLUENT	Before filled pauses, reparanda, and editing terms.
ENDDISFLUENT	After filled pauses, reparanda, and editing terms.

A class-based trigram language model was trained on transcriptions of all the customer utterances in the larger corpus and used to re-estimate the probabilities of the *N* most likely utterances returned by the recogniser. In contrast with approaches where a recognition grammar is not assumed to play a role in the language modeling (e.g., Heeman and Allen, 2001), the recognition grammar in combination with the trigram language model that does the language modelling in the system. Our modelling of disfluencies therefore has two stages: firstly, the disfluency must be present in the speech recognition grammar to allow the possibility of the recogniser returning the disfluency; secondly, the trigram language model approximates the likelihood of the disfluency occurring in a particular context. The trigram language model is the last stage of scoring. The highest scoring utterance at this stage is passed on for further processing.

Disfluencies are removed from the utterance by the Language Repairer, which finds the disfluencies by searching for the special mark-up tokens in the grammar. The Language Repairer performs two simple functions: 1) it removes RESTART tokens and any

words occurring before them; 2) it removes STARTDISFLUENT and ENDDISFLUENT tokens and any words found between them. The result is a single fluent utterance.

The fluent utterance returned by the Repairer is passed to the Analyser for syntactic and semantic analysis. Partial parsing techniques are used, and syntactic analysis is only as deep as is required, usually consisting of just word or phrase spotting. The depth of parsing required is determined by the grammar, which uses tags to mark the phrases in the grammar that have meaning for the application. Doing partial, rather than full, analysis increases speed and robustness, however parser robustness is not used to "skip over" disfluencies (cf. Core and Schubert, 1999; Rosé and Lavie, 2001) since these have been removed by the Repairer. The Analyser performs syntactic and semantic analysis concurrently, returning a meaning representation (MR) for each utterance, in the form of an attribute-value matrix. Only aspects of meaning which are used by the application are represented in the MR. For example, (3) and (4) both produce the MR shown in (5).

(3) I'd like two meat lovers.

(4) Two meat lovers please.

(5)
$$\left[\text{Order} = \left[\begin{array}{c}\text{Pizzaname} = \text{meat\_lovers} \\ \text{Number} = 2\end{array}\right]\right]$$

The Analyser is the last component of the Utterance Processing Subsystem. The MRs it produces are passed on to the Dialogue Processing Subsystem and are used to query the database, control the dialogue, and generate natural language responses.

5. Experiment

In this section we describe an experiment in improving recognition of speech containing filled pauses using the architecture for modelling and processing disfluencies described above. The data consisted of the 238 utterances of pizza orders. We describe first the incorporation of filled pauses into the grammar; secondly the evaluation metric based on language understanding; and, lastly, the results.

5.1. Grammars

A grammar containing no disfluencies was written, and this was later refined by including optional disfluencies at points in the grammar where they were observed to occur frequently in the corpus. The basic structure of the grammar is shown in Figure 2 in Java Speech Grammar Format. (In this format, square brackets indicate optional constituents, vertical slashes indicate that one of several options must be matched, and parentheses are used for grouping purposes.) The grammar was designed to reflect the wording of only the first hundred utterances examined, as a Spoken Language Dialogue System cannot predict all different phrasings users will produce.

The grammar contains semantic tags in curly brackets showing which parts of the utterance have meaning for the application, and how the meaning representation is to be constructed. For example, the "{copy}" tag says that the meaning can be found either in the lexicon (for terminals) or at the previous level of the grammar (for non-terminal). (The "copy" statement is formally equivalent to LFG "↑=↓".) Note that while the grammar is composed of "<discourse_particle>", "<prepizza>", "<pizza>" and "<postpizza>", only the non-terminal "<pizza>" contributes to the meaning since it contains the essential information.

As a result of the analysis of disfluencies described in Section 3, filled pauses were inserted into the grammar in the positions they were likely to occur.

```
<discourse_particle> = yeah | well | .. ;

<prepizza> = we_ll do | we need | [have you] got | .. | like | book us;

<postpizza> = also | please | then | today | .. | thank you ;

<TOPPING> = (ham | mushroom | olives | .. | anchovies) {copy};

<WithoutTopping> = (without | with no ) <TOPPING> {WithoutTopping = !Topping};

<BASE> = (cheesy_crust | deep | .. | thin_crisp) {copy};

<PIZZANAME> = ( barbeque_chicken | .. |supreme | vegetarian) {copy};

<pizza> {new Level} =
[(<number> {copy} |<det> )]
( <PIZZANAME> {copy} [ <BASE> {copy} ]..;

<order> = [ <discourse_particle> ] [ <prepizza> ] <pizza> [ <postpizza> ];
```

Figure 2. Abbreviated fluent grammar in Java Speech Grammar Format.

```
<filler> = STARTDISFLUENT (um|ah|uh|er) ENDDISFLUENT ;
 ..
<order> = [ <filler> ] [ <discourse_particle> ] [ <prepizza> [ <filler> ] ] <pizza> [
<postpizza> ];
```

Figure 3. Example of incorporating filled pauses into a fluent grammar.

For example, the grammar in Figure 3 allows filled pauses at the beginning of an utterance, or after a verb. The experiments used a variety of grammars containing filled pauses in different locations, as shown in Table 2.

Table 2. *Location of filled pauses in grammar.*

Beginning of utterance
After <prepizza>
Before <pizza>
Between <discourse_particle> and <prepizza>

5.2. Evaluation

The evaluation was done at the level of utterance understanding. Word based evaluations such as Word Error Rate (WER) would have been an inappropriate measure of success since the misrecognition of words in a reparandum does not affect the functioning of the system. Similarly, the failure to recognise a filled pause may not be important if the words surrounding the filled pause are recognised correctly.

The Analyser produced attribute value matrices representing the meaning of each utterance returned by the repairer. Fluent transcriptions of each utterance were also processed by the Analyser to produce the correct attribute value matrices. Next, the matrices were reduced to a set of dependencies in a manner similar to (Lenci et. al., 2000). We compared the two sets to obtain precision and recall rates for understanding.

5.3. Results

Since our data consisted of utterances from human-human conversations, the recognition rate was poorer than is expected for utterances spoken to a machine. Nevetheless, the results suggest an improvement in understanding accuracy in all the cases tested (the small number of utterances in the experiment does not permit meaningful tests of significance). The baseline was provided by the grammar with no filled pauses in it, and the best performance was achieved when either one or two filled pauses were optionally allowed to occur at the beginning of an utterance. As mentioned above, this is the location where filled pauses were most frequently observed in the corpus. In this case the precision was 5.1% above the baseline, while recall was higher by 0.6%. The second most frequent location of filled pauses was before an NP referring to a pizza, where the NP is the complement of the verb. Optionally including filled pauses in this position as well resulted in a 2.3% increase in precision, and also a 0.6% drop in recall.

6. Discussion

We now have a Language Repairer which can be used for future applications. The Repairer is a stand alone module compatible with the existing architecture.

In the next stage of our research we plan to test the Repairer with grammars containing speech repairs. We also plan to examine speech repairs and fillers in relation to intonational phrase boundaries and to examine the acoustic cues at the onset and offset of the reparanda. This may lead us to revise our treatment of fillers and consider the category of abridged repairs. We will also look at the automatic inclusion of disfluencies into grammars of fluent language. We plan to use information content of phrases as a guide for including disfluencies, using semantic tags as a guide to information content.

Acknowledgments. We wish to thank our colleagues at Syrinx, in particular Natalie d'Enyar, Juliet Mar and Magdalena Szyma, for their help in selecting the data and for giving us useful comments.

References

Core, Mark G., and Lenhart K. Schubert. 1999. Speech repairs: a parsing perspective. In *Proceedings of the ICPhS Satellite Meeting on Disfluency in Spontaneous Speech*, Berkeley, CA, July.

Estival, Dominique. 2000. The Syrinx Spoken Language System. In *Proceedings of the RIAO'2000 Conference,* Vol.3, 33–34, Paris, April.

Heeman, Peter A, and James F. Allen. 1999. Speech Repairs, Intonational Phrases, and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogue. *Computational Linguistics,* 25(4):527–571.

Heeman, Peter A., and James F. Allen. 2001. Improving Robustness by Modeling Spontaneous Speech Events. In Jean-Claude Junqua and Gertjan van Noord (eds), *Robustness in Language and Speech Technology*, 125–152. Dordrecht: Kluwer.

Lenci, Alessandro, Simonetta Montemagni, Vito Pirrelli, and Claudia Soria. 2000. Where opposites meet. A Syntactic Meta-scheme for Corpus Annotation and Parsing Evaluation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 625–632, Athens, Greece.

O'Shaughnessy, Douglas. 1999. Better Detection of Hesitations in Spontaneous Speech. In *Proceedings of the ICPhS Satellite Meeting on Disfluency in Spontaneous Speech*, Berkeley, CA, July.

Pakhomov, Sergey, and Guergana Savova. 1999. Filled Pause Distribution and Modelling in Quasi-Spontaneous Speech. In *Proceedings of the ICPhS Satellite Meeting on Disfluency in Spontaneous Speech*, Berkeley, CA, July.

Rosé, Carolyn Penstein, and Alon Lavie. 2001. Balancing Robustness and Efficiency. In Jean-Claude Junqua and Gertjan van Noord (eds), *Robustness in Language and Speech Technology,* 239–269. Dordrecht: Kluwer.

Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. In *Proceedings of the International Conference on Spoken Language Processing*, Vol. Addendum, 11–14, Philadelphia, PA, 3-6 October.

Idiosyncratic Fillers in the Speech of Bilinguals

Caroline L. Rieger
University of British Columbia, Canada

Abstract
This paper introduces a never before described strategy used by bilinguals to fill hesitation pauses. This strategy proved so unique that it was given the name 'idiosyncratic filler.' It describes a filler type that is produced unusually often by one individual when hesitating. It is usually a particular lexical filler that is used as often as or more often than all other lexical fillers combined. Idiosyncratic fillers are as flexible as, but more 'prestigious' than quasi-lexical fillers and they are used by bilinguals in their non-native language as an over-generalization and to avoid the incessant production of 'uhs' and 'uhms.'

1. Introduction

'Disfluencies' or 'self-repairs' are umbrella terms which cover more or less the same phenomena, but depending on the researcher and the field of study fillers are either included or excluded. In many studies on disfluencies or self-repairs, fillers do not receive much attention. In fact, conversation analysts have not investigated the role of fillers, although they do recognize that fillers are self-repair strategies. Mostly, they are recognized as repair initiators or indicators [13].

However, fillers do not necessarily need to be part of another self-repair, such as a false start. They can be clear indicators of a self-repair that lacks a hearable repairable, indicators for a word search, or another move in cognitive planning. The word(s) following a filler must be understood as the repairing segment because they constitute the found word(s) or construction. These newly found word(s) or constructions 'repair' the search. Fillers are thus always part of self-repair strategies, and as such they deserve more attention from conversation analysts who seem to consider them a special case requiring no further analysis. As a case in point, Fox, Hayashi and Jasperson [5] do recognize that fillers belong to the same category as other self-repairs, but do not analyze them. They also play no role in Fox and Jasperson's [4] typology of self-repair in which fillers are completely ignored.

Not only conversation analysts, but also psycholinguists who work in the field of self-repair, do not always recognize fillers as part of the disfluency or self-repair 'family.' Bear, Dowding, Shriberg and Price [2], who have developed a labeling system for all types of self-repair, do not in all instances label quasi-lexical (or lexical) fillers. Lexical fillers are even more often ignored by researchers in the study of self-repair. Most of the time they are not mentioned at all, let alone analyzed. Lickley [9] believes that their inclusion in the category of self-repair, which he calls disfluency, is controversial.

The inclusion of lexical fillers (and quasi- or non-lexical fillers, for that matter) is only questionable if we consider the form of self-repair[1] alone; however, when concentrating on the function of self-repair – dealing with some kind of trouble in spontaneous speech – then fillers are clearly a part of this category. They mostly function to gain time and not lose the floor while searching for a word, structure or organizing the remainder of the turn.

Lexical fillers are special in the sense that they often fulfill more than one function at the same time. In addition to 'playing for time' they can fulfill social, interactional, discourse, and symbolic functions, such as engaging the addressee, yielding the floor, asking for feedback, stressing the content of an utterance, making it sound more friendly, and the like [7, 8, 10, 14, 16]. Therefore, lexical fillers are often analyzed under the heading of discourse markers, in which their role as fillers is sometimes not recognized or barely mentioned. As a case in point, even Schiffrin [14] devotes only a few paragraphs in a 350-page book on discourse markers to their role as fillers or 'place-holders' (p. 76).

When fillers are analyzed, either by conversation analysts, sociolinguists or psycholinguists, they are often placed under a different heading, namely 'hesitation phenomena' or 'filled pauses' (e.g., [12]). Some researchers do not acknowledge that they are part of the larger category of 'self-repair' or 'disfluency' (cf. discussion in [10]).

