DiSS Conference

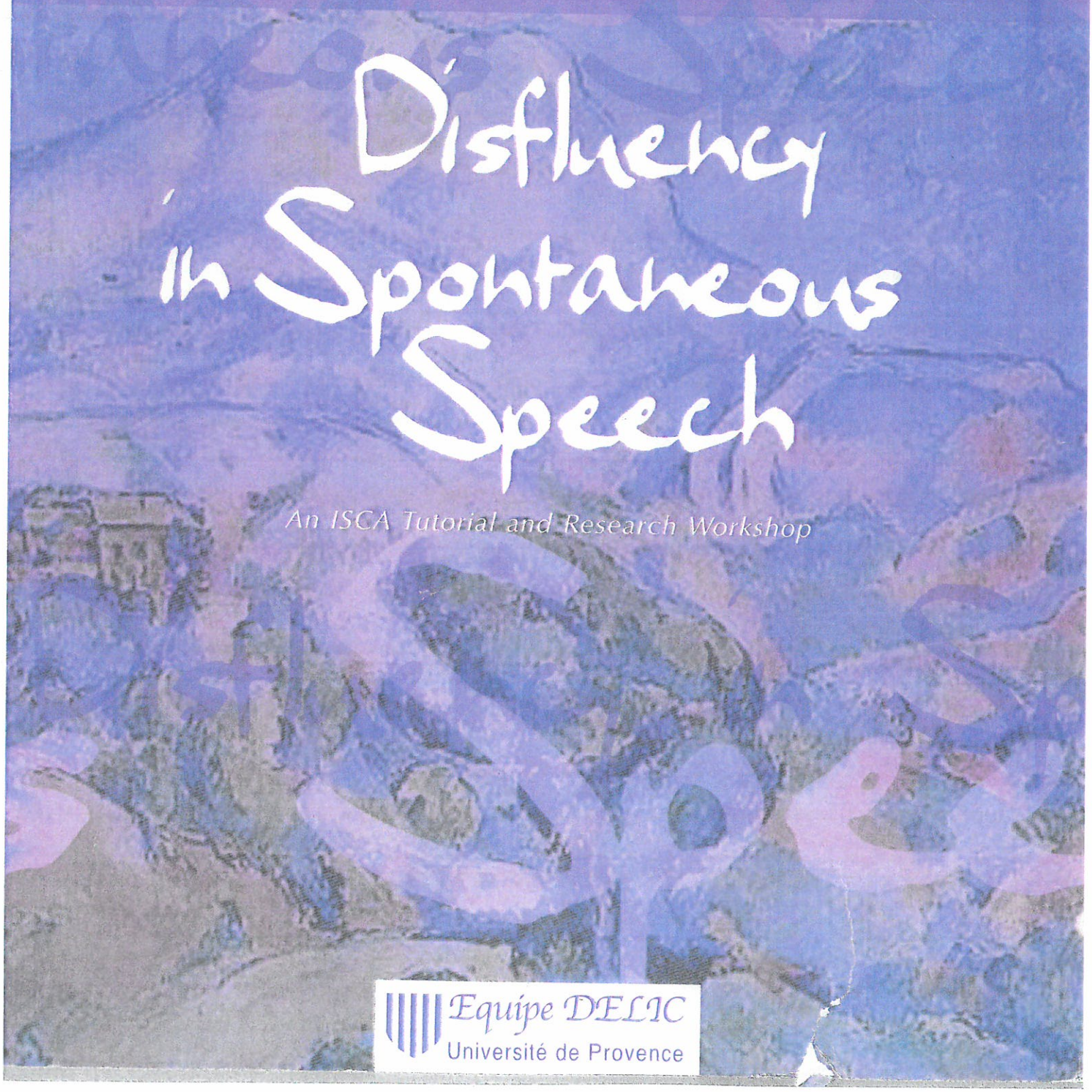
Aix-en-Provence

10, 11 & 12 September

2005

Disfluency in Spontaneous Speech

An ISCA Tutorial and Research Workshop



Equipe DELIC
Université de Provence

Disfluency in Spontaneous Speech

Proceedings of

DiSS'05

An ISCA Tutorial and Research Workshop

Aix-en-Provence, France
September, 10-12, 2005

Equipe DELIC

Université de Provence

COMMITTEES

Program committee

Martine Adda-decker, LIMSI, France

Jens Allwood, Göteborgs Universitet, Sweden

Estelle Campione, Université de Provence, France

Maria Candea, Université Paris III, France

Eugene Charniak, Brown University, USA

Martin Corley, University of Edinburgh, UK

Yasuharu Den, Chiba, Japan

Robert Eklund, TeliaSonera Sweden

Fernanda Ferreira, University of Massachusetts at Amherst, USA

Rob Hartsuiker, Ghent University, Belgium

Peter Heeman, CSLU, Oregon, USA

Robin Lickley, Queen Margaret University College, Edinburgh, UK

Sieb Nooteboom, Utrecht University, Netherlands

Marc Swerts, Tilburg University, Netherlands

Elizabeth Shriberg, SRI International and International Computer Science Institute, USA

Shu-Chuan Tseng, Institute of Linguistics, Taipei, Taiwan

Jean Véronis, Université de Provence, France

Åsa Wengelin, Göteborgs/Lunds Universitet, Sweden

Organising committee

Estelle Campione, DELIC, Université de Provence, France.

Sandrine Henry, DELIC, Université de Provence, France.

Sandra Teston, DELIC, Université de Provence, France.

Jean Véronis, DELIC, Université de Provence, France.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 7-11.

Characteristics of final part-word repetitions

Jan McAllister[*] *& Mary Kingston*[**]

[*] University of East Anglia, United Kingdom
[**] Norwich Primary Care Trust, United Kingdom

Abstract

In an earlier paper, we have described final part-word repetitions in the conversational speech of two school-age boys of normal intelligence with no known neurological lesions. In this paper we explore in more detail the phonetic and linguistic characteristics of the speech of the boys. The repeated word fragments were more likely to be preceded by a pause than followed by one. The word immediately following the fragment tended to have a higher word frequency score than other surrounding words. Utterances containing the disfluencies typically contained a greater number of syllables than those that did not; however, there was no reliable difference between fluent and disfluent utterances in terms of their grammatical complexity.

1. Introduction

Final part-word repetitions (FPWRs) involve the repetition of one or more sounds from the end of a word. When FPWRs have been reported in adults, the disfluency seems to have been associated with neurological impairment [4,5,6,7,8,9]. Stansfield [5] described four adults with learning disabilities who produced FPWRs, prolongations and blocks, although the majority of their repetitions were word-initial. Lebrun and Van Borsel [4] reported various disfluencies in the speech of a 17-year old girl with Down's syndrome, including word-final consonant repetition. Word-final repetitions have been observed (among other disfluencies) in the speech of adults with right-hemisphere brain injuries [8,9].

In the majority of studies that have described FPWRs in children, the repetitions have involved single word-final consonants [1,2,3,4]. In normally-developing children aged younger than three years, the disfluent behaviour was transient, lasting only a matter of months, with spontaneous recovery [1,2,3]. Lebrun and Van Borsel [4] reported the case of an eight-year-old with average IQ who repeated word-final consonants while reading (in addition to producing more typical stuttering-like disfluencies). The boy was aware of his disfluencies, which were often accompanied by grimaces.

Van Borsel, Van Coster and Van Lierde [10] reported the case of a nine-year-old Dutch-speaking boy, T, whose FPWRs were not confined merely to single consonants. As well as repeating individual consonants (e.g. *nooit t t*), the boy repeated word-final consonant clusters (*lucht cht*) and vowels (*toe oe*), whole syllables (*diamanten ten*), and word-final nucleus-coda sequences (*stap ap*). The repetitions occurred only on content words and mainly when the boy was speaking spontaneously (monologue or dialogue), although one instance was observed when he was reading aloud. The repetitions did not occur during singing or when the boy repeated words spoken by the researcher. In addition, the boy produced a relatively high number of 'broken words' in which 'phonation or airflow is stopped within a word' [11]. The boy in this study had sustained left fronto-parietal brain damage due to a fall from a window when he was aged 3 years 10 months, and small sub-cortical lesions were evident on MRI scans taken when he was nine years old.

McAllister & Kingston [12] reported FPWRs in the speech of two seven-year-old boys, E and R. These subjects differed from the majority of older children described in earlier reports of this behaviour in that their disfluency was not associated with neurological problems. Their general language abilities were normal for their ages, as indicated by their performance on a standard clinical test. Both boys produced similar disfluencies in reading, sentence repetition and spontaneous speech.

A range of normal non-fluencies (e.g. filled and unfilled pauses, single and multiple word repetitions, and revisions) occurred with typical frequency in the speech of the boys, but in addition, they produced broken words and FPWRs. Examples of the latter were

a. *I'm just* [ʌst] *wondering* [ɪŋ] *why* [ai] *is all that going on there* (produced by R)
b. *I don't think* [ŋk] *we got the whole way through those.* (E)
c. *They thought the Nutcracker was about* [ʊt] *to attack them* (E)
d. *There are only* [i] *three* [i] *little pod things.* (E)
e. *And then she can* [n] [n] *lift things without touching them* (E)

Neither child appeared aware of the presence of the repetitions or of any other disfluencies in their speech; they produced no visible sign of increased muscle tension and no apparent avoidance strategies.

The repeated fragments occurred after monosyllables and polysyllabic words, and after function words and content words. They occurred following words at all sentence positions (initial, medial and final), though predominantly sentence-medially. Both boys had been observed on occasion to produce two or three iterations of the repetition (e.g., example e above), though only E did so during the recordings.

Describing the phonological form of the repeated fragments, McAllister & Kingston noted that each child was following an individual but highly predictive set of rules which determined the form that the fragment would take. Neither child repeated complete word-final syllables. R almost always repeated the rime (i.e., syllabic nucleus and coda)[1] of the word preceding the fragment, for example, 'scientist [ist]', 'home [əʊm]', 'party [i]'. For E, the form of 91% of the repetitions could be predicted by a more complex set of rules: When the last syllabic nucleus of the word consisted of a diphthong, he repeated the second vowel of the diphthong plus the coda, if any (e.g., 'out [ʊt]', 'say [i]', 'Yugioh [ʊ]'). Otherwise, he repeated the last syllabic element of the word, that is, the nucleus if the word ended in a vowel (e.g. 'army [i]', 'more [ɔ]'), or the coda alone when it ended in one or more

consonants ('off [f]', 'think /ŋk/'. Both boys disregarded the number of syllables in the word and its morphological structure when applying their rules. In keeping with the overall consistency of these forms, when the same word occurred on several occasions with word-final disfluency, the repeated fragment took the same form; for example, there were five occasions when R produced the word *card* with a FPWR, and each time, he said [kɑːd kːɑ]; and E produced 'out' as [ɑʊt ʊt] twice.

In this paper, we describe some further characteristics of FPWRs produced by R and E. One analysis is concerned with the pause pattern in the context of the repeated word fragment. McAllister & Kingston [12] noted that an audible pause always occurred between the target word and the repeated word fragment which followed it; for example, E produced the utterance *But I don't think* [pause] *–nk we got the whole way through those*. By contrast, pauses occurred relatively infrequently between the repeated fragment and the following word. Figure 1 shows a typical example, from the utterance *she can also* /ʊ/ *send these green* /n/ *slicing disks*. This pause pattern is investigated in more detail below.

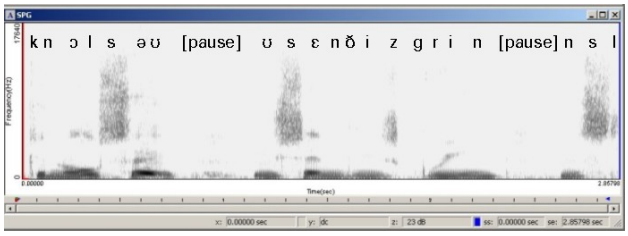


Figure 1: Spectrogram showing typical pause pattern in an utterance containing FPWRs

There is disagreement about whether FPWR should be classified as stuttering. As far as more generally-accepted forms of stuttering are concerned (e.g. word repetitions, word-initial sound repetitions, prolongations, blocks), much is known about the loci or linguistic environments in which these tend to occur. Bernstein Ratner [13] pointed out that although 'stuttered moments' sound qualitatively distinct from normal disfluencies, they share many of their distributional and situational characteristics. She also noted that although some researchers had focused their attention on the characteristics of the current word (e.g. the word whose initial sounds are repeated, or which immediately follows a block), the psycholinguistically relevant event might actually occur later in the sentence. In her list of factors that precipitate stuttering, Bernstein Ratner includes word frequency (words of lower frequency are associated with stuttering), lexical class (content word versus function word; adults stutter more on the former and young children on the latter), and syntactic factors (e.g. more stuttering on complex than on simple structures) [14].

Little is known about the loci of FPWRs. Van Borsel et al [10] noted that T, the boy whom they studied, produced the disfluencies only on content words, while McAllister & Kingston reported that R and E repeated the ends of both content words and function words. For T [10], E and R, FPWRs appeared irrespective of the number of syllables or morphological composition of the words they followed. However, other information about the loci of FPWRs is lacking. In this paper, we will explore the syntactic and lexical environments of the repetitions produced by R and E.

2. Method

2.1. Participants

R, a right handed, monolingual speaker of English with no known sensory or neurological impairment, was aged between 7 years 9 months and 8 years 1 month during the period when the data described below were collected. He was born after a normal full term pregnancy, the fourth of five children. The pattern of disfluency described by McAllister & Kingston [12] was first observed by a speech and language therapist when he was aged 7 years 2 months, as part of a routine checkup following earlier referrals for mild phonological delay and mild disfluency in the form of some monosyllabic word and initial part-word repetitions. Both of these problems had resolved completely by the time the FPWRs were observed.

E, a right-handed monolingual speaker of English with no known sensory or neurological impairment, was aged 7 years 6 months and 7 years 10 months at the times of data collection. He was born after a normal, full-term pregnancy, the youngest of three children. His speech and language development had been completely unexceptional apart from the emergence, when he was approximately five years old, of the disfluency pattern described by McAllister & Kingston [12].

Neither child had ever suffered any illness other than common childhood ailments, and routine hearing and vision screening yielded normal results. There had never been any educational concerns, and both performed at or above the expected level for their ages in national tests in English, mathematics and science taken by English schoolchildren at age seven.

2.2. Procedure

Two samples of spontaneous speech were recorded from each child. The children were recorded in free one-to-one conversation with the second author, an experienced specialist in disorders of fluency. She guided the conversation onto topics which she knew would be of interest to them, such as Pokemon and Yu-Gi-Oh game cards, of which there were examples available in the room, and children's TV programmes. The first recordings lasted approximately 30 minutes and the second approximately 15 minutes in each case.

2.2.1. Transcription, coding and measurement

Transcription: All the recordings were orthographically transcribed by the first author and checked by the second author.

Classification of disfluencies: Once the authors had arrived at an agreed transcription, they listened independently to the recordings to identify instances of disfluency. They compared their analyses and listened carefully to the recordings to resolve any disagreements as to classification. Disfluencies were classified using a system based on that of Van Borsel et al. [10], as follows. *Interjections* were hesitation noises such as *uh* and *um*. An example of a *single-word repetition* is *At the end of that episode the the memory was Kaiba*. An example of a *multi-word repetition* was *He didn't fuse them when he when he had them out*. *Revisions* were self-interrupted utterances containing a correction such as *They thought the whole - the two armies were joined together*. *Incomplete*

phrases were self-interruptions that did not contain a correction. *Prolongations* were defined as speech segments which had greater than expected duration given their linguistic and phonetic context. *Blocks* were defined as inappropriate stoppages of airflow or voice. *Initial part-word repetitions* involved the repetition of one or more segments from the start of a word, e.g. *The grownups ha-had their eyes closed*. *Broken words* involved cessation of airflow or phonation within a word, e.g. *I can't te-ll all of it*. *FPWRs* involved the reiteration of some portion of the end of the word (see examples a-*e* above).

Pause pattern: Pause durations in FPWRsn were measured at two locations – between the preceding word and the repeated fragment (pre-fragment pause), and between the fragment and the word following it (post-fragment pause). For example, in the utterance *I'm just* [ʊst] *wondering*, a pause between *just* and [ʊst] would be a pre-fragment pause, and a pause between [ʊst] and *wondering* would be a post-fragment pause.

Word frequency: One possible reason for the occurrence of a FPWR is that the speaker is experiencing a problem with the formulation of an upcoming part of the utterance, e.g. lexical access of a low-frequency word. It might be predicted that words following the fragment would be of lower frequency that words preceding the fragment. To examine this hypothesis, we used the MRC Psycholinguistic Database to calculate the word frequency of the three words preceding the fragment and the three words following the fragment. We labelled the words as shown in the following example (for the utterance *And then she can* [n] [n] *lift things without touching them*) :

Word -3: then
Word -2: she
Word -1: can
Word +1: lift
Word +2: things
Word +3: without

Syntactic complexity and utterance length: This analysis followed the procedure described by Melnick & Conture[15], who found that utterances containing stuttered disfluencies were longer and more grammatically complex that those containing no such disfluencies. Based on Melnick & Conture, utterances were defined as "a string of words that (a) communicates an idea; (b) is set apart by pauses; (c) is bound by a single intonation contour". Twenty-five utterances containing FPWRs were randomly selected from the speech of each child, along with 25 utterances which contained neither FPWRs nor broken words. Utterance length was calculated by counting the number of syllables in the utterance, excluding the FPWRs themselves. Grammatical complexity was calculated by counting the number of clausal constituents (subject, verb, object, complement and adverbial) that the utterance contained.

3. Results

3.1. Frequency of disfluency types

Since preliminary analysis indicated a similar distribution of disfluency types in each spontaneous speech sample, in this section the two are combined for each child. Across the two samples, R spoke 4062 syllables and E spoke 4058 syllables. Frequencies of the disfluency types defined in 2.2.1. is shown in Table 1.

Table 1: Frequency of disfluency types per 100 syllables in the speech of R and E.

Disfluency type	R	E
Interjection	0.06	1.15
Single-word repetition	0.22	0.54
Multi-word repetition	0.02	0.19
Revision	1.46	1.86
Incomplete phrase	0.37	0.31
Prolongation	0.00	0.02
Block	0.00	0.00
Initial part-word repetition	0.11	0.28
Broken word	0.26	0.69
Final part-word repetition	1.15	2.47
TOTAL	3.64	7.49

3.2. Pause pattern

Twelve items were considered unsuitable for analysis and were omitted, because they involved either overlapping speech or multiple iterations of the repetition; this analysis was therefore based on 96 items for E and 49 items for R. Durations of the pauses is shown in Table 2.

A 2-way ANOVA (Child x Pause Position) was conducted on the measurements for all words. Overall, pre-fragment pauses were reliably longer than post-fragment pauses ($p < 0.0001$), and R's pauses were longer than E's pauses ($p < 0.0001$); the variables did not interact. As was noted above, although pre-fragment pauses occurred in every case, post-fragment pauses were relatively unusual. When the analysis was confined to just those less typical items where there was a post-fragment pause, the pre-fragment pause was still dependably longer than the post-fragment pause ($p = 0.0180$); the means for the two children differed reliably, and the interaction was non-significant.

Table 2: Durations of pre-fragment and post-fragment pauses.

	R		E	
	Pre	Post	Pre	Post
All words	599	152	436	158
Words with no post-fragment pause	558	-	706	-
Words with a post-fragment pause	699	523	601	476

3.3. Word frequency

The mean frequencies of the words preceding and following the repeated fragments is shown in Figure 2. A 2-way ANOVA (Child x Word Position) was conducted. The difference between the children was insignificant, as was the interaction. There were significant differences among the frequency values for the different word positions relative the the repeated fragment; specifically, the word preceding the

fragment was of reliably lower frequency than the other words, and the word following the fragment was of higher frequency, but the other values did not differ significantly from each other.

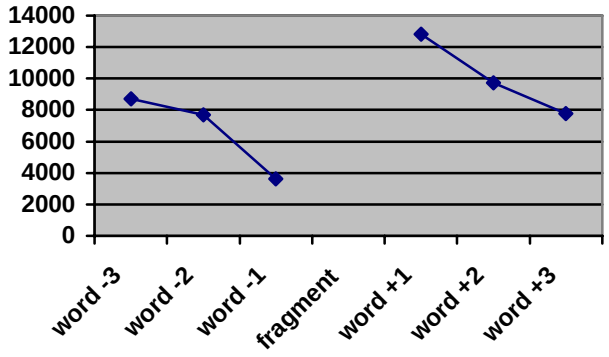


Figure 2: Frequencies of words preceding and following the repeated fragment

Contrary to the hypothesis that the disfluency might reflect a processing problem associated with the relatively low frequency of a word following the fragment, the mean frequency of word immediately following the fragment is higher than the mean frequency of the words preceding the fragment.

3.4. Syntactic complexity and utterance length

Scores for syntactic complexity are shown in Table 3. Though the means suggest a tendency for utterances containing FPWRs to be more complex,in a 2-way ANOVA (Child x Fluency), all differences failed to reach statistical significance.

Scores for utterance length are shown in table 4. A2-way ANOVA (Child x Fluency) indicated that E's utterances were reliably longer than R's ($p = 0.0003$) and that utterances containing FPWRs were longer than fluent utterances ($p = 0.0178$); the interaction was non-significant.

Table 3: Syntactic complexity of utterances as indicated by mean number of constituents

	R	E
Containing FPWRs	4.8	5.76
Without FPWRs	4.36	5.04

Table 4: Mean length of utterances (syllables)

	R	E
Containing FPWRs	11.28	16.20
Without FPWRs	8.28	12.96

4. Discussion

McAllister & Kingston [12] had already established that the FPWRs produced by E and R could occur regardless of the preceding word's lexical status (content versus function word), length (monosyllable or polysyllable) or morphological status (monomorphemic or polymorphemic). The form of the repeated word fragments could be predicted according to phonological rules specific to each child.

In the analyses reported here, we have described some further characteristics of the children's speech. The repeated fragments were always preceded by a pause lasting on average about 600 msec; in the majority of cases, no pause occurred after the fragment, but when it did, it tended to be shorter than the pre-fragment pause. In the speech of these children, the occurrence of the disfluency did not appear to be related to lexical access difficulties inherent in the processing of an upcoming low-frequency word, because the words following the fragment are if anything of higher frequency than the words preceding it. It is tempting to suggest that the frequency pattern that is observed may reflect some structural aspect of the sentences (e.g., that the high frequency of words at position +1 may reflect their status as function words, which might occur at the start of new phrases). This notion deserves further investigation, although it should be noted that the analysis of grammatical complexity does not strongly support the idea that the disfluencies might be associated with difficulties in encoding more syntactically complex sentences. On the other hand, the occurrence of FPWRs was associated with the encoding of phonologically lengthier sentences.

FPWRs may or may not be a form of stuttering; however, in terms of explaining the behaviour within a model of speech production, perhaps such classification is unimportant, since some researchers have sought to explain both stuttering and other forms of disfluency using similar frameworks. Kolk & Postma [16] have grounded their Covert Repair Hypothesis of stuttering within Levelt's [17] theory of speech production. Within this framework, the repeated word fragments might arise at the phonetic spellout stage because of anomalous persistence of activation of rimes (for R) or of the final syllabic constituent (for E). Levelt's account of phonetic spellout is motivated in part by research into sound substitution errors; in the light of E's treatment of items containing diphthongs, it is interesting to note Levelt's comment that substitution errors in which the components of diphthongs are split, though rare, do occur. At least one theory of stuttering [18] proposes that one factor in such disfluency could be a failure on the part of speakers to respect the integrity of syllables.

It is, however, perhaps premature to speculate about the mechanisms through which FPWRs arise. Before a convincing explanation can be provided, a more thorough qualitative and quantitative description of the phenomenon needs to be produced.

5. Acknowledgements

Thanks to R and E, and to Sally Wynne.

6. References

[1] Rudmin, F. (1984). Parent's report of stress and articulation oscillation in a pre-schooler's disfluencies. *Journal of Fluency Disorders, 9*, 85-87

[2] Camarata, S. M. (1989). Final consonant repetition: A linguistic perspective. *Journal of Speech and Hearing Disorders, 54*, 159-162

[3] Mowrer, D. E. (1987). Repetition of final consonants in the speech of a young child. *Journal of Speech and Hearing Disorders, 52*, 174-178.

[4] Lebrun, Y., & Van Borsel, J. (1990). Final sound repetitions. *Journal of Fluency Disorders, 15*, 107-113

[5] Stansfield, J. (1995). Word-final disfluencies in adults with learning difficulties. *Journal of Fluency Disorders.*
[6] Rosenbek, J. (1984). Stuttering secondary to nervous system damage. In R. F. Curlee & W. H. Perkins (Eds.), *Nature and Treatment of Stuttering: New Directions*. San Diego: College-Hill.
[7] Rosenfield, D.B., Viswanath, N.S., Callis-Landrum, L., Didanato, R., & Nudelman, H.B. (1991). Patients with acquired dysfluencies: What they tell us about developmental stuttering. In Peters, H.F.M., Hulstijn, W. Starkweather, C.W. (eds.), *Speech motor control and stuttering*. Amsterdam: Elsevier.
[8] Lebrun, Y., & Leleux, C. (1985). Acquired stuttering following right-brain damage in dextrals. *Journal of Fluency Disorders, 10,* 137-141.
[9] Ardila, A., & Lopez, M. (1986). Severe stuttering associated with right hemisphere lesion. *Brain and Language, 27,* 239-246.
[10] Van Borsel, J., Van Coster, R., & Van Lierde, K. (1996). Repetitions in final position in a nine-year-old boy with focal brain damage. *Journal of Fluency Disorders, 21*, 137-146.
[11] Guitar, B. (1998). *Stuttering: An Integrated Approach to its Nature and Treatment*. Philadelphia: Lippincott, Williams & Wilkins.
[12] McAllister, J. & Kingston, M. (in press). FPWRs in School-age Children: Two Case Studies. *Journal of Fluency Disorders.*
[13] Bernstein Ratner, N. (1977). Stuttering: a psycholinguistic perspective. In R. F. Curlee & W. H. Perkins (Eds.), *Nature and Treatment of Stuttering: New Directions*. San Diego: College-Hill.
[14] Brown, S. (1945). The loci of stutterings in the speech sequence. *Journal of Speech Disorders, 10,* 181-192.
[15] Melnick, K. & Conture, E. (2000). The systematic and nonsystematic speech errors and stuttering of children who stutter. *Journal of Fluency Disorders, 25,* 21-45.
[16] Kolk, H. & Postma, A. (1997). Stuttering as a covert repair phenomenon. In R.F. Curlee & G.M. Siegel (Eds) *Nature and Treatment of Stuttering: New Directions* 2nd Edition. Boston: Allyn & Bacon.
[17] Levelt, W. (1989). *Speaking: From Intention to Articulation.* Cambridge, Mass.: MIT Press.
[18] Wingate, M. (2000). *Foundations of Stuttering*. San Diego: Academic Press.

Notes

1. The nucleus is the peak of prominence in a syllable, and usually consists of a vowel; the onset is the set of consonants, if any, preceding the nucleus, and the coda is the set of consonants, if any, following the nucleus. The nucleus and coda together form the rime.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 13-16.

A comparison of disfluency patterns in normal and stuttered speech

Timothy Arbisi-Kelm & Sun-Ah Jun

Department of Linguistics, UCLA, Los Angeles, CA, USA

Abstract

While speech disfluencies are commonly found in every speaker's speech, stuttering is a language disorder characterized by an abnormally high rate of speech aberrations, including prolongation, cessation, and repetition of speech segments [5]. However, despite the obvious differences between stuttered and normal speech, identifying the crucial qualities that identify stuttered speech remains a significant challenge. A story-telling task was presented to four stutterers and four non-stutterers in order to analyze the prosodic patterns that surfaced from their spontaneous narrations. Preliminary results revealed that the major difference between stutterers' and non-stutterers' disfluencies —aside from the total number—is the type of disfluency and the context affected by the disfluency. Disfluencies in both groups included *prolongation, pause* and *cut*, but stutterers' disfluencies also include *repetition* and combinations of the three (e.g., *cut* followed by *pause*). In addition, stutterers' disfluencies were accompanied by more prosodic irregularities (e.g. pitch accent on function words, creating a prosodic break with degraded phonetic cues) prior to the actual disfluency than non-stutterers' disfluencies, indirectly supporting the overvigilant self-monitoring hypothesis [14, 16, 17].

1. Introduction

Previous stuttering literature has defined disfluencies as the "repetitions or prolongations in the utterance of short speech elements, namely sounds, syllables, and words of one syllable" [18], but the variety of disfluency types have not been translated into a contemporary model of linguistic intonation. Word break patterns are a potentially revealing source of data in stuttered speech, since pauses and prolongations have been claimed to be evidence of processing delays, whether they be early (e.g., conceptual or lexical) or late (e.g., phonological or articulatory). Furthermore, unnatural break patterns may also provoke deviations from normal prosodic patterns, which can in turn result in the loss of information crucial to the intonational meaning of an utterance.

As shown in a number of earlier studies, metrical stress appears to be a trigger of disfluencies in stuttered speech [8, 13]. Arbisi-Kelm [1] manipulated the intonation structure of various sentence types adopting the Autosegmental-metrical model of intonational phonology proposed by Pierrehumbert and her colleagues ([3, 11, 12]) and asked adult stutterers to read the sentences. He transcribed the prosodic patterns of utterances using the English ToBI conventions [2], and found that, with the exception of focused nuclear pitch accent, disfluency patterns follow the hierarchical metrical scale of prominence: that is, stuttered disfluencies are a function of metrical prominence, predictable by properties of intonation [2, 3]. Since pitch accented syllables bear a greater degree of prominence than non-pitch accented syllables, it is predicted that they should attract a higher rate of stuttering disfluency. Consequently, disfluencies should have disruptive effects on the intonation structure in which they occur.

Though some believe that stuttered disfluencies are distinct from that of normal speakers [9, 10], others believe that the difference is simply a matter of degree [4]. Vasic and Wijnen [16] also support the latter view. Adopting Levelt's theory [6] that both overt speech and an internal speech plan are monitored, they claim that the difference between stutterers and non-stutterers are in the degree of self-monitoring. That is, stuttering is a "direct result of an overvigilant monitor". They believe that repairs made often introduce disfluencies rather than prevent them ([17], cited in [14]). Recently, Russell, Corley, & Lickley [14] found that stutterers are more sensitive than non-stutterers in perceiving the disfluencies made by other people, supporting the self-monitoring hypothesis.

In this paper, we will examine the disfluency patterns in normal and stuttered speech and compare the types of disfluencies as well as the effect of each disfluency types on the tonal context and the phrasing. The results will suggest whether the difference in the disfluencies produced by the stutterers and non-stutterers are qualitative or quantitative.

2. Method

2.1. Subjects

Four age-matched adult persons who stutter (stutterers) and four age-matched normal speakers (non-stutterers) were selected to participate in a storytelling task. Participants were recruited from the greater Los Angeles area, through local universities and stuttering support groups. Stuttering level of each participant was moderate-severe, as determined by assessments provided by licensed speech language pathologists.

2.2. Procedure

Subjects were seated in a quiet room, each for a single session of approximately one hour. Subjects wore a head-mounted SM10A Shure microphone, with the signal passed to a Marantz portable cassette recorder (PMD222). Instructions were simply to narrate the picture book, "Frog Where Are You?" (Mayer 1969), as if sharing the story with someone for the first time. This procedure is chosen because it allows subjects to produce spontaneous and natural-sounding utterances delivering the same story, while using a similar or same set of lexical items for the characters and objects shown in the picture. In order to facilitate the creation of a narrative structure, subjects were instructed to peruse the book before the task and form a general idea of the story.

2.3. Data Analysis

All data samples were audiotaped, sampled at a rate of 11025 Hz, and stored digitally. Using the *PitchWorks* signal analysis software program (SCICON R&D), data files were

coded following the ToBI (Tones and Break Indices) transcription conventions for English intonation and juncture (Beckman & Ayers [2]). For tones, five pitch accents (H*, L*, L+H*, L*+H, H+!H*) plus downstepped H tones (e.g., !H*, L+!H*, L*+!H) were used as well as phrase accents (H-, L-, !H-) and boundary tones (H%, L%). For break indices, the existing ToBI labelling for the juncture between words was used: '1' for the default, phrase-medial, word boundary, '3' for the juncture corresponding to an Intermediate phrase boundary, and '4' for the juncture corresponding to an Intonation phrase boundary.

The disfluency labelling in English ToBI is done in two tiers -- break indices (BI) and miscellaneous (misc) tiers. The labelling in the BI tier includes the p diacritic in conjunction with a break index 1, 2, or 3. '1p' (an abrupt cutoff before an actual repair), '2p' (a hesitation pause or prolongation of segmental material without having an intermediate phrase tone), and '3p' (a hesitation pause or a pause-like prolongation where there is an intermediate phrase accent in the tone tier). The misc tier is used when the disfluency spans over a word and is difficult to locate its exact timing. In this case, both the beginning and the end of the event is marked by the diacritics '<' and '>', respectively after the disfluency word such as 'disfl< ... disfl>' or 'repair< ... repair>'.

Since the current disfluency labelling does not cover the variety of disfluent types occurring in stuttered speech, we extended the disfluency diacritics as shown in (1) and labelled in the break index tier after the break index number corresponding to the different degrees of juncture due to this disfluency. For example, when a part of a word is cut, '0c' is used, when a phrase-medial word boundary is cut, '1c' is used, and when a word at the end of an intermediate phrase is cut, '3c' is used. Similarly, '1ps' is used to label a disfluent pause after a word boundary, and '3pr.ps' is used to label a disfluent prolongation followed by a pause after an ip boundary.

(1) cut (c)
  prolongation (pr)
  pause (ps)
  repetition (t)
  combinations of these:
    cut followed by pause (c.ps)
    prolongation followed by pause (pr.ps)
    prolongation followed by cut (pr.c)
    prolongation followed by cut followed by pause (pr.c.ps)

3. Results and Discussion

3.1. Prosodic phrasing and frequency of disfluencies

Labellings have been done for three speakers in each group, but data from two speakers in each group have been analyzed so far and are reported here. Table 1 shows the total number of words, the total number of Intermediate phrases (ip), and the total number of disfluencies in the story narrated by four speakers: sm1 and sm2 are stutterers and cm1 and cm2 are non-stutterers. As expected, stutterers have significantly more number of disfluencies than non-stutterers. All speakers (except for sm1) used a similar number of words in narrating the story. Speaker sm1 used more words to describe the pictures, thus producing a greater number of disfluencies, pitch accents, and ips than the other speakers. However, the average number of words and pitch accents per ip was slightly smaller than those of the other stutterer, sm2, suggesting that sm1's phrase would sound choppier than sm2's. But, since sm1's average number of disfluencies per ip was slightly smaller than that of sm2 (0.89 vs. 1.14 per ip), sm1 may not necessarily be a more severe stutterer than sm2.

On the other hand, the other stutterer, sm2, though he has a large number of disfluencies, seems not much different from cf1, one of the control speakers, in terms of the number of words and pitch accents per ip. This suggests that stutterers may not be qualitatively different from non-stutterers in phrasing words and producing prominent words within the phrase. Rather, it seems that there is a continuum of fluency. Stutterers have fewer words in a phrase than non-stutterers. This may have happened because the disfluency interrupts producing a longer phrase.

Table 1: Total number of words, intermediate phrases (ip), disfluencies, and pitch accents (PA) in the story produced by two stutterers (sm1, sm2) and two non-stutterers (cm1, cf1), and the average and standard deviation of disfluencies and PAs in an ip.

	sm1	Sm2	cm1	cf1
# words	1148	535	492	489
# ip	341	150	109	126
Avg. wd/ip	3.4	3.6	4.5	3.9
#disfl.	314	168	8	9
Avg. disfl./ip	0.92	1.14	0.07	0.07
sd of disfl	0.97	1.01	0.29	0.26
#PA	535	288	272	232
Avg. PA/ip	1.56	1.92	2.39	1.83
sd of PA	0.88	1.07	1.13	0.93

3.2. Types of disfluencies

Though phrasing and pitch accent data seem to suggest a continuum of fluency between stutterers and non-stutterers, the types of disfluency data show a qualitative difference between these two groups. Table 2 lists the types of disfluencies produced by each speaker. It is shown that disfluency types differ between the stutterers and non-stutterers in two ways. First, while *prolongation*, *pause*, and *cut* types of disfluency occur in both groups, *repetition* occurs only in stutterers' speech. Second, the combination of *prolongation*, *pause*, and *cut* was observed mainly in the stutterers' speech, though not as frequently as the single type disfluency. In non-stutterers' speech, the combination was found only once by Speaker cm1 (i.e., prolong-pause). However, data show that among the types of disfluency, the proportion of *prolongation*, *pause* and *cut* is similar between stutterers and non-stutterers. For both groups, *prolongation* or *pause* was the most common and *cut* was the least common.

Table 2: Types of disfluency (single and combinations) for each speaker. In each type, the top row shows a raw number and the bottom row shows the percentage.

Disfluency types	sm1	sm2	cm1	cf1
Prolong (pr)	73	67	2	5
	24%	40%	25%	56%
Pause (ps)	115	39	4	3
	36%	23%	50%	33%
Cut (c)	46	22	1	1
	15%	13%	12.5%	11%
Repetition (t)	11	4	0	0
	4%	2.4%		
Cut-pause (c.p)	7	2	0	0
	2.2%	1.2%		
Prolong-cut	4	2	0	0
(pr.c)	1.3%	1.2%		
Prolong-pause	52	26	1	0
(pr.ps)	16.6%	15%	12.5%	
Prolong-cut-	0	1	0	0
Pause pr.c.ps)	0%	0.6%		
Total	314	168	8	9

3.3. The effect of disfluency on the adjacent contexts

Stutterers' and non-stutterers' speech also differed qualitatively in how the disfluency affected the adjacent tonal contexts and phrasing. Table 3 shows the types of disfluency and their effects (behaviors) and the frequency of each behavior for each speaker. Since the frequencies of each disfluency type produced by non-stutterers are very small, a direct comparison between groups cannot be made. The table is here to provide an idea of how each disfluency type affects the tonal pattern, duration, and phrasing.

In general, the stutterers' disfluencies disturbed the phrasing and tonal context in more various ways than those by the non-stutterers. For example, for both groups, disfluent *prolongation* often created a downstepped pitch accent on the following word and created an intermediate phrase (ip) boundary by extending f0 rise or fall on the prolonged word. However, in stutterers' speech, disfluent prolongations often created pitch accent on the prolonged word and rushed the following words to compensate for the time lost, but this was not the case with non-stutterers' speech. Similarly, for the existence of disfluent *pauses*, both groups often placed an ip boundary at the pause location and produced downstep-like pitch accent on the following word, but stutterers tended to produce pitch reset on the post-pause word and often did not fully produce pre-boundary cues such as phrase-final lengthening or phrase accent. Finally, when disfluent *cuts* were made, the phrasing was rarely affected by speakers of both groups. But one stutterer (sm1) produced pitch reset 28% of the time and the other stutterer (sm2) produced unintended pitch accent on the following word 25% of the time (probably to reinforce failed articulations).