The present study concentrates on fillers and here in particular on a phenomenon that has never before been presented or discussed, namely idiosyncratic fillers in the L2 (non-native) speech of highly fluent bilinguals. Previous studies of hesitation strategies in L2 conversations have focused on beginning L2 learners or speakers. They found that beginners tend to leave their hesitation pauses unfilled making their speech highly disfluent [17]. Native speakers, however, use a variety of fillers to fill their hesitation pauses, such as the lengthening of sounds, quasi-lexical fillers (uh, uhm), lexical fillers (well, you know etc.), and repetitions [10]. Bilinguals also use a variety of fillers in their native as well as in their non-native language. In addition, they tend to use idiosyncratic fillers in their L2.

2. Idiosyncratic fillers

The term 'idiosyncratic filler' has been created after the transcription and a first skimming of sixteen 25-minute conversations produced by English-German bilinguals, made this necessary. The researcher noted that most subjects used one particular lexical filler unusually often in their non-dominant language conversation to fill hesitation pauses or other gaps in her or his turn. This filler was named 'idiosyncratic filler' not only because it is distinct and different for most of the participants employing it, but additionally on account of the fact that it is a noticeable device that gives their conversation a unique, individual mark or style because of its dominance among all other fillers. The

[1] It has to be noted that some researchers believe fillers to have the same form as other self-repairs. Shriberg [13] shows that fillers have the same surface structure as other self-repairs. The present study does agree with her point of view.

idiosyncratic filler is usually but not exclusively a particular lexical filler that is used as often as or more often than all other lexical fillers combined. For each individual, the idiosyncratic filler may be a different filler.

3. Subjects and data collection

The subject group consists of eight bilinguals, four females and four males who use English and German on a daily basis. For all participants, the dominant language is their first language. None of them had grown up as a bilingual, but had learned the L2 initially in a school setting and developed it subsequently in an immersion situation. Gordon, Henry, June, Lauren, and Sue, are native speakers of English while Isabel, Sven, and Werner are native speakers of German.

For every subject, four different conversations of twenty-five minutes each were videotaped in an experimental setting. The speakers engaged in two dyadic speech events - one in English and one in German - with a same-gender partner first and with an opposite-gender partner thereafter. The first two events were taped consecutively. A week later, the last two were recorded consecutively as well.

The data collection yielded about 210 minutes of English and 210 minutes of German conversational data. An overview of the data set is given in Table 1.

Table 1: Overview of data set

Data set	English	German	Total
Words	31,333	31,028	62,361
All self-repairs	2,873	3,516	6,389
All fillers	1701	2402	4103
Idiosyncratic fillers	182	345	527

4. Data preparation

The recorded data were carefully transcribed and divided into units. The clause or a modified clause was chosen as the basic unit. The main coding process started once the data had been divided into units. The information needing to be coded were all elements of self-repair although for this paper only the self-repair type 'filler' more specifically 'idiosyncratic filler' is relevant.

In order to make their contributions comparable and, hence, accessible to quantitative analysis, all the participants' contributions – i.e., the number of their self-repair types and subcategories – in all four conversations were standardized by mathematical manipulation. They were all multiplied by a different factor – depending on the number of words uttered – to adjust the size of their conversational contribution to the largest conversational contribution which was made up of approximately 3,500 words. This allowed the researcher to compare each subject's self-repairs in a quantitative manner.

5. Results and analysis

The quantitative analysis revealed that for all eight subjects their use of fillers changes depending on the language they speak. It has been observed that

- bilingual speakers tend to use more quasi-lexical fillers in their non-native language than in their native language;
- bilingual speakers, in general, tend to use more fillers in their non-native language than in their native language;
- bilingual speakers tend to use a particular lexical filler, a so-called idiosyncratic filler, with unusually high frequency in their non-native language; and
- bilingual speakers tend to use more lexical fillers in their non-native language than in their native language.

Most participants – Sue, Isabel, June, Lauren, Gordon, and Sven – have developed a unique mark or conversational style in their non-native language by using a particular filler noticeably more often than others or even all other lexical fillers combined. Sven displays this behavior in his L2 as well as in his L1. Table 2 shows how many idiosyncratic fillers these six participants used in their two L1 and their two L2-conversations.

Table 2: Idiosyncratic fillers

Idiosyncratic fillers	L1	L2	χ^2
Sue	0	78	77.5*
Lauren	0	209	209.0*
June	75	377	201.9*
Isabel	0	189	188.8*
Gordon	0	66	65.5*
Sven	55	79	4.4*

Legend: Data standardized for 3,500 words per conversation. All frequencies have been rounded to the nearest integer. $\chi 2$ for $p \leq .05 \geq 3.84$ (1df). An asterisk indicates a significant chi-square value.

Table 3 presents the total number of all lexical fillers produced by these six subjects. The total number of lexical fillers includes the idiosyncratic fillers of all participants, except for June's because her idiosyncratic filler is non-lexical since it consists of the lengthening of sounds. The comparison of Tables 2 and 3 demonstrates how dominant the usage of idiosyncratic fillers is, and we comprehend that its production gives these bilinguals' conversations a unique style.

Since it has been stated that the idiosyncratic filler might be a different filler for each individual, it is essential to introduce each participant's unique hesitation strategy. For Sue this is the German filler 'also.' In her German or L2-conversations she even has two specific fillers that she employs almost exclusively. They are 'ja?' and 'also,' however, 'also' is the one that she uses most often and is thus referred to as her idiosyncratic filler. 'Also' is a filler which can be used either at the beginning (example (1)), in the middle (example (2)), or at the end (example (3)) of a clause or turn-constructional unit.

(1) Sue: also ich kann immer fragen [...]
(also) I always have the opportunity to ask someone [...]

(2) Sue: aber reden und also Alltagssprache ... das war viel besser [...]

but speaking and (also) everyday speech ... that was a lot better

(3) Sue: was be- bedeutet bis jetzt? ☺ also
what does until now me- mean? ☺ (also)

Table 3: Total of lexical fillers

Lexical fillers	L1	L2	χ^2
Sue	132	295	61.9*
Lauren	143	259	33.7*
June	44	17	11.8*
Isabel	162	249	18.4*
Gordon	72	127	15.5*
Sven	176	155	1.4

Legend: Data standardized for 3,500 words per conversation. All frequencies have been rounded to the nearest integer. $\chi 2$ for $p \leq .05 \geq 3.84$ (1df). An asterisk indicates a significant chi-square value.

Like Sue, Lauren also employs the German filler 'also' almost exclusively as lexical filler in her L2-conversations. 'Also' can thus be described as her signature filler. Lauren makes use of this idiosyncratic filler most often in initial positions and she frequently uses it in the middle but only seldom in end positions. Sometimes she uses more than one 'also' in a single turn-constructional unit as illustrated in example (4) where she uses three. The first one appears at the beginning, the second one in the middle, and the last one at the very end.

(4) Lauren: ja also ss- ss- die die haben auch ganz gute Leute bekommen also Mike und und Brendan also
yeah (also) they they have also gotten very good people (also) Mike and and Brendan (also)

June is the only participant who does not produce a lexical filler as her signature filler, but the lengthening of vowels and sound combinations. In fact, these lengthenings are so dominant in her L2 conversations that they consist of more than three-quarters of all fillers combined. They can occur in lexical or quasi-lexical items in any position of a clause or turn-constructional unit. It is thus the most flexible filler. Moreover, it is noteworthy that numerous turn-constructional units contain several lengthened vowels, as illustrated in example (5).

(5) June: j=a e=h ihre Mu=tter kommt aus=s Irland [...]
yea=h u=h her mo=ther is from=m Ireland [...]

In her English or L2-conversations, Isabel makes use of one specific filler, namely 'so,' more than twice as often than of all others combined. This is her idiosyncratic filler and she uses it either at the beginning, the end, and sometimes in the middle of a clause or turn-constructional unit. Isabel occasionally uses more than one 'so' in one utterance as can be seen in examples (6) and (7). In the first one, she uses it twice at the beginning of a turn-constructional unit and in the latter she lengthens 'so' and uses it at the beginning as well as at the end. It is not unusual for Isabel to 'sandwich' her turn-constructional unit between two utterances of 'so.'

(6) Isabel: so so that's different [...]

(7) Isabel: s=o it's something in between s=o

Gordon also employs an idiosyncratic filler in his L2-conversations. It is the German filler 'da,' that he uses very frequently. He uses it most often in the middle position of a clause or turn-constructional unit, and he frequently uses it at the end but only seldom in initial positions. Sometimes Gordon uses more than one 'da' in one turn-constructional unit, as shown in example (8), where the second 'da' could be regarded as a location adverb. However, the researcher claims that it is not an adverb otherwise Gordon would have to use 'hier,' (here) because he is talking about the same university at which he is at the time the conversation took place.

(8) Gordon: und wir wir beide eh da w-wir arbeiteten da an der Uni [...]
and we we both uh (da) w-we worked (da) at the university [...]

Sven is the only participant who uses idiosyncratic fillers in his L2 and in his L1-conversations. It is this 'I mean' and 'ne?.' However, he produces more idiosyncratic fillers in his English conversations compared to his German ones. His L2 signature filler can be employed in all three positions, but he mostly uses it in initial and end position. Example (9) illustrates the latter case.

(9) Seven: I don't know why.. I mean

The German 'ne?,' on the other hand, is comparable to the English filler 'right?.' They both tend to appear at the end of turn-constructional units as can be seen in example (10)

(10) Seven: also .. so hab ich 's zumindest gehört ne?
well .. at least that's the way I heard it (ne?)

Henry does not make use of a particular lexical filler considerably more frequently than of other lexical fillers. However, he is the subject who uses the most 'uhs' and 'uhms' in his L2-conversations compared to his L1-conversations, namely 561 compared to 299. Hence, one could argue that quasi-lexical fillers are his signature or idiosyncratic filler.

Like Henry, Werner does not use a particular lexical filler notably more often than other lexical fillers. Moreover, he does not hesitate more frequently when he speaks his non-native language English. Werner, a native speaker of German who has lived in an English-speaking country for many decades, is clearly the most proficient of the eight subjects with a native-like command of both languages and he might therefore not display this particular behavior. Idiosyncratic fillers could be characteristic of very proficient bilinguals who have not or not yet achieved native command of their L2.

6. Discussion

Hesitation pauses and other gaps, like self-repairs in general, may occur at any given moment in a conversation. Therefore, fillers appear in different utterance positions. While the two most versatile filler types, namely the quasi-lexical fillers and the lengthening of sounds can be employed in any position, to some lexical fillers not all positions are available. Fillers like the English 'yeah' and 'okay' are most often found in initial positions where in addition to playing for time they express agreement or acknowledgement or they serve as an uptake or any other link to what the previous speaker has said. A number of fillers, such as 'right?,' 'ja?,' and 'ne?' typically

occur at the end of turn-constructional units. Usually, they also fulfill several functions. They are solicitors of agreement, brief response, and/or attention, and they engage the addressee [10].

A number of fillers can be used in all positions. The German 'also,' 'na,' 'da,' and 'so' and the English 'so,' 'you know,' 'I mean,' and similar expressions, such as 'I think,' 'I guess,' 'I believe,' and their German equivalents belong to this class. In addition to fulfilling the function of a filler they may simultaneously have other functions which vary for different positions. In initial position, they play introductory roles and/or create links between what has just been said and what is about to be said. In middle positions, they can emphasize parts of the utterance or solicit understanding and sympathy. This is especially true for 'you know' and 'I mean,' which can fulfill this particular function in any of the three available positions. In end positions, these versatile lexical fillers often engage the addressee, frame, or emphasize what has just been said [10].

It is apparent that even though the particular filler that each subject 'chose' as his or her idiosyncratic filler is not the same, they all belong to the same class of lexical fillers that can occur in any position. Exceptions to this rule are Sven's German filler 'ne?' which he uses in his L1 and which might therefore be a different phenomenon than the one described here. It can be claimed that in their non-native language six out of eight participants use a particular, very flexible filler unusually often. Apparently they have found a more elegant way than simply using 'uhs' and 'uhms' to deal with a high frequency of hesitation pauses. Certainly, what idiosyncratic fillers have in common with quasi-lexical fillers is the fact that both can be used in any position of the utterance, however, the former are more 'prestigious' than the latter.

Quasi-lexical fillers are often thought of as negative [6] and undesirable flaws. Too many quasi-lexical fillers in one's speech apparently convey the image of an ill-educated, disorganized person [10]. Conversely, idiosyncratic fillers like all lexical fillers reflect favorably on the speaker since their additional social, discourse, and interactional functions make a person's speech more friendly and engaging. It is likely that bilinguals - as individuals with an excellent linguistic and metalinguistic awareness [1] - share this opinion and try to avoid the production of too many 'uhs' and 'uhms.' One way to by-pass the usage of too many quasi-lexical fillers is through the production of other fillers such as lexical fillers or sound-stretching.

While lexical fillers are more prestigious because of their additional functions, for the same reason they are also more complex and thus more difficult to acquire than quasi-lexical fillers. And since the usage of fillers is not taught in the second/foreign language classroom [11], L2 learners and users have to acquire their correct usage without formal training. In the process, they are likely to use acquisition strategies which have been observed for first and second language acquisition of grammatical forms, such as overgeneralization [3]. The unusually frequent production of one particular lexical filler certainly constitutes an overgeneralization of its semantic and pragmatic meaning. Once L2-speakers have 'found' a lexical filler which they can use in all three positions they tend to employ it almost exclusively thereby neglecting the usage of a variety of lexical fillers but also avoiding the production of yet another quasi-lexical filler.

To conclude, the researcher would like to argue that the usage of idiosyncratic fillers in the L2-speech of bilinguals is both a strategy to avoid excessive production of undesirable 'uhs' and 'uhms' as well as an overgeneralization of one particular filler, hence, obviously a characteristic of an interlanguage at the discourse level.

Acknowledgements
This research was supported by Social Sciences and Humanities Research Council of Canada Award 752-98-1241. Please address correspondence to carolin@interchange.ubc.ca

References

[1] Baker, C., & Prys Jones, S. (1998). Encyclopedia of bilingualism and bilingual education. Clevedon: Multilingual Matters.
[2] Bear, J., Dowding, J., Shriberg, E., & Price, P. (1993). A system for labeling self-repair in Speech. SRI Technical Note 522, 1-9.
[3] Edmondson, W. (1999). Twelve Lectures on Second Language Acquisition. Foreign Language Teaching and Learning Perspectives. Tübingen: Narr.
[4] Fox, B., & Jasperson, R. (1995). A syntactic exploration of repair in English conversation. In P. Davis (Ed.), Descriptive and Theoretical Modes. The Alternative Linguistics, (pp. 77-134). Amsterdam: John Benjamins.
[5] Fox, B. A., Hayashi, M., & Jasperson, R. (1996). Resources and repair: a cross-linguistic study of syntax and repair. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), Interaction and grammar (pp. 185-237). Cambridge: Cambridge University Press.
[6] Fox Tree, J.E. (1999). Between-turn pauses and ums. Proceedings of the ICPhS (pp. 15-18), San Francisco: ICPhS
[7] Hänni, R. (1980). What is planned during speech pauses? In H. Giles, W. P. Robinson, & P. M. Smith (Eds.), Language: Social psychological perspectives (pp. 321-26). Oxford: Pergamon Press.
[8] Levinson, S. C. (1983). Pragmatics. Cambridge: Cambridge University Press
[9] Lickley, R. J. (1994). Detecting disfluency in spontaneous speech. Doctoral dissertation, University of Edinburgh.
[10] Rieger, C. L. (2000). Self-repair strategies of English-German bilinguals in informal conversations: The role of language, gender and proficiency. Doctoral dissertation, University of Alberta.
[11] Rieger, C. L. (2001). German and English Conversational Fillers and Suggestions for their Teaching. Manuscript (under review).
[12] Rose, R. L. (1998). The communicative value of filled pauses in spontaneous speech. M.A. thesis, University of Birmingham
[13] Schegloff, E. A., Jefferson, G., & Sachs, H. (1977). The preference for self-correction in the organization of repair in conversation. Language, 53, 361-382.
[14] Schiffrin, D. (1987). Discourse markers. Cambridge: Cambridge University Press.
[15] Shriberg, E. E. (1994). Preliminaries to a theory of speech disfluencies. Doctoral dissertation, University of California, Berkeley.
[16] Stenström, A.-B. (1994). An introduction to spoken interaction. London: Longman.
[17] Temple, L. (1992). Disfluencies in learner speech. Australian Review of Applied Linguistics, 15, 29-44

Disfluencies in Writing – are they Like in Speaking?

Åsa Wengelin

Department of Linguistics and Phonetics
Lund University, Lund
asa.wengelin@ling.lu.se

Abstract

This paper presents a study of disfluencies in written language production. Texts from ten university students are compared to data from people who almost never use writing, namely adult dyslexics and to texts from people who communicate in writing under real-time constraints every day, namely deaf whose main use of writing is text telephone conversations. This paper investigates which types of disfluencies occur in writing, where they occur and their durations. Further, this paper investigates how different text types and the specific characteristics of deaf and dyslexic writers influence the distribution of disfluencies. The results are discussed in relation to earlier work on disfluencies in speaking.

1. Introduction

While spoken communication is typically processed under strict on-line constraints where sender and receiver are located at the same point in time and often also in place, written language has until quite recently mainly been associated with communication where sender and receiver are located at different points in time and place. This is probably one reason why real-time processing phenomena such as pauses, hesitations, repetitions and repairs have mainly been associated with the spontaneous production of spoken language. Another reason is that until quite recently there have been no tools for 'recording' writing and therefore only the final products have been available while recordings of spoken language are always 'on-line'.

However, the use of computers has meant substantial changes for the study of writing. First, they have led to new and more on-line like uses of written communication, like email and chat. These are not studied in this paper. Second, repairs can now be recorded and pauses can be measured automatically by means of on-line logging. This paper presents studies made by means of automatic corpus analyses of on-line recordings of writing.

2. Written language production

2.1. On-line studies of writing

Before 1980 when Gregg and Steinberg [1] published their pioneering book *Cognitive processes in writing*, very little research had been done on writing as a cognitive process. Most earlier studies of writing had addressed pedagogical issues, but with Hayes and Flower's model [2], published in the book mentioned above, the interest started to focus more on the cognitive process of writing, and a new paradigm which attended to the writer's strategies to reach an acceptable solution to a rhetorical problem was developed. Since then there has been an immense increase in studies of written production, mostly by means of think aloud protocols, planning notes , retrospective reports and manual measuring of pause durations. The development of automatic on-line logging of writing, by means of computers has meant substantial changes for the study of written language production. However, there are still few on-line studies of writing around, and many insights could be gained from studies of spoken language.

2.2. Disfluencies and production conditions

In studies of spoken language disfluencies (df:s) have been studied from several different views. Many have been psycholinguistically orientated and in these df:s are often thought to be indicative of the mental processes underlying speech generation while others have focussed on how df:s are processed by the listener (e.g. [3]). A common view of disfluencies is that they indicate problems in the planning of utterances. However, it has also been argued that most disfluencies in spoken language production are not mistakes but signals for coordinating the speakers with their addressees (e.g. [4]), and, that both intraindividual and interindividual perspectives have to be taken into account (e.g. [5]). Most of these studies have been concerned with face to face dialogues. However, the typical writing situation is still the production of a monologue with no addressee present. The sender has plenty of time to plan, encode and revise the message before sending it. Therefore, the interactive function of dislfuencies can hardly be dominant in writing. There are also studies of disfluencies in spoken monologues. In [6] for example, it is shown that there is a system in how speakers produce filled pauses and that these often carry information about larger-scale topical units. Similar patterns could be expected to occur also in writing.

3. Methods

3.1. Subjects and Data Collection

The data used for the current analyses relate to ten Swedish university students, eleven Swedish adults with reading and writing difficulties, and nine congenitally deaf subjects. All subjects were all well acquainted with writing on a computer. The first language of the deaf writers is Swedish sign language, why, their writing my be considered as L2 writing. The data collection was made within the current research programme "Reading and Writing Strategies of Disabled Groups" and comprised 5 production tasks: a picture story task ("Frog where are you?") based on a wordless picture story booklet by the American artist Mercer Mayer [7], a narrative task with a pre-set topic ("I was never so afraid"); a route direction task; a job application task; and a "letter to the editor" type of argumentative discourse task. All writing activities were computer-logged by means of the computer tool "ScriptLog" [8].

3.2. Scriptlog

Scriptlog is a simple editor that keeps a record of all events on the keyboard (i.e., the pressing of alphabetical and numerical keys, cursor keys, the delete key, space bar etc, and mouse clicks), the screen position of these events and their temporal distribution. From a Scriptlog record, you can then derive not only the finally edited text from a writing session, but also the "linear" text with its temporal patterning, pauses and editing operations and log file with detailed information about each keystroke. Examples 1 and 2 below shows a fragment of a finally edited text (with english translations in italics) and a correspondent linear fragment.

(1) "ALDRIG HAR JAG VARIT SÅ RÄDD"
I WAS NEVER SO AFRAID
Aldrig har jag varit så rädd som när jag senaste gången rökte i min fars bil.
I was never so afraid as the last time I smoked in my fathers car.

(2) <START><95.92>ALDRIG HAR GA<2.93>G <DELETE4>JAG VARIT SÅ RÄDD<4.82>2<DELETE>2<DELETE> <4.25>"<4.98><MOUSE(29,0)><2.18>" <5.73> <MOUSE(1,30)><3.20><CR><CR> <154.22> Aldrig har jag varit så rädd<15.07> <DELETE20> r jag varit så rädd som när jag rökte i min fars bil sist<5.20> <DELETE4> <6.53><DELETE21> senaste gången rökte i min fars bil<2.47>.<CR>

Example 1 shows the heading and the first sentence of a narrative written by a control subject. Example 2 shows the linear text as follows: First the start button is pressed. After that the writer waits 95,92 seconds before starting to write - perhaps thinking about how to start, or planning the plot. Then she writes ALDRIG HAR GA and pauses for 2.93 seconds. She writes another G and a space, and then notices that the last word was wrong. She deletes four tokens (the space and GAG) and writes JAG (I) instead. And so on. <MOUSE(29,0)> means that the writer moves 29 tokens backwards in the text and <CR> means Carriage Return (new line).

3.3. The Corpus

The corpus consists of 149 edited texts containing altogether 39527 words. Table 1 shows an overview of how the words are distributed over the three groups. However, as was mentioned above the finally edited texts don't show everything that has been written but only those parts which the writers have chosen to keep. Therefore the total number of keystrokes it has taken to produce these texts and the final numbers of characters left in the texts are also shown. The table further presents the number of pauses and editings which have been used for the analyses in this paper. The frog story is excluded from many analyses and the numbers in italics show the relevant numbers for the other four texts taken together.

Table 1: Overview of the corpus

Group	keystr. lin	keystr. fin	words fin	pauses	ed
Norm	115337	97314	16911	7023	
excl frog	*66734*		*9290*	*2576*	*1167*
Dys	77963	61292	11854	12461	
excl frog	*46695*		*6720*	*4815*	*1212*
Deaf	73894	57108	10762	4311	
excl frog	*33445*		*4749*	*1095*	*690*
Sum	267194	215714	39527	23795	
excl frog	*146874*		*20 759*	*8486*	*3069*

4. Analyses

Three types of disfluencies were found. As expected unfilled pauses are the most common but also repairs are also frequent. Only quantitative analyses of these, which could be made automatically are presented. The pros and cons of this are discussed in section 5. However, in the texts produced by the deaf subjects a phenomenon that resembles filled pauses also occur. The analyses of these three types of df:s are presented in the following sections.

4.1. Unfilled pauses

4.1.1. Pause definition

By a pause, is here meant a transition time between two keystrokes which is longer than what can be expected for merely finding the next key. To make a pause a writer has to interrupt his/her typing considerably longer than the normal transition time between two keystrokes. Therefore pause criteria should optimally be individually tailored according to each subjects' individual typing speed.

However, like in some earlier studies of writing (e.g. [9]) it was stipulated that all transition times longer than two seconds should count as pauses. This is more than twice the median transition time between two letters for our slowest subject (0.817 sec). The subjects in the three groups type with different speeds. The deaf are in general the fastest and the dyslexic group the slowest.The advantage of the two second criterion is that it is quite safe to assume that transitions counted as pauses really are pauses. The disadvantage is that some pauses of the faster writers may not be included.

Pauses are categorized according to which type of *microcontext* they occur in.Each transition between two keystrokes is a possible pause location and a pause could for example occur in the transition between two lower-case letters as just mentioned, or between a space and a letter etc. Table 2 describes the notation used to describe microcontexts. According to this 'a*a' describes a pause that occurs between two letters, and a_*a describes a pause that occurs between a space, which follows a letter, and a letter etc.

Table 2: Notation used to describe microcontexts

character	description
a	a letter
_	a space
*	a pause
D	a deletion
. and ,	major and minor delimiters

4.1.2. Overall pause frequencies (pauses > 2 sec)

Table 3 shows the frequency of pauses in subcorpora of the three groups, with and without the frog story.

Table 3: Pause frequencies

Group	Pause/word		Pause/keystroke	
	Incl frog	Excl frog	Incl frog	Excl frog
Norm	0.359	0.305	0.051	0.041
Dys	0.895	0.753	0.135	0.108
Deaf	0.378	0.320	0.050	0.040

First, we observe that for all groups pauses in writing are more frequent than all disfluencies taken together in speaking. See for example [10] for an account of df:s in Swedish dialogues. Second, the pause frequencies were compared both for group and for text types. There was a significant effect for both ($p < 0.0001$). All groups make more pauses when writing the frog story than when writing the other texts. The explanation for this is probably that the subjects look at the pictures a lot. The frog story is therefore excluded from the rest of the pause analyses. Further, the dyslexic subjects make many more pauses than the other two groups . This was expected, both due to their writing problems and to their lower writing speed.

4.1.3. Pause lengths

First, the overall pause lengths were compared for group and text type. There was no effect for group alone and only a weak effect for text type. However, there was a significant effect for group and text type together (p=0.0075). The deaf subjects made much longer pauses in the "letter to the editor" texts than in the other text types. It is difficult to find a good explanation for this, but perhaps they had little experience of writing argumentative discourse.

Second, the lengths of pauses were compared for different microcontexts. Table 4 shows the results of this analysis.