Another noticeable influence of disfluency on prosody was to produce pitch accented function words and rising pitch accent on the disfluent word (e.g. L+H*), creating an incorrect information structure or meaning of contrast. This pattern was found more often in stutterers' speech than non-stutterers' (though we need more non-stutterers' data to confirm this). In sum, the errors in tone and duration and various unorthodox ip patterns found in stutterers' speech suggest that the grouping of pitch accents was jeopardized because of tune-text alignment deficit (i.e., improperly anchored PAs).

Finally, data show that, in stutterers' speech, different types of disfluency were found *before* the target word (e.g. disfluency on the function word before the following content word) as well as on the target word itself. For stutterers, half of the dislfuency data was found before the target word, but for non-stutterers, almost all disfluencies were found during the target word. The target word was always a pitch accented word, and stutterers produced disfluencies before producing the prominent word, i.e., pitch accented target word. This created more disruptions in the prosodic contour and in the listener's understanding of the information structure, resulting in the perception of more disfluent speech. This may be interpreted as evidence that stutterers monitor their speech in advance to produce the target word, and their early repairs result in more disfluencies, thus supporting the self-monitoring hypothesis.

In sum, the data suggest that the disfluencies produced by stutterers and non-stutterers are qualitatively as well as quantitatively different. The data also suggest that analyzing disfluent utterances prosodically (in terms of tonal patterns, breaks, and phrasing) provides valuable data to the studies of disfluent speech.

4. References

[1] Arbisi-Kelm, Timothy. 2004. An Intonational Analysis of Disfluency Patterns in Chronically Stuttered Speech. A poster presented at the ASA.

[2] Beckman, Mary & Gayle Ayers-Elam. 1994. Guidelines for ToBI Labelling.

[3] Beckman, Mary. E. & Janet. Pierrehumbert. 1986. Intonational Structure in Japanese and English, *Phonology Yearbook* 3: 255-309

[4] Bloodstein, O. 1970. Stuttering and normal nonfluency: A continuity hypothesis. British J. of Disorders of Communication. 1970, 30-39.

[5] DSM-IV-R. 1994. Diagnostic and statistical manual of mental disorders: IV. Washington, D.C. : American Psychiatric Association.

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

[7] Mayer, M. 1969. Frog, Where are You? New York: Dial Books.

[8] Natke, U., Grosser, J., Sandrieser, P., & Kalveram, K.T. 2002. The duration component of the stress effect in stuttering. Journal of Fluency Disorders Volume 27, Issue 4 , 305-318.

[9] Perkins, W. H. 1990. What is stuttering? *J. of Speech and Hearing Disorders*, 55:370-382.

[10] Perkins, W. H. 1995. *Stuttering and science*. San Diego, CA: Singular Publishing Group.

[11] Pierrehumbert, J. 1980. The Phonology and Phonetics of English Prosody, Doctoral dissertation, MIT.

[12] Pierrehumbert, J. & M. E. Beckman. 1988. *Japanese Tone Structure*, Cambridge, MA: MIT Press.

[13] Prins, D., Hubbard, C., & Krause, M. 1991. Syllabic Stress and the Occurrence of Stuttering. JSHR 34, 5, 1011-1016.

[14] Russell, M., Corley, M. & Lickley, R. 2005. Magnitude Estimation of disfluency by stutterers and nonstutterers. In R. J. Hartsuiker, R. Bastiaanse, A. Postma, & F.

Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech*. Hove (East Sussex): Psychology Press.

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

[16] Vasic, N. & Wijnen, F. 2005. Stuttering as a monitoring deficit. In R. J. Hartsuiker, R. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech.* Hove (East Sussex): Psychology Press.

[17] Wijnen, 2000. Stotteren als resultaat van inadequate spraakmonitoring [Stuttering as the result of inadequate speech monitoring]. *Stem-, Spraak- en Taalpathologie*, 9.

[18] Wingate, M. E. 1964. A Standard Definition of Stuttering. Journal of Speech and Hearing Disorders. November 29: 484-9.

Table 3: The effect of disfluency on the adjacent contexts for each speaker.

Disfl patterns	Behavior	sm1	sm2	cm1	cf2
Prolongations	force Pitch Accent (PA) on prolonged word	38	62	0	0
	create timing compensation errors (i.e., rushed after disfluency)	8	7	0	0
	force ip break	64	23	0	1
	force downstep on following word	19	18	1	0
	force pitch reset	3	2	0	0
	force continuation f0 rise/falls	32	48	0	1
	no effect on ip	6	9	3	0
Pauses	force PAs on following word	2	19	0	1
	create timing compensation errors	7	1	0	0
	force ip break	118	27	3	1
	force downstep on following word	8	11	1	0
	force pitch reset	6	3	0	0
	degrade pre-boundary cues (onset closure=block)	46	1	0	0
	no effect on ip	23	10	3	2
Cuts	force PAs on following word	6	7	0	0
	force ip break	9	5	0	0
	force downstep on following word	1	2	0	0
	force pitch reset	18	3	0	0
	no effect on ip	29	9	2	1
Repetitions	force ip break (before or after boundary)	3	1	0	0
	force downstep on same word	1	0	0	0
	force pitch reset	4	0	0	0
	no effect on ip	13	7	0	0
PA on function words		27	62	1	1
Disfluency on target (Nuclear Pitch Accent/Pitch Accent)		**206**	**67**	**8**	**8**
Disfluency NOT on target		**108**	**101**	**0**	**1**

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 17-20.

Extracting the acoustic features of interruption points using non-lexical prosodic analysis

Matthew P. Aylett

ICSI, UC Berkeley, USA and
CSTR, University of Edinburgh, UK

Abstract

Non-lexical prosodic analysis is our term for the process of extracting prosodic structure from a speech waveform without reference to the lexical contents of the speech. It has been shown that human subjects are able to perceive prosodic structure within speech without lexical cues. There is some evidence that this extends to the perception of disfluency, for example, the detection interruption points (IPs) in low pass filtered speech samples. In this paper, we apply non-lexical prosodic analysis to a corpus of data collected for a speaker in a multi-person meeting environment. We show how non-lexical prosodic analysis can help structure corpus data of this kind, and reinforce previous findings that non-lexical acoustic cues can help detect IPs. These cues can be described by changes in amplitude and f0 after the IP and they can be related to the acoustic characteristics of hyper-articulated speech.

1. Introduction

Human subjects respond to prosodic structure without necessarily understanding the lexical items which make up the utterance. For example event-related brain potential (ERP) studies have shown a reliable correlation with phrase boundaries when utterances are made lexical nonsensical, either by humming the words, or by replacing them with nonsense words [9]. The use of prosodically rich pseudo speech for artistic purposes (such as R2D2 in star wars, and The Teletubbies amongst others) reinforce these findings. This effect, of apparently understanding prosodic structure without lexical cues, extends to the human perception of disfluency. Lickley [7] showed that human subjects could recognise interruption points, the boundary between disfluent and fluent speech, in low pass filtered speech where no lexical cues were present.

Non-lexical prosodic analysis (NLPA) attempts to mimic this human ability of non-lexical prosodic recognition. Initially, interest in NLPA was motivated largely by the objective of improving automatic speech recognition (ASR) technology, for example, by pre-processing the speech to find syllables [5] or prosodic prominence [3]. However, improvements in statistical modelling in ASR meant that, often, the speech recogniser itself was best left to model prosodic effects internally. Recently, there has been a renewed interest in NLPA techniques in order to address the problem of recognising, segmenting, and characterising very large spontaneous speech databases. Tamburini and Caini [10] point out that identifying prosodic phenomena is useful, not only for ASR and speech synthesis modelling, but also for disambiguating natural language and for the construction of large annotated resources. In these cases, the ability to recognise prosodic structure without lexical cues has two main advantages:

1. It does not require the resource intensive, and language dependent, engineering required for full speech recognition systems.
2. It can offer a means of modelling the human recognition of prosodic structure which in turn could lead to an improved understanding of human speech perception and production.

The ability of human subjects to recognise interruption points (IPs) without lexical information raises the question of whether NLPA can do as good a job. Although previous work has looked at this problem in some depth (e.g. [4], [7]), NLPA offers the prospect of a structured analysis that could be carried out automatically over very large speech databases. In addition, the presence of previous detailed studies allows us to validate the overall approach.

The non-lexical detection of IPs is also of interest from the perspective of determining dialogue structure. Recent work suggests that disfluency patterns could be used to signal the speakers' cognitive load [1] and thus might be used to determine areas in dialogue involving complex concepts, ideas or planning.

We will first describe in more detail the corpus of speech we analysed and the IP phenomena. Next, we will present the details of the NLPA we applied to this corpus followed by results for a set of acoustic features which may cue the non-lexical perception of IPs. Finally, we will discuss limitations with the approach and possible future work.

2. Corpus and disfluency coding

Our data was selected from the ICSI meeting corpus [6]. This consists of 75 dialogues collected from the regular weekly meetings of various ICSI research teams. Meetings in general run for under an hour and have on average 6.5 participants each recorded on a separate acoustic channel. The speech is segmented into spurts, defined as periods of speech which have no pauses greater than 0.5 seconds.

The data we present here is taken from 4 speakers taken from 10 dialogues. Disfluencies are coded as part of the dialogue act coding [2], where interruption points are shown as a hyphen in the speech transcription. In order to avoid complexity caused by multi-speaker interaction and multiple disfluencies, we looked only at phrase boundaries and IPs where:

- The same speaker continued speaking after the interruption point or phrase break
- No other speakers were speaking within 0.5 seconds of the break
- There was at least 0.5 seconds between any breaks.

Pause duration is the clearest acoustic cue of a prosodic break and can be used to disambiguate between IPs and phrase boundaries with some success. In general, the longer the pause, the more likely the break is a phrase boundary. However there are plenty of examples of phrase boundaries followed by a short pause. An interesting question is whether we can disambiguate between these phrase boundaries and IPs followed by a similar short pause. In order to concentrate on this problem, we limited the analysis to IPs and phrase

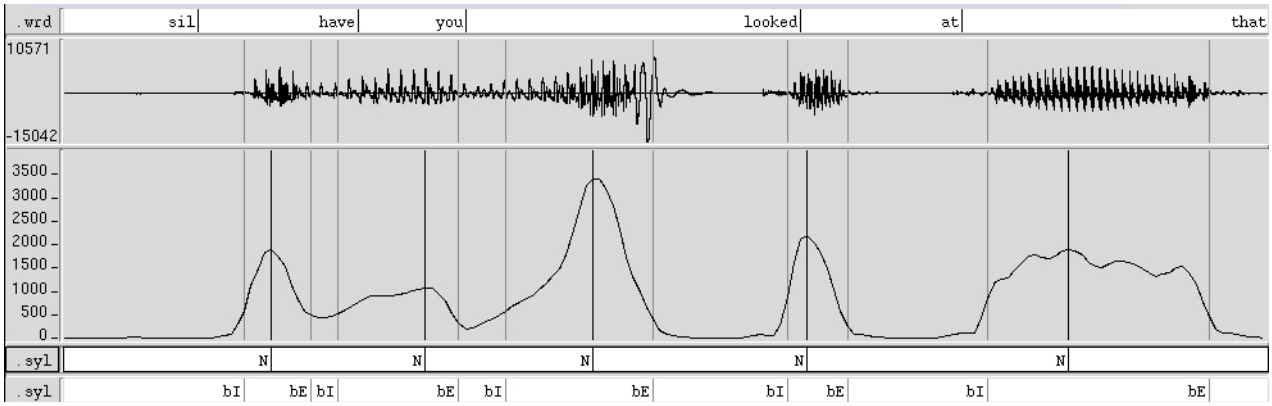


Figure 1: Automatic syllable detection. The lexical contents are shown at the top followed by the waveform, band pass (300-900Hz) energy of the speech and at the bottom labels assigned for syllable nuclei (N) and their initial boundary (bI) and end boundary (bE).

boundaries where the automatic aligner did not insert a following pause.

In this study disfluencies are categorised as repetitions:
"right to my... my right"
substitutions:
"I don't suppose you've got the balloons... the baboons?"
insertions:
"parallel with the ravine... the word ravine"
and deletions:
"oh no what... the line stops at the flagship"

The three dots in the above examples mark the interruption point (IP) (which may or may not be followed by a pause).

3. Non-lexical prosodic analysis

Any acoustic feature can be used to characterise prosody without lexical input. However, a good starting point is features which are reasonably ubiquitous, cross linguistic and have been shown to be sufficient for much human interpretation of prosodic structure. On this basis, amplitude and fundamental frequency are clear starting points. The syllable is a typical means of structuring this acoustic information. Within prosodic theory prominence is associated with syllables, in particular syllable nuclei. Therefore, a first step in any NLPA is syllable extraction. Howitt [5] reviews many of the current algorithms for segmenting speech into syllables. If we evaluate these algorithms in terms of how well they predict the syllable boundaries compared to those produced by human segmentation (or even by autosegmentation), they typically perform rather poorly. However, for NLPA we are not attempting to segment speech, our intention is rather to characterise the prosodic structure. Given that much of the perceived amplitude and pitch change occurs across the syllable nucleus, finding the extent of the nuclei is more important than determining the syllable boundaries. In fact, most simple syllable detection algorithms will find 80% of the syllable nuclei and the syllables they typically miss are unstressed, short syllables, which tend to carry much less prosodic information. In addition, Tamburini and Caini [10] found that the duration of nuclei correlates closely to the overall syllable duration and therefore the syllable nuclei duration can be used to measure the rate of speech as well as assessing prominence.

On this basis, we extracted syllable nuclei as suggested by Howitt [5]. This involves band pass filtering speech between 300-900 Hz and then using peak picking algorithms to determine the location and extent of nuclei. For these experiments we used a simpler peak picking algorithm than the modified convex-hull algorithm [8] described by Howitt [5] and used by Tamburini and Caini [10].

Figure 1 shows an example of the results of the syllable extraction algorithm we applied. The top shows the lexical contents of the speech, followed by a waveform. Below the waveform is the energy of the band pass filtered speech. The labels below the band pass filtered speech show the syllable nuclei (black line) and the extent of the nuclei (grey lines). The process for determining these nuclei is as follows:

1. Remove large portions of silence from the data and divide the speech into spurts - continuous speech with less than 0.5 seconds gap. Allow 0.1 seconds of silence before and after each spurt.
2. Band pass filter the speech between 300-900 Hz.
3. Examine the distribution of the energy for the speaker across the data and set a threshold for syllable energy at the 65th percentile.
4. Find the maximum points in the region. A maximum point has a previous and subsequent lower value with a number of equal values in between. Order the points by amplitude and go through the list picking syllable nuclei providing a previous nuclei has not already been picked within a range of 0.1 seconds.
5. Set the boundaries as equidistant between nuclei in the same voiced region otherwise to the threshold edge of the region.
6. Extract f0 values, using the entropics get_f0 program, for the start, centre and end of the syllable nuclei.
7. Smooth the resulting f0 contour and interpolate values across unvoiced regions.

We can assess the prominence of each syllable either based on amplitude and duration (sometimes described as stress prominence [10]), or the f0 variation over the syllable nucleus (sometimes described as accent prominence [10]). Phrase boundaries are assessed both on the basis of pauses, determined by a simple threshold silence detector, and boundary f0, taken as the f0 at the edge(s) of the surrounding syllable nuclei.

4. Analysing IP boundaries with NLPA

Previous work, Lickley [7] and Hirschberg et al [4], has shown a number of interesting acoustic features which can be associated with IPs. All the features occur after the IP with no discernable acoustic cues before the IP. Both [7][4] found a tendency for increased amplitude after the IP, higher f0 and longer duration. These are all correlates of stressed syllables and also of hyper-articulated speech. Hirschberg et al [4] describe these acoustic features as cues for *corrected speech* and describe a machine learning approach for classifying corrected speech on the basis of these features. This was then applied to reduce recognition error from 25% down to 16%. However, the extent to which these utterances contained IPs was not reported. In Lickley [7], human judgments of low pass filtered speech utterances show a significant, although far from consistent, effect across materials. Human subjects tended to

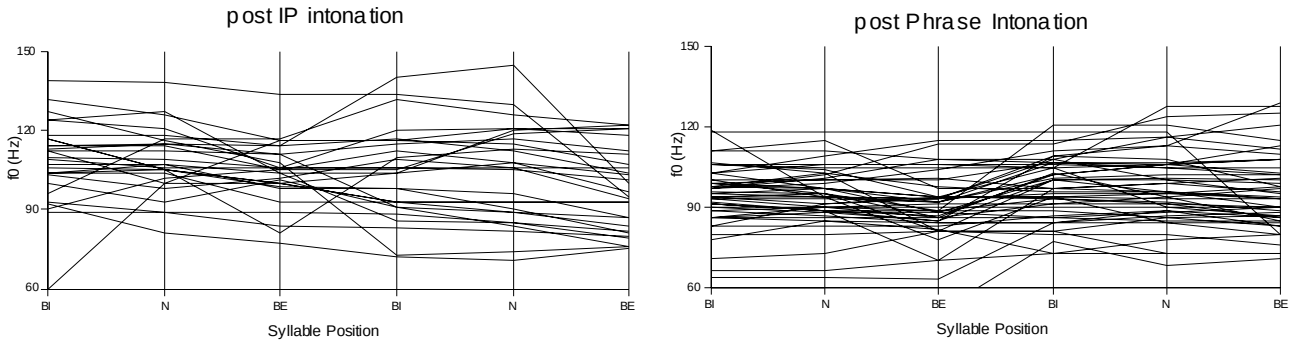


Figure 2: F0 across the two syllable nuclei following both IPs and phrase breaks where no or minimal pause cues are present. (BI - initial boundary of syllable nucleus, N - centre of syllable nucleus, BE - end boundary of syllable nucleus).

misclassify disfluent utterances as fluent utterances more then visa-versa with the best group of human subjects correctly classifying 34% of disfluent utterances as disfluent. In addition, the significant effect in this study appeared to be dominated by the presence and differences of pause durations rather than other acoustic cues.

In this data, as stated earlier, only boundaries with pauses not discernable to the aligner where examined. We compared the results for IPs and normal phrase breaks. As in [7] [4], we looked at acoustic cues in the form of f0 variation, syllabic nucleus amplitude and syllabic nucleus duration after the boundary point.

5. Results

We began by looking at the f0 change across the two syllables to the right of the boundaries. Shown in figure 2 are six f0 points. These values are taken from the first two syllable nuclei found with NLPA subsequent to the phrase or IP boundary for a single male speaker. It is interesting to note a lack of a homogeneous f0 structure in either IP or for PH (Phrase conditions). However, differences are clearly present between both groups. F0 in the IP case tends to be higher and varies more throughout the two syllables.

On the basis of this plot we chose three f0 features to examine statistically: the f0 before the boundary, the f0 following the boundary and the variance of the f0 across the two syllables following the boundary. F0 values were normalised with on the basis of the mean and standard deviation of each speakers voiced f0 values. In addition, we combined the log of the raw amplitude of the first following syllable with the log of the duration of its nucleus by multiplying the factors together to give an overall prominence factor. Thus short, high energy syllable nuclei where regarded as having similar prominence to long, lower energy syllable nuclei.

An independent t-test grouped by IP and phrase boundary (PH) is shown in Table 1. All factors except post boundary f0 variance are significant with an appropriate Bonferroni correction. If we examine the cell means in Figures 3 and 4 the results are in line with previous published results. We see higher initial f0 values for after IPs and more prominence caused by amplitude and duration.

If we use these factors in a discriminant analysis, we find we can categorise 58.7% of the data (58% with cross validation), see Table 2. Even given the absence of pause data these results are poor and suggest significant by-speaker variation. If we analyze the speakers individually we see a varying but similar pattern for pre/post f0 and prominence as reflected in Figure 3 and 4. However the pattern of post boundary f0 variance is very different across speakers. Figure 5 shows the normalized post boundary f0 variance for all four speakers. If we look at subject *me011* (the subject whose f0 is plotted in Figure 2) we see that f0 varies more after the IP than the phrase break ($p < 0.05$ - NS with Bonferroni

Table 1: Independent t-test for acoustic cues following IPs and Phrase Boundaries.

Acoustic Feature	t	df	Sig. (2-tailed)	Bonferroni correction
f0 pre boundary	*4.466*	*1,173*	*0.000*	*0.000*
f0 post boundary	*7.091*	*1,173*	*0.000*	*0.000*
f0 variance post boundary	*-0.403*	*1,198*	*NS*	*NS*
prominence post boundary	*3.181*	*1,154*	*0.002*	*0.008*

Table 2: Results of discrimant analysis using acoustic cues

Discriminant Analysis		Classification	
		PH	IP
Original	PH	*355*	*301*
	IP	*203*	*341*

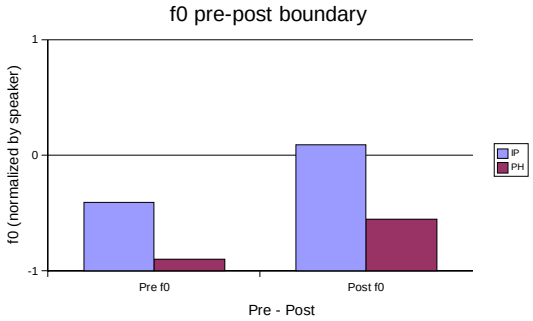


Figure 3: F0 across IP boundary and phrase boundary (PH).

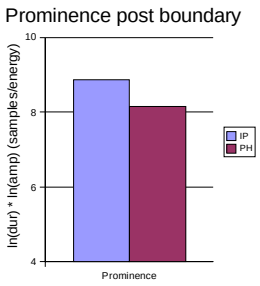


Figure 4: Prominence -nucleus ln(amplitude)ln(duration)) - after IP and phrase boundary (PH).

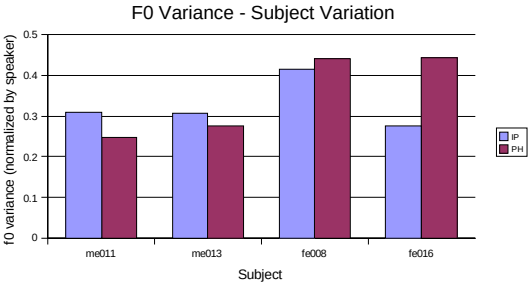


Figure 5: Post boundary f0 variance by speaker

correction). In complete contrast, subject *fe016* shows the reverse ($p < 0.005 – p < 0.05$ with Bonferroni correction). The success of discriminant analysis using these factors also varies from 58% to 69% across speakers. The *fe* denotes a female speaker whereas *me* a male speaker. It is possible that some of this variation is associated with the sex of the speaker although a much larger sample of speakers would be required to explore this possibility.

These results are for the IPs and phrase breaks where no pause had been inserted by the aligner. Pause duration is still, by far, the best indicator of phrasing. This raises the question of to what extent pause determined non-lexically might be useful for classifying IPs and phrase breaks.

In order to address this we examined the boundaries and determined pauses based on the band pass energy being below the 25th percentile. Results suggest the aligner has a tendency to absorb short pauses. The mean pause calculated in this way for boundaries with no pause discernable to the aligner was 45 milliseconds. If we add this acoustic pause factor to our discriminant analysis we see a jump from 58% to 68% success of classification.

6. Conclusion

Results show that NLPA can be used for characterising disfluency. Furthermore, that it would seem to perform as well, or better, than human subjects given the same task. Perhaps the most interesting feature of the work is that NLPA offers a non-lexical structure for dealing with timing. Using the syllable nucleus we can implicitly scale f0 contours which might allow a more structured approach to characterizing intonation non-lexically. Although the prominence feature presented in this work is perhaps an over simplification of the perceptual effect of duration and amplitude, it does allow a starting point for an improved system. Similarly it would be an interesting idea to replace the f0 variance with a more perceptually based model of accentedness.

Results for four speakers suggest that a great deal of inter-subject variation makes it hard to produce a general model of these factors. This is further complicated by treating all IPs as the same when previous work (i.e. [7]) has shown that there are differences between the typical acoustic effects of, for example, repetitions as opposed to deletions.

The results for pauses determined acoustically and non-lexically offer a salutary lesson against depending blindly on automatically aligned results for corpora analysis. A simple non lexical acoustic analysis can be taken quite far: it requires less resources, is less language dependent and arguably less prescriptive.

However, the success of NLPA depends largely on the autosyllabification process. Overgeneration of syllables and overestimation of syllable nuclei, for example, caused by liquids or nasals, can present a significant problem in terms of aligning f0 contours with the output. In future work we will evaluate the syllabification algorithm quantitatively against state-of-the-art automatic alignment. Preliminary results suggest the current NLPA matches 65% of syllables from the alignment. However over 50% of the syllables missed are schwa nuclei which reinforces the idea that NLPA might do a better job at finding prosodically *pertinent* syllables than automatic alignment. The system appears to have around a 10% false alarm rate and, as expected, these are very much associated with nasals and roticization.

The IP analysis reinforces findings from previously published work. The results for automatic disambiguation (especially given the lack of pause information) are promising. However, in order to really test how useful these factors are for discrimination, we must also see to what extent they can tell any boundary (syllable/word) from an IP. In addition, as pointed out by Hirschberg et al [4], different speakers have different characteristics in terms of hyper-articulation. On this basis further work requires the analysis of many more subjects.

7. Acknowledgements

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811 publication).

8. References

[1] Bard, E., Lickley, R.J. & Aylett, M.P. 2001. Is Disfluency Just Difficult? *Proceedings of DISS 01, ISCA Tutorial and Research Workshop*. Edinburgh.

[2] Dhillon, R., Bhagat, H., Carvey, H. & Shriberg, E. 2004. Meeting Recorder Project: Dialog Act Labelling Guide. *Technical Report TR-04-002*. ICSI.

[3] Hironymous, J.L., McKelvie, D. & McInnes, F.R. 1992. Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition. *ICASSP '92 Proceedings.* San Francisco. California, pp225-227.

[4] Hirschberg, J., Litman, D. & Swerts M. 1999. Prosodic Cues to Recognition Errors. *ASRU-99.*

[5] Howitt A.W. 2000. *Automatic Syllable Detection of Vowel Landmarks.* PhD Thesis, MIT.

[6] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. & Wooters, C. 2003. The ICSI Meeting Corpus. *Proceedings ICASSP-03.* Hong Kong.

[7] Lickely, R.J. 1994. *Detecting Disfluency in Spontaneous Speech.* PhD Thesis, University of Edinburgh.

[8] Mermelstein, P. 1975. Automatic segmentation of speech into syllabic units. *JASA.* 58(4) pp880-883.

[9] Pannekamp A., Toepel, U., Alter, K., Hahne, A. & Friederici, A.D. 2005. Prosody-driven Sentence Processing: An Event-related Brain Potential Study. *Journal of Cognitive Neuroscience.* 17(3) pp407-421.

[10] Tamburini, F. & Caini, C. 2005. An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech. *International Journal of Speech Technology.* 8 pp33-44.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 21–25.

Prosodic cues of spontaneous speech in French

Katarina Bartkova

France Telecom R&D, Lannion, France

Abstract

Disfluencies, when present in speech signal, can make syntactic parsing difficult. This difficulty is increased when machines are involved in communication and when speech devices rely on automatic speech recognition techniques. In order to improve automatic speech parsing and thus speech comprehension, methods have been proposed to filter disfluencies out from the speech signal. Attempts have been made to use prosodic parameters to improve such a filtering. However, before introducing prosodic parameters into automatic speech recognition processes, it would be useful to investigate whether disfluencies can be characterized in a prosodic way and whether their prosodic cues would be representative enough to be used in automatic systems. The aim of this study was to examine to which extent prosodic parameters would be able to characterize disfluencies in French. Word repetitions, filled and silent pauses and speech repairs were described in a prosodic way using statistical analyses of their prosodic parameters. These analyses allowed simple prosodic rules to be formulated. The efficiency of the prosodic rules was evaluated on the task of filled pauses, word repetitions and hesitation detections.

1. Introduction

Disfluencies, frequently present in spontaneous speech, cause ill-formed and longer sentences which are thus harder to process for natural language understanding. A listener must break the continuous speech stream down into component parts, then build a syntactic structure and determine the meaning that the speaker intended to convey. According to Bailey & al. [3] non-word disfluencies can affect the syntactic parsing and make the syntactic reanalysis more difficult. They also suggest that utterances with disfluencies are parsed differently from those without disfluencies. Thus, although disfluencies have often been viewed as pragmatic phenomena, they can affect language comprehension. According to Bortfeld [6] disfluency is strongly task related and only slightly age related: older speakers produce only slightly higher disfluency rates than young or middle-aged speakers. The task relation of disfluency was confirmed by Arnold et al. [2] who claim that disfluencies occur more often during references to things that are discourse-new, rather than given. Asp [1] found a higher amount of disfluencies in "open" speech acts.

In speech recognition, researchers have investigated techniques to minimise the impact of non-speech events on the performance of the speech recognition systems. R.C. Rose [9] used careful manual labelling of some disfluencies and background events as an additional level of supervision for the training of the acoustic models and the statistical language models. Honal et al. [7] experimented a noisy-channel model which automatically corrected disfluencies using manually transcribed spontaneously spoken speech. Hutchinson et al. [10] presented a similar approach by removing fillers and reparanda and thus transforming utterances into fluent ones. Baron [4] and Shriberg [11] used prosodic parameters to locate disfluencies in spontaneous speech in English. In both studies, prosodic parameters proved to befor disfluency detection.

The aim of the present study is to describe speech disfluencies in French according to their location and prosodic features. This study tries to answer the question whether the prosodic parameters of disfluencies are typical enough to characterise them in a consistent manner. The disfluencies studied here were silent and filled pauses, word repetitions and speech repairs. In order to detect regularities in their prosodic parameters, the values of selected phone durations and the F0 slopes were statistically analysed.

2. Overview

2.1. Speech Data Base Used

The speech data base which was used here consisted of more than 1080 telephone messages in French, left by clients during a survey dedicated to the analysis of clients' satisfaction for phone services. The data base contained 55180 words, i.e. on average each message contained 54 words. The data base was manually transcribed including the annotation of non-speech noises such as inspirations, laughter, background noises… as well as interrupted words, interrupted sentences and filled pauses. This orthographic transcription was used to automatically align the words and their phonetic transcriptions with the speech signal. Thus, prosodic events became accessible such as silent pause occurrences, silent and filled pause durations, F0 values and vowel durations as juncture cues.

2.2. Investigated prosodic parameters

In order to characterize the disfluencies at the prosodic level, F0 patterns and vowel durations were investigated. The vowel durations and the F0 values were measured on the last vowel of the part of the speech signal under consideration. The last vowel from which prosodic parameters were extracted was different from the final schwa like vowel, which can occur in French after each uttered consonant event when not present in the spelling form. The parameters of the schwa vowel were used exclusively when it was the only vowel in the word.

An F0 slope was calculated for each vowel under consideration as the difference between the F0 measured at the end and at the beginning of the vowel, normalized according to the vowel length. In order to facilitate the comparison of the F0 slopes, they were grouped into 5 categories according to their value: flat slope (ranging from -0.3 to +0.3), mid-high slope (ranges from +0.3 to +1.5), high-high slope (higher than 1.5), mid-low slope (between -0.3 and -1.5) and finally low-low slope (lower than -1.5).

To facilitate the vowel length statistical analysis, histograms were calculated grouping the vowel durations into 50 ms intervals.

Silent pauses were categorized according to their duration into three categories: short pauses (shorter or equal to 150 ms), long pauses (longer than 250 ms) and mid-long pauses (situated between short and long pauses).

The measures of the prosodic values (phone durations and F0 slopes) were carried out on the last syllable of the word as the major prosodic cues of the syntactic structuring in French are

located on the last syllable of the word. Martin [9] showed that low and high F0 slopes alternate on the syntactic junctures in the reading style. As far as the vowel duration is concerned, according to Bartkova [5] they are lengthened on the syntactic boundaries as a function of the boundary depth and the right consonantal context. However, the length of the last vowel, located on the syntactic junctures, can remain relatively independent from the other phone durations of the same prosodic group.

3. Disfluency analysis

As mentioned above, the speech disfluencies analyzed in this study were word repetitions, filled and silent pauses and word repairs.

3.1. Word repetition

Word repetitions are frequently encountered in spontaneous speech. Although word repetition was not explicitly labelled in the data base, it was easily retrieved automatically. A word repetition is most of the time the repetition of only **one word 73 %** (688 cases), a sequence of **2 words 23 %** (156 cases) and a sequence of **3 words 4 %** (25 cases) (3 words repetitions are the longest ones found in our data). A same word can be repeated several times in a row (up to 5 times in the corpus studied here). Some word repetitions (especially repetitions of adverbs) can be used for stylistic purposes, to highlight the meaning of the repeated word. The adverb repetitions were kept in our statistics only when separated by long pauses for in such cases the successive adverbs could be considered as being interrupted by a hesitation and were perceived more as a repetition than a stylistic use. Speakers hardly ever repeat lexical words. Repetition mostly concerns grammatical words such as prepositions, articles, auxiliary verbs, pronouns, conjunctions...

One word repetitions

79% of one word repetitions were repetitions of a grammatical word. The most frequently repeated word (18% of all one word repetitions) was the article "**de**". 19% of word repetitions contained adverbs and only 3% contained lexical words.

Two word repetitions

Like in the case of one word repetitions, structures containing exclusively grammatical words such as articles, pronouns, auxiliary verbs, prepositions and conjunctions, constituted the majority, that is 65% of cases. 17% of repeated structures contained adverbs and 18% of cases contained lexical words (besides grammatical words).

Three word repetitions

The number of three word repetitions was very low in our corpus (only 25 repetitions contain 3 words). From these repetitions 32 % contained lexical words. In most of the cases (80%) some pauses (filled or silent) occurred between the constituents of the repetition.

3.1.1. Prosodic parameters in word repetition

The prosodic parameters investigated in the word repetitions were F0 patterns, vowel durations as well as filled and silent pause locations. The vowel durations and the F0 slopes were measured on the last vowels of the repeated words or of the repeated word groups. This way, for example, when a repetition contained 3 words, it was then the last vowel of the third word in both parts of the word groups, which was considered.

F0 pattern

As illustrated in Figure 1, most of the F0 patterns were flat on both last words of the repeated sections. The F0 movement was slightly smaller on the first word than on the second one on which the number of low-low and high-high movements were higher than on the first word. In French the syntactic junctures are marked with an F0 movement containing a clearly downward (at sentence final position) or upward (at non-final position) patterns. However, speech repetition contains a different pattern with a very slight (flat) or a moderate (mid-low) movement of the F0. Such a flat F0 slope is a prosodic cue of an unfinished speech sequence.

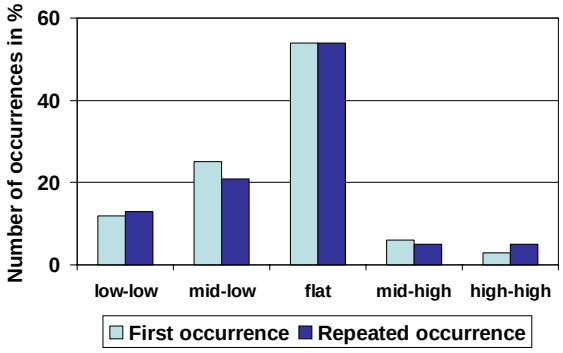


Figure 1: F0 patterns in word repetitions. Number of occurrences as %

Vowel duration in the last syllable

Vowel durations were measured only on the last syllable since the lengthening of the last syllable was one of the prosodic cues of the prosodic boundary. A long final word syllable, accompanied with a clear F0 movement, signaled the presence of a prosodic parsing.

As illustrated in Figure 2, in word repetitions, the first word occurrence contained a stronger vowel lengthening than the repeated word. The hesitation, when present in word repetitions, was expressed by a longer last syllable (and a flat F0 pattern). On the other hand, almost 80% of all the vowel lengths measured on the last part of the repetition sequence were shorter or equal to 100 ms. A short vowel length in this position clearly indicated the absence of a prosodic boundary.

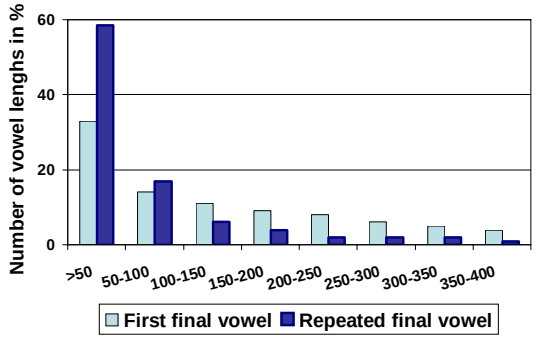


Figure 2: Vowel duration in one word repetitions

Filled pauses in word repetition

In word repetitions the use of filled pauses concerned only 17% of messages which corresponded altogether to 179 filled pauses. Of these filled pauses 26% were placed before the repeated words, only 9% after the repeated words and the remaining 65% between the words of the repetition. Multiple filled pauses in a repetition rarely occurred (only 17 cases that is 9%).

A filled pause can be stuck to the previous word (26%) stuck to the following word (33%) stuck to the previous and following words (19%) but most of the time (75% of the cases) filled pauses were separated from the surrounding words (previous or following or both) by silent pauses.

Filled pauses when non-separated by a silent pause from the surrounding words, followed the final consonants but also the final vowels of the previous word. When a consonant ended a word, the filled pause was perceived as a lengthened schwa vowel and when a vowel ended a word, its timber was "neutralized" into a schwa one. The mean value of the filled pause length in speech repetition was 355 ms with a large standard deviation (251 ms).

Silent pauses in word repetition

70% of word repetitions contained silent pauses. The silent pause mean value was 190 ms and its standard deviation was very large (289 ms). Word repetitions containing silent pauses started with a silent pause in 24% of cases; they ended with a silent pause in 21% of cases and contained silent pauses inside the repetition in 55% of cases. 44% of the word repetitions contained more than one silent pause. 60% of all pauses present in word repetitions were short pauses. The amount of intermediate (mid) pauses reached 26% and of long pauses 14%. The number of pauses was correlated with the number of repeated words: the higher the number of words in a repetition was and the higher the number of pauses encountered in it. In three word repetitions 80% of all the pauses occurred between the words of the repeated sequences. In our corpus the highest number of silent pauses encountered in three words repetition was 7.