Table 4: Mean Pause length (pauses > 2 sec)

Context	Norm		Deaf		Dys	
	n	mean len	n	mean len	n	mean len
a_*a	506	5.2	213	4.8	1152	4.7
a*a	101	3.3	93	5.2	736	4.1
a*_a	252	5.6	187	4.2	579	5.2
._*a	112	6.9	55	5.1	184	6.4
a*.	103	7.6	90	3.9	180	7.5
.*_a	51	6.7	34	5.3	57	11.6
,_*a	35	6.7	8	7.0	18	5.4
a*,	72	4.3	26	4.0	52	10.7
,*_a	4	7.1	3	3.7	5	4.6
a*D	259	7.5	65	4.4	500	8.0
D*D	0	–	39	7.9	214	8.1
D*a	98	6.1	58	3.7	500	6.0

The table is divided into four parts. The first part deals with microcontexts that are likely to occur in the beginning of, within and the end of a word. The second deals with microcontexts which are likely to occur in the beginning and end of a sentence, and the third deals with microcontexts associated with units which are delimited by a comma. The fourth part deals with microcontexts associated with deletions and will be discussed in section 4.3. The table shows how many pauses of each context occur in the corpus and their mean length. The table shows that for the control group and the dyslexic group the mean pause lengths tend to be longer for the sentence related pauses than for the word related pauses. However, surprisingly this effect is significant only for the dyslexic group (p=0.0001). There are several possible explanations for this. First, it may be due to the pause criterion. If shorter pauses were included for the faster typists the means may look different. Notice that the differences in amounts of pauses are much higher for the word related pauses than for the sentence related pauses. Another explanation may be that the pause categories need to be more fine-grained. Perhaps we would see another pattern if we also marked phrase boundaries in the texts. The deaf group show an even more even distribution than the other two groups. It is impossible to tell if this is due to any special characteristic of the group or if the same explanation holds for the deaf as for the normal writers. Nothing specific was found for the comma related contexts.

4.1.4. Pause frequencies in certain microcontexts

Let us now turn to the frequencies in the same microcontexts. Table 5 shows for each context the percentage of pauses in each microcontext and how many of the subjects in each group had included this type of context in their texts at all. For example only eight normal subjects appear to have written `a*`, and 36.6% of these microcontexts cooccurs with a pause > 2 seconds.

Table 5: Proportions of pauses > 2 sec in certain contexts

Context	Norm		Deaf		Dys	
	mean perc	no subj	mean perc	no subj	mean perc	no subj
a_*a	7.1	10	4.8	9	18.9	11
a*a	0,3	10	0.5	9	3.4	11
a*_a	2.8	10	4.3	9	9.2	11
._*a	31.3	10	21.4	9	55.1	11
a*.	22.3	10	23.8	9	52.7	11
.*_a	12.4	10	17.1	9	19.6	11
,_*a	11.1	10	32.1	8	25.7	9
a*,	36.6	8	27.0	8	84.7	9
,*_a	2.5	8	14.7	8	3.3	9
a*D	26.4	10	13.8	9	50.8	11
D*D	0.0	10	1.4	9	5.8	11
D*a	11.3	10	10.0	9	54.1	11

Although pause lengths didn't increase with unit size, the sentence related microcontexts appear to cooccur with pauses much more often than the word related pauses. This result appears to hold for all groups, and perhaps it could partly explain the results of the length analysis. Notice the very high percentage of pauses between a letter and a comma for the dyslexic group. One possible explanation for this may be that they often use comma instead of full stop.

4.2. Filled pauses?

As has been mentioned earlier, filled pauses are not expected to occur in written monologues. However, in the texts of the deaf group, something which resembles filled pauses occur. They use series of dots ranging from two to five full stops in a row. 392 occurrences of the microcontext `.*.` were found in their texts. The corresponding numbers for the normal writers were 12 and 0. These series appear as units. There are no pauses within them. Eight of the nine deaf subjects use these patterns.

A possible explanation for this result could be that it is a text telephone convention and that they are influenced by that, since their main use of writing is in text telephone conversations.

4.3. Repairs

There are different strategies to make repairs in writing. Obviously they don't begin with an explicit editing phrase like they sometimes do in speaking. The two main repair strategies is to start deleting from the current character insertion point and to move to another point in the text and make a deletion and/or an insertion. More than 90% of all repairs in this corpus involved deletions. By measuring the editing range (how much is deleted) and the editing distance (how far the writer moves to get to the repair point) we can get an idea of how a writer makes repairs.

Table 6 shows the mean editing frequency, the mean editing range, the proportion of editings with a range of only one letter and the mean editing distance for the three groups.

Table 6: Deletions: frequency and range

Group	Ed/key	Ed/word	Range	1-let	Dist
Norm	0.014	0.136	3.262	55%	4.39
Deaf	0.017	0.165	3.408	44%	0.13
Dys	0.021	0.188	3.101	54%	3.92

Like for pause frequencies we observe that the frequencies of repairs are higher in writing than in speaking. All of these, except the editing distance were compared for group and for text type. There was no difference between the groups for any of these variables , but an effect for text type was found for editing frequency (p=0.0128). For some reason, all groups made more repairs in the route descriptions. It is worth noticing that about 50% of the repairs are changes of only one letter. These are due to either spelling problems or typos. Concerning the editing distance, the variance between the subjects is too large to make any reliable statistical analyses. Many subjects have total editing distance which is zero. Among the ten normal writers six subjects have an editing distance > 0 in 16 texts altogether. For the dyslexic subjects the correspondent results are nine subjects in 25 texts and for the deaf only two subjects in four texts.

Let us finally look at how pauses and repairs interact (see table 4). While there were no great differences between the groups in how long pauses they made in word, sentence and comma related microcontexts, we find that the deaf subjects appear to make shorter pauses than the other two groups before and after repairs. These results agree with their short editing distance. However, within repairs both deaf and dyslexics make quite long pauses while the normal writers don't make any pauses longer than two second at all. Perhaps this resuts could be explained by their writing problems for the dyslexics and L2 situation for the deaf.

5. Summary and conclusion

To sum up unfilled pauses apppear to be the most frequent df in writing. Further, pauses and repairs appear to be more frequent in writing than in speaking. Considering the different production conditions between spoken and written language this was an expected result. As in spoken language production sentence boundaries cooccur with pauses more often than word boundaries. These results holds for all groups and could therefore be considered as quite robust. Perhaps it could be suggested that while the inter-individual functions of df:s are different in speaking and writing, the intra-individual are quite similar.

Concerning the specific characteristics of the the two groups with disorders, the dyslexic subjects are more disfluent than the normal writers. This was expected, considering their spelling problems. They have a higher pause frequency than the other two groups in general, and specifically a high proportion of pauses within words. Surprisingly they were the only group who showed a significant difference between the pause-lengths associated with sentences and words. It is suggested that this result is an artefact of the high common pause criterion and the fact that no intermediate constituents were analysed. A manual tagging for phrases and clauses within the sentences are suggested as further work. A methodology for setting individual pause criteria, related to each writer's typing speed is also needed.

The deaf subjects on the other hand are not specifically disfluent and their writing process is quite linear. They are fast typists and don't often move far in the text to make a repair. Interestingly, a specific characteristic of their text is that a phenomenon, which resembles filled pauses, sometimes occur in their production. These results could perhaps be explained by their frequent use of text telephone, where writing is used under on-line production conditions, and it is suggested that the on-line/off-line distinction is at least as important as the modality distinction for the language production process. A way of investigating this issue further would be to make on-line recordings of text telephone conversations.

6. References

[1] Gregg, L.W. and Steinberg, E.R., Eds, Cognitive processes in writing, Hillsdale: Lawrence Earlbaum Associates Inc. 1980.

[2] Hayes, J.R. and Flower, L.S. "Identifying the organisation of the writing process" , In: Cognitive processes in writing, Gregg, L.W. and Steinberg, E.R., Eds., Hillsdale, NJ: Lawrence Earlbaum Associates, 1980, pp 3-30.

[3] Lickley, R. and Bard, E. "On not recognizing disfluencies in dialogue", Proceedings International Conference on Spoken Language Processing, Philadelphia, PA, Vol 3, 1996, pp 1876-1879,

[4] Clark, H. "Speaking in time", Proceedings of the ESCA workshop on dialogue and prosody 1999, pp. 1–6.

[5] Allwood, J., Nivre, J. and Ahlsén, E. "Speech Management — On the Non-Written Life of Speech", Nordic Journal of Linguistics, Vol 13, 1990, pp 2–48.

[6] Swerts. M. Filled pauses as markers of discourse structure", Journal of Pragmatics,Vol 30, 1998, pp 485–496

[7] Mayer, M., Frog, where are you?, New York: Dial Press, 1969

[8] Strömqvist, S. and Malmsten, L. "ScriptLog Pro User's manual", Gteborg University, Dept of Linguistics, 1998

[9] Spelman Miller, K. "Academic writers on-line: investigating pausing in the production of text", Language Teaching Research, Vol 4, No 2, 2000 pp. 123–148.

[10] Eklund, R., "A comparative analysis of disfluencies in four Swedish travel dialogue corpora", Proceedings of Disfluency in Spontaneous Speech Workshop, Berkely, Californa, 1999, pp 3–6.

The Usage of Fillers at Discourse Segment Boundaries in Japanese Lecture-style Monologues

Michiko Watanabe

The University of Tokyo, CREST/JST
watanabe@gavo.t.u-tokyo.ac.jp

Abstract

We examined whether fillers (filled pauses) in a Japanese lecture appeared more frequently after discourse segment boundaries (DSB) than after other sentence boundaries. Contrary to our hypothesis that fillers occur more often after DSB than after other sentence boundaries, the frequency of fillers in the first phrase after DSB did not differ statistically from that after other sentence boundaries. The location of fillers in the first phrase after DSB and after other boundaries did not show any clear difference, either. However, the types of fillers at the initial position of the first phrase after two kinds of boundaries were different; sentence initial 'eto' appeared exclusively at DSB. This result indicates that sentence initial 'eto' may help highlighting DSB, but not other types of fillers. Other kinds of fillers ('e', 'ma', 'ano', 'sono') seem to be mainly concerned with planning units of the utterance that are smaller than a sentence.

1. Introduction

1.1. General

Fillers are such utterances as 'uh' and 'um' in English and 'ano' and 'eto' in Japanese. They have neither clear grammatical function nor semantic meaning, but are commonly observed in spontaneous speech. Among Japanese fillers referred to in previous studies are 'a', 'ano', 'de', 'desune', 'e', 'eto', 'kono', 'ma', 'n', 'nto', 'nanka', 'sono', 'sodesune', 'nante iimasuka (what should I say)', and word final vowel lengthening ([1]-[5]). However, what are considered as fillers vary among researchers.

Research on fillers can be grouped into four categories; 1) one which regards fillers as defects in communication and tries to find out ways to decrease them ([6]); 2) one which tries to find out correspondence between occurrence of fillers and speaker's mental processes or emotional states ([7]); 3) one which admits functions as discourse markers in fillers ([8]); 4) one which tries to find out acoustic characteristics of fillers to allow automatic detection for effective speech recognition ([9]). While there is much research on English fillers, not many studies have been conducted on Japanese fillers except from the forth point of view above. The present research is mainly relevant to the second and the third groups.

[4] and [5] are among the few studies on Japanese fillers relevant to the second and third groups. In this literature the authors claim that interjections and responses are relevant to the speaker's mental operations such as input, output, search, registering and editing of information. Each item, according to them, corresponds to a different kind of processing. By using one of these devices, speaker can not only monitor and help his own processing, but also inform a listener of his mental states, and thus, keep smooth communication. They claim that fillers are concerned with output processes. They have certain features in common, but each of them has functionally different aspects. As for 'eto' and 'ano', according to them, they both hint that the speaker has trouble in output processes and needs more time to continue his speech. However, they differ in that 'eto' is uttered when the speaker is searching for knowledge or conceptualizing ideas using his knowledge, while 'ano' is uttered when he has trouble finding suitable forms for content. Their claim is based on such observations that one utters 'eto', but not 'ano', when one is engaged in calculation, and that one utters 'ano', but not 'eto', when one tries to remember names of things or persons that one knows, or when one asks for a favour politely. Although their view appears correct, their model needs to be tested more empirically and elaborated. It would also be necessary for functions of other fillers to be included in the model.

We have investigated university lectures and speech on academic conferences to find out frequent fillers and their distributions, supposing that different types of fillers have different functions if there is distributional difference ([10], [11]). What we have found so far is as follows;

1. The most frequent fillers in these monologues were 'ano', 'e', 'eto', 'ma' and 'sono'; they covered about 90 % of all the fillers.
2. Among these fillers 'e', 'eto' and 'ma' tended to appear at stronger syntactic boundaries such as sentence- and clause-boundaries more often than 'ano' and 'sono'. 'E 'was most frequent at clause boundaries, and 'ano' after a topic particle 'wa'. 'Sono' hardly ever occurred at sentence boundaries.

These findings indicate that 'e', 'eto' and 'ma' tend to be concerned with planning larger units of an utterance than 'ano' and 'sono'.