3.2. Pauses

Speech can be perceived as highly disfluent even when no "loud" disfluency cues are present in it. That happens when the speech signal contains a high number of pauses, especially when the pause locations are not congruent with the syntactic parsing. Pauses are vital in speech production as the speaker must breathe but they are also necessary for the listener for they provide time to decode the speech stream. In fluent speech pauses are situated on syntactic boundaries signalling the syntactic parsing of the speech or are used for stylistic purposes to enhance the word meaning. However, from the data studied, it appeared that pause occurrences in the spontaneous speech did not always follow syntactic parsing.

3.2.1. Silent pauses

12719 speech internal pauses were detected in our data. On average a pause followed every 4th word.

As previously explained, pauses were grouped into 3 categories: short, intermediate (mid) and long pauses. The quantity of **short pauses** (shorter or equal than 150 ms) was **54%** (6920), **21%** (2684) of all the pauses detected were pauses having an **intermediate length** (from 150 to 250 ms) and **24%** (3115) of all speech internal pauses were **long pauses**, longer than 250 ms.

Figure 3 illustrates the number of pauses as a function of the number of words preceding the pause. It appeared for all three categories of pauses that their location after one single word accounted for the highest number of occurrences. Such occurrences of silent pauses suggested that speakers used them as a talk preparation gap and they were not necessarily situated on syntactic boundaries. This was confirmed by F0 pattern values measured before the pauses and reported in Figure 3. As previously recalled, in French the F0 pattern on prosodic boundaries has a clearly downward or upward movement. Yet a great amount of F0 slopes, measured at the vicinity of pauses, is a flat one (with a very slight F0 movement). Therefore it can be supposed that pauses preceded by a low number of words and by a last syllable having flat F0, occur where hesitation is present.

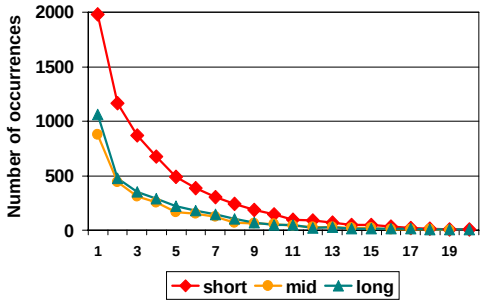


Figure 3: Number of pauses as a function of the number of preceding words

As seen in Figure 4, there were less short pauses when the F0 movement had an upward direction, nonetheless they were slightly more frequent when the F0 movement had a downwards direction. This was due to the fact that a clearly upward movement of the F0 encoded a prosodic parsing which required most of the time long or intermediate pauses.

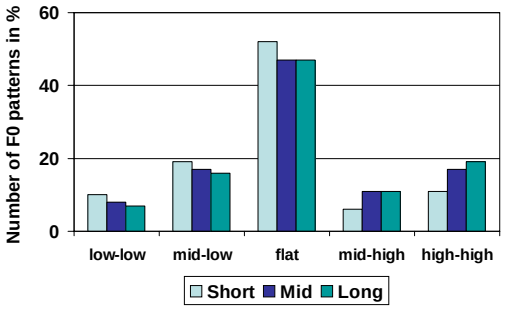


Figure 4: F0 patterns as a function of the following silence pause length.

Figure 5 represents mean vowel lengths as a function of the following pause and the F0 pattern. There was a positive correlation between the pause length and the vowel duration measured in the syllable before the pause. That means that the vowel length was longer when the word was followed by long pauses than when it was followed by intermediate or short pauses. However, when the F0 pattern was high, then the vowel duration was lengthened no matter how long the following pause was: in this place an occurrence of a prosodic boundary could be supposed.

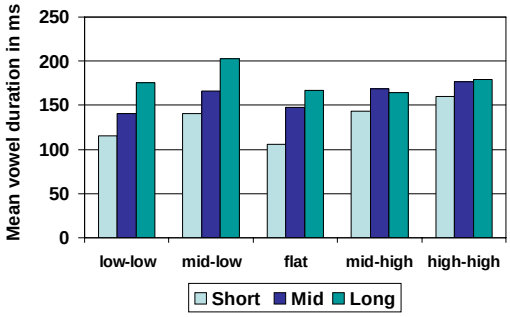


Figure 5: Vowel length when followed by a pause as a function of the F0 pattern and the silent pause length.

In spontaneous speech, even syntactically strongly related words can be separated by pauses. In fact, in our corpus 24% of all pauses occurred between words with strong syntactic links such as pronouns + verbs (je suis – I am) or articles and nouns (la chaise – the chair). Among these pauses 70% were short pauses. In these cases the dominant F0 pattern (60% of cases) was a flat one. Nevertheless the downward pattern was also rather frequent, with 19% of mid-low patterns and 11% of low-low patterns. This was probably due to the fact that grammatical words in French are often uttered with falling F0. One can speculate about the maintenance of this typical F0 pattern of function words despite pause occurrences.

3.2.2. Filled pauses

The number of filled pauses in our corpus was 2871. The filled pauses could be separated from the surrounding words by a silent pause (in 32% of cases). They could be attached to the preceding word (in 68%) as a very long schwa like vowel uttered after a final consonant. They could also be perceived as a neutralized part of the preceding vowel timber when attached to a previous last vowel of a word.

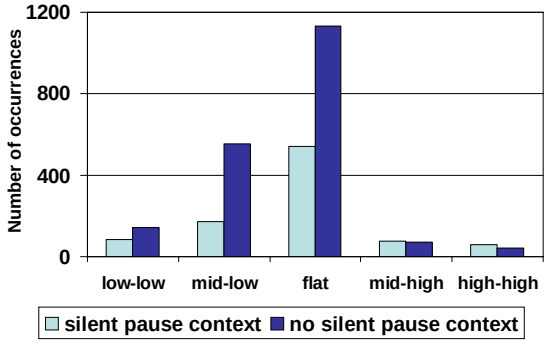


Figure 6: F0 pattern in filled pauses.

The prosodic cues of the filled pauses were very robust. The main cue was its duration (the mean duration was equal to 350 ms) and its F0 movement which was mainly flat. Figure 6 illustrates the prosodic cues of filled pauses followed or not by silent pauses. Although a lot of their F0 patterns were flat, the downward F0 pattern, especially a moderate one (mid-low), was also quite frequent.

3.3. Speech repairs

Speech repair can be a false start containing an interrupted pronunciation of a word or a correction of a wrong word. A simple correction is used mainly for grammatical words when a word has a counterpart considered as the correct one in a given context, such as articles which can be definite or indefinite or masculine or feminine. A lexical word is seldom corrected by adding simply the right word. In fact it is either interrupted or it is corrected by lexical means: such as "la chaise, pardon, la table" (the chair / I mean the table).

Only a very low number of word pronunciations (204) were interrupted in our data base. Speech repairs carried out as a simple word repetition were used only for grammatical words and their number was low (180 cases found in the data). When a speech repair occurred then most of the time (80% of the cases in our data) it was separated by a silent pause. These pauses were in 40% of cases long pauses, in 36% of cases short pauses and in 23% of cases intermediate pauses.

3.4. Other spontaneous speech events

Beside the spontaneous speech events previously discussed in this study, there were others that could be characterized by their prosodic parameters. For example, in spontaneous French, the word "**quoi**" (what) is frequently used by speakers as a loud full stop. In fact, this word had no other role here that to bring down the F0 when its value was too high for a sentence final position. Their prosodic characteristics were therefore very typical. Its prosodic characteristic was a falling F0 (in 80% of cases studied here) and short vowel duration (in 75% of cases shorter than 100 ms).

4. Tests

Prosodic parameters should be used to confirm or refute the solution of the recognition system and to signal disfluency cues (hesitation, filled pauses…). As a preliminary experiment, an assessment of the observations discussed above was carried out on the same corpus and under the same conditions as the statistical study: the phonetic forms of the words were aligned with the speech signal and no speech recognition was carried out.

In order to test how efficient prosodic parameters could be in detecting disfluencies in spontaneous speech, prosodic rules were formulated and their impact tested. The formulation of the prosodic rules was based on the observations of normalized histograms representing F0 slope values and the vowel duration values (see Fig. 7). The slope categories were the same as in the previous part of the study (LowLow, MidLow, Flat, MidHigh, HighHigh) but their values were normalised separately on each recording.

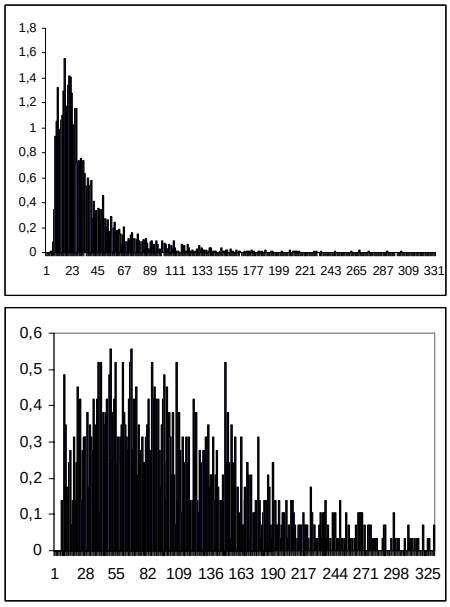


Figure 7: Normalized histograms for word last vowel durations in ms (above) and filled pause durations in ms (under) when the slope is flat.

Tests were conducted to verify how general simple expert-defined rules were and how efficiently they could detect disfluencies in speech. Prosodic rules were used to confirm or not filled pauses. Moreover a flag indicating a hesitation was positioned when the values of the prosodic parameters were considered as conveying a hesitation A hesitation was automatically detected as present when the final syllable was very long (followed or not by a pause) and when it was very short and followed by a pause. In the first case the hesitation was expressed by the syllable length and in the second by the

occurrence of the pause. The correctness of the hesitation flag positions was checked manually, by listening to the speech signal.

The following evaluations were carried out:

Filled pauses: 60% of filled pauses present in the alignments were validated as filled pauses using prosodic parameters about. When the occurrence of a silent pause preceding the filled pause was taken into account, the validation of filled pauses reached 72%.

Hesitation: 33% of the last syllables of content words and 11% of function words were detected as conveying a hesitation.

The accuracy of the hesitation flag positions was verified manually on 10% of our data. It was found that 83% of flags were correctly set. As the transcription of the corpus did not contain information about hesitation (other than filled pauses) therefore it was impossible to estimate the number of missed hesitations.

Word repetitions: No typical prosodic cues were observed in this study for word repetitions. The only prosodic cue detected was hesitation conveyed most of the time by the last syllable of the fist part of the repeated word sequence.

This preliminary evaluation was very promising. However further work is to be conducted using speech recognition results and also other types of data such as data from man-machine speech driven "dialogues". On the other hand appropriate modelling procedure is to be used in order to refine the disfluency decision thresholds.

5. Discussion

The present study analyses the prosodic cues of disfluencies in French spontaneous speech. A very long vowel duration in a last syllable of a word is a reliable disfluency cue. However a flat F0 slope and a short vowel duration in a last syllable of a word followed by a silent or a filled pause can also be considered as a strong prosodic indicator of disfluencies. A small number of words separated by pauses can be another indicator of places in speech where hesitations or unease occur. The filled pauses are characterized in prosodic terms with a long duration and a flat or a slightly downwards F0 movement.

When expert defined simple prosodic rules are used to test how efficient prosodic parameters in disfluency detection are, it appears that they are very reliable in filled pause detection and efficient in hesitation detection which is often present in speech repetitions. Thus, prosodic parameters can be used in speech recognition to yield confidence indication about the occurrence of disfluencies.

6. References

[1] Asp Annika & Decker Anna 2001. Designing with speech acts to elude disfluency in human-comuter dialogue systems, Working Papers 49, 2-5 Lund University.

[2] Arnold Jennifer E., Fagnano Maria & Tanenhaus Michael Disfluencies Signal Theee, Um, New Information, Journal of Psycholinghistic Research, Vol. 32, No. 1, January 2003

[3] Bailey, Karll G.D., & Ferreira, Fernanda. 2001. Do Non-Word disfluencies Affect Syntactif Parsing?, DISS'01, Edinburgh, Scotland, UK pp. 61-64

[4] Baron Don, Shriberg Elizabeth & Stolcke Andreas 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues, Icslp 2002.

[5] Bartkova, Katarina & Sorin Cristel, 1987. A model of segmental duration for speech synthesis in French, Speech Communication 6, pp 245-260.

[6] Bortfeld Heather, Leon Silvia D., Bloom Jonathan E., Schober Michael F. & Brennan Susan E. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender, Language and speech, 2001, 44 (2), pp. 123-147.

[7] Honal Mathais & Schultz, Tanja 2003. Correction of Disfluencies in spontaneous speech using a noisy-channel approach, EuroSpeech 2003;

[8] Hutchinson Ben & Pereira Cécile 2001. Um, One Large Pizza. A Preliminary Study of Disfluency Modelling for Improving ASR, DISS'01, Edinburgh, Scotland, UK, pp 77-80;

[9] Martin Philippe 1981, Pour une théorie de l'intonation, *L'intonation de l'acoustique à la sémantique*, Klincksieck Paris pp. 234-271.

[10] Rose R.C & Riccardi G.1999. Modeling Disfluency and Background events in ASR for a natural language understanding task, ICASSP 1999;

[11] Shriberg Elizabeth, Batte Rebecca & Stlcke Andreas 1997, A prosody-only decision-tree model for disfluency detection, Eursopeech 1997.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 27-32.

A quantitative study of disfluencies in French broadcast interviews

Philippe Boula de Mareüil, Benoît Habert, Frédérique Bénard, Martine Adda-Decker, Claude Barras, Gilles Adda, Patrick Paroubek

LIMSI-CNRS, Orsay, France

Abstract

The reported study aims at increasing our understanding of spontaneous speech-related phenomena from sibling corpora of speech and orthographic transcriptions at various levels of elaboration. It makes use of 9 hours of French broadcast interview archives, involving 10 journalists and 10 personalities from political or civil society. First we considered press-oriented transcripts, where most of the so-called disfluencies are discarded. They were then aligned with automatic transcripts, by using the LIMSI speech recogniser. This facilitated the production of exact transcripts, where all audible phenomena in non-overlapping speech segments were transcribed manually. Four types of disfluencies were distinguished: discourse markers, filled pauses, repetitions and revisions, each of which accounts for about 2% of the corpus (8% in total). They were analysed by utterance", speaker and disfluency pattern types. Four question were raised. Where do disfluencies occur in the utterance? What is the influence of the speakers' status? And what are the most frequent disfuency patterns?

1. Introduction

This empirical study aims at expanding our knowledge of spontaneous speech-related phenomena from "sibling" resources of audio and written documents, such as available in TV archives or for parliament debates: instances of close, *bona fide* or trustworthy transcriptions, which are used for quotations, and even for legal purposes. The same content being both spoken and written, the comparison between either means of communication may then contribute to improve the modelling of so-called disfluencies (filled pauses, repetitions, etc.) in automatic speech recognition — to make transcriptions readily readable — and in speech processing, in applications like subtitling. Note that along this article we use the default term "disfluency" for the sake of convenience, regardless of its negative connotation which we do not assume. The mainstream terminology is questionable: if some phenomena may be described as production errors, as a "pollution" of the signal for automatic speech processing, others may help conceptualisation and contribute to fluency. Nonetheless, we dared not write "(dis)fluency" out of respect for the workshop title.

Written language and spoken language differ in many respects [6; 4; 9; 12; 3]. First, written language has a vocation for being persistent — it remains on a medium, unlike spoken words, which spring out in an ephemeral way and fly away. It is not tied to the same physiological constraints as spoken language, which uses the same organs as breathing. Whereas a lapse of time separates reading from writing, the simultaneity of spoken communication allows untimely interruptions and overlaps in the speech flow. In the former process, which is threefold, one can distinguish the writing activity (the dynamic nature of which may be traced in drafts), the written document (the result validated by the writer, which may also be anonymous) and the reading mechanism (a decoding step). In everyday conversation as in more supervised interviews, which are typically face to face, word tuning, talking and listening are synchronous: hence hesitations, repetitions and overlapped speech.

Some lexical and syntactic uses are well-established characteristics of written French: e.g. *car* ("for", "because"), inversion of verbs and subject pronouns in questions, while the drop of the *ne* in the discontinuous negation *ne… pas* is reminiscent of spoken language. These are only scattered examples; wide-coverage usage-based studies are better suited. Spoken corpora have known a considerable and unprecedented development for over a decade. The existence of large parallel spoken/written corpora now offers new prospects to answer the following question, which is crucial in linguistics: what characterises "spoken style" and "written style"? More generally, what is a "style"? — the "movement of the soul" according to Cicero, the "face of the soul" according to Seneca [5]. Labov [8] distinguishes between casual speech (in ordinary conversations) and careful speech (in an interview situation). Additional degrees of more or less colloquial speech could be considered. Yet, it is unnecessary here: our study definitely deals with careful speech, since it is based on TV shows in which journalists ask questions which politicians or representatives of civil society are bound to answer. We are thus faced with dialogues in which the interlocutors' roles are clearly established and accepted by both parts. They are at least partially prepared and they are public, since they are broadcast. The interaction is asymmetrical, since it is guided, and the situation is rather formal.

We have been interested in a corpus composed of audio files, and press-oriented transcripts, provided by the French Institut National de l'Audiovisuel (INA). We then added precise orthographic transcripts. The information that speech bears is incomparably richer than the one conveyed by the corresponding transcript. In particular, the quite limited typographic means we have at our disposal (punctuation, expressive capitalisation, orthographic stretching) can hardly express attitudes and emotional states. More generally, prosody and voice quality are badly or not at all indicated by typography. But other speech phenomena can be transcribed orthographically: filled pauses, splutters, slips of the tongue, and self-repairs (revisions), repetitions (of function words especially) and all these "little words" typical of spontaneous speech (discourse markers) such as *enfin, bon, ben, eh bien* ("well"), *donc, alors* ("so"), etc. In our corpus, they were reported and labelled in precise transcriptions, but are often missing in press-oriented transcriptions, which tend to render the message linear. This allows us to measure the distance between the two, which is the main goal of this quantitative study. Where do disfluencies occur in the "utterance" (intuitively defined by non-linguist transcribers)? What is the impact of the speakers' status? And what are the most frequent disfuency patterns? These are questions we will attempt to answer.

2. Corpus and transcription guidelines

This study makes use of 9 hours of *L'Heure de Vérité* ("The Hour of Truth"), a French TV show recorded a dozen years ago. In each one-hour show, a major personality from either political or civil society (e.g. charities) is interviewed by at most 3 journalists and a chairman, who is the same in the 9 shows. The journalists prepare their questions (most of them are to be expected), and the answers are not casual speech (some of them are "caned" answers, prepared answers to obvious questions). On the other hand, the chairman who leads the debates, makes sure that beforehand determined topics are stuck to and watches over the schedule, often interrupts the interviewee and the current interviewer. This configuration favours disfluencies, and speech overlaps are frequent. Only part of the numerous disfluencies reveals information about the planning problem of the speaker; the rest corresponds to a "struggle for speech" between interlocutors, even though journalists do not "jump in" haphazardly [16; 14].

For each show, we have both the audio and a press-oriented transcript (TPress). The latter is intended to be rather close to the audio while keeping to implicit conventions: it lies somewhere in between written text and exact transcript. As a matter of fact, most disfluencies are discarded. We consequently produced an exact transcription (TExact) for the audio data, with all audible phenomena, and in particular disfluencies. Speech recognition was particularly helpful because it precludes the unconscious filtering of disfluencies. It is often difficult to distinguish hesitations, for instance, from the pronunciation of a final schwa. With the help of the LIMSI system, first in its standard version, then in an "informed" version (i.e. taking TPress into account in the lexicon and the language model), we generated an automatic transcription (TReco) [1]. We took advantage of a modified version of Transcriber (http://sf.net/projects/trans/) to align time-codes and to display coloured mismatch zones between TPress and TReco (about 15% of the archive corpus, which were a priori made up of disfluencies) [2]. The coupling of the "informed" TReco and the *bona fide* TPress can then be regarded as a transcription draft.

In order to label disfluencies, we followed the LDC (http://www.ldc.upenn.edu/Projects/MDE/,) metadata annotation guidelines, adopted in the Rich Transcription evaluations conducted by NIST [11]. We chose these conventions because they fit some of our purposes (i.e. providing readable transcriptions), and represent the result of a vast discussion. LDC metadata annotations cover fillers (filled pauses, discourse markers, explicit editing terms, asides and parentheticals), edit disfluencies (repetition, revisions, restarts and complex disfluencies), and sentence-like units (statement, question, backchannel and incomplete sentence).

With some adaptations to French and simplifications, we distinguished and annotated filled pauses (FP), discourse markers (DM), repetitions (RP) and revisions (RV). Transcribed as *euh*, FPs were labelled automatically.

DMs may have either a simple filler role or a real discourse structuring function. According to the Geneva school terminology of discourse linguistics, DMs may be consecutive (e.g. *alors*, *donc* "so"), counter-argumentative (e.g. *mais* "but") or re-evaluative (e.g. *enfin* "well") [15]. We are aware that discourse markers may have several functions and mean different things; they do not have exactly the same disfluency status as filled pauses. But it is not always straightforward to interpret their precise role. The conjunction *et* ("and"), in particular, may also be used to structure the dialogue, to begin speaking, to avoid a stigmatised *euh*, to link two utterances or to prevent from being interrupted. The same happens with idioms such as *je crois que* ("I believe that"), which may be mere habits or verbal tics for some speakers, and which are difficult to consistently annotate.

RPs cover:

- repetitions of words (possibly truncated and/or interrupted by another speaker), where the left-most term(s) are marked up (e.g. *(RV le) le* "the the");
- emphatic repetitions, strengthening a statement;
- discontinuous repetitions (after parentheticals).

Note that only really "disfluent" repetitions were considered in the LDC metadata annotation guidelines, excluding emphatic or distant repetitions.

Finally, RVs involve word fragments, words or short chunks that are abandoned, without necessarily being corrected. Unfinished sentences, resulting from an interruption by another speaker, do not fall into this category. Nevertheless, albeit rare, complex cases exist, where RVs can include DMs, which in turn can include RPs and FPs. Disfluencies are particularly numerous at the borderline of overlapped speech sequences and within them. Only "clean" speech was so far systematically marked, because overlapped speech is quite difficult to handle. After discarding 24 minutes of overlapped speech, we have an amount of 7:18 of speech (88,056 words): 30 minutes by interviewee, 51 minutes for the chairman, 25 minutes for 3 recurrent journalists, 5 minutes for the other 6 journalists.

For disfluency annotation, a customised version of Transcriber was used. It allowed a quick annotation through contextual menus and a coloured display of the various disfluency types, similar to what LDC proposed for their own annotation scheme. The new disfluency annotation tags were embedded into the initial XML transcription files.

- Example of annotation:
  (RP ça veut dire) ça veut dire , par exemple , que quand on gagne le SMIC , **(DM eh ben)** bien évidemment non , on perdra pas *(RP de de)* de revenus parce qu' on peut pas , (FP euh) quand on gagne le SMIC , ***(RV perdre un)*** perdre son revenu.
- Press-oriented counterpart:
  Ça veut dire, par exemple, que quand on gagne le SMIC, bien évidemment non, on ne perdra pas de revenus parce que l'on ne peut pas, quand on gagne le SMIC, perdre son revenu.
- English translation:
  It means, for example, that when you earn minimum wage, of course not, you won't loose incomes because you can't, when you earn minimum wage.

3. Results

Our annotation enabled us to classify the words involved in disfluencies into DMs (2.5%), FPs (1.9%), RPs (2.3%) and RVs (2.2%). The proportions, computed with respect to the total number of words in the corpus, are relatively well balanced. It turns out in Table 1 that interviewers produce more filled pauses and repetitions, while interviewees produce more discourse markers and revisions. A test of comparison of two proportions reveals that each difference is significant with $\alpha = 0.05$:

$(p_1 - p_2) / \sqrt{p(1-p)(1/n_1 + 1/n_2)}$ > 1.96 in absolute value

where

n_1 is the number of words uttered by journalists,
n_2 is the number of words uttered by interviewees,
p_1 is the proportion of disfluent words uttered by journalists,
p_2 is the proportion of disfluent words uttered by interviewees,
$p = (n_1 p_1 + n_2 p_2) / (n_1 + n_2)$.

This difference may be due to the difficulties journalists meet, when they try to interrupt their interlocutor, while interviewees try to build a real argument.

3.1. Overall distribution

The occurrences of disfluencies can be studied as a function of the "utterance" length, the speaker's status (role, authoritativeness) and their context-dependency (whether some sequences of disfluencies are more prone to appear together, in a given order).

Table 1: Interviewees' and interviewers' disfluencies (DM = discourse marker; FP = filled pause; RP = repetition; RV = revision).

Speaker	words	%DM	%FP	%RP	%RV	%dis.
Brauman	8,174	1.5	1.2	2.2	3.6	8.4
de Robien	7,589	4.7	1.8	1.0	1.9	9.4
Delors	7,462	3.2	0.6	3.1	3.4	10.4
Voynet	7,177	4.0	2.5	1.8	1.7	10.0
Pasqua	5,385	1.4	0.9	1.6	1.5	5.4
Diouf	4,809	0.4	0.8	1.9	2.2	5.6
Brittan	4,806	4.4	4.3	5.8	3.5	18.0
Pinay	4,006	1.6	0.7	3.3	3.3	8.8
Chevènement	3,842	3.8	2.8	1.4	0.9	8.9
Lamassourre	2,729	0.6	1.1	0.9	0.6	3.2
Total inter-viewees	55,979	**2.8**	**1.6**	**2.1**	**2.2**	**9.1**
de Virieu	10,184	1.6	2.3	1.8	1.2	7.0
Duhamel	7,175	1.8	2.2	2.8	1.2	8.0
Colombani	4,818	2.2	1.3	2.3	2.9	8.7
du Roy	3,706	1.3	4.3	2.5	2.7	10.8
Diop	1,904	1.8	2.0	1.9	1.4	7.5
Tesson	1,270	2.4	2.0	2.5	5.8	12.7
Giesbert	886	5.6	2.2	4.6	3.1	16.0
Laffon	809	1.9	2.6	2.2	0.6	7.3
d'Orcival	743	3.2	1.1	2.2	2.4	8.9
English	622	3.7	3.5	4.2	2.3	13.7
Total inter-viewers	32,117	**2.0**	**2.4**	**2.4**	**1.9**	**8.7**
Total	88,056	**2.5**	**1.9**	**2.3**	**2.2**	**8.9**

For almost all speakers, the longer the utterance, the lower the percentage of disfluent words, as is apparent in Figure 1, where average rates for "utterances" of less than 12 words and more than 16 words are displayed. This observation is most likely due to the speech communication situation. As established by [17], disfluencies occur at the beginning rather than at the end of utterances (see Figure 2). We interpret this fact by a higher difficulty to start a rather than to continue a formulation.

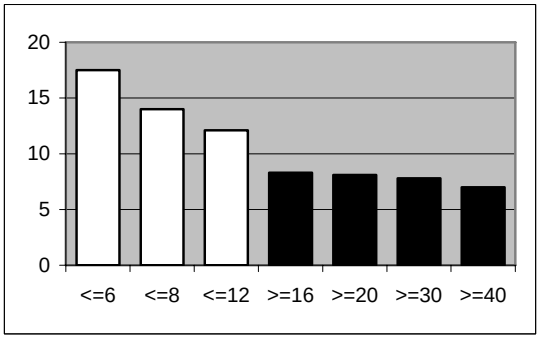


Figure 1: Percentage of disfluent words as a function of the "utterance" length (in words).

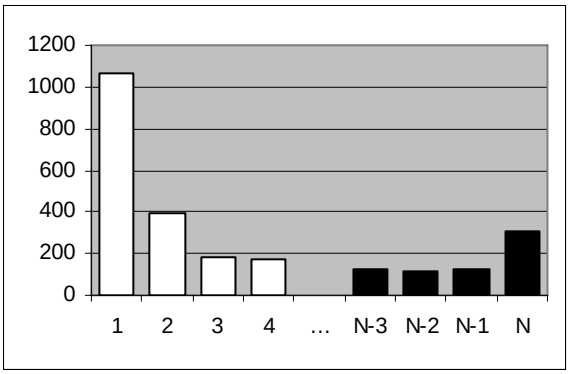


Figure 2: Distribution of disfluencies as a function of their position in the "utterance", from 1 (1st word) to N (last word).

In addition to the disfluency location, the content and context of appearance of DMs, FPs, RVs and RPs were investigated. The 1,568 FP occurrences all correspond to *euh* or its variants. As far as the other types of disfluencies are concerned (DMs, RPs and RVs), their distribution in terms of lexical items or idioms seems to follow Zipf's law (see Figure 3). Further details will be presented in the next subsections.

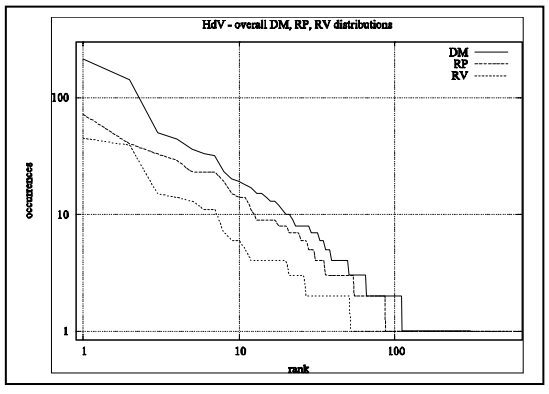


Figure 3: Zipf's distribution of lexical items or idioms involved in the DM, RP and RV categories.

3.2. Discourse markers

In DMs, which represent more than a quarter of all disfluencies, we find expected conjunctions, adverbs and

interjections (see Table 2): *et* ("and"), *alors* ("so"), etc. Either the former or the latter is the most frequent or the second most frequent DM for each speaker; but the other item in the leading pair is quite variable. For instance, the conjunction *mais* ("but") is the 4th most frequent DM, while it never appears in the leading pair of any speaker; and inversely, the phrase *je crois que* ("I believe that") is not so frequent, but appears in the heading pair of two speakers (see Tables 3 and 5). Interestingly, these two speakers are interviewees: not only are DMs speaker-specific, they also depend on the interviewer/interviewee position. Journalists are more inclined to use impersonal fillers (e.g. the interjection *hein* for Virieu). As for the interviewees who produce many DMs, they resort to a wide range of different expressions.

Table 2: DM morphosyntactic classes — *donc* ("therefore") is counted among conjunctions even though its distributional behaviour distinguishes it from other traditional conjunctions.

Category	%DM	Example
"conjunction"	27	*et, mais, donc*
adverb	25	*alors, enfin, d'ailleurs*
complex	17	*oui eh bien écoutez*
verb "phrase"	17	*je crois que*
interjection	12	*eh bien, ben, hein, bon*
pronoun	2	*moi*

To sum it up, we can distinguish three broad types of discourse markers: structuring (e.g. *alors*), position (e.g. *je crois que*) and interaction (e.g. *hein*). Each subtype represents about one third of all DMs, even though the latter — which shows that we are answering an interlocutor, that we try to convince him or that we agree with him — are somewhat fewer. As for the position subtype, its overuse by interviewees is particularly obvious with a speaker like Voynet.

Table 3: The 2 most frequent DMs for each speaker and their ratio within the DM class for the speaker.

Interviewees			Interviewers		
Brauman	*ben, et*	37%	Colombani	*alors, et*	33%
Brittan	*et, je crois que*	66%	de Virieu	*alors, hein*	39%
Chevènement	*et, hein*	29%	Diop	*alors, donc*	64%
Delors	*et, je pense que*	24%	Duhamel	*et, alors*	47%
de Robien	*et, eh bien*	49%	du Roy	*alors, et*	61%
Diouf	*et, moi*	50%	Tesson	*et, moi*	37%
Lamassourre	*et, je crois*	38%	English	*alors, et*	33%
Pasqua	*et, alors*	33%	Giesbert	*alors, bon*	43%
Pinay	*et, moi*	35%	Langelier	*et, alors*	62%
Voynet	*je crois que, alors*	45%	d'Orcival	*alors, et*	50%

3.3. Filled pauses

FPs can be found almost anywhere. More precisely, 35% of FPs occur at a sentence boundary indicated by a full stop (14%) or at a major phrase boundary indicated by a comma (21%), with respect to the TPress punctuation. For the remaining 957 FPs, Table 4 gives the distribution of the most frequent left and right contexts, considered independently. Even in the middle of a sentence, FPs frequently precede a determiner or a preposition; they rather follow a conjunction or a preposition. This asymmetry suggests that filled pauses (at least transcribed as *euh*) are avoided within noun phrases, especially between a determiner and a noun. In this situation, other mechanisms such as final lengthening or repetitions are preferred.

Table 4: FPs' most frequent left and right contexts; RVs' most frequent right contexts.

FP Left context		FP Right context		RV right context	
word	# (%)	word	# (%)	word	# (%)
que	40 (4.2)	*de*	53 (5.5)	*d'*	34 (4.7)
et	27 (2.8)	*la*	41 (4.3)	*l'*	30 (4.1)
pour	26 (2.7)	*des*	38 (4.0)	*la*	29 (4.0)
de	21 (2.2)	*les*	33 (3.4)	*vous*	25 (3.4)
avec	19 (2.0)	*l'*	26 (2.7)	*de*	23 (3.2)
à	13 (1.4)	*le*	23 (2.4)	*on*	21 (2.9)
qui	12 (1.3)	*un*	21 (2.2)	*le*	19 (2.6)

3.4. Repetitions and revisions

RPs and RVs exhibit some features in common: first, they both involve 1 or 2 words on average, and there is a high correlation (0.8) among speakers between their numbers of RP and RV occurrences. Speakers who produce many repetitions also tend to make many revisions. Second, if we look at the most frequent RPs and RVs, we can only see monosyllabic function words: *de* ("of", 72 RPs + 45 RVs), *le* ("the/him", 40 RPs + 39 RVs), etc. For all speakers, in the first two places and in the same order, we have very frequent French words. The form *le* is by far more often a determiner than a pronoun, even though nothing prevents a subject pronoun such as *je* ("I") from being one of the most repeated or revised words (see [7]). Most words are shared between RPs and RVs in Table 5, which is not surprising according to the following interpretation: in the process which consists of looking for words, a bootstrap word such as the masculine singular article *le* in French (or *the* pronounced as [i:] in English) may be repeated if it agrees grammatically with what follows, and may be corrected otherwise. The fact that there are more masculine nouns than feminine nouns in French (16k vs. 12k in the BDLEX dictionary [13]) does not seem to be sufficient to explain why *le* outweighs *la* in both RPs and RVs. By contrast, the conjunction *et* ("and") hardly lends itself to revisions, and we only find it among RPs.

Inspection of the right part of Table 4 shows that the most frequent words that follow RV-labeled words are *d'* ("of") and *l'* ("the"): precisely the shortened forms of the most frequently revised words. This means that the most frequent repairs are of the form *de d'*, before a word beginning with a vowel. We then have *la* (more frequent than *le*), which is in keeping with what we have just seen in the previous paragraph. Next, the presence of *vous* ("vous") or *on* ("we") is striking, since these personal pronouns are absent from Table 5: they really represent syntactic breaks, following abandoned phrases. The part-of-speech mismatch between the reperandum and the repair could be an objective criterion to label restarts, which we consider as RV subtypes. Levelt's [10:499] assertion, according to which "speakers tend to preserve the original syntax in the repair", deserves to be quantified.

Table 5: Most frequent words involved in disfluencies (DMs, RPs and RVs) — numbers of occurrences and percentages of the disfluency type they represent.

DM			RP			RV		
word	#	(%)	word	#	(%)	word	#	(%)
et	214	(9.8)	*de*	72	(4.3)	*de*	45	(2.2)
alors	141	(6.5)	*le*	40	(2.4)	*le*	39	(1.9)
je crois que	50	(2.3)	*et*	33	(2.0)	*à*	15	(0.7)
mais	44	(2.0)	*je*	29	(1.7)	*que*	14	(0.7)
donc	36	(1.6)	*un*	23	(1.4)	*la*	13	(0.6)
eh bien	33	(1.5)	*à*	23	(1.4)	*les*	11	(0.5)
hein	32	(1.5)	*les*	23	(1.4)	*je*	11	(0.5)

Content words may also be involved in repetitions and revisions, and are more affected by truncation phenomena than are function words. This is unsurprising, since they are far more often polysyllabic. In our annotation scheme, truncation phenomena are split into RPs and RVs, but they only represent 0.4% of our corpus.

3.5. Disfluency patterns

So far, we have considered what happens within and around disfluency markups. To finish with, let us regard disfluencies as single events. Apart from isolated disfluencies which dominate**,** the most frequent patterns of immediately consecutive disfluency labels are RV FP (53 occurrences), DM FP (47 occurrences), FP RV (46 occurrences), FP RP (45 occurrences) and DM RP (33 occurrences). Once more, a certain asymmetry is noteworthy in their order of appearance, chiefly between DM FP (47 occurrences) and FP DM (21 occurrences). More generally, the patterns DM + other disfluency appear twice as much as the patterns other disfluency + DM (98 vs. 50 occurrences). A possible explanation is that DMs, are often used to start a message (133 occurrences in position 1 vs. 65 occurrences in position 2, in disfluent zones which involve at least two disfluency labels) owing to their structuring and filler role. Once an interruption point is met, discourse markers are less employed in the editing term and the subsequent repair. Longer patterns exist, such as FP DM RP (5 occurrences), even though they are fewer: e.g. nous nous trouvons confrontés `(FP euh)` , **`(DM disons)`** , *`(RP à des)`* à des incohérences ("we are confronted to, let's say, to inconsistencies"). More data is required to extensively address disfluency patterns.

4. Conclusion and future work

Aligning press-oriented and automatic transcriptions diminishes the cost of an exact transcription, and enables the use of natural language processing (NLP) tools. It allowed us to examine a large speech corpus: several hours of French broadcast interviews. Four types of spontaneous speech-specific phenomena were analysed: discourse markers, filled pauses, repetitions and revisions (accounting for 8% of the corpus). They were sorted out by "utterance", speaker and pattern types.

Despite the size of our corpus, the conclusions we draw should be related to its genre, that of broadcast interviews, and would benefit from a comparison with conversational speech. With this end in view, the probabilities of discourse markers such as *je crois que*, *je pense que* ("I think that") were considered and compared to what is obtained in other corpora of fine-grained transcriptions in French — Broadcast News (3.6M words) and Telephone Conversational Speech (1M words). We notice that for interviewees we are close to the value estimated in conversational speech, whereas for journalists we are even below the value estimated in BN.

The corpus *bona fide* and fine-grained transcriptions were enriched with morphosyntactic tags. This information will be used in future work. Also, a prosodic study is planned, on function word final lengthening (phonemes longer than 300 ms) and speech rate in fixed expressions such as *je crois que* ("I believe that"): the latter is indeed quite frequent in almost all interviewees, but it seems to have been labelled as a discourse marker only when it is pronounced quickly.

In the near future, we also plan to study the relationship between disfluencies and turn taking, their position within sentence-like units (SUs) as well as the influence that struggle for speech has on disfluencies. Finally, this type of analysis would arguably improve by being related to the study of eye movements and body gestures, since we have video recordings at our disposal.

5. Acknowledgements

We are indebted to the INA Research and Experimentation Directorate (http://www.ina.fr/) for the *L'Heure de vérité* corpus (audio and video files) and its *bona fide* transcription. INA plays the role of public archive for audio and video resources in France.

6. References

[1] Adda-Decker, Martine *et al*. 2003. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. *Proc. DISS*, 5–8 September 2003, Göteborg, pp. 67–70.

[2] Barras, Claude *et al*. 2004. Automatic audio and manual transcripts alignment, time-code transfer and selection of exact transcripts. *Proc. LREC*, 24–30 May 2004, Lisbon, pp. 877–880.

[3] Biber, Douglas. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge, Cambridge University Press.

[4] Blanche-Benveniste Claire *et al.* 1990. *Le français parlé, études grammaticales*. Paris, Éditions du CNRS.

[5] Fónagy, Ivan. 1983. *La vive voix. Essais de psycho-phonétique*. Paris, Payot.

[6] Hagège, Claude. 1985. *L'homme de parole*. Paris, Fayard.

[7] Henry, Sandrine & Bertille Pallaud. 2003. Word fragments and repeats in spontaneous spoken French. *Proc. DISS*, 5–8 September 2003, Göteborg, pp. 77–80.

[8] Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, University of Pennsylvania Press.

[9] Léon, Pierre. 1993. *Précis de phonostylistique.* Paris, Fernand Nathan.

[10] Levelt, Willem J. M. 1989. *Speaking. From Intention to Articulation.* Cambridge, Massachusetts: MIT Press.

[11] Liu, Yang *et al*. 2005. Structural Metadata Research in the EARS Program, *Proc. IEEE ICASSP*. 18–23 March 2005, Philadelphia, pp. 957–960.

[12] Morel, Marie-Annick & Laurent Danon-Boileau. 1998. *Grammaire de l'intonation. L'exemple du francais*, Paris, Éditions Ophrys.

[13] Pérennou, Guy & Martine de Calmès. 1987. *BDLEX, base de données lexicales du français écrit et parlé*. Toulouse, Travaux du laboratoire CERFIA.

[14] Plauche, Madeleine & Elizabeth Shriberg. 1999. Data-Driven Subclassification of Disfluent Repetitions Based

on Prosodic Features. *Proc. ICPhS*, 1–7 August 1999, San Francisco, vol. 2, pp. 1513–1516.

[15] Roulet, Eddy *et al*. 1991. *L'articulation du discours en français contemporain*. Berne, Peter Lang.

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

[17] Shriberg, Elizabeth. 2001. To "Errrr" is Human: Ecology and Acoustics of Speech Disfluencies. *Journal of the International Phonetic Association* **31**(1), pp. 153–169.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 33-37.

Disfluency phenomena in an apprenticeship corpus

Jean-Leon Bouraoui & Nadine Vigouroux

*IRIT, Toulouse, France

Abstract

This papers presents a study carried out on an apprenticeship corpus. It features dialogues between air traffic controllers in formation and "pseudo-pilots". "Pseudo-pilots" are people (often instructors) that simulate the behavior of real pilots, in real situations.

Its main specificities are the apprenticeship characteristic, and the fact that the production is subordinate to a particular phraseology.

Our study is related to the many kinds of disfluency phenomena that occur in this specific corpus. We define 6 main categories of these phenomena, and take position in regard to the terminology used in literature. We then present the distribution of these categories. It appears that some of the occurrences frequencies largely differs from those observed in other studies. Our explanation is based on the corpus specificity: in reason of their responsibilities, both controllers and pseudo-pilots have to be especially careful to the mistakes they could do, since they could lead to some dramas.

The remainder of our paper is dedicated to the more deepen study of a disfluency class: the "false starts". It consists of the beginning utterance of a word, that is not achieved. We show that this category consists of several sub-categories, of which we study the distribution.

1. Introduction

It's beyond doubt that disfluencies do occur very frequently in everyday conversations. Many studies are devoted to these phenomena, and to their various manifestations. The majority of this studies is carried out on corpus made of everyday life productions.

But one can wonder what would be the manifestations of disfluencies in a corpus made of very specialized and constraint language. This is a very important question, since answering to it would give us very important hints on disfluencies. If one find that they occur in the same way than in everyday language, that proves that some "universal" classes of disfluencies do exist. If it didn't, then it would be very interesting to study the nature of their differences, and their cause. In any case, this entails some additional knowledge on the disfluencies. Beyond the single evident theoretical interest, it also offers a wide range of practical applications, notably in automatic speech recognition and understanding.

The work we present in this paper is precisely devoted to a study carried out on a corpus such as we described above. In the first part, we present in details its characteristics. We also explain the methodology we used to transcribe and annotate it. The second part of this article relates to the precise description of the phenomena we sought, and the results we obtained.

2. Description of corpus

2.1. Characteristics of controllers – pseudo-pilots communication

The formation of the Air Traffic Control (ATC) controllers includes theoretical teachings, but also consists of a lot of training sessions. These sessions are made of communication between air-traffic controllers being formed and "pseudo-pilots operators" (that is, people simulating real pilots).

The aim of the exercises is to train apprentice controller activities, and then to evaluate them. It consists of managing several planes that are in a controlled area, for example by assigning them a given speed and/or position. Two languages are used: French and English (French being the majority); all the speakers are French native speakers. The exercise conditions are as near as possible from real environment: controllers work with screen giving the radar position of virtual "planes"; the air traffic is simulated by several persons assuming the role of one or many pilots. Some background noises (overlapping conversations, sounds emitted by microphones, etc.) also occur.

The utterances produced by the controller, as well as the pilots' ones, must respect a phraseology [1]. It describes, for example, the way the speaker must pronounce the plane call signs, or the order that the different components of a message/an utterance have to follow. Two speakers can't speak at the same time, due to technical limitations: the audio channel is only assigned to one speaker. During the formation step, the phraseology is not always strictly respected (neither in real work conditions), though its general guidelines are kept. However, its learning and mastering is also aimed by exercises.

An instance of a simple order that an air controller can formulate to a pilot is: "D T C climb level 9 0": we find, first, the call sign of the pilot's plane ("D T C"), and then the order itself. More complex utterances can also occur, composed of a sequence of simple orders. For a complete description of the French call signs and orders, see [5].

The use of the phraseology entails that the lexico-syntactic schemas are limited in number. Moreover, due to the restrictive task, and to the apprenticeship property, it is obvious that, from a linguistic point of view, the corpus differs a lot from a more "traditional" one. We hypothesize that this may probably influences the phenomena that appear, in comparison to some less constraint tasks, such as daily conversations or train timetable reservation for instance.

To conclude, it is important to note that the spoken dialogues are actually spontaneous speech. We insist on that point, because the important role played by the phraseology could make one think that all utterances are already planned. It's not true since both controllers and pilots do not know what will happen, and consequently what has to be said. The phraseology only set up a general framework for utterances; what is actually said depends on the dynamic interaction between a given controller and pseudo-pilot.

2.2. Transcription and annotation methodology

We transcribed dialogues as well as annotated them according to some specifications ([3] and [4]). These authors made a distinction between the orthographic transcription and annotation, which corresponds to an interpretation (at semantic, dialogic levels, etc.) of the orthographical string.

Specifications were defined, firstly to determine elements that has to be transcribed, and secondly to obtain homogeneity of transcriptions in case where several annotators processed the tasks. They consist essentially of rules to follow to transcribe technical ATC items such as call signs, speeds, etc. It also gives instructions to transcribe phenomena such as hesitations, pauses, or accentuations. While transcribing the formation corpus, we believed that this specification wasn't sufficiently fine grained to mark out specific phenomena. Consequently, we contributed to it by creating other classes of phenomena necessary to transcribe, and by refining existing one with sub-categories. Indeed, we considered the fact that the annotator could possibly not have access to the recordings, or not have time to refer to it for a given detail. So, it is necessary to spot any phenomenon that could be interpreted as a marker for a language act, and accessible only via recordings hearing.

It appears that, by doing this, we reach beyond the framework of "raw information" given by specifications, since this decision is based upon an interpretative classifying activity. However, we thought that if it wasn't done during the transcription, the annotator would miss some interesting phenomena.

We used Transcriber 1.4.2 to carry out the transcription.

2.3. Description of corpus

The recordings were made on July 2001 at the ENAC (*Ecole Nationale d'Aviation Civile*; in English: National School of Civil Aviation) from Toulouse.

They were sampled at 16 kHz (16 bits). A DAT (Digital Audio Tape) was used. For recording reasons, the speech signal quality sometimes suffers from saturation or noises such as interferences. However, it stays intelligible.

We present the main features of the corpus in table 1 below.

Table 1: Main characteristics of our corpus.

Length	Number of speakers	Number of "exchanges"	Number of speech turns	Number of words
36h50mn	16 (distributed in 2 groups)	2 019	11 427	76 306

3. Disfluency phenomena

3.1. Terminological considerations

In literature, many different words are used by various authors to refer to a same disfluency phenomena. This is why it is very important to present which terminology we use here, and the phenomena it designates.

We describe here 6 main classes of disfluencies. Whenever it is necessary, we give an example of the phenomenon they correspond to.

- Hesitations: this term only designates the interjection "euh", which corresponds to "er" in English. It is usually considered to point out a moment's thought on what has to be said next. According to some terminologies (notably [7]), it belongs to the category of "filled" pause. Example:
  maintenons niveau 1 0 0 Poitiers Amboise euh Lacan
- Repeated words: we gave a slightly restrictive definition to this one. We called "repeated words" any situation where a word (or a group of words) appears at least two times consecutively. We do not take into consideration any repetition of a disfluency phenomenon, such as hesitations, fragments, etc. Example:
  station station calling euh repeat your callsign
- False starts: one of the word which has the most various meanings according to authors. We use it to refer to the utterance of the word that does not come to an end. It is worthy to note that in our own terminology, a false start always corresponds to a fragment of word that can be identified. The knowledge of the phraseology helps a lot for this identifying task. Let's see for example the following example; we put the false start within brackets:
  speed euh 200 Kts [mak] euh minimum

The context (both of the previous utterances and of the situation) and phraseology help to understand that the speaker first began to utter "maximum". He realized that he was wrong, and stopped the production ("mak"). Finally, he said the correct word: "minimum".

- Fragment: contrary to "false starts", "fragments" refers to a sound (usually a single phoneme) that can't be identified as a part of a word, or that clearly does not belong to any lexicon. We do not include in this category physiological sounds (breathes, cough, etc.). Example (the fragment is within brackets):
  due to [ou] due traffic euh descend level 9 0
- Lengthening: a lengthening occurs when the production of a sound (usually a phoneme) lasts more than usual. According to some authors, it also belongs to "filled pause" category. While transcribing and annotating the corpus, we took as minimum value for lengthening 20 cs (centiseconds), as many authors in literature.
- Long pause: it is pause (a silent one) that lasts more than 20 cs. We only take into account pauses that occur during a given speaker's speech turn (and not, for example, between two speech turns from different speaker).

Now, we see in the next section the distribution of those phenomena, and do some comparisons with results obtained in similar studies. As we'll see, some results differ a lot from the average observed on others corpus.

3.2. Distributions of phenomena

The Figure 1 displays the distributions of the phenomena described in section 3.1. The number in bold corresponds to the total number of occurrences in the corpus; the percentage is computed in regard to this total number. The sector corresponding to the word repeats does not actually appear on the graphic since it is below 1%.

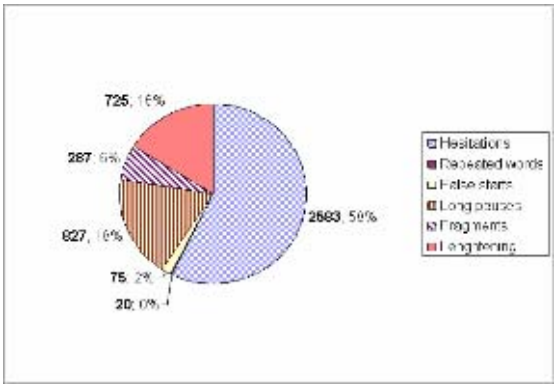


Figure 1: Distributions of the various disfluencies in our corpus.

This distribution calls for many commentaries. In this paper, we will proceed to a detailed comparison with some others studies. Here is a short description (nature of the task, number of words, etc.) of each spontaneous speech corpus on which these studies are based:

- [2]: the work presented in [2] is based on 13 tales orally told by children. It lasts 70 minutes and 25 seconds;
- [6]: based on a 1 000 382 words corpus (various spoken situations; 794 different speakers);
- [7]: based on a corpus lasting 54 minutes, and comprises 8500 words. The 10 different speakers talk about their job or their memories;
- [8]: this thesis is based on a corpus consisting of negotiations (in English language) of merchandise transport by train. It comprises 52 000 words.

As we see, these studies are based on some very different corpus, whether in task or in length. This diversity will thus constitute a valid basis of comparison with our own corpus.

We will now present the comparisons for the categories of disfluencies that we defined. Of course, since any study does not cover all the disfluencies, we will only present those that concerns a given phenomena, or which categorization is close from our. These comparisons would need to be deepen, since in most of the cases, the categories we defined are more or less slightly different from those from other studies. But it will give us a good overview of the specificities of our corpus.

Table 1: comparison for repeated words

Name of the study	Our corpus	[2]	[7]	[8]
Number of repeated words	20	110	141	256

- Repeated words: the most surprising result concerns the number of word repeats, as it appears in table 2.
  Indeed, it appears that the frequency of word repeats are always considerably higher than in our corpus. How to explain such a difference? First, remind that we do not take into account repetitions of any phenomena of disfluencies, even false starts. But that do not explain the huge difference of number. We think that the main explanation is the very nature of the corpus. Our hypothesis is that, in the ATC situation, the speakers (both controller and pilot) can not afford to produce any ambiguity or problem that could affect the comprehension of the utterance. Also, the time necessary to produce an utterance is not extensible: the speaker must not spend too much time in hesitation or other pauses (filled or silent). As we will see next, this hypothesis is confirmed by the fact that for all of the disfluencies that we defined, there is always less occurrences in our corpus (proportionally to the size of the corpus of comparison).

Table 3:

Name of the study	Our corpus	[2]	[8]
Number and/or percentage of hesitations (in regards to the total number of words)	2583 3..38%	554	3512 6.75%

- Hesitation: as we saw in table 3 below, there is also much less hesitations in our ATC corpus than in other studies. There is admittedly 544 occurrences in [2], but this corpus is 70 minutes long, whereas our is 35 hours long. So, there is proportionally more occurrences in the corpus used by [2]. However, one can notice that the difference seems to be overall lesser than what we observed for repeated words.
- False starts and fragments: among the studies we chose to compare with, [6] is the only one whose categorization is the closest from our, resorting to what we call "false starts" and "word fragment". It also present detailed statistics about their distribution. As the authors do not distinguish between "false starts" and "word fragment", we will add up the occurrences of both phenomena that appear in our corpus. The result is a total of 362 occurrences, i.e. 0.47% of the total number of words. [6] reports a total of 6094 occurrences of "word fragments" for about 1 000 000 words (approximately 0.6%). Thus, the distribution in our corpus of this twofold category is quite close of the one observed in [6], contrary to what occurs for the others categories. But this result might be due to the fact that this twofold category do not exactly match with the one defined by [6].

Table 4: Comparison for lengthening

Name of the study	Our corpus	[2]	[7]
Number and/or percentage of lengthening (in regards to the total number of words)	725 0.9%	284	669 (including "euh") 7..9%

- Lengthening: again, as we see in table 4, the frequency of what we called lengthening is lower in our corpus than in the other ones.
- Long pause: the table 5 shows that the specificity of our corpus is a little less pronounced than for the other categories of disfluencies. But, here again, we notice that there is less "long pauses" than in other corpus.

Name of the study	Our corpus	[2]	[7]
Number and/or percentage of long pauses (in regards to the total number of words)	827 1.08%	147 1	318 3.74%

Many pages would be necessary to exhaustively examine the different phenomena, the differences observed with other study, and their causes. In the framework of this paper, we will only focus on false starts. They are the object of the next section.

4. A study on "false starts"

Why focusing on false starts? As we saw, they are far from being the most frequent disfluencies in our corpus. All the same, we think they are worth of interest, for two main reasons. First, they are special cues on the "work of formulation" (we take up here the expression used in [2]). Notably, they can show, in some case, the word that the speaker has in mind. Thus, they help to base some hypothesis on the nature of the problem. Secondly, though in little number, they manifest themselves in different ways that are interesting to identify.

In the first section, we present the different kinds of false starts we found in our corpus. Then, we present the distribution of those categories.

4.1. The different types of false starts

First able, we found out that two kinds of false starts do exists They differ according to their function in the production of the

utterance. The first one do not have any visible function. Example:

route Lacan [amboi] Amboise Balon Limoges

It is probably useless to seek out any function in this category. This kind of false starts is only a mark of "the work of formulation". 29 occurrences of false starts, that is to say 39% of the total number belong to this category.

To the contrary, the second type plays a role of correction of a mistake that was about to be done. Let's see for example the following example:

[mike] Paris 124 decimal 05 Littoral M C

This concerns 46 occurrences, i.e. 61% of the total number of false starts.

This last category can give us some precious hints on the behavior of the speaker and the causes of his errors. Consequently, they deserve a deeper analysis. Many works on disfluencies (for instance, [2], [6], [7], [10]) carry out their study by analyzing the distribution of a given disfluency phenomena according to the lexico-syntactic category of the word it affects. For now, we prefer not to do so. Indeed, our corpus specificities require a specific linguistic characterization. For instance, it is difficult to make a distinction between "function words" VS "lexical words".

In the meantime, we do a typology of the false starts according to the "role" of the word or group of words they affects (except for a category). By "role", we mean the function assumed in regard to the phraseology. We define the following sub-categories:

- Errors on a "word": we quote the term "word" for he designates commands or order (such as "climb", "request", etc.) but also call signs ("Britair 452" for example). Here is a typical example:

  climbing for level 1 7 0 [mak] euh minimum D M C

- Errors on utterance organization: they occur when a word (or words group) does not appear at the position in the utterance that is requested by phraseology. Example:

  [poi] Absie Poitiers Balon Reson Britair B X

  In this example, the speaker begins to say "Poitiers" at the start of the utterance. But he realize that the name of this town must be said after "Absie". This explains why he stops the first production of "Poitiers".
- Errors on the language used: when the speaker talks in an other language than the requested one. Example:

  P I [vite] speed 2 1 0 Kts

Here, the speaker is supposed to speak English. Or, he begin the production of "vitesse" (the French word for "speed"). This is why he stopped before the end of the production.

- Errors of pronunciation: contrary to the previous ones, this category is not linked to a problem in regard to the phraseology. It appears when the speaker use the correct word, at the proper position, but mispronounce it. For instance, in the following example, the speaker mispronounced the word "Littoral":

  It's [lio] Littoral

As we proceeded for the disfluencies, we will now present the distribution of these categories.

4.2. Distribution of repair false starts

The figure 2 below shows the distribution of the categories described above. As in figure 1, the numbers in bold correspond to the number of occurrences, and the percentage is computed in regard to the total number of repairs false starts.

Most of the errors concerns incorrect "words"; the second most frequent category is the incorrect "word" position. We see there a confirmation of one of our main hypothesis: most of the errors are directly linked to the most unusual (in regard to everyday language) sides of the phraseology and of the task. Thus, "words", such as we defined them, are often call signs, i.e. complex sequences of letters and numbers. It is obvious that they are difficult to handle, especially for an apprentice controller. This leads to a high cognitive load, that itself generate some troubles of production. The same reasoning can be applied to errors related to position of "words".

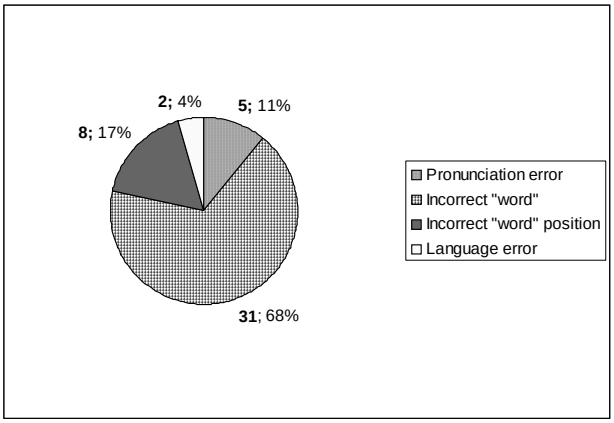


Figure 2: Distributions of the categories of false starts.

5. Conclusion

The corpus on which we lead our study presents many differences with more usual corpora. Indeed, it results from an apprenticeship task. Besides, it requires the use of a phraseology, what entails very restricted lexico-syntactic schemas.

However, it contains many various disfluency phenomena. We showed that there is some differences between the distribution observed in our corpus, and those on other corpus. This is especially true for word repetitions. They seem to be a phenomena that is very sensitive to the task. We explain these differences, on one hand by the phraseology, and on the other hand by the fact that controllers have to be very careful not to produce ambiguities that could have disastrous consequences. We also seen that, more generally, there is less disfluency phenomena in our corpus. We explain this by the same reasoning that we use for word repetitions.

We also studied the specific case of "false starts". They sometime assume the function of corrective instances, or are just some cues for the formulation of utterances. We showed that they can be seen as evidences that the huge cognitive load induced by the apprenticeship is responsible for errors.

Of course, this analysis needs to be deepen. We account to do this from three perspectives. First, we will lead a study of the linguistic properties of the corpus. As we saw, they are very specific, and to compare with others study, it is necessary. We can also take into account the speech rate of

speaker, which is faster than in usual dialogues, and the differences according to the language used.
Of course, we also plan to further deepen the study of the different disfluencies, notably by seeing the correlations between them, as well as with problems that can arise, such as corrections and languages troubles.
The third perspective is to determine the precise nature of the relation between disfluencies and the cognitive load that we attributed to the apprenticeship task. For this, we have at our disposal a corpus made of recordings of dialogues between controllers and pilots that are in real situations. We will apply the same methodology of annotation and study to this corpus. It will permit us to compare the rate of disfluencies in the two different corpus, and to check our hypothesis on apprenticeship influence on their manifestations and frequency.

6. Acknowledgements

This study is funded by the CENA. We thank Philippe Truillet for the records of the exercises, and the information he gave us about it. We are also very grateful to Gwenael Bothorel, who helped us a lot for the transcription/annotation task.

7. References

[1] Arrêté du 27 juin 2000 relatif aux procédures de radiotéléphonie à l'usage de la circulation aérienne générale. *J.O n° 171 du 26 juillet 2000*, p. 11501.

[2] Candéa Maria. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. Étude sur un corpus de récits en classe de français.* Ph.D. thesis, Université Paris III (Sorbonne Nouvelle).

[3] Coullon I., Graglia L., Kahn J. & Pavet D. (2001) *Définition détaillée du document type (DTD) pour le codage sous XML des communications VHF en route – VOCALISE Trafic CRNA / France 2000*. CENA internal report.

[4] Coullon I. & Graglia L. 2000. *Spécifications de la base de données pour l'analyse des communications VHF en route, CENA internal report.*

[5] Dourmap Loic & Truillet Philippe. 2003. Interaction vocale dans le contrôle aérien : la comparaison de deux grammaires contextuelles pour la reconnaissance des indicatifs de vol, *CENA internal report*, NR03-669.

[6] Henry Sandrine & Pallaud Bertille. 2004. Word fragments and repeats in spontaneous spoken French, *Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop*, 5-8 Septembre 2003, Göteborg University, Suède, p. 77-80.

[7] Henry Sandrine, Campione Estelle & Véronis Jean. 2004. Répétitions et pauses (silencieuses et remplies) en français spontané, *Actes des XXVèmes Journées d'Etude sur la Parole (JEP'04)*, Fès (Maroc), p. 261-264.

[8] Kurdi M.- Z. 2003. *Contribution à l'analyse du langage oral spontané.* Ph.D. Thesis, Université J. Fourier, Grenoble, France.

[9] Nakatani C. H. & Hirschberg J. (1994), A corpus–based study of repair cues in spontaneous speech, *Journal of Acoustical Society of America*, 953, p. 1603-1661.

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

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 39-42.

Improvement of verbal behavior after pharmacological treatment of developmental stuttering: a case study

Pierpaolo Busan[*], *Giovanna Pelamatti*[*], *Alessandro Tavano*[**,°], *Michele Grassi*[*], *Franco Fabbro*[**,°]

[*] Psychology Department, University of Trieste, Italy
[**] Scientific Institute "E. Medea", Italy
° University of Udine

Abstract

Developmental stuttering is a disruption in normal speech fluency and rhythm. Developmental stuttering usually manifests between 6 and 9 years of age and may persist in adulthood. At present, the exact etiology of developmental stuttering is not fully clear. Besides, the dopaminergic neurological component is likely to have a causal role in the manifestation of stuttering behaviors. Actually, some studies seem to confirm the efficacy of antidopaminergic drugs (haloperidol, risperidone and olanzapine, among others) in controlling stuttering behaviors. We present a case of persistent developmental stuttering in a 24-year-old adult male who was able to control his symptoms to a significant extent after administration of risperidone, an antidopaminergic drug. Our findings show that the pharmacological intervention helped the patient improve on a set of fluency tasks but especially when the tasks involved the uttering of content words. Our results are discussed against the current theories on the cognitive and neurological basis of developmental stuttering.

1. Introduction

The World Health Organization classifies stuttering as a disruption of the normal speech rhythm, whereby the subject knows exactly what he or she wants to say but is unable to utter the intended words and sentences fluently [17]. This definition remains valid. However, in the last 10 years some experimental studies have suggested that the neurological component may play a more important role in this disorder.

Developmental stuttering is characterized by behaviorally evident sound and syllable repetition at the beginning of words, phrases, and sentences, sound prolongation, interruptions, silences or sound blocks, facial spasms and muscular tensions in the oro-facial district during speech.

Developmental stuttering usually begins in childhood between the age of 6 and 9 years and affects around 5% of the child and adolescent population to different degrees of severity. In most cases, spontaneous remission of the symptoms does occur but one percent of developmental stutterers are still affected by this problem in adulthood. Finally, male subjects are more affected than females [1].

A genetic component of the disturbance is suggested by Ambrose, Cox & Yairi [1] and Yairi, Ambrose & Cox [19].

Wu, Maguire, Riley, Lee, Keator, Tang, Fallon & Najafi [18] found a strong activation of the dopaminergic neurons in left caudate and in left amygdala in a group of stutterers vs. a control group. Such stronger activation was evident in the left insula and in the hearing cortex, too. A greater than normal dopaminergic activity in the left basal ganglia might induce a lower activity of the speech circuits in the dominant hemisphere and could partly explain the physiological mechanisms of stuttering. Actually, a finding common to many studies on the physiology of stuttering is the lower and different pattern of cerebral activation of the left hemisphere language cortex, combined with a stronger activation of the homologue areas of the right hemisphere [3]. These results allowed to hypothesize [9] that developmental stuttering might be considered as a mainly neurological dysfunction, and more specifically a dopaminergic dysfunction of the basal ganglia. This hypothesis is indirectly confirmed by the efficacy of antidopaminergic drugs such as haloperidol [11], risperidone [8] and olanzapine [6] which proved useful in controlling the stuttering symptoms. These drugs are principally presynaptic antagonists of dopaminergic receptor D2, which is largely present in human basal ganglia.

The exact functioning of antidopaminergic drugs is not completely clear [13], and it is impossible to generalize the results of these studies to suggest the efficacy of the pharmacological treatment of stuttering.

An indirect confirmation of the dopaminergic theory of developmental stuttering comes from Parkinson's Disease (PD). PD is characterized by a degeneration of dopaminergic neurons in the basal ganglia, and sometimes its first symptoms include speech difficulties [7].

Many researchers suggest that the neurotransmitters' balance as a whole (and not just a quantitative variation of a single neurotransmitter) is fundamental for the smooth execution of speech [15]. Actually, some studies report on the utility of SSRI (paroxetine) in the treatment of stuttering [4].

Finally, developmental stuttering closely resembles Tourette's Syndrome, a tic-disorder with an important obsessive-compulsive component. The resemblance between the two syndromes is evident when we analyze the secondary behaviors like repetitive or prolonged eye-blinks, jaw blocks and jaw tremors, or abnormal head and arms movements associated to dysfluencies in stuttering and typical of Tourette's Syndrome, as shown by Mulligan, Anderson, Jones, Williams & Donaldson [10].

On the other hand, but not necessarily in opposition with dopaminergic theory, Vasic and Wijnen [16] proposed a psycholinguistic theory about etiology of stuttering. The authors suggest that stuttering depend on an excessive attentive threshold level for speech.

2. Method

We present a case of developmental stuttering persisting into adulthood in a 24-year-old Italian male who, after administration of the antidopaminergic drug risperidone, could successfully reduce the symptoms of stuttering.

Risperidone is an "atypical" antipsychotic drug: its prolonged assumption causes a lower incidence of extrapiramidal adverse reactions and a lower incidence of Tardive Dyskinesia because risperidone is a D2 and 5HT-2a

antagonist. Therefore, the serotoninergic antagonism can help controlling the adverse reaction of the dopaminergic block.

2.1 Behavioral assessment

Initially, developmental stuttering was confirmed by means of the Stuttering Severity Instrument [14]. After obtaining the patient's informed consent to the treatment, including the possible risks, we obtained all the most important treatment-related biological health parameters like heart and liver functioning. Then, we have decided to operate as follows:

1. we established the patient's baseline level of stuttering on all experimental behavioral tasks;
2. 0.5 mg/d risperidone was administered for a six-week period as suggested by the literature [8];
3. At the end of the first treatment period, all experimental behavioral tasks were re-administered to investigate the efficacy of treatment;
4. The first treatment period was followed by a six-week washout period, while repetition of all experimental behavioral tasks came immediately after;
5. A second six-week drug intake period followed (same dose), and the administration of all experimental behavioral tasks was successively repeated;
6. Finally, during a washout period of 12 weeks the long-terms effects of the drug were explored and successively all experimental behavioral tasks were repeated.

2.2. Behavioral investigation measures

The following behavioral measures were administered:

- Stuttering Severity Instrument [14]: at baseline and at the end of all treatment and washout periods, a conversation sample of 150 words and a reading sample were recorded. The percentage of stuttering, the mean of the longest three blocks and the subjective evaluation of the secondary behaviors related to stuttering were computed. The degree of stuttering according to a graded scale was defined.
- Measures of verbal fluency at baseline and at the end of all treatment and washout periods: content word production, content word repetition and nonword repetition on the basis of the Italian versions of the FAS [2] and BAT [12] tests, originally designed for the assessment of aphasia deficits. We individually analyzed each test by calculating the percentage of stuttered syllables against the total number of syllables.
- Finally, secondary behaviors associated to stuttering were explored based on Mulligan & colleagues [10], at baseline and at the end of all treatment and washout periods: a subject's phone conversation of ca. 400 words with a familiar person was recorded. The video-analysis was later run at zero volume to avoid the influence of stuttered speech. The "yes or no" head movements were not counted. However, later it was decided to count all movements that were highly frequent at baseline like repetitive eye-blink, sustained left eyes, jaw jerking and sustained low head movements.

2.3. Experimental hypotheses

With reference to the available theories about the dopaminergic etiology of stuttering, we expected an improvement on language performance after the first pharmacological treatment period and a worsening of these results after the first washout period, and finally a further improvement after the second pharmacological treatment period.

3. Results

3.1. The Stuttering Severity Instrument

On the Stuttering Severity Instrument (SSI), after the baseline measures, the subject scored 34, with a percent value of 90-96, corresponding to a severe degree of stuttering.

After the first treatment period, the degree of stuttering decreased to moderate with a scale score of 22 and a percentile value between 24-40. This shows a strong improvement which was maintained during the first washout period. Stuttering was moderate with a mild worsening up to a scale score of 25 and percentile measures of 56-66. During the second drug period, the patient obtained the best results with a scale score of 18, a percent value of 5-11 and mild stuttering. Finally, after the 12-week period moderate stuttering was observed, with a scale score of 23 and 24-40 percent values (Fig. 1).

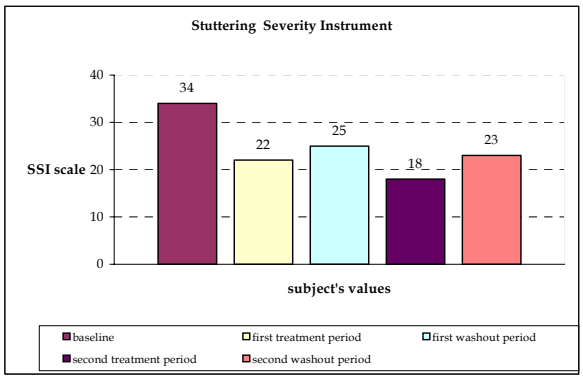


Figure 1: Scale scores on the Stuttering Severity Instrument.

The subject indicated a feeling of unprecedented easy speaking, with no muscular and jaw tensions and a new ability to manage the blocks among the major subjective sensations.

3.2. Verbal Fluency Analysis

Each of the three tasks were entered into the Logistic Regression Test. Then, all task results were compared by a simple Variance Analysis.

3.2.1. Content word production

At baseline, a Stuttered Syllable (SS) versus Total Syllable (TS) ratio of 35.13% was obtained. During the two treatment periods the SS/TS ratio was 8.24% and 6.77%, while during the washout periods it amounted to 7.39% and 7.02% (Fig. 2).

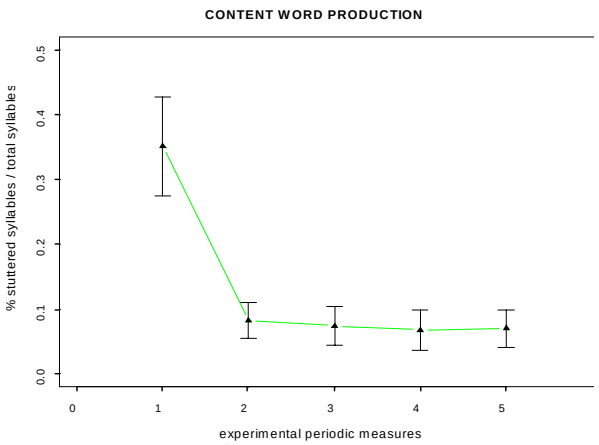


Figure 2: Content word production. 1-baseline; 2-first treatment period; 3-first washout; 4-second treatment period; 5-second washout.

The Logistic Regression Test results support all these findings in all comparisons between baseline and treatment and washout periods (z=-7.06, z=-6.73, z=-6.60, z=-6.93; p<0.001).

3.2.2. Content word repetition

On this task the patient scored 17.86% at baseline. The two six-week treatment periods demonstrate a significant improvement as shown by Logistic Regression (z=-2.86, z=-2.86; p<0.01), with 1.19% in both cases.

Significance is confirmed by the statistical analysis after the washout periods too (z=-2.85, z=-2.71; p<0.01), with SS/TS ratios of 2.38% and 3.57% (Fig. 3).

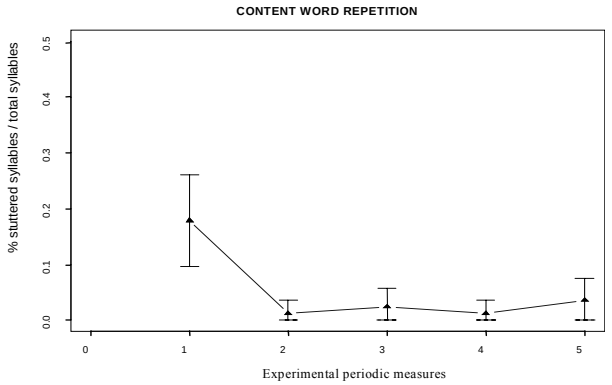


Figure 3: Content word repetition. 1-baseline; 2-first treatment period; 3-first washout; 4-second treatment period; 5-second washout.

3.2.3. Nonword repetition

On this task which is useful to determine verbal fluency, the SS/TS ratio was 9% at baseline. Statistical analysis reveals that the only two significant comparisons concern the treatment periods (z=-2.14, z=-2.14; p<0.05) with 1% of SS/TS. No significant variations were found between washout periods and baseline scores, with percentages of 6% and 3% (Fig. 4).

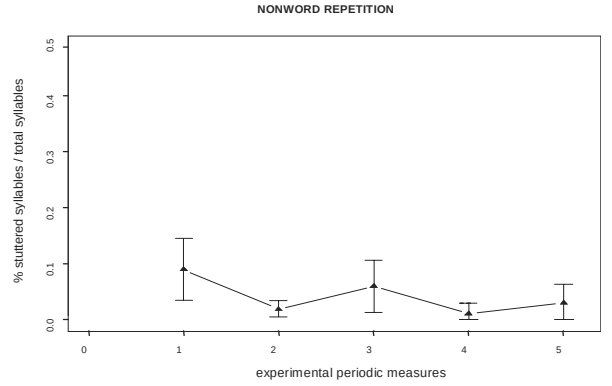
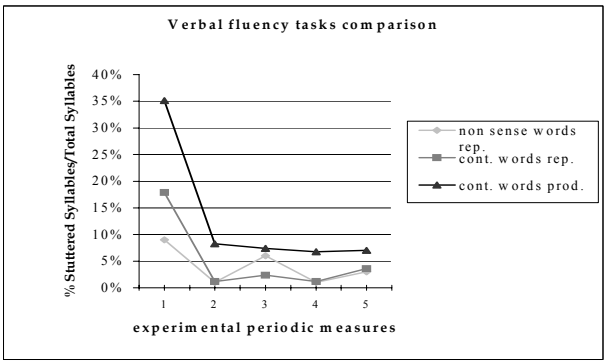


Figure 4: Nonword repetition. 1-baseline; 2-first treatment period; 3-first washout; 4-second treatment period; 5-second washout.

3.2.4. Comparison between all verbal fluency tasks

An Analysis of Variance was performed to compare all three verbal fluency tasks with the five treatment periods and verify whether the drug could have different effects on these tasks.

The statistical results are significant (F= 4.69; p<0.05) and indicate a probable different effect of risperidone on the



various verbal fluency samples (Fig. 5), where probably content word tasks obtained the most important fluency gain.

Figure 5: Verbal Fluency tasks comparison. 1-baseline; 2-first treatment period; 3-first washout; 4-second treatment period; 5-second washout.

3.3. Secondary behaviors associated to speech

A final statistical analysis was run on involuntary movements associated to normal and stuttered speech.