1.2. Fillers and Discourse Segment Boundaries

As we mentioned in the previous section, some researchers have investigated whether fillers convey information about discourse structure. Swerts *et al.* ([12]) found that, in Dutch, phrases right after major discourse boundaries contained more fillers than those after minor boundaries, and that fillers after stronger breaks tended to occur at phrase-initial position, while those after weaker breaks at phrase-internal position. They also pointed out that 'um' tended to occur phrase-

initially, whereas 'uh' phrase-internally. They concluded that fillers seemed to carry information about discourse segment boundaries, and that difference in types might reflect different planning processes.

Their research motivated us to investigate into occurrence of Japanese fillers at discourse segment boundaries. Fillers must share some features in common across languages. If fillers are relevant to speech planning, it is reasonable to suppose that they appear more frequently at discourse segment boundaries than at minor boundaries, because the speaker is supposed to do discourse level planning as well as more local planning there.

As is mentioned in the previous section, 'e', 'eto' and 'ma' tended to appear at deeper syntactic boundaries than 'ano' and 'sono', and seem to reflect planning a larger unit of the utterance. As a discourse segment is usually larger than a sentence, it is likely that the former group of fillers tends to occur more frequently at discourse segment boundaries than the latter. As for 'ma', distribution was divergent among speeches on academic conferences ([11]). Therefore, we focused our attention mainly on 'e' and 'eto' in the former group in the present research.

The hypotheses tested in this research were as follows;

1) Fillers appear more frequently in the vicinity of discourse segment boundaries than at other sentence boundaries.
2) Frequent location of fillers in a phrase after discourse segment boundaries differs from that in a phrase after other sentence boundaries.
3) 'E' and 'eto' tend to appear more frequently at discourse segment boundaries than 'ano' and 'sono'.

2. Method

2.1. Material

We used an excerpt from a university lecture as material. The lecture was about international law. It is a part of larger corpus of lectures. We chose this lecture, because the speaker was a native speaker of Tokyo Japanese, and his speaking speed was about an average of the lectures in our corpus.

The lecture was recorded on a DAT (Sony TCD-D100) using a microphone (Sony ECM-717) in a large audience room in university. Forty-one minutes of speech from the beginning of the lecture was taken for analysis.

The excerpt was transcribed in Japanese orthography. The transcription was divided into smaller units bounded by perceptual pauses by the author. The number of units amounted to 1527. Rough average duration of a unit with a following pause is 1.6 seconds. Hereafter, this unit is called an 'inter-pausal-unit' (IPU).

The material contained 682 fillers in total. This means that the speaker uttered a filler every 3.6 seconds on average. Kinds and numbers of fillers which appeared in the sample are shown in Table 1.

Table 1: Kinds and numbers of fillers, and the ratio of each type of fillers in the sample

Fillers	eto	e	ano	sono	ma	Others	Total
Frequencies	13	177	261	65	90	76	682
%	2	26	38	10	13	11	100

2.2. Procedures

Two people participated in segmenting the text. Both were university lecturers. They were instructed to segment the script based on the speaker's purpose of utterance. They were told not to divide an IPU except when they were definitely sure that there was a boundary in it. They were also instructed to write down a purpose of each segment. An example of a segmented text (part of another lecture) was shown before they started segmenting.

2.3. Method of analysis

Labeler A divided the text into 93 discourse segments, and labeler B, 87 segments. The Kappa value of inter-raters' agreement on locations was .63, which we regarded good enough to base our further discussions on.

We call the locations where both labelers marked boundaries 'discourse segment boundaries' (DSB), and each agreed segment a 'discourse segment' (DS). We got 60 DSB. 53 out of 60 DSB accorded with sentence boundaries, and seven with clause boundaries. As most of them were sentence boundaries, we decided to compare distribution of fillers after sentence boundaries of DSB with that after sentence boundaries which neither of the labelers marked as DSB. We call the latter sentence boundaries 'non-discourse segment boundaries' (NDSB). We got 56 NDSB, out of 140 sentence boundaries in total.

First, we examined what percentage of the first IPU after DSB and NDSB included fillers. Second, we investigated what kinds of fillers appeared at initial and non-initial positions in the first IPU after DSB and NDSB. We also examined kinds and frequencies of parts of speech that occurred right after DSB and NDSB.

3. Result

Table 2 shows numbers and percentages of the first IPU after DSB and NDSB that included fillers. There was no statistical difference between these percentages. The first IPU after DSB did not include fillers more often than those after NDSB.

Table 2: Numbers and percentages of the first IPU after DSB and NDSB that included fillers

	IPU with fillers	IPU without fillers	Total
DSB	25 (47%)	28 (53%)	53
NDSB	26 (46%)	30 (54%)	56

Table 3 shows numbers and ratios of fillers at IPU initial and IPU non-initial positions after DSB and NDSB. There is no statistical difference between these ratios.

Table 3: Numbers and ratios of fillers at IPU initial and non-initial positions after DSB and NDSB

	IPU initial	IPU non-initial	Total
DSB	11(44%)	14(56%)	25
NDSB	15(58%)	11(42%)	26

Table 4 shows kinds and numbers of fillers at sentence initial positions right after DSB and NDSB. 'Eto' was the most frequent filler after DSB. All the sentence initial 'eto' appeared at DSB, and none of them at NDSB. 'Eto' occurred more often than 'ano' and 'sono' at DSB, as we had expected. However, contrary to our hypothesis, 'e' occurred only once at DSB. It occurred more often at NDSB.

Table 4: Kinds and numbers of fillers at sentence initial positions after DSB and NDSB

	eto	e	ano	sono	ma	Total
DSB	6	1	3	0	1	11
NDSB	0	6	2	1	6	15

Table 5 shows kinds and numbers of fillers that appeared at non-sentence initial positions in the first IPU after DSB and NDSB. As we can see from the table, there is no clear difference between the distributions of fillers in the two locations.

Table 5: Kinds and numbers of fillers appearing at non-sentence initial positions in the first IPU after DSB and NDSB

	eto	e	ano	sono	ma	Total
DSB	0	0	8	3	3	14
NDSB	0	1	6	3	1	11

Table 6 shows frequencies of parts of speech including fillers that appeared at sentence initial position after DSB and NDSB. Figure 1 shows their ratios. χ^2 analysis revealed that there was statistical difference in the distribution of fillers, connectives and others (including nouns, demonstratives and adverbs) ($\chi^2 = 30.4$, df = 2, $p < .00$). There was statistical difference between the ratios of connectives and others. However, there was no statistical difference between the frequencies of fillers after DSB and NDSB.

From Figure 1, it is clear that connectives occurred significantly more often at DSB than at NDSB. 75% of the sentences after DSB started with a connective, whereas those starting with a connective after NDSB were 29%.

Table 6: Frequencies of parts of speech (including fillers) appearing at sentence initial position after DSB and NDSB

	Filler	Connective	Others			Total
			Noun	Demonstrative	Adverb	
DSB	11	40	0	0	2	53
NDSB	15	16	11	11	3	56

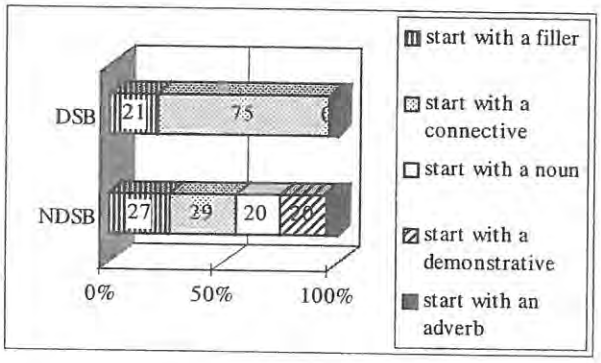


Figure 1: Ratios of parts of speech including fillers that appear at sentence initial position after DSB and NDSB

Table 7 shows the details of connectives that appeared at sentence initial position after DSB and NDSB. 'De' (abbreviated form of 'sorede') was by far the most frequent connective particularly after DSB. About a half of the sentences after DSB (27 out of 53) started with 'de', whereas sentences starting with 'de' after NDSB were 16% (9 out of 56).

Table 7: Connectives at sentence initial position after DSB and NDSB

	DSB	NDSB
de (and)	27	9
sorede (and so)	1	
sorekara (and then)	4	
dakara (so)	4	4
tokorode (by the way)	4	
soredemo (even so)		1
tokoroga (however)		1
aruiwa (or)		1
Total	40	16

4. Discussion

Our hypothesis, 1) Fillers appear more frequently in the vicinity of discourse segment boundaries than at other sentence boundaries, was not supported by the result. IPU

right after DSB did not contain fillers more often than those after NDSB. This means that, unlike Dutch, difference in quantity of fillers near boundaries cannot be a marker of DSB in Japanese.

As for the frequent location of fillers, while fillers tended to appear more often at phrase initial position after major boundaries, and phrase internally after minor boundaries in Dutch, we had no such difference in Japanese. Therefore, hypothesis 2) was not supported by the result, either. However, when we looked at types of fillers at phrase initial position after boundaries, 'eto' appeared exclusively after DSB. This means that sentence initial 'eto' can be a marker of DSB.

As for hypothesis 3), 'eto' appeared more frequently at DSB than 'ano' and 'sono', but not 'e'. 'E' hardly ever occurred at DSB. This indicates that 'e' is hardly ever concerned with planning a DS. Based on these results, our hypothesis 3) should be modified to "'eto' tends to appear more frequently at discourse segment boundaries than 'ano' and 'sono'".

The discrepancy of the results of research on Dutch and Japanese fillers may be attributed to differences in the unit of analysis. Swerts et al. based their analysis on 'prosodic phrases', while we based ours on IPU. However, the difference in the first element after boundaries remains to be explained, because a difference in the unit of analysis did not play any role here.

Another possible explanation for the discrepancy of the results is that it derives from difference in speech samples. Although both samples were monologues, Swerts et al. used description of paintings, while we used lectures. Our speech material may well be less spontaneous, because lectures are usually planned in advance and often rehearsed. With more spontaneous speech, results may be different.

5. Conclusions

The present research tested the three hypotheses;

1) Fillers appear more frequently in the vicinity of discourse segment boundaries than at other sentence boundaries.
2) Frequent location of fillers in a phrase after discourse segment boundaries differs from that in a phrase after other sentence boundaries.
3) 'E' and 'eto' tend to appear more frequently at discourse segment boundaries than 'ano' and 'sono'.

Hypothesis 1) was not supported by the result. There was no quantitative difference in frequencies of fillers at the two boundaries. Hypothesis 2) was not supported, either. We did not find any clear difference in the two kinds of phrases. However, frequent types of fillers at initial positions were different. Sentence initial 'eto' exclusively appeared at DSB. Therefore, 'eto' is the only filler that can be a marker of DSB. Hypothesis 3) was true only with 'eto'. 'E' did not appear more frequently at discourse segment boundaries than 'ano' and 'sono'.

From these findings, it is likely that, among the four fillers investigated here, only 'eto' may convey information about DSB, but not the other fillers. The others seem mainly relevant to planning units of an utterance that are smaller than a sentence. Japanese fillers as a whole do not seem to have a lot to do with DSB marking, and discourse level planning in the present research.

The next step will be to increase the quantity of speech samples. Examining more spontaneous speech will help to find out whether Japanese fillers have little relevance to DSB marking and discourse level planning. As most fillers seem to be concerned with planning units of an utterance that are smaller than a sentence, it may be more reasonable to examine the usage of fillers at more local level to establish a comprehensive model of function of each type of fillers.

6. References

[1] Kawamori, K., and Shimazu, A. (1996) "On the form and function of Japanese discourse markers" *Technical Report of IEICE: NCL*96-5, pp. 27-32.

[2] Koide, K. (1983) "Hesitation" in *Lecture Series of Japanese Expression 3: Expressions in Speech* pp. 81-87. Mizutani, Y. (Eds.) Chikumashobou

[3] Nakagawa, S., Kobayasi, S. (1995) "Phenomena and acoustic variation on interjections, pauses, and repairs in spontaneous speech" *Journal of the Acoustical Society of Japan* 51 (3), pp. 202-212.

[4] Takubo, Y. and Kinsui, S. (1997) "The function of responses and interjections in discourse" *Speech and Grammar*, pp. 257-279. Spoken Language Working Group (Ed.) Kuroshioshuppan

[5] Sadanobu, T. and Takubo, y. (1995) "The monitoring devices of mental operations in discourse – a case of 'eeto' and 'ano(o)'-", *Gengo Kenkyu.*, 108: 74-93.

[6] Christenfeld, N. (1996) "Effects of a metronome on the filled pauses of fluent speakers" *Journal of Speech and Hearing Research* 39 (6) pp.1232-1238.

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

[8] Schiffrin, D. (1987) *Discourse markers*. Cambridge: Cambridge University Press.

[9] Goto, M., Itou K., and Hayamizu, S. (1999) "A real-time system detecting filled pauses in spontaneous speech" *Information Processing Society of Japan SIG Notes* 99 (64) pp.9-16.

[10] Watanabe, M. and Ishi, C. T. (2000) "The Distribution of Fillers in Lectures in the Japanese Language" *Proceedings of the 6th International Conference on Spoken Language Processing*. Beijing : Volume3 167-170.