As previously said, only the movements that were more present at baseline were explored: repetitive eye-blink, sustained left eyes, jaw jerking and sustained low head movements (Tab. 1).

Table 1: Most relevant involuntary movements associated to speech classification.

musc. districts	Base-line	ther. per. 1	Wash-out 1	ther per. 2	Wash-out 2
Rep. eye-blinks	56	19	11	15	24
Sust. left eyes	12	4	12	2	7
Jaw jerk.	33	8	12	3	20
Sust. head mov.	16	9	15	10	14

A Two Proportions Test with single comparisons between consecutive treatment and washout periods was applied, with baseline as referent. Data interpretation is not simple, but significant results were found for repetitive eye-blinks between the second treatment period and the second washout period (χ^2=3.19; p<0.05). This suggests that the increase in involuntary movements during the second washout period is caused by the lack of risperidone. Sustained left eyes movements were significant for all comparisons (χ^2=12, p<0.0005; χ^2=17.14, p<0.00005; χ^2=4.44, p<0.05). Thus, it seems that risperidone is effective when it is taken, but not during washouts. Jaw-jerking results suggest the effectiveness of the drug only during the second treatment period, since these data are significantly different from the two washout periods (χ^2=6.99, p<0.005; χ^2=19.29, p<0.000001). Finally, sustained low head movements show significant differences between the two treatment periods and the first washout (χ^2=6, p<0.01; χ^2=4.57, p<0.05). This indicates the effectiveness of risperidone during therapy and the duration of the effects (only for this specific type of involuntary movement) after the second washout (Fig.6).

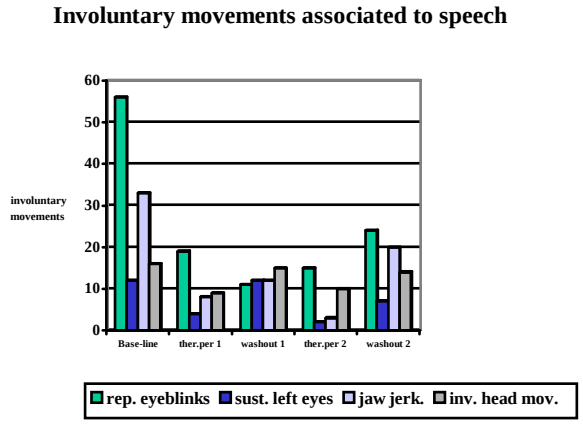


Figure 6: involuntary movements associated to speech

4. Discussion

Our findings support the hypothesis that pharmacological treatment with risperidone can help manage the typical dysfluencies of developmental stuttering. In fact, the Stuttering Severity Instrument scores demonstrate that under therapy stuttering changed from severe to moderate during the first treatment period and to mild after the second treatment period. The worsening after the washout periods is minimal because stuttering does not increase beyond the moderate level.

Verbal fluency measures suggest that both content word production and repetition and nonword repetition improved under therapy. These findings seem to confirm the findings that adult stutters produce more dysfluencies on content words [5]

Regarding the analysis of the secondary behavioral components of stuttering, the total amount of involuntary movements tends to diminish from baseline to the second treatment period but it increases sensibly after the second washout period. Thus, in this instance treatment with risperidone was effective only during its assumption, while it had a more enduring effect on verbal fluency measures and SSI.

These results allow us to confirm Mulligan and colleagues' [10] theory, which defines stuttering as a tic disorder. We could consider an excess of typical motor activity in the child, as a positive predictive factor for the development of stuttering behaviors. This, in turn, confirms the relevance of the neurological and motor component of stuttering. We may assume that risperidone can influence a cognitive component and/or a motor component of verbal fluency.

In conclusion, during treatment periods risperidone seems to be effective in controlling stuttering symptoms like dysfluencies and involuntary movements associated to stuttered speech. Most importantly, risperidone was well tolerated by the subject.

Therefore, our results confirm a possible dopaminergic etiology of stuttering that can be considered a psychomotor disorder with an important neurological component determining its manifestation.

5. References

[1] Ambrose N.G., N. J. Cox & E. Yairi 1997. The genetic basis of persistence and recovery stuttering. *Journal of Speech, Language and Hearing Research*, Vol. 40, No. 3, pp. 567-580.

[2] Benton A.L. & K. Hamsher 1989. Multilingual aphasia examination. Iowa City, IA: AJA Associates.

[3] Braun A.R., M. Varga, S. Stager, G. Schulz, S. Selbie, J.M. Maisog, R.E. Carson & C.L. Ludlow 1997. Altered patterns of cerebral activity during speech and language production in developmental stuttering. An H2(15)O positron emission tomography study. Brain, Vol. 120, Pt. 5, pp. 761-784.

[4] Costa D. & R. Kroll 2000. Stuttering: an update for phisicians. Canadian Medical Association Journal, Vol. 162, No. 13, pp. 1849-1855.

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

[6] N. Lavid, D.L. Franklin & G.A. Maguire 1999. Management of child and adolescent stuttering with olanzapine: three case reports. Annals of Clinical Psychiatry, Vol. 11, No. 4, pp. 233-236.

[7] Leder S.B. 1996. Adult onset of stuttering as a presenting sign in a parkinsonyan-like syndrome: a case report. Journal of Communication Disorders, Vol. 29, pp. 471-478.

[8] Maguire G.A., G.D. Riley, D.L. Franklin & L.A. Gottschalk 2000. Risperidone for the treatment of stuttering. Journal of Clinical Psychopharmacology, Vol. 20, No. 4, pp. 479-482.

[9] Maguire G.A., G. D. Riley & B.P. Yu 2002. A neurological basis of stuttering?. Neurology, Vol. 1, pag. 407.

[10] Mulligan H.F., T.J. Anderson, R.D. Jones, M.J. Williams & I.M. Donaldson 2003. Tics and developmental stuttering. Parkinsonism and Related Disorders, Vol. 9, pp. 281-289.

[11] Murray T.J., P. Kelly, L. Campbell & K. Stefanik 1977. Haloperidol in the tratment of stuttering. British Journal of Psichiatry, Vol. 130, pp. 370-373.

[12] Paradis M. 1999. Test per le afasie in un bilingue. EMF Edizioni Bologna.

[13] Rang H.P., M.M. Dale & J.M. Ritter 2001. Farmacologia. Casa Editrice Ambrosiana.

[14] Riley G. 1980. Suttering Seveirity Instrument for children and adults. Cs. Publications.

[15] Schreiber S. & C.G. Pick 1997. Paroxetine for secondary stuttering: further interaction of serotonine and dopamine. Journal of Nervous and Mental Disease, Vol. 185, No. 7 pp. 465-467.

[16] Vasic N, F. Wijnen 2001. Stuttering and speech monitoring. In DISS' 01, pp. 13-16.

[17] World Health Organization. 1977. Manual of the international statistical classification of disease, injuries and causes of death. Geneva, WHO, Vol. 1, International Classification of Disease.

[18] Wu J.C., G. Maguire, G. Riley , A. Lee, D. Keator, C. Tang, J. Fallon & A. Najafi 1997. Increased dopamine activity associate with stuttering. Neuroreport, vol. 8, No. 3, pp. 767-770.

[19] Yairi E., N. Ambrose & N. Cox 1996. Genetics of stuttering: a critical review. Journal of Speech and Hearing Research, Vol. 39, No. 4, pp. 771-784.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 43–46.

Pauses and hesitations in French spontaneous speech

Estelle Campione & Jean Véronis

Equipe DELIC
Université de Provence, Aix-en-Provence, France

Abstract

In traditional terminology, *silent* and *filled pauses* are grouped together, whereas hesitation lengthening is put into a separate category. However, while these various phenomena are very often associated, there have been few studies on how they interact. We analyzed an hour of spontaneous speech to show that silent and filled pauses operate in a totally different way, and that contrary to common belief, silent pauses by themselves never serve as hesitation markers, but only do so when coupled with other markers—mostly syllabic lengthening and filled pauses. These last two hesitation markers have similar acoustic and articulatory characteristics; they are also distributed and function alike.

1. Introduction

There are traditionally two types of pauses (see for example Duez [10]): *silent* pauses, in which all vocal production ceases, except for possible respiratory noises, and *filled* pauses (or pauses containing sound), consisting of quasi-lexical tokens (*euh* in French, *er* and its nasal variant *erm* in English).

By grouping both of these very different acoustic and articulatory phenomena under the same term of *pause*, one makes the implicit hypothesis that they have the same function. However, the likelihood of this is very low. It has been known for a long time (see for example Boomer [2]) that silent pauses have a double role. Some silent pauses are *demarcative*, and appear at the junction of speech segments, which they help structure and parse. Others are *hesitation* pauses, caused by the sporadic difficulties the speaker encounters during "searching and encoding" mental operations (Barik [1]) or the "formulating work" (Morel & Danon-Boileau [17]) inherent to speech production. Conversely, it seems that filled pauses are only used for this second role: they are used as a conventional signal by the speaker to signal that he/she is not done speaking and to prevent interruptions during the time required for building the next part of the speech (see Clark & Clark [7], etc.).

It is worthy to note that the role of hesitation signals, corresponding to filled pauses, is also that of some types of syllabic lengthening (generally affecting a vowel at the end of a word), that we will call *hesitation lengthening*. Their properties are similar to those of filled pauses (see Guaïtella [14], Candea [6]), to the point where they can be detected using the same algorithms (see Goto, Itou, & Hayamizu [13]). Current terminology thus groups under the same term of *pause* two phenomena that are acoustically and functionally very different, whereas it puts hesitation lengthening and *er/erm* in different categories, even though they share similar properties and function alike. We group hereafter hesitation lengthening and *er/erm* under the same term *filled pauses*, as done by Goto et al. [13].

There are many studies on silent pauses in the literature (see for example Zellner's state of the art [20]). However, as demonstrated by a recent thesis (Candea [6]), there are far fewer studies on sentence planning markers in general, and on filled pauses and on lengthening in particular (which is probably due to the fact that most phonetic studies have been for a long time devoted to "laboratory speech", at the expense of spontaneous oral speech, as noted by Cutler [8], Duez [11] and others). In any case, there are almost no studies on how these phenomena interact.

This paper aims at giving a precise study of how silent pauses, filled pauses and hesitation lengthening interact, based on a study of a corpus of French spontaneous speech. We show that silent pauses by themselves never serve as hesitation markers, and need another sentence planning marker, most often a type of lengthening or an *euh* (or a combination of both) to play that part.

2. Corpus

The corpus used in this study consists of 8500 words and 54 minutes of French spontaneous speech, involving 10 different speakers (5 male and 5 female). It is a subset from the *Corpus de référence de français parlé* (Spoken French Reference Corpus) recently recorded by our team [9][1], which consists in 136 recordings of ca. 15 minutes each, involving speakers from 40 different locations covering the France map (36 hours of speech). The corpus is sampled according to age and education levels, and speech genres (public, private and professional speech). Recordings have been made in a quiet room, using minidisk recorders. Disfluency phenomena (hesitations, repetitions, false starts, *euh*, etc.) have been carefully transcribed with several independent verifications. Syllable lengthening and intonation were not marked in the initial transcription.

We have selected our sub-corpus in order to balance sexes, age groups and education levels. Five minutes segments were extracted from the original recordings, in which the interviewed speaker was speaking without interruption. A more detailed description and the transcribed corpus itself are available in Campione [3][2].

Disfluency phenomena have been once more verified in the sub-corpus, which was augmented with syllable lengthening and intonation markup. Syllable lengthening was entirely done manually (with the help of a signal editor), and intonation markup was obtained semi-automatically with careful manual verification, according to the method described in Campione & Véronis [5].

3. Silent pauses

[1] Available on-line at:
http://www.up.univ-mrs.fr/veronis/pdf/2004- presentation-crfp.pdf

[2] Available on-line at:
http://www.up.univ-mrs.fr/delic/theses/resume-campione.html

3.1. Tagging

Silent pause transcription is a very difficult exercise when done entirely manually, and we have noticed that most linguists, even highly competent ones, tend to miss many silent pauses, especially when they are coupled with other phenomena (such as hesitation or syllable lengthening). Despite the multiple verification of the original corpus, some silent pauses were still missing (this difficulty confirms Candea's observations [6]). Since, in addition, we needed carefully-measured duration times, silent pauses were detected using a program that calculates the fundamental frequency[3] and isolates voiceless segments. We applied a threshold of 200 ms, in line with past studies (Candea [6]). Shorter silent pauses, whose existence and importance have been underlined in the literature (Hieke, Kowal & O'Connell [15]), were added manually afterwards, with no lower limit (we have found silent pauses as short as 60ms). We have stressed elsewhere the dangers of applying arbitrary thresholds when studying silent pauses (Campione & Veronis [4]), since they can lead to considerable baises in the results.

We then corrected all of the silent pauses using a signal editor: pauses that were not correctly detected (which corresponded in general to voiceless stops) were deleted, and those that had not been detected were added (including those below the initial threshold of 200 ms), and the boundaries of those that were correctly detected were adjusted if needed.

The corpus contained 1375 detected potential silent pauses and 1163 actual silent pauses after correction. The distribution of silent pause lengths is highly skewed, approximately following a log-normal law, with a geometric mean at 496 ms (Campione & Veronis [4]). We categorized silent pauses in three groups, according to the tri-modal behavior described in the same study:

- short (< 200 ms)
- medium (200-1000 ms)
- long (> 1000 ms)

These three types of silent pauses are respectively marked ^, + and ++ in the examples throughout this paper.

3.2. Demarcative role

It is largely accepted by psycholinguists that speech production is based on a planification-execution cycle that results in a series of relatively short units (named discourse segments hereafter), separated by silent pauses (Fromkin [12] ; Levelt [16] ; etc.). Silent pauses are needed by speakers for them to plan their wording, and by listeners for processing the speech. Contrary to hesitation pauses, demarcative pauses play an important part in speech structure, and are probably an important factor in the correct parsing of utterances by listeners.

We tagged all silent pauses that were of a demarcative nature in the corpus, as perceived by two independent experts. These pauses are easy to detect because there are many different converging cues (intonation, vowel duration and quality, syntax, etc.). The presence of a rising or falling intonation (detected automatically and then corrected manually—see above) was a determining factor. The example below shows the type of segmentation that was obtained (speech segments are separated by ||). We provide a literal translation below each example:

ben je travaille dans un pressing ↗ ++ || on fait pas que le pressing on fait aussi la blanchisserie ↗ + || plus la blanchisserie d'ailleurs ↘ + || les draps les nappes la restauration ↗ ++*

well I work at a dry-cleaner's ↗ ++ || we don't do just dry-cleaning we also do laundry ↗ + || and laundry as a matter of case ↘ + || sheets tablecloths catering ↗ ++

Table 1 shows how silent pauses are distributed. 71% percent of silent pauses are demarcative.

Table 1: Types of silent pauses

Length	Demarcatives	Non-demarcative	Total
Short	0	24	24
Medium	673 (70%)	293 (30%)	966
Long	148 (86%)	25 (14%)	173
Total	821 (71%)	342 (29%)	1163

Short silent pauses are never demarcative, which seems to justify the thresholds of 200 ms used in some studies. We agree with Morel & Danon-Boileau [17] for whom silent pauses that are below this threshold do not have a defined iconic value. They often are of a respiratory nature and appear where the intonative cohesion is blatant:

donc c'est ^ la calandre qui travaille aussi

thus it's ^ the drying machine that works as well

4. Filled pauses

4.1. Tagging

Hesitation lengthening and *euh* (that we refer both to as filled pauses in this paper, as said before) are characterized by a continuous vowel lasting much more than the norm and with a constant vocalic quality, and are associated with a flat or slightly falling fundamental frequency (F_0) curve (Guaïtella [14]). These characteristics seem to be common to many languages (Quimbo, Kawahara & Doshita [18]; Goto et al. [13]).

Aside from the time required for the task, tagging filled pauses in the corpus is relatively easy provided the latter is listened to carefully with the help of a signal editor. Hesitation lengthening is easily set apart from types of lengthening pertaining to syntactic structure, to stress, etc., which generally have a rising or falling intonation contour, at least in French (see Vaissière [19]). Hesitation lengthening often occurs on function words, or at positions that are not syntactic or discourse breaking points. *Euh* is an easy to recognize *quasi-lexical* token (it is listed in dictionaries), and human annotators almost never disagree on its presence. Only a few cases of words ending with a schwa are borderline: lengthening of the schwa or progressive change to an *euh*? Neither listening, nor intonation curves, nor sonagrams seem to provide any decisive answer to this question. One might wonder whether this "hesitation" on hesitation is not indicative that hesitation lengthening and *euh* have the same function. We use a colon to note hesitations in the form of syllabic lengthening:

*voilà alors hein ^ on **ne:** ++ il est il est difficile d'aborder **la:** + la question du métier sans parler un petit peu des origines*

[3] Developed by Robert Espesser, as well as various tools including the signal editor used in this study.

*Well then eh ^ we do **not:** ++ it's difficult to tackle **the:** + the question of one's job without talking a little about the origins*

4.2. Filled pause sequences

In 12% of cases, filled pauses are part of a complex sequence containing several types of lengthening, several *euh*'s, or a combination of both (possibly mixed with one or more silent pauses). The following excerpt provides an example of a particularly long combination (it is interesting to note that the next sequence is a word fragment, followed by another filled pause:

*euh Beaune est une **euh la : la : la : euh le :** cé- le : cépage de : euh la ville de Beaune je veux dire*

*er Beaune is **er : the : the : the : er the :** ty- the : type of vine of : er the city of Beaune I mean*

There are 679 filled pauses in all, 591 of which are separate sequences. Table 2 shows the number of pauses of each type.

Table 2. Types of filled pauses

Type	N	sub-type	N
lengthening	230	*simple*	216
		complex	14
euh	323	*simple*	305
		complex	18
combination	38		
Total	**591**		

5. Interaction study

Silent and filled pauses can hardly be studied separately:

- 380 of the 1163 silent pauses of our corpus occur next to a filled pause (33%).
- Conversely, 344 of the 591 sequences of filled pauses are next to a silent pause or contain a silent pause (58%).

This shows the importance of the interaction between these two types of pauses.

5.1. Two types of filled pauses

We think it is important to separate filled pauses (or sequences) into two functionally different types. Filled pauses that are within a speech segment are the most common type (478 out of 591 sequences, or 81%). They mark an interruption that may or may not be followed by a repetition and/or a repair:

*euh Beaune est une **euh la : la : la : euh le :** cé- le : cépage de : euh la ville de Beaune je veux dire*

*er Beaune is **er : the : the : the : er the :** ty- the : type of vine of : er the city of Beaune I mean*

However, in 113 out of the 591 sequences (19%), filled pauses occurred at the beginning of a segment:

*enfin bon voilà euh le dé*but de mon voyage ça a été ça ++ || **euh** et après bon ben après les choses se sont mises en place hein*

*well then er the start of my trip went well ++ || **er** and after well after things got into place eh*

In such cases, the interruption does not result from a difficulty in setting up the end of the segment in a lexical or syntactic sense, but is rather a way of "filling in" during the time it takes to make up the rest of the speech, and of preventing others from interrupting. This phenomenon is quite frequent since it affects 133 out of 822 segments (14%). These 113 cases are split into two groups:

- 97 sequences starting with *euh*;
- 16 sequences starting with a type of lengthening.

At least in our corpus, these types of lengthening are all monosyllabic words, mainly connectives such as *and*, *but*, *then* (11 out of 16 cases). The rest is/are function words that introduce phrases (*of*, *where*, *the*).

5.2. Role of silent pauses

As shown above, short silent pauses are never demarcative. Among the 318 non-demarcative silent pauses left, 289 (91%) are associated with a filled pause. In a large majority of cases (257 out of 289 or 89% of pauses), a filled pause comes before the silent pause:

*on l'ap- au départ on faisait euh le **le:** + euh le: ma*cérer le: le: **le poulet:** + le poulet bien sûr qui est issu aussi de la Bourgogne puisque **euh** + le poulet de Bresse*

*at the beginning we made er the **the:** + er the: macerate the: the: **the chicken**: + the chicken: of course that comes from Burgundy since **er** + the chicken from Bresse*

In only 32 cases (11%), there is a sudden interruption and the filled pause occurs immediately after:

et il a apprécié + euh ce Corton blanc

*and he appreciated + **er** this white Corton*

These cases, which seem to occur after a major intonation change and/or a stressed syllable, should be studied in detail.

Let us take a deeper look at the 29 cases of non-demarcative silent pauses that are not associated with a filled pause. They can be broken down as such:

1. In 13 cases, the speaker hesitates, but the silent pause is in fact associated with another sentence planning marker:

- word fragment (8 cases)

  *et puis là on vit **au jour le j- + au jour** le jour*

  *and then now we live **one d- + one day** at a time*

- repetition (5 cases)

  *donc s- nous avons un rôle **de + de** soutien de marché*

  *thus we have a role **of + of** supporting the market*

2. In 11 other cases, the silent pause is associated with a *discourse marker*:

- onomatopoeia (*pff*, etc.) (2 cases)

alors que moi ça me dérange pas du tout au contraire: ***pssff*** *+ je trouve que chacun a ses limites*

whereas that does not bother me at all on the contrary: ***pssff*** *+ I think everyone has their limits*

- particle (*ben*, *hein*, etc.) (3 cases)

  elle apprendra comme moi sur le tas ***hein +*** *de toute façon*

  she'll learn as I have on the fly ***eh*** *+ anyway*

- focal stress (6 cases)

  j'avais ***+ te*llement*** *soif de cette liberté-là*

  I longed + ***so much*** *for this type of freedom*

This last phenomenon is quite interesting and is worthy of a more in-depth study using a larger set of data. We hypothesize that it helps reinforce how well the stress is perceived and allows the speaker to fully reload his/her pulmonary capacity before or even after the stress (we mark the stress with a star after the stressed syllable):

3. Only 5 out of the 1163 silent pauses of the corpus remain unassociated with any other cue:

elle va se sentir dans ***un + endroit*** *de confiance donc*

she'll feel in ***a*** *+* ***secure*** *place then*

In these 5 cases there is a syntactic cohesion of the segment (there even is a liaison in the example above). There is no intonative discontinuity at the point of the silent pause. It would be worthwhile to find more examples and make a minute analysis of this phenomenon from an acoustic and syntactic aspect. Nevertheless, one conclusion that can be drawn is that there is no perceived hesitation in the cases we observed. These pauses act exactly like short pauses, and in fact they do not last long since their duration is below the (geometric) mean (200 to 470 ms).

6. Conclusion

Based on a subset of the *Corpus de référence du français parlé* lasting about one hour and involving 10 speakers (5 male and 5 female), the present study shows that a silent pause by itself never serves as a hesitation or sentence planning marker. It has that function only when coupled with other markers, mainly filled pauses (syllabic lengthening and the quasi-lexical item *euh*). Other cues are also associated with silent or filled pauses, such as word fragments, repetitions, or quasi-lexical items like *well*, *eh*, *pff*, etc. An extensive study on the phenomena associated with sentence planning should be conducted.

7. References

[1] Barik, H. C. 1968. On defining juncture pauses: a note on Boomer's "Hesitation and grammatical encoding". *Language and Speech, 11*, 156:159.

[2] Boomer, D.S. 1965. Hesitation and grammatical encoding. *Language and Speech, 8*, 148:158.

[3] Campione, E. 2001. *Étiquetage semi-automatique de la prosodie dans les corpus oraux: algorithmes et méthodologie.* Thèse de doctorat, Université de Provence, Aix-en-Provence.

[4] Campione, E., & Véronis , J. 2002. A Large-Scale Multilingual Study of Silent Pause Duration. In B. Bel & I. Marlien (Eds.), *Proceedings of the Speech Prosody 2002 conference* (pp. 199-202). Aix-en-Provence: Laboratoire Parole et Langage.

[5] Campione, E., & Véronis, J. 2004. Semi-automatic tagging of intonation. In G. Sampson & D. McCarthy (Eds.), *Readings in Corpus Linguistics* (pp 462-473). London: Continuum.

[6] Candea, M. 2002. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané.* Thèse de doctorat, Université Paris III.

[7] Clark, H. & Clark, E. 1977. *Psychology and Language*. New York : Harcourt, Brace, Jovanovich.

[8] Cutler, A. 1998. The recognition of spoken words with variable representations, *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech* (pp. 83-92). Aix-en-Provence, France.

[9] DELIC. 2004. Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé*, 18, 11-42.

[10] Duez, D. 1982. Salient pauses and non salient pauses in three speech style. *Language and Speeh*, 25(7), 11:28.

[11] Duez, D. 1998. The aim of SPoSS, *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech* (pp.VII-IX). Aix-en-Provence, France.

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

[13] Goto, M., Itou, K., & Hayamizu, S. 1999. A Real-time Filled Pause Detection System for Spontaneous Speech Recognition, *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech '99)* (pp. 227-230). Budapest.

[14] Guaïtella, I.1991. *Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée.* Thèse de doctorat, Université de Provence, Aix-en-Provence.

[15] Hieke, Kowal, & O'Connell, D. C. 1983. The trouble with "articulatory" pauses. *Language and Speech, 26*, 203:214.

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

[17] Morel, M. A., & Danon-Boileau, L. 1998. *Grammaire de l'intonation. L'exemple du français.* Paris : Ophrys.

[18] Quimbo, F. C. M., Kawahara, T., & Doshita, S. 1998. Prosodic analysis of fillers and self-repair in Japanese speech, *Proceedings of the International Conference on Spoken Language Processing (ICSLP).* Sydney, Australia.

[19] Vaissière, J. 1991. Rhythm, accentuation and final lenghtening in French. In J. Sundberg & L. Nord & R. Carlson (Eds.), *Music, Language, Speech and Brain*. Macmillan Press.

[20] Zellner, B. 1998. *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de doctorat, Université de Lausanne, Lausanne.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 47–51.

Inter- and intra-language acoustic analysis of autonomous fillers

Maria Candea[1], Ioana Vasilescu[2], Martine Adda-Decker[3]

[1] Paris 3 – EA1483, 13 rue de Santeuil, bur.431, 75005 Paris, [2]LTCI-ENST, 46, rue Barrault, 75634 Paris cedex 13, [3]LIMSI-CNRS, bat. 508, BP 133, F-91403 Orsay cedex

Abstract

The present work deals with autonomous fillers in a multilingual context. The question addressed here is whether fillers are carrying universal or language-specific characteristics. Fillers occur frequently in spontaneous speech and represent an interesting topic for improving language-specific models in automatic language processing. Most of the current studies focus on few languages such as English and French. We focus here on multilingual fillers resulting from eight languages (Arabic, Mandarin Chinese, French, German, Italian, European Portuguese, American English and Latin American Spanish). We propose thus an acoustic typology based on the vocalic peculiarities of the autonomous fillers. Three parameters are considered here: duration, pitch (F0) and timbre (F1/F2). We also compare the vocalic segments of the fillers with intra-lexical vowels possessing similar timbre. In this purpose, a preliminary study on French language is described.

1. Introduction

Among various hesitation or "edition" phenomena, the one we analyze here is widely encountered in world's languages, i.e. the insertion at any moment within spontaneous speech of a long and stable vocalic segment, defined as a type of filler. The role of this item is "to announce the initiation of what is expected to be a […] delay in speaking" [1]. Such elements have no lexical support and are hence distinguished from the lengthening of a vocalic segment belonging to a particular lexical item (most often a function word). Most of the studies conducted on large spontaneous speech corpora have focused on English or French [2], [3], [4], [5], [6], [7], even if recent description can be found in other languages (see for example [9] or the proceedings of the DiSS03 workshop [8]).

We address here the question whether the autonomous fillers are carrying universal acoustic characteristics or language-specific information. They occur frequently in spontaneous speech, i.e. about five percent in spontaneous corpora, and this proportion can increase according to the spontaneous speech communication situation. We are also interested in the modeling problem of these phenomena in a language identification context. The question is then whether autonomous fillers (such as *uh/um/er* in English and *euh* in French) deserve language-specific models or whether a language-independent filler model is more appropriate.

The vocalic segment of autonomous fillers is generally lengthened. This segment can occur alone or surrounded by additional segment as nasal coda in English (*um*) and represent in our terminology the *vocalic support* of the filler.

More precisely, we study in this paper the vocalic peculiarities of autonomous fillers in several languages, i.e. the realization of a central vs. non-central timbre of their vocalic support. In previous studies [10], [11] we observed acoustic differences among the vocalic supports of the multilingual fillers. We also conducted perceptual experiments in order to test listeners' capacity at differentiating languages from isolated autonomous fillers without any context [12].

In the following section we describe the corpus and the methodology adopted. Section 3 is dedicated to the inter-language analysis of the acoustic characteristics of fillers. It will be followed by intra-language study carried on French (section 4). Finally, section 5 will summarize the current findings.

2. Corpus and methodology

A multilingual broadcast corpus has been gathered for the following eight languages: standard Arabic, Mandarin Chinese, French, German, Italian, European Portuguese, American English and Latin American Spanish. French and Arabic are French DGA resources, partially available via the ELDA linguistic resources agency. English, Spanish and Mandarin are excerpts from LDC Hub4 corpora. German, Portuguese and Italian BN data are resources acquired within various European FP5 LE projects (OLIVE, ALERT) or purchased from ELDA. The audio data correspond either to news data which is mainly prepared speech, or news-related shows containing more spontaneous speech specific items. From this multilingual corpus, a subcorpus of autonomous fillers has been extracted semi-automatically for the eight languages under consideration: fillers, which have been located automatically in aligned speech, are listened to and selected if the selection criteria are met.

Filler extraction is based on duration and autonomy criteria. 200ms has been considered as the minimum duration threshold. Items considered in this study as autonomous fillers are isolated from the speech context by silences in order to avoid lengthened words. Finally, 57 to 1889 occurrences per language have been selected for both genders (see table 1 below). The size of the present corpus allows exploring the questions mentioned in introduction. However some languages are less represented and as a general observation data from female speakers are less abundant. Current work conducted by the authors focuses on the size of the database on hesitations, which is progressively increased.

Table 1: Number of occurrences of hesitations per language (male and female speakers).

Language	Nb. of occurrences (M+F)
Arabic	246
Mandarin Chinese	89
French	1889
German	458
Italian	57
European Portuguese	64
American English	532
Lat. Amer. Spanish	93

A supplementary corpus has been extracted in French in order to conduct the intra-language analysis from about 6 hours various types of broadcast speech. In addition to the 1889 fillers (1509 from male speakers, 380 from female speakers) it contains also other intra-lexical vocalic segments with similar vocalic qualities, i.e. 1718 [ə-œ] (954 male, 754 female) and 1114 [ø] (923 male, 291 female). The intra-lexical vocalic segments have been extracted via the LIMSI speech alignment system. A duration criterion has been employed, i.e. we selected intra-lexical vocalic segments superior to 40 ms. As a general observation we can notice a higher representation of male compared with female speakers. The vowel proportion reflects the gender representation observed in broadcast-type corpora (about 70% male vs. 30% female speakers). Besides, the number of occurrences per vocalic timbre illustrates also the frequency of the analyzed segments in French. Indeed, the vowel [ø] is 10 times less frequent than [ə-œ].

The PRAAT software[1] has been used to extract the acoustic parameters comprising fundamental frequency (F0) and the first two formants (F1, F2).

3. Inter-language analysis: acoustic features of fillers in eight languages

Three parameters have been considered: the duration, the F1/F2 characteristics of the fillers' vocalic segments and the pitch (F0). Whereas pitch and duration are mainly useful to localize the fillers in the speech flow, the F1/F2 parameters potentially contain more language specific characteristics.

The F0 and the duration of the fillers do not show significant differences among the eight languages confirming previous findings [1], [2], [3], [6], [7], i.e. fillers are significantly longer than intra-lexical vocalic segments (see below French example) and have a flat F0 contour. The behavior of the two parameters seems to be identical across the eight languages. This results answer affirmatively to the first question mentioned above, i.e. duration and pitch tend to be universal criteria.

In return, the acoustic analysis of F1/F2 peculiarities of the fillers' vocalic segments reveals language-dependent characteristics.

The figures 1 and 2 below provide the mean values for F1 and F2 per language.

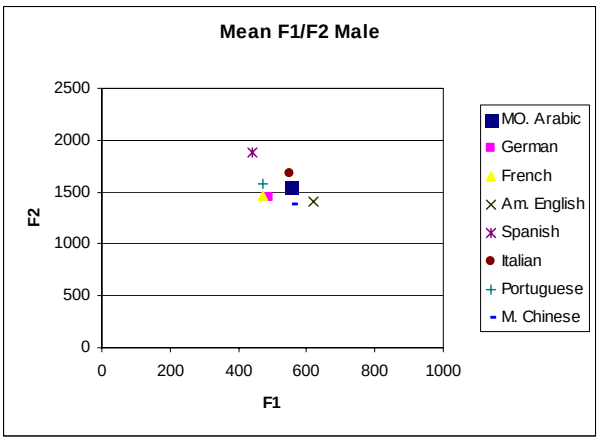


Figure 1: F1/F2 distribution of vocalic segments of autonomous fillers: all languages (mean values for male speakers)

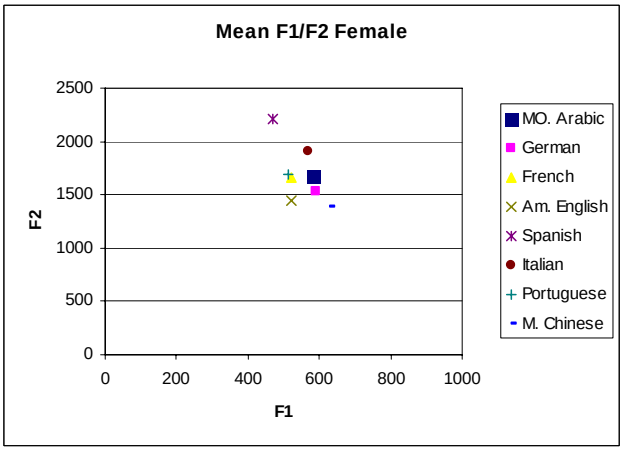


Figure 2: F1/F2 distribution of vocalic segments of autonomous fillers: all languages (mean values for female speakers).

Preliminary results strengthen the hypothesis of timbre differences across languages of the vocalic support of the autonomous fillers. Indeed, the central position does not seem to be a universal realization. These results tend to show that different languages analyzed here admit various vocalic realizations. The realizations can be central [ə] and/or correspond to other vocalic quality. We can hypothesize that the vocalic supports are vowels of the system. Spanish employs thus a mid closed vowel [e] and English makes use of low central vowels. Italian so far is the only language of the corpus with both central [ə] (which is not part of the Italian vocalic system) and non-central vocalic supports, i.e. the front mid open vowel [ɛ]. More data from other languages are needed to consolidate these hypotheses. Finally, the observed differences are not uniquely in terms of vocalic timbre. Language-specific features can be observed in the segmental structure of the fillers. French, for example, prefers a vocalic segment as filler realization, whereas English prefers vowels followed occasionally by a nasal coda consonant [m], which confirms observations made by [1] and [7]. In Portuguese as well, more complex diphthongized segments can be found.

[1] www.praat.org

To conclude, for some languages the vocalic support of the fillers might be a segment exterior to the vocalic system of the language (i.e. Italian in our corpus). However, all the eight languages seem to accept as fillers' vocalic support at least one of the vowels of their vocalic system. The vowel generally exhibits a timbre close to a quite central position. However the central position does not seem to be universal "rest position", but rather a language dependent realization.

In order to evaluate the relationship between the vocalic support of the filler and the vocalic system of the language we conducted a preliminary analysis on the French language. We compared thus the vocalic timbre of *"euh"* with the closest vowels of the system [ə], [ø] and [œ].

4. Intra-language analysis: vocalic support of French filler *"euh"* vs. vocalic system

As for the vocalic support of the fillers, the same parameters have been considered for the intra-lexical vowels of the system: duration, vocalic timbre (F1/F2) and pitch (F0).

Duration analysis confirms previous observation made by [2], [4], [5]. The distribution of the duration for fillers exceeds significantly the duration of intra-lexical segments as shown in Figures 3, 4 and 5 below. The duration criterion adopted here avoids fillers below 200ms. Fillers shorter than 200ms definitely exist in the spontaneous speech. However, as the extraction has been conducted automatically, we selected a threshold high enough to eliminate the potential confusion with intra-lexical segments.

In order to evaluate the amount of fillers potentially eliminated by the selected threshold, we proceeded to a listening of a 45 minutes speech sample and we manually extracted fillers shorter than 200ms. It appears that 14.7% of the fillers show a duration between 150 and 200ms and 11.6% show durations inferior to 150ms. This observations support the hypothesis that fillers are mainly longer than 200ms. The listening experiment confirmed though that by selecting a 200ms threshold we eliminated about 25% of real fillers from our analysis.

We can notice thus that duration of a very large part of the fillers varies from 200 till 650 ms whereas intra-lexical segments rarely extend beyond 200ms. Fig. 4 and Fig. 5 show differences in the duration distribution for intra-lexical vowels. The duration distribution for the vowels [ə-@] (among them schwa segments are suppressible in speech) focuses close to the minimum segment duration whereas duration for [ø] has a broader distribution with a peak around 80ms. Schwa segments are suppressible in speech and belong to non accented syllables. The duration is thus the shortest among the analyzed segments, the exception being the lengthened schwa (i.e. the realization as fillers). The [ø] segments can occur in accented syllable as well, and the duration shows a more important variability in terms of realization.

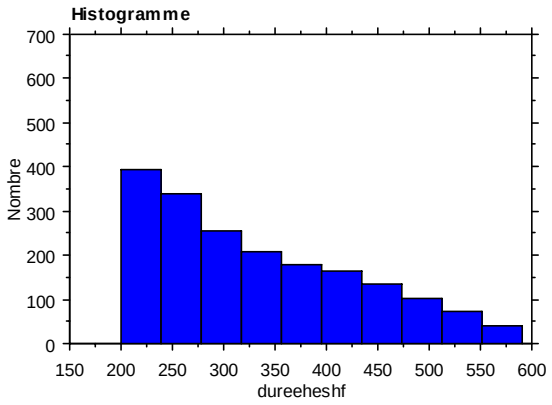


Figure 3: Duration above 200ms vs. nr of occurrences distribution for fillers (male/female)

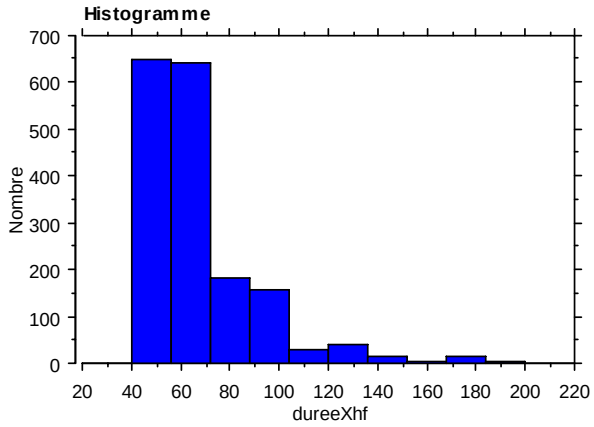


Figure 4: Duration vs. nr of occurrences distribution for intra-lexical [ə-œ] (male/female).

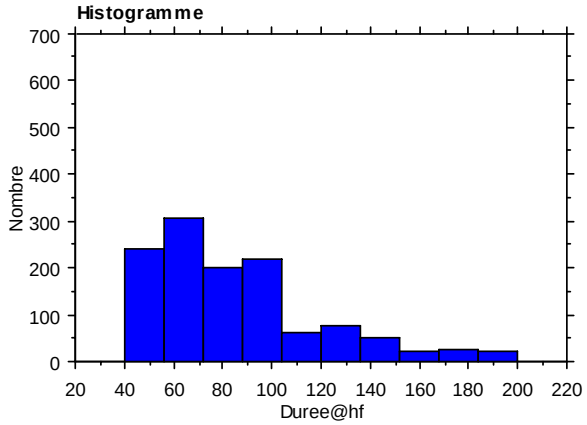


Figure 5: Duration vs. nr of occurrences distribution for intra-lexical [ø] (male/female).

Concerning **timbre** analysis of fillers and intra-lexical vowels, mean F1/F2 measures are shown in Table 1. The measured values are compared with reference values provided by [13] and by [14] for the French intra-lexical vowels.

Table 2: F1/F2 mean values and for male and female speakers of vocalic support of filler *"euh"* and of intra-lexical segments [ə-œ] and [ø] (Hz).

	F1 male/female	F2 male/female
Fillers	470/523	1464/1659
[ə-œ]	404/413	1421/1675
Fant [13]	*500/550*	*1450/1650*
Gendrot&al.[14]	*400/437*	*1444/1659*
[ø]	382/430	1465/1666
Fant [13]	*400/450*	*1450/1650*
Gendrot&al.[14]	*375/417*	*1465/1677*

Values presented in Table 2 do not show any notable difference in terms of back/front distribution, i.e. values on F2 for both filler and intra-lexical segments are analogous. On the F1 axis (open/close) a difference is observed in particular for male speakers, in vocalic support of the fillers vs. intra-vocalic segments: the first ones are more open. This difference could be related to the duration, as the intra-lexical vocalic segments are significantly shorter than the fillers. Among the intra-lexical vocalic segments, [ə-@] are more open than [ø]. However, these differences are not statistically significant (independent t-test).

Mean values calculated for F1 for the intra-lexical vowels are similar to the values observed by [14], for male and female speakers. Mean values for F2 are similar to values observed by both [13] and [14].

We considered as well the F3 in order to evaluate if a distinction could be made in terms of rounded/non-rounded opposition. We compared thus the measurements automatically completed by [14]. They reach mean values of 2500 Hz for both [ə-œ] and [ø], consequently we do not have evidence of a timbre difference of the analyzed vowels.

Finally, measurements on the **pitch (F0)** have been calculated for both vocalic segments of the fillers and intra-lexical vowels. Table 3 shows mean values and standard deviations in F0 for the three types of segments. F0 differences among vowels are not significant (independent t-test) and the current data do not allow at concluding on the fillers mean F0 peculiarities compared with the intra-lexical vocalic segments.

Table 3: F0 mean values and standard deviation for F0 (male and female speakers) for vocalic support of the filler *"euh"* and for intra-lexical segments [ə-œ] and [ø] (Hz).

	F0mean Male/Female	St.Dev. Male/Female
Fillers	155/179	97,3/68,5
[ə-œ]	114/192	48,6/38,3
[ø]	144/219	45,4/56

This preliminary result suggests that F0 distribution for the fillers correspond to a larger "ambitus" than for the intra-lexical segments. This variability is suggested by the standard deviation values globally higher for the fillers. These remarks concern more male than female speakers. However, if we listen to the fillers providing high F0 values (>300Hz), they seem to correspond to detection errors of PRAAT more than to extreme articulations. In addition, we could notice that number of fillers show both a low perceived F0 and an irregular voice quality (i.e. vocal fry). Most of the F0 detection errors stem from these type of segments and in a further work we intend to compute male F0 mean without these potential "erratic" segments.

In order to get an overall impression of the irregular voice quality in the production of the fillers, we compared the number of F0 detection errors for *"euh"*, [ə-œ] and [ø]. They correspond to the "undefined" values provided by Praat for the speech samples for which the software could not compute the F0. These detection errors concern more often the fillers than the intra-lexical vocalic segments. Detection errors for [ə-œ] and [ø] represent 3% of the production of the male speakers and 0,5% of the female speakers. In return, for the fillers, F0 detection errors represent 11,5% of the productions of the male speakers and 8,2% of the female speakers. These findings corroborate the hypothesis of an unstable voice quality in the production of the fillers, which could be either vocal fry, creaky or breathy. Besides, these observations confirm previous remarks made by [6] and [7] about the American English: the intra-lexical vowels are produced more often with a modal voice quality than fillers. Finally, differences in detection errors for male vs. female speakers suggest that women might tend to pay more attention to the degree of control of their "disfluent" productions than men. Our further studies will consider more deeply the voice quality aspect of the fillers compared with "fluent" speech. We hypothesise that fillers might be produced with a limited articulator and/or airflow.

5. Conclusions

In this paper we presented inter- and intra-language acoustic analysis of autonomous fillers. The question which has been addressed is whether the fillers possess universal acoustic characteristics or if they are language-specific phenomena. The current work answers partially to the question.

Fillers are frequent phenomena in spontaneous speech and show regular patterns in terms of duration, fundamental frequency and vocalic timbre which tends to be central for most languages considered in this study. Whereas F0 and duration are similar among languages, language-specific acoustic characteristics can be observed in terms of vocalic quality and segmental structure of the fillers. Among the eight languages analyzed in this paper, non-central vocalic timbre characterize at least two of them, Spanish and to a lesser extend, English.

The comparison of the vocalic support of French fillers with the closest intra-lexical vocalic segments [ə], [ø] and [œ] allowed at observing differences in terms of duration, pitch and voice quality. Differences on the F1 axis were also observed, but they are not statistically significant.

Interesting questions still remain open. They concern for example the relationship between the vocalic timbre of the fillers and the vocalic systems. The comparison of the fillers' vocalic supports and the intra-lexical vowels will be developed as well on other languages of the multilingual corpus in order to determine the relationship between the fillers and languages' vocalic system.

Another aspect related to the universal characteristics of the fillers concerns the relationship between the filler and the so-called "articulatory rest position". The question addressed is whether this position exists and whether the fillers are close to it and thus they might represent universal realizations across languages. The preliminary results presented in this paper seem to exclude the universal timbre of the fillers. In addition, a recent study on English and French articulatory

settings [16] suggests that a clear correlation between "hesitation vowel" and the "articulatory rest position" could not be proved. The "articulatory rest position" itself seems to be significantly different for each language. Such studies are still in progress and may provide interesting information about an eventual influence, in a given language, of the global speech posture and articulatory rest position on vocalic fillers.

Finally, further studies should consider other aspects allowing at describing fillers in the context of the so called "disfluencies phenomena" which characterize the spontaneous speech. It would be thus interesting to observe those aspects which differentiate fillers from vocalic lengthening. The relationship between vocalic fillers such as *huh, hum, euh* and language word hesitation (as, for example in Japanese [9]) would be as well examined.

6. Acknowledgements

This research has been carried out in the context of the MIDL (Modélisations pour l'IDentification des Langues) project, supported by a CNRS interdisciplinary program and involving several French laboratories (LIMSI-CNRS, LTCI-ENST, CTA-DGA, LPP Paris3 and EA 1483 – Paris3). Its aim was to bring together linguistic and computer engineering knowledge in order to increase our knowledge in human language identification and to contribute to the domain of automatic language identification.

The authors want to thank Cédric Gendrot (ILPGA, Paris III) for his help in the acoustic analysis of the corpus.

7. References

[1] Clark H.H., Fox Tree J.E. 2002. Using uh and um in spontaneous speaking, *Cognition* 84, 73-111.

[2] Adda-Decker et al. 2003. A Disfluency study for cleaning spontaneous automatic transcripts and improving speech language models, *DISS'03, Göteborg, Sweden (Papers in Theoretical Linguistics* 90: 67-70).

[3] Shriberg, E., Bear, J., Dowding, J. 1992. Detection and correction of repairs in human-computer dialog, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Delaware.

[4] Candea, M. 2000. *Contribution à l'étude des pauses silencieuses et phénomènes dits "d'hésitation" en français oral spontané.* PhD thesis, University of Paris 3-Sorbonne nouvelle.

[5] Guaïtella, I. 1991. Hésitations vocales en parole spontanée: réalisations acoustiques et fonctions rythmiques, *Travaux de l'Institut de Phonétique d'Aix*, vol.14: 113-130.

[6] Shriberg, E. 1999. Phonetic consequences of speech disfluency, *ICPhS'99*, San Francisco.

[7] Shriberg, E., 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies, *Journal of the International Phonetic Association*, 31/1.

[8] Eklund R. editor, 2003. Disfluencies in Spontaneous Speech, DiSS workshop, *Proceedings of DISS'03, (Papers in Theoretical Linguistics 90*), Göteborg.

[9] Watanabe, M. 2003. The constituent complexity and types of fillers in Japanese, *15th ICPhS*, Barcelone.

[10] Clerc-Renaud J. Vasilescu I., Candea M., Adda-Decker M. 2004. Etude acoustique et perceptive des hésitations autonomes multilingues, *XXVes JEP*, Fès Morocco.

[11] Vasilescu I., Candea M., Adda-Decker M. 2004. Hésitations autonomes dans 8 langues : une étude acoustique et perceptive, *Workshop MIDL04*, Paris F.

[12] Vasilescu I, Candea M., Adda-Decker M., 2005. Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages, *Interspeech 2005* Lisboa, Portugal.

[13] Fant, G., 1973. *Speech sound and features*, MIT Press, Cambridge, USA.

[14] Gendrot C., Adda-Decker M., 2004. Analyses formantiques automatiques de voyelles orales: évidence de la réduction vocalique en langues française et allemande, *Workshop MIDL04*, Paris France.

[15] Calliope, 1989. *La parole et son traitement automatique*, Paris, Masson ed.

[16] Gick B., Wilson I., Koch K., Cook C., 2004. Language specific articulatory settings: evidence from inter-utterance rest position, *Phonetica*, 61 (4), 220-233.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 53-58.

Prosodic parallelism as a cue to repetition and error correction disfluency

Jennifer Cole, Mark Hasegawa-Johnson, Chilin Shih, Heejin Kim, Eun-Kyung Lee, Hsin-yi Lu, Yoonsook Mo, Tae-Jin Yoon

University of Illinois at Urbana-Champaign, USA

Abstract

Complex disfluencies that involve the repetition or correction of words are frequent in conversational speech, with repetition disfluencies alone accounting for over 20% of disfluencies. These disfluencies generally do not lead to comprehension errors for human listeners. We propose that the frequent occurrence of parallel prosodic features in the reparandum (REP) and alteration (ALT) intervals of complex disfluencies may serve as strong perceptual cues that signal the disfluency to the listener. We report results from a transcription analysis of complex disfluencies that classifies disfluent regions on the basis of prosodic factors, and preliminary evidence from F0 analysis to support our finding of prosodic parallelism.

1. Acoustic-prosodic correlates of disfluency

Disfluency occurs in spontaneous speech at a rate of about one every 10-20 words, or 6% per word count [17], yet this interruption of fluent speech does not generally lead to comprehension errors for human listeners. Recent research has shown that important cues to disfluency can be found in the syntactic and semantic structures conveyed by the word sequence, and in the phonological and phonetic structures signaled by acoustic features local to the disfluency interval. These cues identify the components of the disfluent region --- the reparandum (REP), edit phrase (EDT), and alteration (ALT) --- and their junctures. Work on automatic disfluency detection has shown that the most successful approach combines both lexical and acoustic features, with explicit models of the lexical-syntactic and prosodic features that pattern systematically with disfluent intervals [1,6].

Of the acoustic-prosodic correlates of disfluency, the post-reparandum pause (filled or unfilled) has been studied the most extensively. Nakatani & Hirschberg's [12] detailed acoustic and classification studies examine duration, F0 and energy, and also report unusual patterns of lengthening, coarticulation, and glottalization near the interruption point of a disfluency. In this paper we examine the nature of prosodic correlates of disfluency in the characteristic patterns of F0, duration and energy that identify and distinguish among various types of disfluency involving word repetition and error correction.

There are distinct types of disfluency that can be characterized in terms of their form and function. Shriberg [16, 17] classifies the disfluencies of the Switchboard corpus into six categories: filled pause ("uh" and "um"), repetition (of one or more words, without correction), substitution (repetition of zero or more words, followed by the correction of the last word in the disfluent interval), insertion, deletion, and speech error. Other work identifies abandonment (fresh start) disfluencies, in addition [6,11,18]. These distinct types of disfluency may be caused by different psychological processes. Levelt [9] suggests that corrections of a single word may result from monitoring of the phonetic plan, while corrections that involve repair or abandonment of an entire phrase may result from monitoring of the pre-syntactic message. Clark & Fox Tree [3] and Clark & Wasow [4] propose a different psychological account for filled pause and repetition disfluencies. In these accounts filled pauses like "uh" and "um" are phonological words that are used by the speaker to signal a delay in the preparation of the upcoming speech. Repetition disfluencies occur when the speaker makes a premature commitment to the production of a constituent, perhaps as a strategy for holding the floor, and then hesitates while the appropriate phonetic plan is formed. The continuation of speech is marked by "backing up" and repeating one or more words that precede the hesitation, as a way of restoring fluent delivery. Henry & Pallaud [7] support the findings of Clark & Wasow [4] by demonstrating that morphological, syntactic, and structural features strongly differentiate repetition disfluencies from word fragment disfluencies. Clark & Wasow [4] note that repetition disfluencies are four times as common as repair disfluencies; they suggest that a small number of repetition disfluencies may be "covert repairs" [9], but that most repetitions are more closely related to filled pause disfluencies than to speech repairs.

The acoustic-prosodic features that serve to cue disfluency vary according to the type of disfluency. Levelt & Cutler [10] observe a contrastive emphasis on the repair segment of an error-correcting disfluency, manifest in increased F0, duration and amplitude. Shriberg [15] and Plauché & Shriberg [13] find that F0 contours, word durations, and the distribution of pauses serve to differentiate among three types of repetition disfluencies. Shriberg [15] describes repetition disfluencies that signal covert repair as having a characteristic reset of the F0 contour to a high, phrase-initial value at onset of the alteration. Similarly, Savova & Bachenko [14] propose an "expanded reset rule," according to which "alteration onsets are dependent on both reparandum onsets and reparandum offsets," echoing the observation of Shriberg [15] that when speakers modify the duration of a repeated word in a repetition disfluency, "they tend to do so in a way that preserves intonation patterns and local pitch range relationships."

In our study of prosody and disfluency in the Switchboard corpus of conversational telephone speech, we observe parallelism in the prosodic features of the REP and ALT phases as characteristic of repetition and error correction disfluencies. Highly similar F0 patterns express a parallel intonation structure that cues the relationship between the REP and ALT for the majority of repetition and error correction disfluencies we have observed. We propose an extended typology of repetition disfluencies in this paper, based on prosodic comparison of REP and ALT. Section 2 describes the methods of our transcription study of disfluency in Switchboard, and section 3 presents frequency data on five sub-categories of repetition and error correction disfluency that are prosodically distinguished based on a comparison of the prosodic features of the REP and ALT intervals. Section 4 reports on preliminary quantitative evidence from F0 data that support our analysis based on perceptual transcription.

2. Method

2.1. Corpus

Switchboard is a corpus which consists of 2500 spontaneous informal telephone conversations [5]. We selected 70 sound files from those conversations, representing 58 different speakers. Within each file we used a random process to excerpt a two minute sound segment. These short files were transcribed for disfluency intervals by the authors, all of whom are trained in acoustic phonetics with prior experience in prosodic transcription using ToBI annotation conventions. 3 transcribers labeled disfluencies for the entire two-minute duration of 10 files each (for a total of 60 minutes of speech) and 5 transcribers labeled only for the first talker turn of duration ranging from 3 to 60 seconds in each of 10 files (for approximately 25 minutes of speech). All eight labelers participated in a series of three group training sessions to assure consistency of labeling criteria, and two group sessions were held for the resolution of problem cases raised by individual labelers.

2.2. Labeling Criteria

Disfluencies are classified by their function into two types, hesitation and repair. These functional categories divide into several subtypes based on lexical and prosodic form. Hesitation disfluencies are classified as repetition, lengthening, silent pause and filled pause. Repair disfluencies are classified as error correction and abandonment. Classification was based on lexical, syntactic, and prosodic factors. Lexical factors are the presence of a repeated word, an error-correcting word substitution, or a filled-pause phrase like "um" or "ah". Syntactic criteria were used to identify instances of phrase abandonment followed by fresh restart and to identify the REP-ALT correspondence in error-correction. Prosodic factors were used to identify lengthening, and provided additional evidence for some cases of error correction (with prosodic emphasis on ALT) and abandonment (with truncation of an intonational tune at the abandoned edge). Labeling was done on the basis of listening and visual inspection of the waveform, spectrogram, F0 and intensity contours, using Praat [2]. The disfluency labels were entered on two tiers in the TextGrid associated with each wave file, and disfluency intervals (REP, EDT, ALT) were aligned with the beginnings and endings of the associated word intervals. Table 1 shows the typology of disfluencies by function and form and the labeling conventions used.

Table 1. Typology of Disfluencies and Labeling Convention

Type of Disfluency		**Labeling**			
		1st Tier	**2nd Tier**		
Hesitation	Repetition	hesi-r	REP	EDT	ALT
			REP	ALT	
	Lengthening	hesi-l			
	Silent Pause	hesi-s			
	Filled Pause	hesi-f			
Repair	Error Correction	repair-e	REP	EDT	ALT
			REP	ALT	
	Abandonment	repair-a	REP	EDT	
			REP		

Labels on the first disfluency tier identify the type of disfluency (e.g., hesi-r for Hesitation Repetition), while the components of complex disfluencies were individually segmented on the second disfluency tier. A complex disfluency always includes a reparandum (REP) and an alteration (ALT), and may also include an edit phrase (EDT). Hesitation Repetition disfluency labeling is illustrated in Figure 1. The hesi-l label marks hesitation lengthening that can not be attributed to prosodic phrase-final lengthening based on tonal evidence and perceived disjuncture.

In addition, hesi-s denotes a sentence internal silence that interrupts an otherwise fluent phrase, and hesi-f marks an independent occurrence of filled pause expressions such as "um", "uh". For the repair category, repair-e marks an error followed by a self-correction (e.g. "he can stri- he can swing") and repair-a denotes a semantic and syntactic abandonment of the phrase (e.g. "they you know you can't live in Dallas").

…the kids	instead of	[sil]	instead of	teaching…
	hesi-r			
	REP	EDT	ALT	

Figure 1: TextGrid tiers for a Hesitation Repetition disfluency [Switchboard file: SW03719A.wav]

The EDT label marks the occurrence of filled pauses, silent pauses and editing expressions (e.g., "I mean", "you know") between REP and ALT.

Hesitation Repetition and Repair Error Correction disfluencies, the two disfluency types that have both REP and ALT intervals, were further broken down into five sub-classes based on comparison of prosodic features between REP and ALT. These five sub-classes, listed in Table 2, were proposed on the basis of our earlier exploratory analyses with Switchboard samples; the present study was designed to test the adequacy and acoustic correlates of the proposed classification scheme. Data used in the exploratory analysis were not included in the present study. Prosodic features were assessed on the basis of listening in conjunction with visual inspection of the F0 and intensity contours, spectrogram and waveform. Repetitions in which the ALT and REP were judged to have highly similar prosodic patterns, with identical intonation features in a ToBI transcription, were assigned the label suffix '-same'. An example pitch track from a Repetition-Same disfluency is shown in Figure 2. The '-fp' label was used to label examples where the ALT interval had prosody characteristic of a filled pause: low intensity, and low, flat F0, with reduced consonant or vowel articulations. (This pattern was not observed with Repair Error Correction disfluencies.) The '-ip' label was used to label cases where the REP was perceived as the final word in a well-formed intermediate phrase, based on the F0 contour and perceived disjuncture between REP and the onset of ALT. The '-exaggerated' label was applied to examples in which the ALT displayed a similar but exaggerated version of the prosodic pattern of the REP, typically with increased duration, intensity and higher F0 values. In many cases these examples would receive the same ToBI transcription for REP and ALT, with differences in F0 scaling. Finally, the label '-change' was used for examples where the ALT differed prosodically from the REP in its accentuation (different type or location of accent, or presence vs. absence of accent). Hesitation Repetitions labeled in the change subcategory sounded much like error corrections, where the correction was at the level of pragmatic meaning expressed through accent, rather than at the level of word or syntactic meaning. For all disfluency types, the REP interval was further identified as ending in a word fragment (-frag), or a complete word (-nonfrag), but comparison of fragment and non-fragment tokens is not presented here.

Table 2. Types of Repetition: Prosodic Classification

Hesitation-Repetition	**Repair-Error Correction**
hesi-r-same	repair-e-same
hesi-r-fp	repair-e-fp
hesi-r-ip	repair-e-ip
hesi-r-exaggerated	repair-e-exaggerated
hesi-r-change	repair-e-change

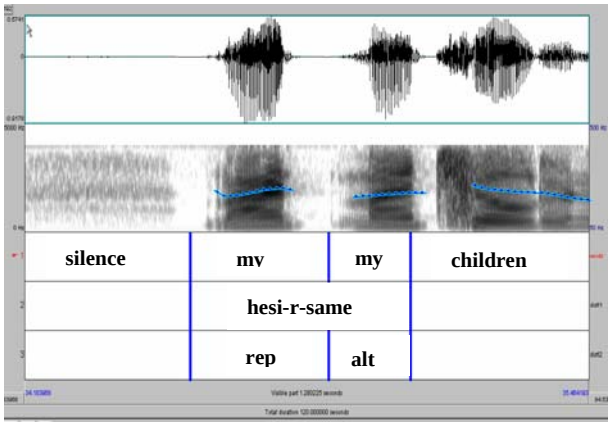


Figure 2. Example of highly similar F0 tracks on REP and ALT (my…my) in Hesitation-repetition-same disfluency: "[sil] my my children…" [Switchboard file: SB03633b]

3. Results

Table 3 provides the number of tokens of each type of disfluency labeled in the corpus, pooling data from all labelers. The most frequent type of disfluency in this corpus is Hesitation, with Silence the most frequent sub-type. Repetitions and Filled Pauses are also frequently occurring Hesitation types. Among Repair disfluencies, Abandonment is fairly common, while error correction and lengthening are infrequent.

Table 3. Distribution of the types of disfluency

		Frequency	Percentage
Hesitation	Silence	299	31.54 %
	Repetition	205	21.62 %
	Filled pause	208	21.94 %
	Lengthening	52	5.48 %
Repair	Abandonment	123	12.97 %
	Error correction	50	5.27 %
Total		948	

Table 4 presents the total number of REP, EDT and ALT intervals, automatically extracted from Hesitation-Repetition and Repair-Error Correction disfluencies in the transcription files. The number of REP and ALT intervals are not equal, due to the occurrence of multiple repetition tokens that contain more than two instances of the repeated word (e.g., "I I uh I tried to…"). For multiple repetitions all but the non-final repetition are coded as independent REP intervals, with the final repetition coded as ALT.

Table 4. Distribution of REP, EDT, and ALT for Hesitation Repetition and Repair Error Correction

	REP	EDT	ALT
Repetition	216	94	205
Error Correcction	51	18	50
Total	267	112	255

The distribution of disfluencies in our corpus over the 10 prosodically-defined sub-classes is shown in Table 5. Repetitions in which REP and ALT have the same prosody (*same*) are the most numerous, and are as frequent as the total of repetitions that mimic filled pauses, cross intermediate phrase boundaries, display exaggerated prosody, or display changed prosody on ALT. These results indicate that while a variety of prosodic patterns are observed over REP and ALT in complex disfluencies, the most common pattern perceived by labelers is that of prosodic parallelism, corresponding to the *same* label, which occurs most frequently with hesitation-repetition disfluencies, but which also occurs as the most frequent pattern with repair-error correction disfluencies.

Table 5. Number of Hesitation Repetition and Repair Error Correction examples by prosodic sub-class.

	same	**fp**	**ip**	**exag**	**change**
Hesi- r	102	22	21	32	28
Repair-e	12	0	4	5	9

The reliability of the labeling scheme was tested by assessing the agreement between pairs of labelers who labeled the same files. Agreement was assessed on a subset of the files labeled for disfluency. Specifically, one or two files from each labeler's bunch were randomly assigned to each of the other labelers for an independent labeling. This second-pass labeling resulted in a set of 59 files labeled independently by two labelers, utilizing all possible labeler pairs. The files labeled in the second-pass labeling were limited to a short interval of between 1.94 – 58.82 seconds, representing the first talker turn of the file, for a total of 1,298 seconds of speech. These files were labeled by second labelers using the same labeling scheme as the first pass labeling. Based on the second-pass labeling of this subset of files, the agreement rate for disfluency type (e.g., hesi-r or repair-a) was 86.82%. When taking into account the subclasses of disfluency in Table 2, the agreement rate among labelers was 85.07 %.

4. F0 Analysis

F0 values were compared between REP and ALT as an empirical measure of intonational similarity. This section describes the method for extracting smoothed F0 contours, time normalization, and a measure of F0 contour difference.

F0 is calculated from short-term autocorrelation and smoothing with Praat [2]. Null values of F0 at the start of a REP or ALT interval, which reflect silence or voiceless segments, are eliminated before aligning initial non-null zero values of REP and ALT. Non-initial frames with null F0 values are discarded in the comparison of REP and ALT F0 contours. Also discarded are any frames in which delta-F0 after smoothing is unexpectedly high or low (change of more than 100 Hz in 10 ms). Four methods of pitch comparison are used in this study: trimmed F0 difference, time-normalized F0 difference, trimmed F0 distance, and time-normalized F0 distance. Trimming and normalization are two methods we use to guarantee that the F0 contours of REP and ALT that we are comparing have the same length. For trimmed F0 analysis, the F0 trajectories of the REP and ALT are compared, where the longer F0 trajectory is trimmed to match the length of the shorter F0. For time-normalized F0, the shorter F0 trajectory (REP or ALT) is time-normalized to match the length of the longer one by using the linear interval interpolation. Because the trimming and time normalization methods do not result in any significant difference in our analyses, only the trimming method is reported here.

The mean F0 difference of REP and ALT is obtained by:

$$\Delta F0 = \frac{\sum_{i,j=1}^{n} (F_0^{(i)} - F_0^{(j)})}{n}$$

Here, *i* is the *i*th sample of REP and *j* is the corresponding sample of ALT, and *n* is the number of samples in the F0 contours (equal for REP and ALT). The F0 difference value is not squared in the equation, because we want to preserve the sign to distinguish cases where REP F0 is scaled higher than ALT from cases which have the opposite scaling relation. We have visually inspected the F0 contour of the REP and ALT sections to be certain that we do not encounter cases where the F0 contours have opposite slopes. In the F0 difference calculation the first F0 values correspond to those of the reparandum (REP) and the second F0 values correspond to those of the alteration (ALT). Thus, when $\Delta F0 > 0$, REP is higher in F0 than ALT and when $\Delta F0 < 0$, ALT is higher in F0 than REP. Figure 3 shows overlaid time-normalized F0 contours for one REP-ALT pair.

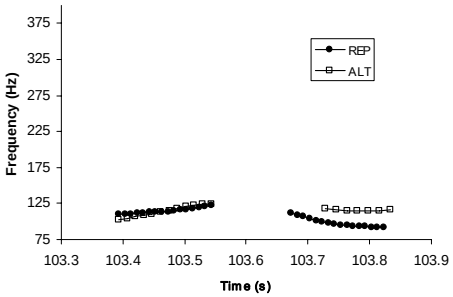


Figure 3. F0 trajectories during the REP (circles) and ALT (squares) segments of a repetition disfluency. Segments are aligned using the time-normalized F0 difference measurement.

Figures 4 and 5 show results of F0 difference comparison that provide evidence for the distinct F0 patterns in our labeling scheme for hesitation repetition and repair error correction, respectively. Both trimmed F0 difference and time-normalized F0 difference show the same trends, thus only box plots of trimmed F0 difference are presented below for the sake of space. Hesi-r-same in Figure 4 and repair-e-same in Figure 5 have mean values close to 0, indicating prosodic parallelism of REP and ALT. Hesi-r-exaggerated and repair-e-exaggerated in Figures 4 and 5 exhibit negative mean values, consistent with our perception that the F0 trajectory of ALT is scaled higher than that of REP, as an exaggeration of the REP F0 contour. The same trend in F0 mean values is shown in hesi-r-ip and repair-e-ip, but in this case the underlying F0 patterns are different than in the exaggerated pattern. An intonational phrase boundary (ip) is perceived at the end of REP, with a pitch reset at the onset of ALT that is responsible for the higher scaling of ALT, especially at the beginning of the ALT interval. The two patterns are further differentiated by the presence or absence of final lengthening and filled or unfilled pauses. The opposite sign (i.e., positive value), for hesi-r-fp is also consistent with our expectation that the ALT in hesi-r-fp functions like a filled pause (e.g., *uh* or *um*), with the characteristically low F0 of a filled pause, scaled lower than the F0 of REP. The effect of hesi-r-change is not strong enough to support our expectation, based on our impression of this sub-class as a kind of prosody repair disfluency, that ALT is scaled higher than REP.

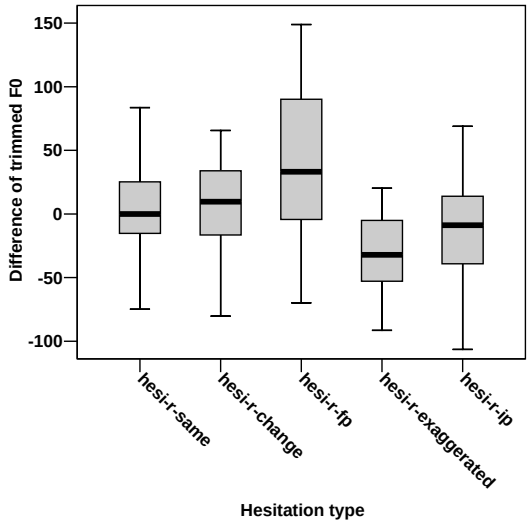


Figure 4. Box plot of trimmed F0 differences between REP and ALT for Hesitation repetition.

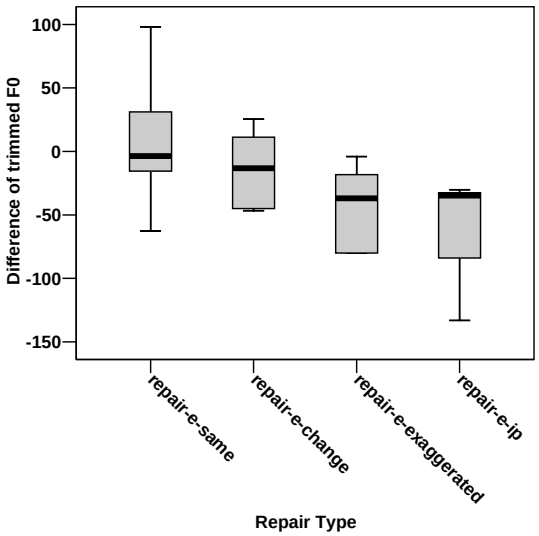


Figure 5: Box plot of trimmed F0 difference between REP and ALT for Repair error correction

The other comparison metric used in this paper is the mean F0 distance metric, which calculates Euclidean distance divided by the number of samples. The F0 difference and F0 distance have the following advantages and disadvantages. F0 difference measures only the difference between average F0 of the REP and average F0 of the ALT; F0 distance, on the other hand, is a measure of the dissimilarity in shape of the two F0 trajectories, when plotted as functions of time. The disadvantage of F0 distance is that it is an unsigned measure: F0 differences may be positive or negative, but all F0 distances are positive. The mean F0 distance of REP and ALT is obtained by:

$$D(F_0) = \sqrt{\frac{\sum_{i,j=1}^{n} (F_0^{(i)} - F_0^{(j)})^2}{n}}$$

Here, as in the mean F0 difference metric, *i* is the *i*th sample of REP and *j* is the corresponding sample of ALT, and *n* is the length of the F0 contours. Division of the Euclidean distance by the square root of the number of samples is done to normalize the effect of sampling length of the tokens in our corpus. (The formula is also known as Root Mean Square (RMS)). As in the difference metric, trimming and time normalization are used for F0 comparison. When D(F0) is close to 0, REP and ALT have similar F0 trajectories, and when the value of D(F0) is large, the F0 trajectories of REP and ALT are different. Thus, when REP and ALT are almost parallel, then the value of D(F0) is expected to be close to 0, and when REP is higher than ALT or ALT is higher than REP, we expect the value of D(F0) to be greater than 0.

Figures 6 and 7 show results of mean F0 distance comparison of hesitation repetition and repair error correction. As noted above, because both trimming and time normalization resulted in similar trends, we present distance of trimmed F0 for hesitation repetition (Figure 6) and repair error correction (Figure 7). Like the results of F0 difference comparison, the F0 distance results confirm our perception of the F0 patterns in the labeling task. In general, hesi-r-same and repair-e-same reveal a small mean distance, near zero in both Figure 6 and Figure 7, while hesi-r-exaggerated and repair-e-exaggerated have the largest mean distance in both Figure 6 and Figure 7. We note that the mean F0 distance of hesi-r-change is also short, indicating that the *change* sub-class exhibits the same prosodic parallelism we predicted and observed for the *same* sub-class.

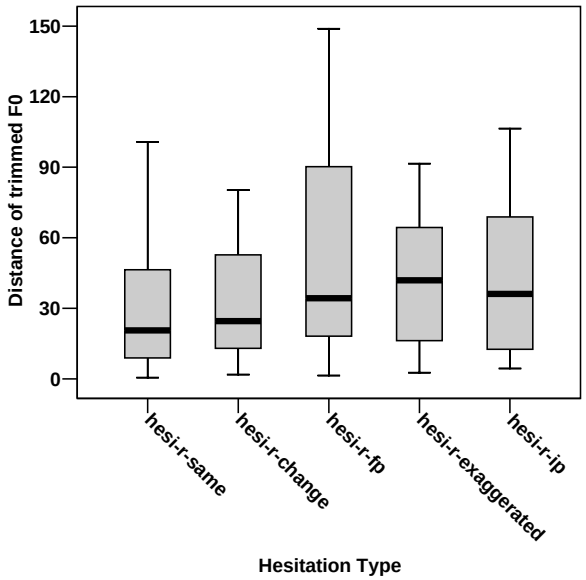


Figure 6: Box plot of trimmed F0 distance between REP and ALT for Hesitation repetition.

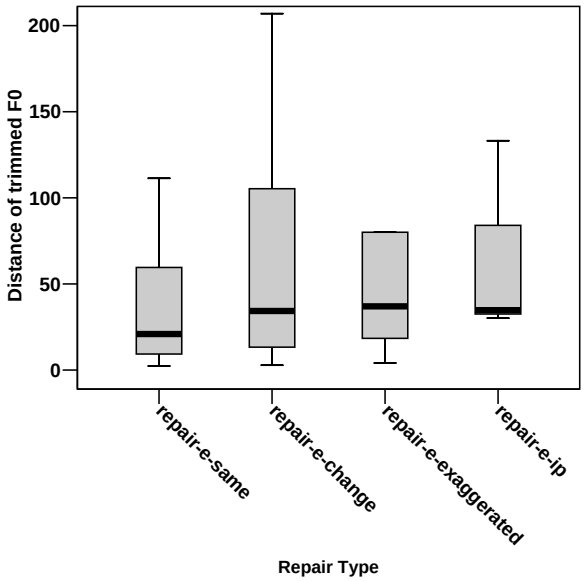


Figure 7: Box plot of trimmed F0 distance between REP and ALT for Repair error correction.

In this section we presented quantitative analysis of F0 comparison using difference and distance metrics on the trimmed and time-normalized intervals of REP and ALT. The overall results of the F0 comparison provide quantitative evidence from F0 measurements for the labelers' perception of the prosodic relationship between REP and ALT in complex disfluencies. Prosodic parallelism is evident in small F0 difference and distance measures for the most frequent sub-classes of hesitation and error correction disfluencies. F0 measures also confirm distinct patterns of F0 relationship between REP and ALT for disfluencies in the *exaggerated* and *ip* sub-classes.

5. Discussion and Conclusion

Our labeling of repetition and error correction disfluencies (Table 5) demonstrated the frequency of five distinct intonational patterns that characterize the prosodic relationship between REP and ALT intervals of complex disfluencies. The most frequent pattern (*same*, with 114 tokens) involved the perceived repetition in the ALT segment of the F0 pattern of the REP segment. The *same* pattern represents almost half of the total number of repetition and error correction disfluencies in this study. The remaining four categories each contain between 1.7 - 12.5% of the total number of complex disfluencies, and represent patterns in which ALT is produced with a low flat F0 (*filled pause*), or patterns where ALT is produced with higher F0 due to pitch reset (*ip*) or higher overall F0 scaling on ALT (*exaggerated*).

Our quantitative measures of F0 provide supporting evidence for the F0 patterns described in our perceptual labeling scheme. Prosodic parallelism of REP and ALT is confirmed by highly similar F0 contours for the largest prosodic sub-class of hesitation and error correction disfluencies. The frequent occurrence of this pattern may provide an important perceptual cue to the listener for the occurrence of disfluency, and may help in the online editing of the disfluency.

Acknowledgment

This research is supported by NSF award number IIS-0414117. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

References

Baron, Don, Elizabeth Shriberg, & Andreas Stolcke. 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. *Proc. ICSLP'02*, Denver, CO, vol. 2, pp. 949-952.

Boersma, Paul & David Weenink. 2005. *Praat: doing phonetics by computer* (version 4.3.04) [Computer Program]. Retrieved March 8, 2005 http://www.praat.org.

Clark, Herbert H. & Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, vol. 84, pp. 73-111.

Clark, Herbert H. & Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology*, vol. 37, pp.201-242.

Godfrey, John J., Edward C. Holliman, & Jane McDaniel. 1992. Telephone speech corpus for research and development. *Proc. the International Conference on Acoustics, Speech, and Signal Processing,* March 1992, San Francisco, CA, pp. 517-520.

Heeman, Peter A. & James F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics,* vol. 25(4), pp. 527-571.