[11] Watanabe, M. (2001) "An analysis of usage of fillers in Japanese lecture-style speech" *Proceedings of the Spontaneous Speech Science and Technology Workshop*. Tokyo, pp. 69-76.

[12] Swerts, M., Wichmann, A. and Beun, R. (1996) Filled pauses as markers of discourse structure. *Proceedings of ICSLP* 96, pp. 1033-1036.

Dialogue moves and disfluency rates

Robin J. Lickley

Department of Theoretical and Applied Linguistics and Human Communication Research Centre
University of Edinburgh, Scotland
robin@ling.ed.ac.uk

Abstract

Many factors conspire to cause speakers to produce hesitations and self-repairs in dialogue. It has been noted that disfluency rates vary between corpora, with different overall dialogue tasks and with different modalities (e.g. human-computer vs. human-human) and between speakers, where they play different roles within a given dialogue.

In this paper, we attempt to account for some of these results by examining the interaction between rates of different types of disfluency and types of utterance (dialogue moves) within one corpus of human-human task oriented dialogues.

We find both that overall disfluency rate varies by dialogue move type, with moves which require more planning producing more disfluency, and that the distribution of disfluency types varies between move types, most notably with complex and negative responses to questions producing more filled pauses than positive replies and other moves.

This work helps us to understand how dialogue structure can account for differences in disfluency rates between and within speech corpora and has implications for research in speech production and perception, discourse studies, dialogue management and automatic speech recognition.

1. Introduction

As if understanding normal spontaneous speech wasn't difficult enough, we have to make it tougher, by throwing in disfluencies with alarming frequency. So, why be disfluent? Two major functions of disfluency are **hesitation** and **self-repair**.

Speakers may hesitate before or during an utterance for any of a number of reasons. Hesitation may take the form of pausing, either via silence, filled pause (e.g. *em uh, um*) or word prolongation or some combination of these, or it may occur as repetition of utterance onset.

- Planning what to say next takes time, whether it be a matter of formulating the whole utterance, or selecting the most appropriate lexical item, searching memory for a word or to answer a question [1], or physically searching for information needed to complete an utterance (e.g. *his phone number is ...*) – speaker pauses while looking up a number).
- If an error has been detected, whether overtly or covertly (i.e. prior to articulation: e.g. [2,3]), it may take time to replan the utterance and produce the repair.
- In dialogue, a speaker may hesitate to ensure that their interlocutor is paying attention [4] and not speaking at the same time or attending to some other task or to ensure that they can be heard when competing with extraneous noise.
- Also within dialogue, the realisation that there is a mismatch between interlocutors' representations of the current dialogue state [5] may necessitate hesitation for replanning purposes.

Speech production is a dynamic act which involves more or less constant planning and self-monitoring on many levels. Spontaneity results in errors, and in the vast majority of cases, speakers detect and correct their own errors, producing self-repairs. Levelt [2] distinguishes two major classes of self-repair:

- in *appropriateness repairs*, the speaker realises that the message needs to be made more specific in some way, for the meaning to be conveyed correctly (e.g. *go **past** - just a short distance past the tree*).
- in *error repairs*, the speaker detects a mistake at some level of production and this needs to be corrected (e.g. *go **path** - go past the tree*).

As with hesitation, speakers may be forced into performing repairs in order to adjust for their view of the hearer's knowledge state (e.g. *turn left at the po- do you know where the post office is?*).

While most disfluencies fall into the categories of hesitation and repair, some cases may simply be habitual. The "RP stutter", for example, multiple repetitions of utterance-initial words is fairly common in certain groups of Southern British English.

Previous work has given us insights into how disfluency rates vary within and between speech corpora.

Within corpora, role, sex, eye-contact and familiarity have been found to have some effects on disfluency rates. We find higher disfluency rates for speakers whose role it is to give instructions compared to their interlocutors [7,8,9]. There is some evidence that male speakers may be more disfluent than females [7,8,9,10,11]. Work on our corpus suggests that there are some effects of eye-contact on disfluency rates, with higher rates of repetitions when people are unable to see each other [7], and effects of familiarity, with speakers who do not know each other producing more disfluency [9].

Uncertainty in answering questions has been found to result in greater use of filled pauses within general knowledge tests [1,12], though no comparison was possible with other types of utterance.

Syntactic complexity has effects on repetition rates, with complex structures more likely to be prefixed by repetition disfluencies than simple structures [13].

Discourse structure has also been found to interact with disfluency rates. In monologues, disfluencies cluster around ideational segment onsets (e.g. [14]). Within dialogues, we have a similar finding, that utterances which commence larger dialogue units contain more disfluencies [9]. Longer utterances have been found to be associated with higher rates of disfluency [15], but not for all corpora [11].

Other cross-corporal differences include higher disfluency rates for human-human than for human-computer dialogues,

higher rates in telephone conversations than face-to-face and lower rates when dialogues are more constrained [11,15].
Various of the works cited above have taken into account different types of disfluency. Some have looked at dialogue structure. This work is the first that attempts an analysis of a large corpus of spontaneous dialogues, taking into account variations in dialogue move types and disfluency types. We ask what kinds of dialogue moves are most likely to contain disfluencies, whether any differences remain when we take into account length of utterance and whether different move types attract different disfluency types.

2. Corpus and methods

The materials for this study come from the HCRC Map Task Corpus (hereafter, MTC) [16], a collection of 128 dialogues between 64 Scottish undergraduates, comprising about 15 hours of spontaneous dialogue and around 150,000 words. In each dialogue, speakers take the role of instruction Giver or instruction Follower, the Giver describing a route through a map to the Follower who has the same basic map but a slightly different set of landmarks to negotiate. The entire corpus has been transcribed orthographically, with individual words and intervals between words time-aligned with the speech signal and annotated at many levels. Importantly, the annotation schemes are in a format (XML) which allows us to make enquiries about interactions between the various levels [17].

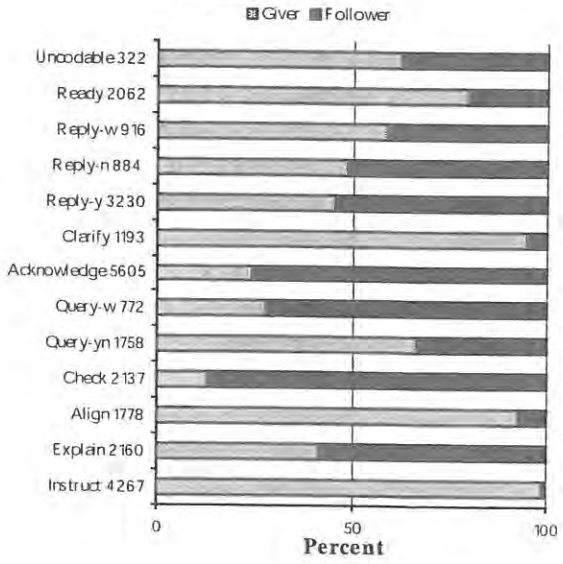
2.1. Move annotation

The MTC is fully annotated for three levels of dialogue structure: **transactions** are subdialogues which form major units in the overall task (typically, in the MTC, a transaction involves the completion of one segment of the route on the follower's map); **games** are components of transactions and comprise sets of dialogue exchanges which achieve a specific sub-goal (e.g. successful completion of an instruction); **moves** are the utterances which make up games (e.g. an instruction game, may consist of the following set of moves: instruct, query, reply, explain, align). Details of the dialogue annotation are given in [18,19].
We focus on *moves* in this paper. Moves fall into the three subcategories of initiation, response and preparation. Initiation moves are commands (INSTRUCT), statements (EXPLAIN) or questions (ALIGN (e.g. *have you got that?*), CHECK, QUERY). Response moves may show that a command has been successful (ACKNOWLEDGE) or provide an answer (REPLY) or an amplified answer (CLARIFY). . "Uncodable" moves are shown in Figure 1, but will be omitted from further analysis in this paper, as will the single type of preparation move (READY).

2.2. Disfluency annotation

The whole MTC is annotated for disfluencies, following the Map Task Disfluency Coding Manual [20]. Disfluencies are labelled for type, (delete, repetition, insertion, substitution, combinations of these with a single interruption point), for length of reparandum (in words), for complexity (embeddings), and for speaker-overlap. Individual words within disfluent reparanda and repairs (c.f. [2]) are labelled for their role within the disfluency and silent and filled pauses and editing expressions associated with the main disfluency types are marked. Annotation was performed against the speech wave-forms and word-level transcriptions, using Xwaves and Xlabel from Entropic and the resulting files were converted into XML.

Figure 1. Move type (and number in whole corpus) by speaker role



2.3. Data extraction

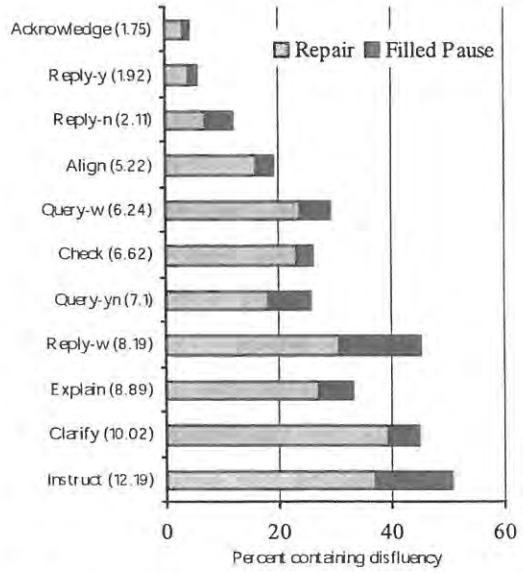
Data for Moves and Disfluency were extracted from the XML version of the MTC via unix-based xml-query tools [21] as well as standard UNIX tools.

3. Results

Figure 1 shows the count and distribution by speaker role for move types in the whole MTC. We exclude uncodable and "ready" move types from further analysis and focus on the results from the remaining 24,697 moves, containing 2,249 deletions, 1675 repetitions, 555 insertions, 712 substitutions, 293 combination types (all of which we will refer to as "repair-type disfluencies" henceforth) and 1491 filled pauses.
The first observation is that all disfluency types occur in all move types. But, as Figure 2 suggests, move types vary considerably (1) as to whether they are likely to contain any disfluency at all and (2) as to their ratio of repairs to filled pauses. 43.4% of the longest moves (**Instruct**) are disfluent, while only 4.2% of the shortest (**Acknowledge**) are. A larger proportion of disfluent **reply-n**, **reply-w** and **instruct** moves have filled pauses than other moves. Of course, these preliminary observations need to be enhanced by taking into account disfluency rate per words and utterance length, and we do so in the analyses that follow.
We take as our dependent variables, disfluency rate per 100 words and report rates for each type of disfluency introduced above, as well as total rate for repair-types, filled-pause rate and overall disfluency rate. In this study, we include in the word count words in reparanda, but not filled pauses. The results reported here are for General Linear Model multivariate analyses with the Move type (the 11 types shown in Figure 2) as independent variable. Post Hoc *Tukey* tests are used to show differences in disfluency rates by homogeneous

subgroups of move types – types within each group do not differ significantly from other group members.

Figure 2. Percent of moves containing repairs and filled pauses, ordered by mean length in words (N).



For the full set of data, rates of all disfluency types vary significantly by move type ($p < .001$). Table 1 shows homogeneous subsets of moves for overall disfluency rates. For this data, replies to wh-questions (**reply-w**) stand out as the most disfluent move type, followed by **clarify** (amplified replies), **instruction** and wh-question (**query-w**) moves. The least disfluent - **acknowledge**, positive replies (**reply-y**) and **align** moves – are amongst those with the shortest mean length in words.

Table 1. Homogeneous subsets of move types for total disfluency rates – cells show mean disfluency rates. (α=.05)

Move Type	Subset					
	1	2	3	4	5	6
Acknowldg	1.74					
Reply–Y	1.93					
Align	2.52	2.52				
Query-YN		3.77	3.77			
Reply-N			4.51	4.51		
Explain			4.66	4.66		
Check			4.81	4.81		
Query-W				5.49	5.49	
Instruct					6.29	
Clarify					6.59	
Reply-W						8.22

That some of the shortest moves are in the subset with the lowest disfluency rate and some of the longest moves are in the higher subsets suggests that there may still be an effect of move length on disfluency rates (although, in fact, there is not a simple correlation between move-length and disfluency rate in this data). So, for more detailed analysis, we take that subset of the data that includes only moves of between 4 and 6 words in length, comprising 4,973 moves with a minimum of 107 moves and a maximum of 819 moves of any one type.

For the length-controlled subset of the data, the overall picture is similar to the full set. All but the numerically smallest type of disfluency (combination) differs significantly for rate by move-type ($p<.05$). For total disfluency rates, the moves fall into 3 homogeneous subsets (Table 2). Three facts are notable in the move-type subgroups. (1) 3 of the 4 most disfluent moves represent either complex or negative replies to questions. (2) 2 of the move disfluent subgroup are predominantly Giver moves (see Figure 1), while the others are role-neutral. (3) With shorter moves removed from the analysis, move-types which for the whole data set are predominantly 1 or 2 words long (**Acknowledge** and **reply-y**) no longer stand out as having vastly smaller disfluency rates.