Henry, Sandrine & Berthille Pallaud. 2003. Word fragments and repeats in spontaneous spoken French. *Proc. DiSS'03*, 5-8 September 2003, Goeteborg University, Sweden, pp. 77-80.

Lendvai, Piroska, Antal van den Bosch, & Emile Krahmer. 2003. Memory-based disfluency chunking. *Proc. DiSS'03*, 5-8 September 2003, Goeteborg University, Sweden, pp. 63-66.

Levelt, Willem J. M. 1989. *Speaking. From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, William J. M. & Anne Cutler. 1983. Prosodic marking in speech repair. *Journal of Semantics*, vol. 2, pp. 205-217.

Liu, Yang, Elizabeth Shriberg, & Andreas Stolcke. 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. *Proc. Eurospeech*, Geneva, Switzerland, pp. 957-960.

Nakatani, Christine H. & Julia Hirschberg, 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, vol. 95(3), pp. 1603-1616.

Plauché, Madelaine C. & Elizabeth Shriberg. 1999. Data-driven subclassification of disfluent repetitions based on prosodic features. *Proc. International Congress of Phonetic Sciences*, San Francisco, CA, vol. 2, pp. 1513-1516.

Savova, Guergana & Joan Bachenko. 2003. Prosodic features of four types of disfluencies. *Proc. DiSS'03*, Goeteberg University, Sweden, pp. 91-94.

Shriberg, Elizabeth. 1995. Acoustic properties of disfluent repetitions. *Proc. International Congress of Phonetic Sciences*, Stockholm, Sweden, vol. 4, pp. 384-387.

Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. *Proc. ICSLP'96*, 3-6 October 1996, Philadelphia, PA, vol. Addendum, pp. 11-14.

Shriberg, Elizabeth. 2001. To `errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, vol. 31(1), pp.153-164.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 59-63.

Promotion of Disfluency in Syntactic Parallelism

Andrew A. Cooper & John T. Hale

Michigan State University, East Lansing, Michigan, USA

Abstract

The development of a disfluency-robust speech parser requires some insight into where disfluencies occur in spontaneous spoken language. This corpus study deals with one syntactic variable which is predictive of disfluency location: syntactic parallelism.

A formal definition of syntactic parallelism is used to show that syntactic parallelism is indeed predictive of disfluency.

1. Introduction

Disfluencies (such as 'uh', 'um', and repairs) are a significant challenge in the implementation of parsers and other tools dealing with spontaneous spoken language.

One avenue to the development of disfluency-robust tools is to treat them probabilistically, an approach that has met with success elsewhere in syntax. This approach requires some knowledge of where disfluencies occur in spontaneous speech.

Syntactic parallelism is a variable which (as we will show below) interacts with disfluency. Intuitively, syntactic parallelism is the reuse of a syntactic structure in proximity.

The results presented here are from the Switchboard corpus.[1]

1.1. Outline

The paper is organized as follow. First, section 2 defines the disfluency types of interest. Then section 3 proposes an explicit definition of syntactic parallelism. Two hypotheses based on this definition are set out. Section 4 presents several methods for evaluating these hypotheses in the parsed section of the Switchboard corpus. Section 5 presents the results of experiments using these methods, and section 6 discusses their significance. Sections 7 interprets the results, considering their implications both for the theory of disfluency and for robust parsing. Section 8 suggests possible directions for future study.

2. Disfluency

Disfluency is generally treated as a phenomenon of spontaneous speech in which the speaker makes agrammatical utterance such as "pauses, fillers ('um's and 'er's), repetitions, speech repairs, and fresh starts" [7]. One element common to all types of disfluency is their potentially-universal distribution[2]—disfluencies can and do occur everywhere.

Because they are not marked in the Switchboard corpus, pauses and fresh starts are not considered here, but fillers, repetitions, and self-repairs are.

2.1. Fillers

Fillers are words which are semantically empty and seem to serve as placeholders in speech. An example from the corpus:

> I'm sure we have to have <u>uh</u> permits

2.2. Edit-type disfluencies

We treat repetitions and self-repairs as subtypes of *edit-type disfluencies*. These are characterized by the speaker's attempt at a fluent utterance and subsequent correction of it. An example from the corpus:

> <u>the first kind of invasion of</u> the first type of privacy

3. Syntactic Parallelism

Linguistic theories differ in their characterizations of syntactic parallelism. The definition proposed here attempts to formalize both proximity (in section 3.1) and structural similarity (in section 3.2).

The concepts used in this definition are: c-command, like category of nodes, and depth.

3.1. Lobes

First, a domain on which to describe the parallelism is required. This domain will be called **lobes**.

> Two phrases are said to be lobes of a parallelism if they are of the same category and one c-commands the other[3].

This definition is the basis of **Hypothesis I**:

> A lobe is more likely to contain a disfluency if the other lobe in the parallelism contains a disfluency.

Hypothesis I was tested using methods described in section 4.1, with results reported in section 5.1 and discussed in section 6.1.

3.2. Parallel Constituents

Having defined the parallelism coarsely, it is possible to describe parallelism between the substructures of one lobe and the substructures of the other.

A constituent (say, X1) **is parallel to** another constituent (say, X2) if:

> A lobe Y1 contains X1 and a lobe Y2 contains X2; and X1 performs a parallel grammatical function in Y1 as X2 does in Y2.

Formalizing the notion of 'grammatical function' is of course controversial. One option is to follow Chomsky [1] in the idea that grammatical function corresponds to configurational position. The **parallel grammatical function** requirement is then approximated[4] by:

[1] The Switchboard corpus consists of 2400 telephone conversations among 543 speakers [4].

[2] Universal, but not uniform—this study finds that the uniform-distribution hypothesis is, in all cases, much less than 0.1% likely to derive the results presented in tables 1, 2 and 5.

[3] A node A is said to c-command a node B if every node dominating A also dominates B.

[4] This approximation is based on the conjecture (unproved to the authors' knowledge) that, in a binary-branching

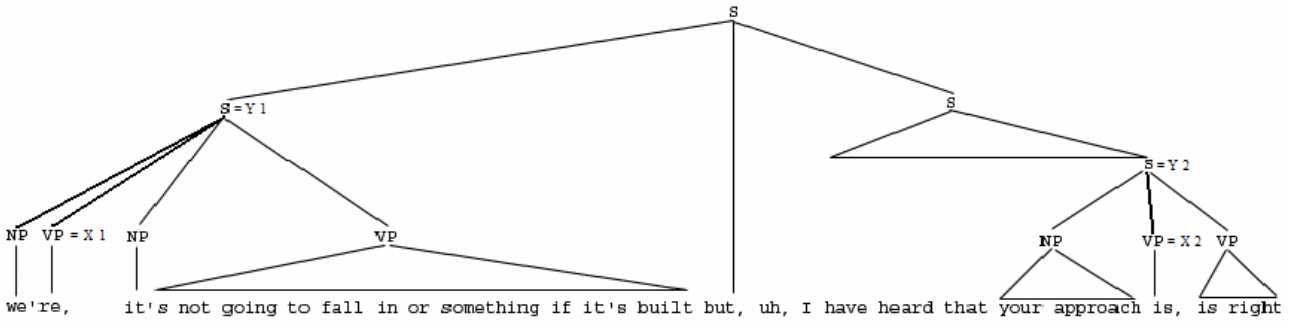


Figure 2: Two parallel S lobes with parallel disfluent VP constituents.

> The same number of nodes intervene between X1 and Y1 as between X2 and Y2.

This conception of parallel grammatical function is the basis of **Hypothesis II**:

> Disfluencies occur in parallel positions.

A (partially-treed) example in which disfluencies occur in parallel positions is given in Figure 2.

Hypothesis II was tested using methods described in section 4.2, with results reported in section 5.2 and discussed in section 6.2.

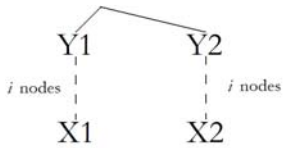


Figure 1: Y1 and Y2 are lobes; X1 and X2 are parallel constituents.

4. Methods

A sample of seventeen conversations was inspected by eye to test Hypothesis I. Only parallelisms resulting from conjunction [5] were counted. The results of this inspection justified further examination.

4.1. Examination of Lobes - Hypothesis I

Hypothesis I, namely that parallel lobes tend to have similar disfluency containment, was examined using `tgrep2`, a tree search program [6]. This examination again considered only conjunction-related parallelism in the sample of seventeen conversations mentioned above.

4.1.1. Editing Treebank

To test Hypothesis I, the Treebank parse files were edited by hand to make conjoined top-level sentences sisters under a common node called 'TOP'. This was necessary to have `tgrep2` recognize top-level conjunctions.

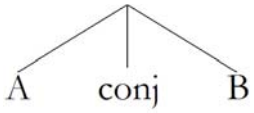


Figure 3: Hypothesis I structure searched for with `tgrep2`.

4.2. Examination of Parallel Constituents - Hypothesis II

Hypothesis II, namely that disfluencies tend to occur in parallel positions, was confirmed using both `tgrep2` and `TIGERSearch`[7], another tree search program.

4.2.1. tgrep2

The `tgrep2` examination of Hypothesis II covered a sample of 78 conversations among 86 speakers[8], constituting 12% of the entire corpus.

For the `tgrep2` examination, the Treebank files were automatically edited to remove the "," tag, which obscured some c-command relations.

The queries used in this examination were designed to consider each node and determine its disfluency status, then find all parallel constituents and determine their disfluency statuses. Since `tgrep2` cannot calculate depths, 4400 queries enumerating every possible depth up to 10 were used.

4.2.2. TIGERSearch

The `TIGERSearch` examination of Hypothesis II covered all 650 Treebank files. Prior editing and enumeration over categories and depths were unnecessary because of `TIGERSearch`'s more powerful query language.

4.3. Statistical tests

The chi-square test of distributional significance and the phi statistic were used to examine the data obtained from both experiments.

The chi-square test measures the likelihood that the axes of a contingency table are independent. Higher values of χ^2 indicate a smaller likelihood of independence.

The phi-statistic measures how much of the variance along one axis of a contingency table is explained by the variance of the other axis. Higher values of ϕ indicate a greater percentage of variation which is so explained.

5. Results

The results described above in general confirmed both research hypotheses.

grammar obeying reasonably formalized X-bar restrictions, it is impossible for two nodes of the same category and depth from a common ancestor to occupy different configurational positions.

[5] Treebank treats conjuncts as sisters.

[6] `tgrep2` was developed by Douglas Rohde of the Department of Brain and Cognitive Science at Massachusettes Institute of Technology, accessible at `http://tedlab.mit.edu/~dr/Tgrep2/`.

[7] `TIGERSearch` is a project at the Universität Stuttgart, accessible at `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch`.

[8] No single speaker participated in more than 6 conversations.

5.1. Hypothesis I

The results of testing Hypothesis I are presented in contingency table 1. Each cell contains the number of parallelisms with lobes of the indicated disfluency status. Here 'disfluent' means that the lobe contains one or more disfluencies, and 'fluent' means that the lobe contains no disfluencies at all.

Table 1: Disfluency status of conjoined lobes, obtained with `tgrep`. A significant distribution.

N=821 df=1	χ^2=61.25, p<.001 ϕ=0.273	Lobe 2	
		Disfluent	Fluent
Lobe 1	Disfluent (Expected) % of total	150 (99.2) 18.3	126 (176.8) 15.3
	Fluent (Expected) % of total	145 (195.8) 17.7	400 (349.2) 48.7

5.1.1. Other results

The majority (68%) of parallelisms due to conjunction were top-level sentential conjunctions. An even stronger majority (91%) of parallelisms with a disfluent lobe were top-level sentential conjunctions.

5.2. Hypothesis II

5.2.1. tgrep2

The results of testing Hypothesis II with `tgrep2` are presented in contingency table 2. Each cell contains the number of parallelisms with lobes of the indicated disfluency status. Here 'disfluent' means that the constituent is the site of a disfluency, and 'fluent' means that the constituent is fluent. 'Constituent 1' is the linearly first constituent.

Table 2: Disfluency status of parallel constituents, obtained with `tgrep` . A significant distribution.

N=20267 df=1	χ^2=461.29, p<.001 ϕ=0.151	Constituent 2	
		Disfluent	Fluent
Constituent 1	Disfluent (Expected) % of total	635 (307.1) 3.1	1242 (1569.9) 6.1
	Fluent (Expected) % of total	2681 (3008.9) 13.2	15709 (15381.1) 77.5

There are potentially several constituents parallel to any given node, with potentially different disfluency statuses. That possibility is excluded here; in order to be counted as 'fluent', a node had to be parallel to no disfluent nodes. Such an exclusion is conservative because it tends to reduce the number of matches with two fluent nodes in favor of matches with one fluent and one disfluent node.

5.2.2. TIGERSearch

The results of the `TIGERSearch` examination are given in contingency tables 3, 4 and 5. Each cell contains the number of pairs of parallel constituents with the indicated disfluency status. 'fluent' and 'disfluent' have the same meanings as above. 'Constituent 1' here is the constituent in the c-commanding node; 'Constituent 2' is the constituent which is c-commanded.

Table 3: Filler-type disfluency status. A significant distribution.

N=111844 df=1	χ^2=899.19, p<.001 ϕ=0.089	Constituent 2	
		Disfluent	Fluent
Constituent 1	Disfluent (Expected) % of total	434 (118.4) 0.4	2745 (3060.6) 2.4
	Fluent (Expected) % of total	3732 (4047.6) 3.3	104933 (104617.4) 93.8

Table 4: Edit-type disfluency status. A significant distribution.

N=111844 df=1	χ^2=286.7, p<.001 ϕ=0.051	Constituent 2	
		Disfluent	Fluent
Constituent 1	Disfluent (Expected) % of total	355 (153.2) 0.3	3179 (3380.8) 2.8
	Fluent (Expected) % of total	4495 (4696.8) 4.0	103815 (103613.2) 92.8

Table 5: Disfluency status (aggregate). A significant distribution.

N=111844 df=1	χ^2=286.7, p<.001 ϕ=0.034	Constituent 2	
		Disfluent	Fluent
Constituent 1	Disfluent (Expected) % of total	789 (541.2) 0.7	5924 (6171.9) 4.9
	Fluent (Expected) % of total	8227 (8474.9) 7.4	96904 (96656.2) 86.6

Contingency table 3 contains the data for filler-type disfluencies alone. Contingency table 4 contains the data for edit-type disfluencies alone. Contingency table 5 contains the aggregate data, excluding matches in which one parallel constituent was the site of an edit-type disfluency and the other was the site of a filler-type disfluency. This omission is conservative, since it biases against matches in which both parallel constituents are disfluent.

6. Discussion

In general, the results confirm the research hypotheses. These results extend Shriberg [9], who found that sentence-initial disfluencies and sentence-medial disfluencies have a high cooccurrence rate.

The results here also show that disfluency in the Switchboard corpus (both filler-type and edit-type) is responsive to a syntactic variable, extending the findings of Fox & Jasperson [3], who argue that repair responds to several syntactic variables.

6.1. Hypothesis I

Hypothesis I was confirmed in conjoined lobes; that is, *conjoined lobes tend to have the same disfluency status*. The ϕ-value reported indicates that about 7% of the variation in disfluency status in conjoined lobes is due to parallelism.

6.2. Hypothesis II

Hypothesis II was confirmed in both experiments testing it; that is, *syntactically parallel constituents tend to have the same disfluency status*. Given the several conservative assumptions made in calculating ϕ-values, it is possible that the effects observed are stronger than reported here.

6.2.1. tgrep2

The tgrep2 examination of Hypothesis II confirmed it. The ϕ-value reported indicates that about 2% of the variation in disfluency status is due to parallelism.

6.2.2. TIGERSearch

The TIGERSearch examination of Hypothesis II confirmed it. The ϕ-values reported indicate that very little of the variation in disfluency when considered in the aggregate is due to parallelism, but about 1% of the variation in filler-type disfluency and edit-type disfluency is due to parallelism.

That different types of disfluency are impacted distinctly by parallelism likely reflects the differences in distribution and characteristics of different types of disfluency observed by Shriberg [9], McKelvie [7] and others.

6.3. Sources of Error

Possible error sources are the original Switchboard corpus and the Treebank annotation scheme. The exclusion of some types of disfluency also introduces some error.

6.3.1. Switchboard

As Shriberg [9] and McKelvie [7] note, the Switchboard's transcriptions include some errors. Shriberg reports that in a small sample, her transcriptions disagreed with the Switchboard's on 25% of disfluent turns. Most of the discrepancies were due to miscategorization. Since this study treats fillers as one category and all other disfluency types annotated in the Treebank scheme as another, it is largely immune to error resulting from mistranscriptions. Only in the rare event that a filler were misheard by the transcriber as an edit or vice versa, or in the rarer case that a disfluency were transcribed when one did not occur would the error introduced be non-conservative.

6.3.2. Treebank

The Treebank annotation scheme introduces some additional sources of potential error.

In particular, the way Treebank treats repairs obscures constituency relations, as the reparandum is treated as sister to the repair.

A similar problem results from Treebank's treatment of sentential adjuncts, which are treated as sisters, again obscuring constituency.

Finally, and most critically, the Treebank is not parsed according to a binary X-bar grammar. This means that the critical conjecture upon which the definition of syntactic parallelism in section 3.2 is based is not strictly applicable to the corpus. In some cases, such as a VP like *give the dog a bath*, where *the dog* and *a bath* are treated as sister NPs, this will tend to falsely increase the number of 'parallel' constituents. However, since less than half of occurrences of each category are disfluent (McKelvie [7] gives a 6% rate of edit-type disfluency in the MAPTASK corpus) such false hits will come overwhelmingly in the fluent-disfluent, disfluent-fluent, and fluent-fluent cells of the contingency tables. Consider the effect of a false hit in each case:

Case I: Fluent-disfluent (bottom-left cell)

A false hit here artificially increases the number of observed hits. Since in each contingency table the expected number of hits is already greater than the number of observed hits, false hits here reduce the contribution of this cell to χ^2. Thus a false hit in case I is a conservative error.

In table 5, for example, the observed number in case I is 8227. If 227 of these were false hits, the true number would be 8000, which is farther from the expected value of 8474.9. The increase in χ^2 due to eliminating the false hits would be 19.36.

Case II: Disfluent-fluent (top-right cell)

As in case I, false hits reduce the contribution of this cell to χ^2. Thus a false hit in case II is a conservative error.

For example, in table 5, if 224 false hits were counted, the true number of observed hits would be 5700, which is farther from the expected value of 6171.9. The increase in χ^2 due to eliminating the false hits would be 26.12.

Case III: Fluent-fluent (bottom-right cell)

A false hit here artificially increases the number of observed hits. In each contingency table the expected number of hits is less than the number of observed hits, so the false hits artificially increase the contribution to χ^2 of this cell, a non-conservative error. But a look at the numbers in the tables suggests that the observed number case III could indeed be much lower and the distribution remain significant.

In table 5, the contribution to χ^2 of case III is only 0.64. Any decrease in the observed value up to 496 hits would not increase the total χ^2 for the table.

The results seen here are in fact stronger than necessary for application to parsing. For left-to-right parsing, one needs only to use a disfluency to predict subsequent disfluencies. Even true hits in case I (a fluent constituent followed by a parallel, disfluent constituent) are irrelevant to this prediction. In the same way, case III is entirely irrelevant to parsing, as it measures only the base rate of disfluency. True hits in case II are the only possible counterexamples to disfluency's predictiveness in parallel syntactic contexts, and the results of section 5 find that they occur at a unexpectedly low rate.

7. Interpretations

These results may be read to argue against the conventional interpretation of disfluency as only a phenomenon of production errors. If this view were correct, one would expect the second lobe of a parallelism to be *more* fluent than the first, since the speaker has already had a chance to work out the difficulties in a particular construction. But this is the opposite of what is observed in the present study. Often the speaker is disfluent in the second lobe, just as in the first.

There are a number of possible interpretations along these lines.

One might be that disfluency is prosodically and pragmatically salient enough that a speaker inserts disfluencies in the second lobe of a parallelism to maintain similarity with a disfluent first lobe. This conclusion would explain the stronger effect of conjunction, a kind of intentional parallelism, on disfluency. It would also explain why fillers, which are more likely to have pragmatic use, are more strongly affected by parallelism.

Another argument might be that speakers simply do not improve their production abilities, even in the short term. The speaker might be in a similar mental state, hence likely to make similar errors, at the second lobe as he was at the first. A production analogue of Steiner's [10] Iteration Model might be appropriate to this interpretation.

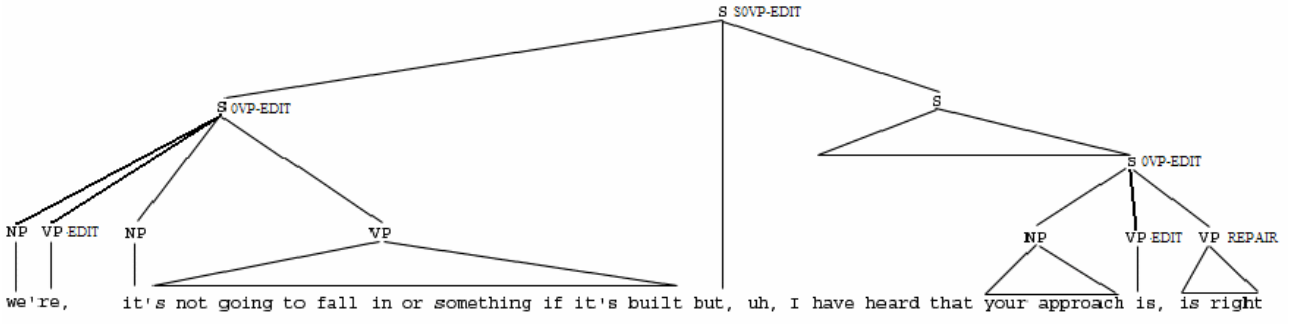


Figure 4: A possible parsing scheme.

8. Further Work

As noted above, the χ^2 statistic is fairly robust to the sorts of error encountered here. The ϕ-statistic does not fare as well, and further investigation with more powerful measures will be necessary to more precisely quantify the effect observed here.

To see how these numbers might be employed, consider the sentence in Figure 2 again. Having parsed the first S and noted that its VP is the site of an edit-type disfluency, the parser could add a 0VP-EDIT feature to that S, indicating that a descendant at depth 0 was an edited VP. Upon reading 'but', the parser would build the top-level S and propagate the feature S0VP-EDIT, indicating that it dominates an S with a zero-depth edited VP descendant. The parser then passes the 0VP-EDIT feature to all S descendants of the top-level S, thus predicting the second disfluent VP and its repair. Figure 4 is a schematic of this parse.

Such an arrangement might be implemented by interacting phrase structure schema:

1. S[0VP-EDIT] → NP VP[EDIT] VP[REPAIR]
2. S[0VP-EDIT] → NP VP[EDIT]
3. S[0VP-EDIT] → NP VP

Rule 1 would have a higher probability than rule 2, representing the greater likelihood of a repaired disfluency. Rule 3, representing case II (disfluent-fluent), would have the lowest probability.

Even if not explicitly coded in the Treebank, such rules may be automatically derivable from extant annotations.

9. Conclusion

Studied with the aid of a formal definition of syntactic parallelism, the Switchboard data suggest that syntactic parallelism is predictive of disfluency. With some further investigation, this predictive power could be incorporated into a parser to enable more efficient and accurate parsing of spontaneous speech.

In addition, the interaction of parallelism and disfluency has theoretical ramifications. It calls into question the treatment of disfluencies as solely the result of production problems. It also provides more evidence that disfluency is partly syntactic in nature.

10. Acknowledgements

Thanks to Dan Jurafsky and the participants in the 2004 Midwest Computational Linguistics Colloquium for their comments on an earlier iteration of this project.

11. References

[1] Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

[2] Clark, Herbert and Jean Fox Tree. 2002. Using uh and um in spontaneous speech. *Cognition*, 84:73-111.

[3] Fox, Barbara, and Robert Jasperson. 1995. *A syntactic exploration of repair in English conversation*. In Davis, Philip W., ed. Alternative Linguistics: Descriptive and Theoretical Modes. pp. 77-134. Amsterdam: Benjamins.

[4] Godfrey, John, Edward Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. *Proc IEEE ICASSP*. volume 1, pp. 517-520.

[5] Koenig, Esther, Wolfgang Lezius, and Holger Voorman. 2004. *TIGERSearch user's manual*.

[6] Marshall, Mitchell, Beatrice Santorini, and Mary Ann Marcinkewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*. 19:313-330.

[7] McKelvie, David. 1998. The syntax of disfluency in spontaneous spoken language. Technical Report HCRC/RP-95, Human Communications Research Centre.

[8] Rohde, Douglas. 2001. *Tgrep2 manual, version 1.11*

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

[10] Steiner, Ilona. 2003. Parsing syntactic redundancies in coordinate structures. *Proc EuroCogSci 03*. p. 443

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 65–70.

Modeling conversational styles in Italian by means of overlaps

Rodolfo Delmonte

Department of Language Sciences - Università Ca' Foscari
Ca' Garzoni-Moro - San Marco 3417 - 30124 VENEZIA

Abstract

Conversational styles vary cross-culturally remarkably: communities of speakers – rather than single speakers - seem to share turn-taking rules which do not always coincide with those shared by other communities of the same language. These rules are usually responsible for the smoothness of conversational interaction and the readiness of the attainment of communicative goals by conversants. Overlaps constitute a disruptive element in the economy of conversations: however, they show regular patterns which can be used to define conversational styles (Ford and Thompson, 1996).

Overlaps constitute a challenge for any system of linguistic representations in that they cannot be treated as a one-dimensional event: in order to take into account the purport of an overlapping stretch of dialogue for the ongoing pragmatics and semantics of discourse, we have devised a new annotation schema which is then fed into the parser and produces a multidimensional linear syntactic constituency representation.

This study takes a new tack on the issues raised by overlaps, both in terms of its linguistic representation and its semantic and pragmatic interpretation. It will present work carried out on the 60,000 words Italian Spontaneous Speech Corpus called AVIP, under national project API - the Italian version of MapTask, in particular the parser, to produce syntactic structures of overlapped temporally aligned turns. We will also present preliminary data from IPAR, another corpus of spontaneous dialogues run with the Spot Differences protocol. Then it will concentrate on the syntactic, semantic and prosodic aspects related to this debated issue.

The paper will argue in favour of a joint and thus temporally aligned representation of overlapping material to capture all linguistic information made available by the local context. This will result in a syntactically branching node we call OVL which contains both the overlapper's and the overlappee's material (linguistic or non-linguistic). An extended classification of the phenomenon has shown that overlaps contribute substantially to the interpretation of the local context rather than the other way around. They also determine the overall conversational style of a given community of speakers with cultural import.

1. Introduction

A distinctive feature of any conversation is the way in which people interrupt each other: ideal conversations would be constituted by a smooth sequence of turn-taking moves which the speakers may predict by means of conventional hints – mainly short pauses or specific intonational contours. However it is a fact that whenever speakers have a communicative goal to attain – and this is what happens in the great majority of conversations – they also want to do it following Grice's Maxims. Conversations should then be guided by principles of efficiency and effectiveness. This is very much so in case conversants have a task guiding their conversation as happened in our dialogues, which include the Italian version of Map Task and the task called Spot Differences. It is conceivable that in this case conversants are put under pressure and their conversational strategies will certainly come to the fore much more clearly than what might happen in more relaxed scenarios. However, as will be shown clearly below, this may only apply to certain communities of speakers and not to others. As reported in the literature the way in which overlaps are used by a given community can vary remarkably from a supportive to a competitive manner. In our case, overlaps will also be used as redundant means for supportive information, notably by Neapolitan and Roman speakers.

This paper presents work carried out at the University of Venice for the characterization of conversational styles in four regional Italian varieties which include, Southern (Pugliese) Italian, Tuscan (Pisano) Italian, Neapolitan (Campano) Italian and Roman (Central) Italian. The specific topic of this paper will be the characterization of overlaps along the lines of what has been done in MATE project and other international projects in progress like the MEETING project. In the AVIP/API dialogues the quantity of overlapping speech is very high, as we shall see below. At an international level, even though everybody agrees on the relevance of the phenomenon, there is no universal agreement on its representation from the linguistic point of view, in particular as concerns syntactic structure both at constituent and functional level.

The definition of OVERLAP in the literature on conversational studies is rather cumbersome: this is due in our opinion to a tendency to mix up pragmatic imports with semantic and sometimes syntactic ones. In some cases, also ethnographic elements have been brought in, where a colonialistic point of view has been used to sanction natives' communicative interchanges as being unruled because of the high number of interruptions and overlaps. In other cases, it would seem that gender imports come into play to differentiate between a competitive and a collaborative use of overlaps and interruptions.

One first argument was the terminological issue related to the function of the overlap, either as an interruption by the intruder speaker or simply as a continuer - a backchannel or a confirmation word.

A second argument brought to bear on the definition of overlap was the outcome of the interruption, where it causes the intruder speaker to usurp the floor or not, and in this second case whether the current speaker was obliged to repair his utterance as a reaction to the interruption, or not.

A third argument was the place at which the overlap occurs, whether it is at a TRP, Transitional Relevance Place in which case also syntactic completion of some kind was involved - this was defined, for lack of a better linguistic basis, whenever a predicate was present. Or in case it is in a Nontransitional

Relevance Place: these two terms also hinged on the semantic completeness of the already uttered turn by the current speaker.

A fourth argument relied on the recognition on the side of the overlapper of the incoming TRP and the irrelevance of the completion of the turn on the side of the current speaker, so that an interruption in the middle of the turn could still be interpreted as collaborative.

A final more decisive factor is the computability of overlaps in terms of its predictability.

As fas as predictability is concerned, data are often conflicting: it would seem that males are more prone to produce interruptions in order to show their desire/tendency to compete, if compared to females who are more oriented towards collaborative and cooperative attitudes in conversations. This could be used to predict conversants' behaviour with respect to their gender/sex: however this is not universally borne out so it is easily contradicted. I will not quote colonialists'-like opinions on the way less educated communities handle conversations if compared to more educated ones. This is also not proven and can also be dismissed on social equity principles.

Sachs et al. introduced the concept of conversational rules and Grice introduced the concept of conversationals maxims: however none of these rules and maxims can be used to predict the actual progress of turn interaction in real dialogues.

The only criterion that can be safely regarded to be useful for such an important issue as computability should in our opinion come from experimental data: and in this regard, they can be derived from syntactic structure or from acoustic/linguistic analysis of FØ movements. As reported in Shriberg et al., these two parameters may be taken to represent possible anchors for an algorithm for overlaps predictability. It would seem that whenever overlaps occur at TRPs they would do so because of turn-change projectability from the current speaker: this is usually linked to the presence of a Boundary Tone; while on the contrary the presence of a High Tone would indicate a less likely projectable TRP, hence a desire by the speaker not to be interrupted by its interlocutor. In turn, the overlapper may simply produce a semantically empty Turn Unit, a backchannel or an affirmative continuer, to express his/her wish for the current speaker to continue. A viable syntactic-semantic definition of TRP can only be formulated on the basis of a treebank available or a parser able to compute any spontaneous dialogue text without having to go through its output manually. In our case we will use the treebank which we semi-automatically built for the Italian National Project on spontaneous speech based on the Italian MapTask and the Spot Differences Dialogues.

Our treebank will not only allow us to derive a precise definition of TRP but also to classify all overlaps accordingly. This is due to a peculiar feature of the constituency which contains a distinct major constituent for overlaps, called OVL, which contains both the overlapped and the overlapping linguistic material in the actual location in the turn in which the interruption took place.

2. Overlaps

Overlaps may be defined as a speech event in which two people speak simulateneously by uttering actual words or in some cases non-words, when one of the speakers, usually the interlocutor, interrupts or backchannels the current speaker. This phenomenon takes place at a certain point in time where it is anchored to the speech signal; but in order to be fully parsed and subsequently semantically interpreted, it needs to be referred semantically both to a following turn and to the local turn where it may produce conversational moves to repair what has been previously said by the current speaker.

One of the distinctive characteristics of naturalistic conversation (in contrast to monolog situations) is the presence of overlapping speech. Overlapping speech may be of several types, and affect the flow of discourse in various ways. An overlap may help to usurp the floor from another speaker (e.g., interruptions), or to encourage a speaker to continue (e.g., back channels), or simply end up just in an attempt at usurping the floor without success. As a preliminary and tentative pragmatic definition we may define an overlap as being normally a physical event that happens in a single time unit in which two or more speakers want to comunicate different and non-coincident comunicative intentions. Exception made for rare cases in which the two or more speakers intended to say the same thing in the same time unit.

Speaker overlaps, are directly observable in our data, since by definition overlaps occur at points of simultaneous speech on more than one of the (individually recorded) channels, besides their explicit indication in the ortho-phonetic transcription thus transliterated into the orthographic transcription. What we are interested in is finding out whether there is any correlation between the onset of overlaps and their possible characterization from the point of view of syntactic structure, which we have proposed to treat by introducing a node of discourse constituency called OVL (overlap), from where the two temporally aligned components of overlapping, the overlappee and the overlapper stretch of speech/text, branch.

Both punctuation and overlap have been discussed in the literature as correlating with prosodic cues. For example, past computational work has discussed prosodic features for sentence boundaries as well as disfluency boundaries. Past work in conversation analysis, discourse analysis, and linguistics has shown prosody to be a useful cue in turn-taking behavior. So we may assume that overlapping can be safely be described also in prosodic terms or lends itself to use prosody as a linguistic correlate to linguistic descriptions.

2.1. Overlaps: why caring about them in the first place?

Why detecting and labeling Overlaps is so important? These are the most important reasons for taking care of them:

- They are very frequent;
- They may introduce linguistic elements which influence the local context;
- They may determine the interpretation of the current utterance;

and for these reasons, they cannot be moved to a separate turn because they must be semantically interpreted where they temporally belong. After moving overlaps to their original temporal position, as a side-effect, some turns are just empty conversational moves because the speaker has already been taking the turn with a previous overlap which may have been followed by a repairing move of the other speaker thus conversationally concluding the communicative exchange.

2.2. Overlaps and Syntax

As said above, overlaps challenge all criteria of linguistic representation which require the input sentence to be mono-dimensional, i.e. to contain the utterance of one single speaker. This fact is semantically essential in order to guarantee the linguistic representation to be interpretable. On the contrary, overlapped linguistic material, i.e. sentences which contain at the same time linguistic material coming from two or more

participants in the dialogue are not only hard to parse: they might also constitute an obstacle to semantic interpretation.

As in most robust parsers, we use a sequence or cascade of transducers: however, in our approach, since we intend to recover sentence level structure, the process goes from partial parses to full sentence parses. Sentence and then clause level is crucially responsible for the right assignment of arguments and adjuncts to a governing predicate head. This is clearly paramount in our scheme which aims at recovering TRPs by referring solely to syntax.

The sequential processors receive the input sentence split by previous processors, which is recursively/iteratively turned into a set of non-sentential level syntactic constituents. Non-sentential level constituents, can be interspersed by heads which are subordinate clause markers, or parentheticals markers. The final output is a list of headed syntactic constituents which comprise the usual set of semantically translatable constituents, i.e., ADJP, ADVP, NP, PP, VC (Verb Cluster). In addition to that, sentence level markers interspersed in the output are the following: FINT, interrogative clause marker; DIRSP, direct speech clause marker; FP, parenthetical clause marker; FC, coordinate clause marker; FS, subordinate clause marker; F2, relative clause marker.

The task of the following transducer is that of collapsing into the corresponding clause the clause material following the marker up to some delimiting indicator that can be safely taken as not belonging to the current clause level. In particular we assume that at each sentence level only one VCluster can appear: we define the VC as IBAR indicating that there must be a finite or tensed verb included in it. VClusters containing non-tensed verbal elements are all defined separately, as follows: SV2, for infinitive VCs; SV5, for gerundive VCs; SV3, for participial VCs.

The second transducer has also two additional tasks: it must take care of ambiguity related to punctuation markers such as COMMA, or DASH, which can either be taken as beginners of a parenthetical or indicators of a list, or simply as separators between main clause and subordinate/coordinate clause. It has also the task of deciding whether conjunctions indicated by FC or by FS are actually starting a clause structure or rather an elliptical structure.

2.3. An example of parsed overlapped dialogue

As to orthographic trascription, the decisions taken in the Italian MapTask was to follow the original transcription schema and conventions: in particular, overlaps are fully marked in the local speech aligned orthographic transcription, by introducing the index of the turn containing the overlapping material, which however is not visible and should be looked up in the following turn. In addition, two #s are introduced at the front of the turn index and at the end of the overlapped speech as shown in the following example:

Dialogue 2.

p1#94: no <sp> cioè sì c'ha<aa> <mh> <sp> una specie di tappo
p2#95: sì #<p1#96> c'ha un ta+ tappo <sp># , sì
p1#96: #<p2#95> di funghetto# <lp> c'ha prima una base un po' altina

Dialogue 2.1

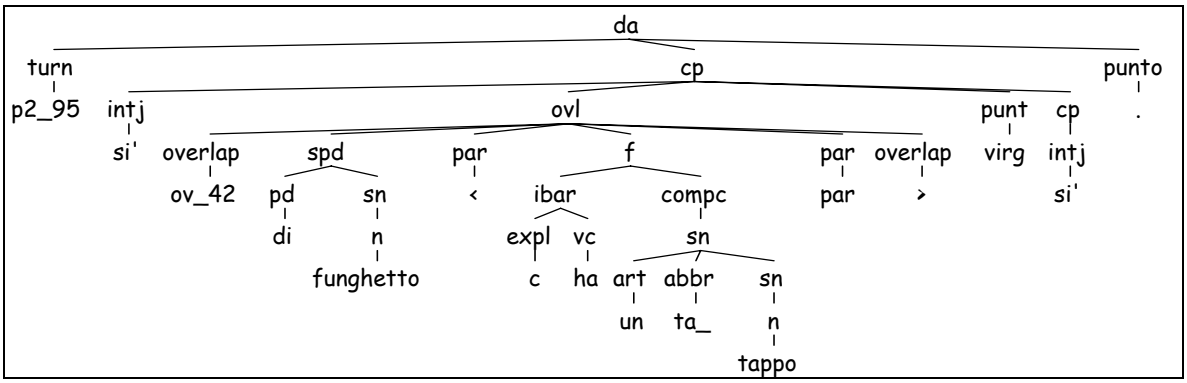
p1_94: no, cioè sì c'ha, una specie di tappo.
p2_95: sì ov_42 di funghetto < c'ha un ta_ tappo - >, sì.