Table 2. Homogeneous subsets of move types for total disfluency rates for moves of 4-6 words in length – cells show mean disfluency rates. (α=.05)

Move Type	Subset		
	1	2	3
Query-YN	2.12		
Align	2.43		
Reply–Y	2.91		
Explain	3.08		
Acknowldg	3.41	3.41	
Check	3.63	3.63	
Query-W	3.66	3.66	
Clarify		5.51	5.51
Reply-N		5.59	5.59
Instruct			6.33
Reply-W			6.47

Figure 3 . Repair and filled pause rates for moves of 4-6 words, ordered by overall disfluency rate.

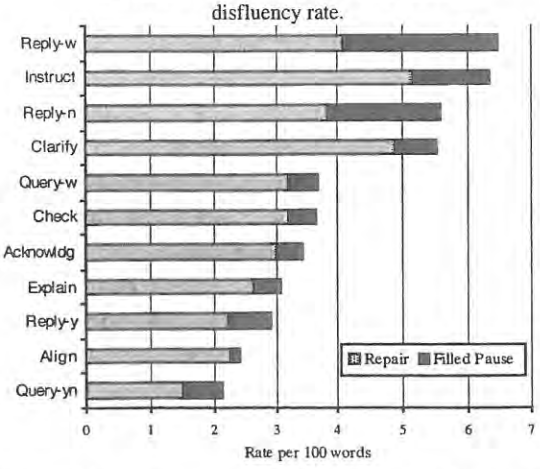


Figure 3 shows overall disfluency rates, split by repair and filled pause for each move-type. Amongst analyses for rates of

different types of disfluency for this subset of the data, all four types of *replies to questions* form part of the most disfluent homogeneous subset (of 5 types) for the rate of *Filled Pauses*. **Reply-w** (2.4) and **reply-n** (1.78) are at the top of the group, with **reply-y** (0.68) and **clarify** (0.65) following **instruct** moves (1.18). **Clarify** and **reply-w** moves also show the highest rates for *disfluent repetition*, and for the full data set, they form their own subset for this dependent variable.
Instruct moves stand out amongst initiate-type moves as having high disfluency rates of all types.

4. Discussion

Clearly, dialogue moves differ in the extent to which they are likely to contain disfluencies, and longer moves tend to be more disfluent than shorter moves. But even when we level out move-length, significant differences in disfluency rates of all types remain between move-types.
Instructions in a route-giving domain entail planning, creativity and introducing new referents and typically involve hesitation and self-correction more than many other types of move.
Answering questions requires time, and speakers appear to use filled pauses and repetitions in order to gain time. The more difficult it is to find and formulate an answer, the more disfluency speakers produce: positive replies (**reply-y**) in the MTC typically confirm the presence of a landmark on the map, which involves less searching than affirming absence (**reply-n**); more complex replies (**reply-w**, **clarify**) may require time to find the answer as well as providing more opportunity for error.
We can begin to explain why instruction givers in the MTC and in [8] are more disfluent than followers. The act of constructing an instruction produces a high rate of disfluencies, the **instruct** move accounts for about 30% of all words in the corpus and is almost exclusively a Giver move (99.2%). **Clarify** moves are also highly disfluent and mostly produced by Givers (94.9%).
We can also see why some corpora have lower disfluency rates than others. In simpler dialogues, like typical human-computer interactions, less scope is allowed for producing complex replies and longer utterances in general are suppressed (they also contain little, if any, speaker overlap, which may contribute to repetition and deletion rates).

5. Acknowledgments

This work was supported by the ESRC main grant to HCRC and by EPSRC Project Grant GR/L50280/01. Thanks to David McKelvie and Amy Isard for help at various stages of data handling and to Matthew Bull and Cathy Sotillo for assistance with disfluency annotation.

6. References

[1] Smith, V. and Clark, H. H., "On the course of answering questions", *Journal of Memory and Language*, 32:25-38.
[2] Levelt, W.J.M., "Monitoring and self-repair in speech", *Cognition*, 14:14-104, 1983.
[3] Postma, A. and Kolk, H., "The covert repair hypothesis: prearticulatory repair processes in normal and stuttered disfluencies", *J. Speech Hearing Res.*, 36:472-487, 1993.
[4] Clark, H. H., "Using Language", Cambridge University Press, Cambridge, 1996.
[5] Pickering, M. and Garrod, S., "Towards a mechanistic psychology of dialogue", Manuscript in preparation, 2001.
[6] Schegloff, E. A., Jefferson, G. and Sacks, H., "The preference for self-correction in the organisation of repair in conversation", *Language*, 53:361-382, 1977.
[7] Branigan, H., Lickley, R., McKelvie, D. "Non-linguistic influences on rates of disfluency in spontaneous speech" Proc. 14th ICPhS, 1999.
[8] Bortfeld, H., Leon, S.D., Bloom, J. E., Schober, M. F. and Brennan, S. E., "Disfluency rates in conversation: Effects of age, relationship, topic, role and gender", *Language and Speech*, 44, 2001.
[9] Bard, E. G., Lickley, R. J. and Aylett, M. P., "Is disfluency just difficulty?", *Proceedings of Disfluency in Spontaneous Speech '01, ISCA Tutorial and Research Workshop*, Edinburgh, Scotland, 2001.
[10] Lickley, R. J., "Detecting Disfluency in Spontaneous Speech", PhD Thesis, University of Edinburgh, UK, 1994.
[11] Shriberg, E. E., "Preliminaries to a theory of speech disfluencies", PhD Thesis, University of California at Berkeley, 1994.
[12] Brennan, S. E. and Williams, M., "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive state of speakers", *Journal of Memory and Language*, 34:393-398, 1995.
[13] Clark, H. H., and Wasow, T., "Repeating words in spontaneous speech", *Cognitive Psychology*, 37: 201-242, 1998.
[14] Greene, J. O. and Capella, J. N., "Cognition and talk: the relationship of semantic units to temporal patterns of fluency in spontaneous speech", *Language and Speech*, 29(2):141-157, 1986.
[15] Oviatt, S., "Predicting disfluencies during human-computer interaction", *Computer Speech and Language*, 9:19-35, 1995.
[16] Anderson, A., Bader, M., Bard, E.G., Boyle, E., Doherty, G., et al., "The HCRC Map Task Corpus", *Language and Speech*, 34:351-366, 1991.
[17] Isard, A., "An XML architecture for the HCRC Map Task Corpus", *Proceedings of Bi-Dialog, 2001*, Bielefeld, Germany, 2001.
[18] Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A., "HCRC Dialogue structure coding manual", *HCRC/TR-82*, HCRC, University of Edinburgh 1996.
[19] Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A., "The reliability of a dialogue structure coding scheme", *Computational Linguistics*, 23:13-31, 1997.
[20] Lickley, R. J., "HCRC Disfluency Coding Manual" *HCRC/TR-100*, HCRC, University of Edinburgh, 1998.
[21] LTG, "LTXML". http://www.ltg.ed.ac.uk/software/xml.

Is disfluency just difficulty?

Ellen G. Bard, Robin J. Lickley, Matthew P. Aylett

Department of Theoretical and Applied Linguistics and Human Communication Research Centre
University of Edinburgh, Scotland
ellen@ling.ed.ac.uk

Abstract

The question addressed by this paper is whether disfluency resembles Inter-Move Interval, a measure of reaction time in conversation, in displaying effects of the overall difficulty of conducting a coherent conversation. Five sources of difficulty are considered as potential causes of disfluency: planning and producing an utterance, comprehending the prior utterance, performing a communicative task, order effects, and interpersonal factors. A multiple regression analysis on simple disfluencies in the HCRC Map Task Corpus shows that planning and production make the major independent contribution to predicting the rate of disfluencies, with interpersonal variables and position in dialogue also contributing significantly. Notably, comprehension variables did not affect either the total rate of disfluency or the rate of individual kinds of disfluencies.

1. Introduction

Many disfluencies are edited errors in speech production. They mark those occasions when speakers have not framed an utterance which satisfies their goals before they begin to speak. Disfluencies are thought to occur when speakers fail to monitor and edit successfully during earlier phases of production [1, 2]. We do not yet know exactly what prevents correct initial formulation or internal self-correction in natural circumstances, but there are many possible culprits among the tasks competing for the speaker's attention. To produce any spontaneous utterance, a speaker must plan, assemble, and articulate a string of words. In dialogue, interlocutors must also comprehend one another's contributions and provide appropriate replies promptly enough to make it plain that they wish to take the floor. In task-oriented dialogue, they must use the interaction to achieve a non-conversational goal. As with any other task, initial attempts at any of these activities in a given setting may prove difficult.

If disfluency is induced by such difficulties, then it should behave like Inter-Move Interval (IMI). Defined as the time, positive or negative, between the offset of one speaker's utterance and the onset of the interlocutor's reply, IMI is a measure of reaction time in dialogue [3]. IMI is longer early in a dialogue, when the interlocutor's utterance is difficult to comprehend, when the interlocutors are having difficulty with the task, when a long utterance follows, and when that utterance begins a larger unit of dialogue. At the same time, IMI is shorter in what might be the more delicate social situation: conversations between persons of different sexes who have just met. Thus, time to begin speaking is sensitive to interpersonal factors as well as to cognitive pressures of various kinds.

There is already evidence that planning and production burdens affect fluency. Disfluencies tend to occur early in utterances, when planning of later stages is incomplete. Disfluencies are more common in longer utterances [4-6], in more complex constituents [4], and when response choices are complex [6].

There is good reason to predict what affects delay to begin speaking will also affect fluency of speech. First, we predict effects *of the prior utterance*. Pressure to hold or take the floor may induce imperfections in planning [5,7]. Since modal IMI, at around 150 msec, with 14% negative, is too low to allow planning to follow listening entirely [3], speakers must begin planning their next utterance while listening to their interlocutor's prior utterance. If normal comprehension processes are also used for self-monitoring [8], then any competition should obstruct production. The effect might be more disfluency at shorter IMIs or more disfluency with longer or more complex prior utterances. Also, the fluency of the prior utterance may be important: cross-speaker syntactic priming [9] or mere difficulty in perceiving disfluent utterances [10] could yield disfluent adjacency pairs. Second, we might predict effects of ***task difficulty*** in general, because competition for attention is likely to affect any process serving an activity as complex as conversation. Third, if task-oriented dialogue comprises as a series of similar problems which have to be solved by communicating, then there may be ***order effects*** within a single such conversation or across a series. Certainly, dialogues and expressions used in them get shorter with time [11]. By building expertise and mutual knowledge, interlocutors effectively narrow the choices they have to make, and these basic conditions should enhance fluency. Finally, it seems likely that ***interpersonal factors***, like the number of sensory channels or the familiarity of the interlocutors should affect delicate processes of planning and feedback [12]. We know that these variables affect the structure of dialogues where sensory channels or communicative links are limited (e.g., [13]).

The difficulty with testing so many predictions, of course, is that the predictors may intercorrelate. The longer utterances earlier in a conversation, for example, may induce disfluencies because they are long, because they are early, or both. To determine what independent contributions are made by each kind of predictor, we use a method similar to the one which found predictors of IMI [3]: we run a multiple regression analysis on all appropriate items from the same coded corpus of task-oriented dialogues, deriving our reported results from the most fully coded subset of the materials.

2. Method

2.1. Corpus

Materials came from the HCRC Map Task Corpus [14] (hereafter MTC), 128 unscripted dialogues in which 32 pairs of Glasgow University undergraduates communicated routes

defined by labeled cartoon landmarks on schematic maps of imaginary locations. Instruction Giver's (hereafter 'IG') and Follower's (IF) maps for any dialogue matched only in alternate landmarks. Participants knew that their maps might differ but not where or how. Players could not see each other's maps. Familiarity of participants (within subjects) and ability to see the interlocutor's face (between subjects) were counterbalanced. Each participant served as IG for the same route to two different IFs and as IF for two different routes. Channel per speaker digital stereo recordings were orthographically transcribed and digitally word-segmented. Like the coding systems described below, the segmentations form part of an XML corpus database.

2.2. Unit of analysis

The word-segmented corpus is framed as a series of Conversational Game Moves [7], turns or parts of turns whose purpose in moving the dialogue forward can be determined by their form and context. Moves are stages of Conversational Games, which are themselves usually stages in completing Transactions, sections of the task which the dialogue serves [15]. Here we used only those Moves which involve a change of speaker and which are likely to be a reply to the previous speaker's Move: we excluded those which began too early to respond to the prior Move (onset preceding offset of the prior speaker's Move by > 1 sec, or onset < 350 msec after prior Move onset) and those which were actually resumptions of an earlier Move by the same speaker (onset < 300 msec after the end of a previous Move by the speaker).

2.3. Disfluency coding

The dependent variables were ***numbers of disfluencies*** of various kinds ***per Move***. Disfluency annotation [16] was performed on the whole corpus using Xwaves/Entropic xlabel. Annotators examined the speech waveform closely and made use of spectrograms where necessary. For each disfluency, individual words were labeled by part and type of disfluency. Disfluency parts [17] are original utterance, reparandum, interruption/filler, repair and continuation.

The current paper omits silent and filled pauses, combination and complex disfluencies, and reports on only simple disfluencies of 4 types. In ***repetitions*** the speaker repeats a string verbatim, with no additions or deletions: e.g. *[we're going] we're going left of the camera shop.* In ***insertions*** the speaker repeats a string and inserts a word or words within the repeated string: e.g. *[go left] go just left of the camera sho.* In ***substitutions*** a word or string is replaced by another with no major syntactic alteration: e.g. *go [left] right of the camera shop.* In ***deletions***, the speaker interrupts an utterance and either restarts without repeating or directly substituting or simply surrenders the floor to the other speaker: e.g. *[you're away f-] right see the wee bit that's jutting out?*

2.4. Predictor variables

2.4.1. Current Move

Three sets of predictors reflect hypotheses about how production tasks encourage disfluency. First, the planning functions are represented by the ***speaker's role*** (because Instruction Givers bear more of the burden of structuring the dialogue), and by the conversational ***boundary*** preceding this current Move. If planning a section of the task or dialogue affects fluency as it affects IMI, then disfluency rates should follow the size of the planned unit, with Transaction-initial Moves (see 2.2) most disfluent (2), Game-initial Moves (1) somewhat less disfluent, and Game-internal Moves (0) least disfluent . Second, the burden of constructing ***referring expressions*** [4] is measured via separate counts for the combinations of New/Given and Shared (on both players' maps) / Unshared (on only one). Finally, as a more general indicator of complexity, ***length in words*** is measured, but omitting any words in the reparanda of disfluent Moves.

2.4.2. Prior Move

Three sets of predictors represent aspects of the prior speaker's utterance which may make the current speaker disfluent. First, difficulty in comprehending a complex set of references to map locations may interfere with the process of production. Hence, numbers of ***referring expressions*** in the prior Move are classed as for current Moves (2.4.1). To test for priming by disfluent structures, prior Move ***disfluency*** is tallied for each kind of disfluency (see 2.3). Finally, ***length in words*** is included.

2.4.3. Difficulty metrics

To reflect difficulty in pursuing the task itself, we use 3 measures. ***Deviation score*** is the mismatch in cm^2 between the model route on IG's map and the route ultimately drawn on the IF's. Major miscommunications yield large deviation scores. ***Drawing*** shows whether the prior Move was followed by an attempt to draw part of the route. Finally, ***Inter-Move Interval*** itself is an indicator of various kinds of cognitive load [3].

2.4.4. Order

To capture effects of practice and of increasing discourse context, 2 order codes were used. ***Conversation*** records which of the 8 dialogues produced by a quad (pair of speaker pairs) is in progress. Conversations 4-8 are second trials with a map on the part of the IG. ***Position*** is the ordinal position of the Move in the dialogue ($\mu = 136.69$, *s.d.* = 110.97).

2.4.5. Interpersonal

These are aspects of the corpus design which affect the social distance between IG and IF. ***Eye-contact*** refers to the presence (0) or absence (1) of a flimsy barrier blocking the line of sight between interlocutors. ***Familiarity*** records whether the pair have just met (0) or are friends (1).

3. Results

Detailed results are reported for the 6882 'response' Moves (see 2.2) of the dialogues coded for the presence of drawing between Moves. The results were essentially the same for the whole corpus, that is all 14389 'response' Moves not containing complex disfluencies (see Figures 1-7).

3.1. Significant contributions

Multiple regression equations using all predictors were prepared with total number of disfluencies per move as dependent variable, and then with number of each individual

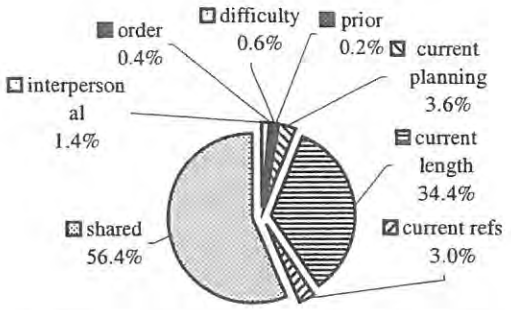


Figure 1.Contributions of groups of predictor variables to the explained variance in total rate of disfluencies

subcategory. Table 1 displays only the significant β-values (standardized regression coefficients) for total disfluency count, since the patterns of results are the same throughout. All the regression equations accounted for significant proportions of the variance in disfluency rates ($p < .0001$), with explained variance in the overall measure (Multiple R^2) nearly 14% (deletions 2.5%, insertions 5.3%, repetitions 8.1%, substitutions 4.3%).

The principal question addressed was whether disfluency behaves like IMI in sensitivity to cognitive load from many sources. As Table 1 shows, it does not. Significant individual predictors are restricted to characteristics of the current utterance and to interpersonal and order effects. Difficulty and prior Move variables do not make significant individual contributions to accounting for the rates of disfluencies. Figure 1 displays the proportion of accounted-for-variance attributable to each group of variables. As expected, a large proportion of the explained variance (> 56%) is shared among the intercorrelating predictors. Of the five groups, the current Move predictors clearly predominate (41%), with length in words alone accounting for over 30% of the variance in total disfluency rate, more than all other groups combined.

3.2. Individual effects

The remaining figures display effects of individual predictor variables on overall disfluency rate. Raw means are used for simplicity of interpretation, but significant β values indicate that trends would be robust even adjusted for effects of other predictors.

Figures 2 and 3 display predicted effects of planning on disfluency. Figure 2 shows that IGs, who usually take responsibility for directing the dialogue, are more disfluent (.218) than IFs (.093) ($\beta = .07$, $p < .01$). Figure 3 shows that there are more disfluencies in Moves which initiate larger constituents of a dialogue, with rates of disfluency rising from Game-internal Moves (.116) to Game-initial (.219) and again to Transaction-initial (.294) ($\beta = .02$, $p < .05$).

Figure 4 and 5 show the predicted effects of current-Move length and referential complexity. As in other corpora [4, 6], longer Moves attract more disfluencies (rising continuously from .063 for single-word Moves to .709 for > 17 words, $\beta = .28$, $p < .01$) (Figure 5). Moves containing more referring expressions (.104 for 0, .248 for 1, .492 for 2 to 5) also attract more disfluencies ($\beta = .28$, $p < .01$).

Figure 6 shows an order effect but clearly not a simple practice effect: Moves later in the dialogue exhibit higher rates of disfluency (.147, .135, .169, .167 for successive quartiles, $\beta = .03$, $p < .05$).

Finally, Figure 7 associates disfluency with interpersonal difficulty: speech to an unfamiliar partner is more disfluent (.174 v .140, $\beta = -.03$, $p < .05$).

Table 1. Significant β-values in multiple regression equations predicting occurrence of disfluencies in 6882 Conversational Game Moves (coded for presence of drawing). df = *23, 6858, p < .0001. (Key: * : p < .05, *: p < .01)*

Type		Variable	All
Interpersonal		Familiarity	- .04*
		Eyecontact	
Order		Conversation	
		Position	.03*
Difficulty		I.M.I. deviation score	
		Drawing	
Prior move	Reference	new shared	
		given shared	
		new unshared	
		given unshared	
	Disfluency	deletions	
		repetitions	
		substitutions	
		insertions	
	Length	length (words)	
Current move	planning	Role	.07*
		boundary	.02*
	reference	new shared	
		given shared	.06*
		new unshared	.02*
		given unshared	.03*
	length	length (words)	.28*
		Multiple R^2	.138
		F	47.64
		n	1065

Figure 2 . Effects of role on disfluency per move

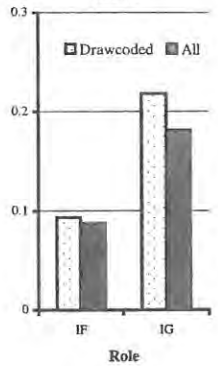


Figure 3. Effects of dialogue unit on disfluency per move

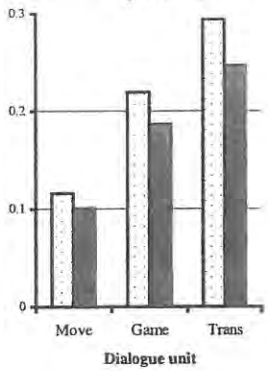


Figure 4. Effect of Move-length on disfluency per move

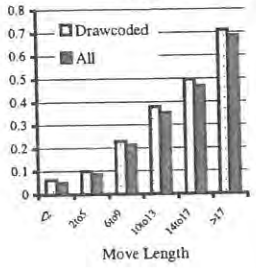


Figure 5. Effect of referring expressions on disfluency per move

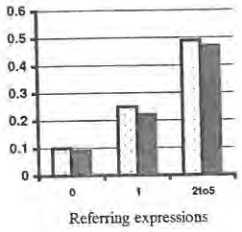


Figure 6. Effect of position in dialogue

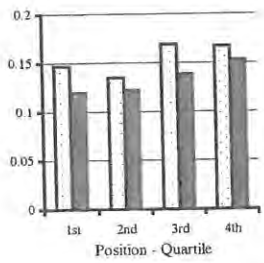
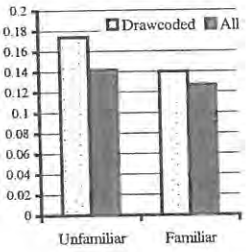


Figure 7. Effect of familiarity



4. Conclusions

The results of the multiple regression analyses of the correlates of disfluency show quite a different pattern from the one observed for IMI. Rather than behaving like a general measure of difficulty affected by interpersonal, practice, comprehension and production processes, disfluency seems to be linked to processes of production, with the greater part of the uniquely explained variance attributable to characteristics of the disfluent Move: length, referential complexity and likely role in larger scale planning of the dialogue.

The smaller contributions of unfamiliarity and position in dialogue could also be construed as difficulty effects: the theory of common ground suggests that framing a satisfactory utterance may be more difficult if the addressee is a stranger rather than a friend. The effect of position in dialogue, here measured by Move number, may be unduly influenced by dialogues which are unusually long because communication is proving difficult..

Yet it is plain that disfluency has a particular area of insensitivity: Even though human language production and comprehension are thought to share components, disfluent output is not associated with any of the current measures difficult or disfluent input. This fact suggests a separation rather than a sharing of processes.

5. Acknowledgments

This work was supported by the ESRC main grant to HCRC and by EPSRC Project Grant GR/L50280/01. Dr Aylett is now with Rhetorical Systems, Edinburgh.

6. References

[1] Blackmer, E., and Mitton, J., "Theories of monitoring and the timing of repairs in spontaneous speech", *Cognition*, 39:173-194, 1991

[2] Postma, A. and Kolk, H. "The effects of noise masking and required accuracy on speech errors, disfluencies, and self-repairs", *J. Speech Hearing Res.*, 35:537-544, 1992.

[3] Bard, E. G., Aylett, M., and Bull, M. "More than a Stately Dance: Dialogue as a Reaction Time Experiment", *Proc. Soc Text and Disc.*, 2000.

[4] Clark, H. H., and Wasow, T., "Repeating words in spontaneous speech", *Cognitive Psychology*, 37: 201-242, 1998.

[5] Maclay, H., and Osgood, "Hesitation phenomena in spontaneous English speech", *Word*, 15:19-44, 1959.

[6] Oviatt, S., "Predicting disfluencies during human-computer interaction", *Comput. Speech Lang.*, 9:19-35, 1995.

[7] Carletta, J. , Isard, A. , Isard, S. , Kowtko, J. , Doherty-Sneddon, G. , Anderson, A., "The reliability of a dialogue structure coding scheme", *Computat. Ling.*, 23:13-31, 1997.

[8] Levelt, W.J.M., Roelofs, A., and Meyer, A. S., "A theory of lexical access in speech production", *Behav. Brain Sci.*, 22:1-45.

[9] Branigan, H., Pickering, M., and Cleland, A., "Syntactic coordination in dialogue", *Cognition,*, 75:813-825, 2000.

[10] Bard, E. G., and Lickley, R. J. "Graceful failure in the recognition of running speech", *Proc 20th Ann. Meeting of the Cog. Sci. Soc.*, 108-113.

[11] Brennan, S., and Clark, H., "Conceptual pacts and lexical choice in conversation", *JEP:LMC*, 22:1482-1493, 1996.

[12] Branigan, H., Lickley, R., McKelvie, D. Non-linguistic influences on rates of disfluency in spontaneous speech. Proc. 14th ICPhS, 1999.

[13] Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S., and Bruce, V., "Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance" *J Exp. Psych.: Applied*, 3:105–125, 1997.

[14] Anderson, A., Bader, M., Bard, E.G., Boyle, E., Doherty, G., et al., "The H.C.R.C. Map Task Corpus", *Lang. and Speech*, 34:351-366, 1991.

[15] .Isard, A., and Carletta, J., *Transaction and Action Coding in the Map Task Corpus.* Research paper HCRC/RP-65. HCRC, U. of Edinburgh., 1995.

[16] Lickley, R.J., HCRC *Disfluency Coding Manual.* Technical Report HCRC/TR-100, HCRC, U. of Edinburgh., 1998.

[17] Levelt, W.J.M., "Monitoring and self-repair in speech", Cognition, 14:14-104, 1983.