Turn 95 contains an overlap which is introduced and erased from the following turn and indexed as shown in 4.1 version of the dialogue: the convention being that the ov_42 index is followed by the overlapper's speech intruding in the overlappee's turn. The material being overlapped then follows the open '<' and the close of the overlap is marked by the closing '>'. In this way the orthography linearizes the bidimensional event of the overlap by keeping the linguistic material within the same turn as adjacent text rather than scattering it in different turns. The ownership of the material by one of the speakers is guaranteed by its local respective position within the boundaries of the overlap: the ov_N starting symbol and the '>' at the end. It is important to notice that the two words are respectively pronounced by a woman and a man, the intruder utters with a rising tone: the implicit communicative intention is that of producing a better indication of the shape of the object currently under discussion and trying to get the other speaker to accept it.

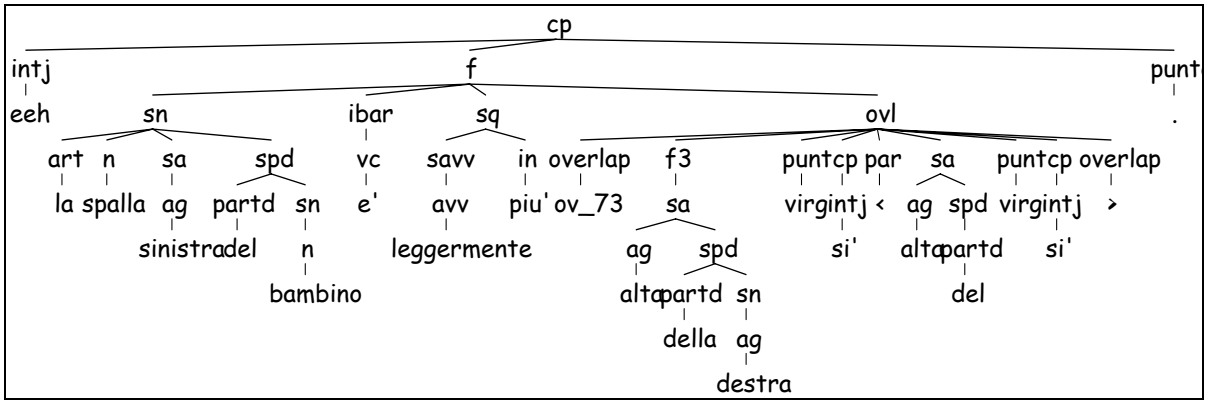
The utterance contains a short pause <sp> right after the overlap which is then followed by an affirmative interjection "sì"/yes: this is a very common feature of overlaps in our corpus, a confirmation is a conversational act reacting to the overlapping material, which however is not present in the current utterance since it has been moved to the following turn. As can be understood by recomposing the overlapping portions of this conversation, what really happens is that the two speakers, Speaker 1 and Speaker 2 are interacting very closely while the description of the scenario is carried on. At the same time at which a certain shape is individuated and properly described a consensus is reached: but this is reached by trial and errors in a continual re-approximation of the task. There are two internal repairs caused by the overlap: the first one is "sì"/Yes as a reaction of Speaker 2 to a first definition of the shape "tappo"/cork, which is however taken only as being suggestive "una specie di"/a kind of, of a better yet to be defined final shape. And in the Speaker 2 turn, the repetition of "tappo" which is intentionally interrupted by recovering the turn role and suggesting the most appropriate shape, "di funghetto"/of a little mushroom.

Dialogue 2.b

da(turn(p2_95),cp(intj(si'), ovl(overlap(ov_42), spd(pd(di), sn(n(funghetto))), par(<), f(ibar(expl(c), vc(ha)), compc(sn(art(un), abbr(ta_), sn(n(tappo))))), par(par), overlap(>)), punt(virg), cp(intj(si'))), punto(.))



.Figure 1: Syntactic Structure for Dialogue 2.b with temporally aligned overlap



.Figure 2: Syntactic Structure for Dialogue 3 with temporally aligned overlap and linguistic material on the right.

The realignment of all turns has given as a result a certain number of empty turns, i.e. all those turns which had been artificially built by simply containing overlapping material which had been already uttered by the current speaker before the previous turn was elapsed.

The need to represent linguistic information related to two speakers in the same syntactic structural representation, which is both semantically and pragmatically strongly intertwined has a lot of theoretical implications.

This implements principles of linguistic representation expressed in previous work of ours, in particular in Delmonte, (1987), where syntactic structure was to interact with semantic and pragmatic structure in order to take into due account phenomena like Contrastive and Emphatic Focus. *Discourse Grammar* which is crucially grafted onto rules of sentence grammar; does not directly relate to unconscious and innate LAD mechanisms but stems and develops on extralinguistic, contextual/situational or pragmatic conditions.

As a matter of fact, no neat division should be drawn between these two theoretical domains, apart from empirical reasons, i.e. in order to reduce interfering factors which do not contribute in an essential way to the construction of an internal grammar. In particular, the realm of performance, being the less studied if compared to competence, contains quite a number of such interfering factors. We might also surmise that a lot of performance (as such describable within a discourse grammar) interferes strongly with competence (Bresnan, 1982: xxiii) leading to an interactive (see Marsley-Wilson, Tyler, 1980), model for discourse understanding, rather than a sequential one.

Interpretation could be triggered independently from sentential material or be determined by the presence of coreferring extrasentential expressions; as a further option, it could be triggered locally by logical operators which in turn may vary their scope according to the presence of extrasentential factors.

- In other words, to allow for feedback to take place between the two levels of grammatical relations, we need discourse level phenomena to be adequately represented by sentence grammar. This is certainly the case with the case we are tackling now: overlaps take place at a discourse level, however their import is deeply grafted into sentence grammar, by conditioning interpretation from taking place.

Consider now another interesting example represented by the following utterance, where the overlapper corrects the current speaker – the overlappee – who, as a consequence of that, drops its utterance and confirms what the overlapper said.

> "eeh, la spalla sinistra del bambino è leggermente più ov_73 alta della destra, sì > alta del, sì <."/the left shoulder of the child is slightly more ov_73 high than the right one, yes > high of the, yes <."

Whose syntactic structure is,

> cp(intj(eeh),f(sn(art(la),n(spalla),sa(ag(sinistra)),spd(partd(del),sn(n(bambino)))),ibar(vc(e')),sq(savv(avv(leggermente)),in(piu')),ovl(overlap(ov_73),f3(sa(ag(alta),spd(partd(della),sn(ag(destra))))),punt(virg),cp(intj(si')),par(>),sa(ag(alta),spd(partd(del))),punt(virg),cp(intj(si')),overlap(<))),punto(.))

3. Overlaps and Conversational Moves

After creating the treebank, overlaps have been organized as follows:

- Overlaps occurring between specifier and head;

- Overlaps occurring in a parallel and unintentional simultaneous way;
- Overlaps which are semantically empty and are computable as backchannels;
- Overlaps at a higher constituency level, i.e. at sentence level or after main predicate and main complement has been computed.

In Table 2. and 3. below we show both absolute and percent values of all overlaps distributed in the four locations where the two tasks have been recorded. As can be easily gathered, Naples is the seat where in absolute terms most overlaps have occurred; it is also the place where in absolute and – together with Rome – in relative terms, most semantically empty overlaps occurred. Rome is the place where in relative terms most semantically empty overlaps occurred; it is also the place where the least number of spec/head overlaps occurred. Bari, is the seat where in absolute terms the least overlaps have occurred: it is also the place where in relative terms we find the least semantically empty; in addition to that, it is the place where the most specifier/head overlaps occurred and the most TRP relatable overlaps occurred. Eventually, Pisa is the seat where most parallel overlaps occurred.

Our proposal will then be articulated as follows:

A. TRP coincides with sentence level or whenever main complements have been parsed and we are left with adjuncts;
B. Non-TRP coincides with all the remaining cases, i.e. when the overlap starts between the specifier and the head of a constituent when still in preverbal position; or else whenever the overlap is positioned at the constituent boundary but the main governing predicate has not been parsed yet.

Case A. thus constitutes the semantically valid option corresponding to a projectable smooth TRP with/without overlap; on the contrary, case B. constitutes the non-semantically viable option where the conversant does not have enough semantic content to project a TRP and simply wants to prevent the current speaker from continuing his turn. If this is so, we will also divide up all Overlaps into two categories: Competitive vs. Collaborative we end up with the following general subdivision,

- Competitive Overlaps – Parallel + Spec/Head
- Collaborative Overlaps – Semantically Empty – Higher than Constituent

3.1. Overlaps and Dropping

Another important indicator of the actual import of an overlap is the relation intervening between an overlap and the completion of the turn by the current speaker. An overlap that also marks a dropping of turn by the current speaker who yields his turn to the overlapper can be computed differently according to whether the overlap takes place at the end of the turn or not. This is due to the fact that in normal conversational interaction speakers would be in the conditions to forecast when the current turn is ending and would produce an overlap past the TRP to speed up the attainment of the communication task.

On the contrary, dialogues by speakers who use overlaps turn internally would contain a lot of cases of Continuing Conversation in presence of Overlaps: these we call Non Dropping Overlaps (hence NDOs). It is a fact that NDOs may only occur in a competitive situation: either within what we defined Parallel Overlap or within a Spec/Head Overlap. We may interpret the occurrence of NDO as an indication of a collaborative attitude between the interactants: in presence of an overlap, people continue speaking. Turns containing more than one Overlap are 98 overall. To these cases of NDOs we add all cases in which speakers alternate short and long pauses with overlaps during a long turn without the overlapper actually usurping the floor.

We computed NDOs for all dialogues and the overall picture we get is that, Bari has the least number of NDOs, Napoli on the contrary has the highest number thus confirming our previous conclusion. Naples conversational style has as a specialty the exploitation of overlaps as a means to make dialogues more communicative, most redundant and least efficient. Nonetheless, this is accepted as a rule by conversants of the same regional variety and is regarded as an effective tool.

Table 1. Disaggregated Overlaps data in all Dialogues in absolute values.

Sites / Overlaps	Overlap Totals	Comp-etitive Overlaps	NDOs	% NDOs	% NDOs wrt. Competitive
Bari	142	64	14	9.89	21.87
Napoli	909	318	275	30.25	86.47
Pisa	264	122	32	12.12	26.22
Rome	189	66	19	10.05	28.78
Totals	1504	576	340	22.6	59.02

Table 2. Disaggregated Overlaps data in all Dialogues in absolute values.

Sites / Overlaps	Overlap Totals	Specifier head Overlaps	Parallel Overlaps	Semantically Empty Overlaps	Higher Constituency Level Overlaps
Bari	142	53	11	25	53
Napoli	909	221	97	333	258
Pisa	264	64	58	70	72
Rome	189	42	24	77	46
Totals	1504	380	190	505	429

Table 3. Disaggregated Overlaps data in all Dialogues in percent values.

Sites / Overlaps	Overlap Totals	Specifier head Overlaps	Parallel Overlaps	Semantically Empty Overlaps	Higher Constituency Level Overlaps
Bari	9.44	**31.2**	7.8	17.6	**31.2**
Napoli	**60.43**	24.31	10.67	**36.33**	26.38
Pisa	17.55	26.12	**21.96**	26.51	27.27
Rome	12.56	22.22	12.69	**40.74**	24.33
Totals	100	25.26	12.63	**33.57**	28.52

4. Overlaps and Prosody

As said above, overlaps computability can be derived from syntactic-semantic structure or from acoustic/linguistic analysis of FØ movements: data coming from these two experimental areas may be taken to represent possible anchors

for an algorithm for overlaps predictability. It would seem that whenever overlaps occur at TRPs they would do so because of turn-change projectability from the current speaker: this is usually linked to the presence of a Boundary Tone; while on the contrary the presence of a High Tone would indicate a less likely projectable TRP, hence a desire by the speaker not to be interrupted by its interlocutor.

To test these hypotheses, we analysed the prosodic content of those overlaps constituting interruption at constituent level and we found a strong correlation with the acoustic signal. We thus analysed all semantically relevant overlaps and classified them by means of ToBI representations. Results are shown in the Table here below where we report data related only to the AVIP/API corpus.

We measures two different set of phenomena: semantically and pragmatically relevant overlapped turns, then phonetically relevant overlapped turns and then compared them with Competitive Overlaps. At first we marked overlapped turns with relevant semantic information, i.e. turns in which contrasting information is expressed by one or both of the speakers; we then marked all turns containing relevant FØ movements, i.e. tones marked H* and H+L*. Overlapped turns with contrasting information are only 1/3 of all turns, with Naples going down to 1/4 and Bari up to 1/2. Then we counted tones in Competitive Overlaps (COs) as defined on the basis of syntactic structure and we called them Relevant Tones (RTs). Eventually we counted those COs containing Relevant Tones (RTs) and these are reported in Tab.4. As can be seen, the prediction expressed by means of syntactic information is also born out by phonetic data. In particular, percent values for relevant tones, i.e. relevant FØ movements in Competing Turns is very high, over 70%. Finally, Competing Overlaps with RTs are half of all competing overlaps: with the notable exception of Neapolitans who do not use intonational cues to drive their competing overlaps preferring to use redundant non competing or collaborative means. On the contrary, Bari speakers accompany their competitive overlaps with phonetic cues in a totally predictable and systematic manner, reaching 100% of all COs.

Table 4: Semantic and Phonetic data of Relevant Tones for Competitive Overlaps in AVIP/API Corpus.

Sites / Overlaps	Overlap Totals	Competitive Overlaps Cos)	Total Relevant Tones (RTs)		No. Of COs with Rts Totals	%
Bari	**142**	64	82	74.54	64	100.00
Napoli	**674**	318	165	69.32	144	45.28
Pisa	**264**	122	72	69.23	80	65.57
Totals	**1080**	576	319	70.57	288	50.00

References

Delmonte R., 2005, Parsing Overlaps, in B.Fisseni, H.C.Schmitz, B. Schroeder, P. Wagner (Hrsg.), Sprachtechnologie, mobile Kommunikation und linguistiche Ressourcen, Sprache, Sprechen und Computer, Bd.8, Peter Lang, Frankfurt am Main, pp.497-512.

Delmonte R. 2003. Parsing Spontaneous Speech, in Proc. EUROSPEECH2003, Pallotta Vincenzo, Popescu-Belis Andrei, Rajman Martin "Robust Methods in Processing of Natural Language Dialogues" , Genève, pp, 16-23.

Ford, C.E. and Thompson, S.A. 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E.A. Schegloff & S.A. Thompson (eds) Interaction and grammar, Cambridge: Cambridge University Press, pp. 134-184.

Sacks, Harvey, Emanuel Schegloff & Gail Jefferson. 1974. 'A simplest systematics for the organization of turn-taking for conversation,' Language 50(4), 696-735.

Schegloff, Emanuel. 1996. 'Turn organization: one intersection of grammar and interaction,' in Ochs, Schegloff, & Thompson (eds.), Interaction and grammar, Cambridge University Press, Cambridge, pp. 52-133.

Shriberg, E.; A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, vol. 2, pp. 1359–1362, Aalborg, Denmark, 2001.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 71-76.

Hesitations and repair in German

Kristy Beers Fägersten

Universität des Saarlandes, Saarbrücken, Germany

Abstract

The occurrence of pauses and hesitations in spontaneous speech has been shown to occur systematically, for example, "between sentences, after discourse markers and conjunctions and before accented content words." (Hansson [15]) This is certainly plausible in English, where pauses and hesitations can and often do occur before content words such as nominals, for example, "uh, there's a … man." (Chafe [8]) However, if hesitations are, in fact, evidence of "deciding what to talk about next," (Chafe [8]) then the complex grammatical system of German should render this pausing position precarious, since pre-modifiers must account for the gender of the nominals they modify.

In this paper, I present data to test the hypothesis that pre-nominal hesitation patterns in German are dissimilar to those in English. Hesitations in German will be shown, in fact, to occur within noun phrase units. Nevertheless, native speakers most often succeed in supplying a nominal which conforms to the gender indicated by the determiner or pre-modifier. Corrections, or repairs, of infelicitous pre-modifiers indicate that the speaker was unable to supply a nominal of the same gender which the choice of pre-modifier had committed him/her to. The frequency of such repairs is shown to vary according to task, with fewest repairs occurring in elicited speech which allows for linguistic freedom and therefore is most like spontaneous speech. The data sets indicate that among German native speakers, hesitations occurring before noun phrase units (pre-NPU hesitations) indicate deliberation of what to say, while hesitations within or before the head of the noun phrase (pre-NPH hesitations) indicate deliberation of how to say what has already been decided (cf. Chafe [8]).

1. Introduction

Pauses in spontaneous speech are naturally occurring phenomena. They establish prosody and flow, as well as facilitate content organization on the part of the speaker (Fromkin [12], Garrett [13], Levelt [18], Mayer, [21, 22], Shattuck-Hufnagel & Klatt [28], Ward [36, 37]) and comprehension on the part of the listener (Brennan & Schober [6], Brennan & Williams [7], Maclay & Osgood [19], Schachter et al [26], Shriberg [29]). Hesitations, on the other hand, might be considered accidental or unintentional pauses and are often regarded as disfluency in speech. Indeed, hesitations may be due to an ongoing cognitive process which forces speech to slow down or even come to a halt (Bock [4], Bock & Levelt [5], Fromkin [12], Garrett [13], Shriberg [30], Van-Winckel [35]). Research on hesitations as a manifestation of disfluency in speech has largely focused on identifying systematicities of hesitations, resulting in categorization of form (Batliner, et al [3], Shriberg [30], Ward [37]) or of location in discourse (Arnold et al [1], Auer & Uhmann [2], Clark & Fox Tree [9], Fox Tree [10], Hansson [15], Makkai [20], Stenström, [32]).

A large majority of previous research has been based on English, which both allows for the establishing and cross-referencing of norms specific to English, as well as encourages comparison with other languages. The starting point for this paper is the observation that, in English, hesitations can –and often do– occur before such "accented content words" (Hansson [15]) as nominals. In other words, hesitations often occur within noun phrase units, for example *after* pre-modifiers such as determiners and adjectives. It is not uncommon, for example, to encounter hesitations in the following contexts, taken from Chafe [8]:

"uh, there's a … man."

"the … the-- … the basic action,"

where hesitations (or pauses) occur after a determiner. Modern English has no grammatical gender and, consequently, no inflections for gender on pre-modifiers. Thus, the hesitations in the examples above are most probably cognitive in nature (as opposed to pragmatic), revealing the speakers' active search for a completion to the noun phrase. In the following example, the meta-language of the filled pause suggests even further that this is the case:

"one of them has a … what do you call those little … um (.85) paddleball?" (Chafe [8])

Unlike English (and to the dismay of its second-language learners), German has retained its inflections and still has a complex grammatical system of gender and case. Pre-modifiers of nouns, for example, must show agreement in gender, for example: *der Hund*, where *der* is the masculine definite article; *ein grosser Hund*, where *ein* is the masculine indefinite article and *–er* is the strong, masculine adjective inflection. Gender is part and parcel of the German noun system and, as such, pre-modifiers should be less prone to being teased apart from their nouns by cognitive processes than they are in English (at least in terms of cognitive processes reflecting lexical decision making; Tseng [34] gives examples of parentheticals in German in pre-NPH). While, Wode [38] did not identify hesitations in German specifically occurring in pre-NPH position, Langer [17] provides examples of pre-NPH hesitation and repair similar to Chafe's [8] examples. The questions at hand, therefore, are the following: what is the distribution of pre-nominal hesitations (pre-NPU vs. pre-NPH) occurring in German, and how are hesitations within the noun phrase unit resolved?

2. Method

2.1. Multi-task experiment set

In order to investigate the distribution of pre-nominal hesitations in German native-speaker speech, a multi-task experiment set was designed. The experiment set consisted of three tasks, detailed below, allowing for different degrees of spontaneity in speech.

The experiments were conducted at Universität des Saarlandes (Saarbrücken, Germany) among university students. Participants in each of the tasks were all native speakers of German and, at the time of the experiment, between the ages of 20 and 23. In addition to age, gender of the participants was noted and, for one particular experiment,

number of semesters of English studied at university level. A total of 26 males and females took part in one or more of the experiments. The participants were not informed of the specific purpose of the experiments, only that each experiment was to investigate linguistic phenomena. All output was tape recorded, with consent from each participant.

2.2. Object description

An object description experiment was designed to create a situation which would be conducive to hesitating due to cognitive processing of the names of familiar and unfamiliar items (cf. Chafe [8]). Participants were shown two index cards: one card with an adjective written on it such as *gross* (big), and one card with a nominal written or pictured. The participants were then asked to provide, as quickly as possible, a noun phase consisting of the words indicated by the words or pictures on the cards, and including the definite article. After receiving instructions, each participant was presented with an example of expected output: when shown cards with the word *gross* and a picture of a dog or the word *Hund*, the participant should say, "*der grosse Hund*." In order to avoid the possibility of phonetic neutralization, participants were asked to provide the definite article (*der, das* or *die*) for each noun phrase as opposed to the indefinite article (*ein, ein* or *eine*).

Following Chafe's [8] experiment in which participants were asked to retell the events of a film which featured "objects that [were] expected to be high in codability and objects with which [it was] expected speakers would have difficulty," the nominals on the cards consisted of familiar objects such as *Hund* (dog) and presumably unfamiliar objects (based on pre-experiment judgments from members of the same speech community) such as a *Strickliesel* (yarn holder) and brandnames such as 'Nutella'. In addition, adjective cards were included for the purpose of creating a longer noun phrase than determiner+noun, which, it was hoped, would allow for more opportunity for hesitating. The complete set of 37 nominals included: ten objects presented as words: *Abszess, Aerobic, Baldachin, Banner, Konklave, Kuvert, Manufaktur, Reklamation, Tuberkel*; ten objects presented as pictures: *Bienenstock, Grammophon, Gravur, Haus, Hund, Monokel, Musterklammer, Pentagon, Strickliesel, Tipi;* and 17 product names: *Blistex, Bounty, Cillit Bang, Colgate, Corvette, Fiesta, General, Golf, Hanuta, Labello, Lenor, Nutella, Toffifee, Touareg, Twix, Urquell, Whiskas.* The ten adjectives included: *blau, dick, gelb, gross, grün, gut, klein, rot, schlecht, weiss*.

Participants in the object description experiment included ten male and ten female German native speakers (20 total participants).

2.3. Translation

A translation experiment was designed to represent an increased degree of spontaneity in speech while still controlling for stimuli and thus creating a situation which would encourage cognitive processing and, presumably, hesitations. Participants were given two minutes to silently read through a short narrative (see below) in English, which they were then to translate into German. When their reading time was over, the participants were told to translate the passage as quickly as possible and as best they could without asking for help with or clarification of vocabulary.

The translation narrative was chosen for its inclusion of English noun phrases which could have several possible translations in German (such as 'party', which could be (*die*) *Party* or (*das*) *Fest*), nominal constructions which would more idiomatically occur as verbalizations in German ('an announcement was made'), or words likely to be unknown to the participants ('proctor'): *Once upon a time, there was a student named John. Most people who knew him considered John to be a disaster. In fact, he hardly had a reputation as a source of pleasure, and his classmates looked for any chance to keep their distance from him. A party was a rare scheduling on his calendar indeed, and his absence was hardly missed.*

One day during the holidays – which were particularly lonely and filled with doubt about his likeability – John got the idea of going on an exchange. He hoped it would provide an escape from his environment and be the end to the failure he usually experienced in social settings.

In order to partake in the exchange, John just needed to pass the end-of-term exam, a minor detail, or so he thought. However, on examination day, an announcement was made that the exam would have to be postponed. The reason for the delay was that the proctor had suddenly cancelled due to a pain in her stomach, and no one else was available to do the job. Unfortunately, John was to leave for his exchange later that day. In the end, he …

The translation narrative ended with an unfinished sentence, which the participants were told to complete in order to conclude the narrative. This last task reflects an attempt to incorporate spontaneous speech as a point of comparison.

Participants in the translation experiment included eight male and five female German native speakers (13 total participants). As the task consisted of translating from English to German, it was necessary that the participants have an adequate proficiency in English, in other words, at least two semesters of English university study.

2.4. Retelling

Like the translation experiment, the retelling experiment was also designed to represent an increase in spontaneity while still controlling for stimulus and input. Participants were given two minutes to silently read a lengthier narrative in German, which they were then to retell in as much detail as possible, also in German. No time limit was imposed on the retelling, nor were the participants instructed to be quick. As in the translation task, participants had no recourse to clarification.

The retelling narrative was chosen for its uncommon juxtaposition of characters and plot –a toxicologist and veterinarian foil a would-be thief– as well as for the high frequency of nominals:

Es war einmal ein junges erfolgreiches Paar. Eines schönen Tages saßen die Veterinärmedizinerin und der Toxikologe in ihrem schönen großen Garten auf der Hollywoodschaukel, als ihnen die Idee kam, zu verreisen. Sofort rannte die Veterinärmedizinerin ins Haus, um den Reisekatalog zu holen. Nach mehrmaligem Durchblättern stellten beide enttäuscht fest, dass die Destinationen ihnen nicht gefielen. Um die Auslandsinvestition dennoch zufrieden stellend zu tätigen, fuhren beide ins nächstgelegene Reisebüro. Dies grenzte unmittelbar an das Internetcafe „Online“. Im Reisebüro empfing sie die Reiseverkehrs-kauffrau mit einem tollen Angebot: „Ich empfehle Ihnen das Gipfeltreffen der Ministerpräsidenten aller Industrieländer im reichen Emirat Dubai“, sagte sie begeistert. Freudig stimmten der Toxikologe und die Veterinärmedizinerin zu, denn nach Dubai wollten sie schon immer. Dort angekommen erhielten sie die Hiobsbotschaft, es sei kein Zimmer gebucht. Sie standen auf der Straße. Plötzlich tauchte neben ihnen der Trickdieb Ranjid auf und versuchte der Veterinärmedizinerin den

Pradarucksack zu entreißen. Geschickt konnte der Toxikologe den Entwendungsdelikt abwehren, indem er den Karategriff anwendete, den er letzte Woche im heimischen Kurs gelernt hatte. Der Dieb nahm Reißaus. Ohne zu zögern rief der Toxikologe im Reisebüro an, um über die Kostenerstattung des Höllentrips zu verhandeln. Am anderen Ende der Leitung versuchte die Reiseverkehrskauffrau den aufgebrachten Mann zu beruhigen. Das unglückliche Paar solle zunächst die Übergangsunterkunft der deutschen Botschaft nutzen. Endlich dort angekommen erhofften sich die Leidgeplagten etwas Ruhe und Erholung. Doch im Zimmer nebenan lief der Monumentalfilm „Spartakus" in einer solchen Lautstärke, dass die Veterinärmedizinerin das Ohrensausen ihres Lebens bekam. Auch das Hypnotikum, dass ihr der Toxikologe verabreichte, half nicht. Sofort fuhr das Paar zum Flughafen und flog mit der Concorde nach Hause. Über das Mysterium dieser ungewöhnlichen Reise dachten die Unglücksraben noch lange nach.

Participants in the retelling experiment included 12 male and 13 female German native speakers (25 total participants).

3. Data analysis

All data were recorded digitally using an Olympus Digital Voice Recorder or a Sony IC Recorder. The recordings were then transcribed with all hesitations represented in orthographic approximations (in the case of filled pauses or hesitations) or as silences, regardless of position. The transcriptions were then checked against the recordings to assure accuracy of identification and position of hesitations. In the data analysis, no distinction was made between pauses, hesitations, fillers, filled pauses, filled hesitations, etc., nor were silences timed. Any interruption in fluency, even meta-linguistic, was considered a hesitation and therefore this term will be used throughout the data analysis sections for the sake of simplicity and consistency.

3.1. Object description

Each of the 20 participants in the object description experiment produced 37 noun phrases, resulting in a total of 740 three-word noun phrases. Due to the design of the experiment, initial hesitations were to be expected; the participants needed time to read or look at the information on the cards and therefore immediate responses were impossible. A time of two seconds was determined sufficient for reading; any extension of this time would therefore be considered a hesitation. In general, the problem of distinguishing reading time from hesitation was solved by the participants themselves, who most often marked the end of their reading time with audible cues such as inhalations, exhalations or utterances like, "Hmm."

As each noun phrase was produced in isolation, all noted hesitations are included in the pre-nominal hesitation data analysis. A total of 226 hesitations were produced, corresponding to an average of 11.3 per participant. Of the total, 158 (70%) occurred before the noun phrase unit (pre-NPU position), while 68 (30%) occurred within the noun phrase unit (pre-NPH). Table 1 shows the distribution of hesitations in real numbers and percentages.

Table 1: Distribution of object description hesitations.

	Hesitations	pre-NPU	pre-NPH	Repairs
Number:	226	158	68	23
% of total :	100%	70%	30%	10% / 34%

At 70%, the amount of pre-NPU hesitations clearly shows a significant (one sample t-test: $p<0.05$) tendency among speakers to approach the noun phrase as a unit, most frequently pausing before the determinatives to identify objects and/or determine gender. However, the 68 hesitations occurring *within* the noun phrase unit (pre-NPH) indicate that some speakers may, in fact, commit to a gender before considering the nature of the nominal. It is interesting to now take a closer look at these pre-NPH hesitations. Only 29% of these hesitations (20 from a total of 68) occurred between the definite article and adjective. These hesitations cannot be considered in terms of linear cognition; in other words, the speaker is most definitely not deciding what to say next since the subsequent (final) two words of the noun phrase are provided. The same conclusion must therefore also apply to the hesitations occurring between the adjective and the head, a clear majority at 71% (48 from a total of 68). Instead, each of these pre-NPH hesitations may more accurately be considered evidence of deliberation over the congruence of the chosen determiner.

Infelicities in grammatical gender did occur: in 110 instances (15% of the total number of noun phrases produced), the original article chosen did not reflect the correct gender of the nominal. However, only 23 of these instances resulted in repairs, corresponding to only 21% of the total number of mistakes. Repairs were always preceded by hesitations; thus, 34% of the pre-NPH hesitations resulted in repairs. Approximately two-thirds of these repairs occurred after the adjective, suggesting that the closer the speaker comes to the head of the noun phrase, the greater the chance is of infelicities being noticed and repaired.

3.2. Translation

Compared to the object description task, the translation task was designed to be more challenging in terms of cognitive demands. The two-minute time limit on the reading of the English text along with the instruction to provide a translation as quickly as possible reduced planning time and encouraged on-line processing. The result was a clear tendency to attempt linear, word-for-word translations, which were often problematic at best, unsuccessful at worst and delivered with uncertainty. The consequent hesitations seemed to be cognitive in nature, reflecting efforts among the participants to understand the text as well as to determine not only what to say in German, i.e., translation equivalents, but also how to say it, i.e., how to frame or structure the text.

The translation experiment yielded a total of 420 hesitations. Distributed among 13 participants, the average number of hesitations per participant is 32.3. Considering the brevity of the text, the total as well as the average number of hesitations would seem to confirm the proclaimed level of difficulty.

Of the total number of hesitations, 197 (47%) were pre-nominal hesitations and therefore included in the data analysis. These pre-nominal hesitations consisted of 78 (40%) pre-NPU hesitations and 119 (60%) pre-NPH hesitations. Table 2 shows the distribution of hesitations in real numbers and percentages.

Table 2: Distribution of pre-nominal translation hesitations.

	Hesitations	pre-NPU	pre-NPH	Repairs
Number:	197	78	119	29
% of total :	100%	40%	60%	15% / 24%

Unlike the hesitations which occurred in the object description task, the majority of pre-nominal hesitations found in the translation task occurred in pre-NPH position, a significant difference not only across tasks, but also locally (one and two sample t-tests: $p<0.05$).

Pre-nominal hesitations accounted for almost half of the total, 420 hesitations. However, 202 other noun phrase units occurred uninterrupted by hesitations – almost twice the number of interrupted, disfluent noun phrases. Thus, like the data from the object description experiment, the translation experiment data also reveal a tendency among German speakers to treat noun phrases as units. This conclusion represents different means of arrival: the tendency is proved by examining both where hesitations do occur as well as where they do not.

The two data sets reveal an opposite distribution of pre-NPU vs. pre-NPH hesitations, and the value differences are indeed significant (two sample t-test: $p<0.05$). One possible reason for this lies in the nature of the tasks and the degree of linguistic freedom they allowed. In the object description task, two of the three words comprising the noun phrase were predetermined. In the translation task, the stimulus was predetermined, but there remained a fair amount of freedom in terms of response. Thus, pre-NPH hesitations stood a greater chance of more successful resolutions.

As observed in the object description data, resolutions to hesitations included corrections, or repairs. The last column of Table 2 shows the number of pre-nominal, more specifically, pre-NPH hesitations which resulted in repairs. Although more than the 10% of overall object description hesitations (see Table 1), the figure of 15% of overall hesitations followed by a repair in the translation task is insignificant (two sample t-test: $p=0.12$). Specific to pre-NPH hesitations, the number of repairs found in the translation data represents a lower percentage (24%) than in the object description data (34%). However, this difference is also insignificant (two sample t-test: $p=0.14$). Thus, in terms of tendencies among Germans to resort to repairs as a resolution to hesitations, no conclusion can be drawn based on the data so far.

3.3. Retelling

The retelling task represented the greatest degree of linguistic freedom within the multi-task experiment set and, as such, elicited speech which most closely represents spontaneous speech. In addition to providing a greater degree of linguistic freedom, the retelling task incorporated participant freedom. The lack of time limit minimized stress while the instructions of retelling the story in as much detail as possible were subject to individual interpretation and standards. It is not surprising, therefore, that the hesitation patterns observed in the retelling data differ considerably from the data acquired from the first two tasks. With a total of 25 participants as well as a longer text than in the translation task, there was clearly a greater opportunity for hesitations to occur. However, the retelling task resulted in the overall least amount of hesitations and, equally notable, the fewest repairs. Table 3 shows the distribution of hesitations in real numbers and percentages.

Table 3: Distribution of retelling hesitations.

	Hesitations	pre-NPU	pre-NPH	Repairs
Number:	135	36	99	4
% of total:	100%	27%	73%	3% / 4%

The shift in hesitation type, in terms of syntactic position, that was observed in the translation data is further established by the retelling data (two sample t-test: $p<0.05$). Clearly, the hesitation pattern originally hypothesized and most frequently found in the object description cannot be confirmed as the default pattern. Instead, the retelling data show that hesitations do occur within the noun phrase and furthermore are resolved felicitously without repair. As the tasks and the data they elicit more closely approach spontaneous speech, the more frequently pre-NPH hesitations occur and the less frequently repairs occur. Only 3% of the total hesitations corresponding to 4% of the pre-NPH hesitations resulted in repair. Although in both real numbers and percentages, the greatest amount of repairs were observed in the translation task, it must be noted that the object description task resulted in the most infelicities which may have been repaired in less time restrictive circumstances, or not have occurred at all in more natural speech situations. There is indication, therefore, of frequency of repairs decreasing by task or increasing in direct proportion to the spontaneity of speech.

The emergent hesitation and repair distribution patterns now call for a hypothesis of what can be concluded about hesitations in the spontaneous speech of German native speakers.

4. Discussion

Unlike speakers of English, who can haphazardly utter pre-modifiers without committing to a specific noun, German speakers must decide on at least a gender of a noun before they can utter pre-modifiers. They may also have to consider predicators or prepositions as well, in order to determine case. Due to the variety of possible grammatical inflections to account for, it was hypothesized that pre-nominal hesitations among speakers of German would occur overwhelmingly before the noun phrase unit, unlike hesitations in English, which often occur within the noun phrase. The experiments were designed to provide ample opportunity to observe hesitation patterns in the speech of German native speakers by presenting participants with cognitively challenging tasks and demanding real-time linguistic output.

One of the most important findings of this pilot study is that German native speakers cannot always immediately produce fluent, grammatical speech. While arguably obvious as well as applicable to speakers in general, it is a point that needs to be made explicit so as to provide a framework for interpreting hesitations. In the object description task, the participants were given two out of three noun phrase elements and, in spite of this, needed to pause or hesitate an average of 11.3 times per 37 noun phrases. Such hesitations are most certainly cognitive and not a product of experiment design: the words which most frequently caused hesitations appeared in text form, not as pictures.

Hesitations within the noun phrase were rarest in this experiment due first to the restricted context, and second to the fact that, in essence, this task tested the participants' knowledge of grammatical gender. Gender is part and parcel of German nominals; one is rarely decided independently of the other. However, hesitations within the noun phrase did occur, quite possibly as a result of the instructions to be quick to provide the noun phrases. Fewest hesitations were to be found between the definite article and adjective, which can be attributed to the fact that adjectives were provided.

Data from the translation task show a pattern of hesitation distribution different from the pattern established from the object description task. Thus, the original hypothesis is challenged, namely, that pre-nominal hesitations mostly occur in pre-NPU position. Instead, the translation data indicate that hesitations pattern in much the same way in German as in English; that is, hesitations occur before nominals, within noun phrases. In German, this distribution increases the opportunity for ungrammaticalities and, indeed, almost a quarter of the pre-NPH hesitations preceded a repair. It is not only difficult to believe but also contrary to personal observation and communication that German native speakers

repair with such frequency. However, the recorded translation data reveal that the hesitations are often filled with meta-linguistic mumblings about how to proceed or what a particular word means. Thus, it would seem that, perhaps due to the time pressure, the majority of participants attempted a word-for-word translation and, in so doing, adopted the hesitation behavior (i.e., pre-NPH hesitations) common to the source language, English. As in the object description task, the repairs indicate an awareness of as well as intolerance for grammatically infelicitous pre-modifiers. On the other hand, the translation task was not as linguistically restrictive as the object description task, resulting in much fewer errors, none of which was left unrepaired.

The data from the retelling task further disproved the original hypothesis, establishing pre-NPH hesitations as the most common. Despite the fact that the retelling task provided the greatest opportunity for speech production and involved the most participants, hesitations were least frequent in this task. It is important to point out that the retelling task allowed for the greatest degree of spontaneity in speech among the participants and was also without time pressure. It is furthermore interesting to note that, although the text was chosen for its use of uncommon characters and bizarre events, which the participants were told to retell in as much detail as possible, many of the retellings had details omitted or were significantly different than the original narrative. Unlike the translation task, in which the participants could consult the narrative text in front of them, the retelling task did not allow the option of consulting the text. Thus, different capabilities in remembering details or varying standards of accuracy contribute to both variation between original and retold texts as well as great inter-participant variation.

Because of the reduced restriction on output and a lack of time pressure, the retelling task elicited speech that is much more similar to natural, spontaneous speech than the speech elicited from the other two experiment tasks. This data set is also conspicuous in the low number of repairs, suggesting that the naturally occurring spontaneous speech of German native speakers is also low in repairs. Hesitations occurred most frequently in pre-NPH position. However, this does not necessarily challenge the suggestion that German native speakers treat pre-modifier+noun clusters (i.e., noun phrases) as single units. It may rather provide support for the claim that pre-NPH hesitations reflect a deliberation of how to say what is already chosen since hesitations tended to be followed by gender felicitous nominals. The hesitations can thus be considered both pragmatic and cognitive, in that they allow the speaker to hold the floor while planning further ahead. While hesitations in German pattern much like hesitations in English, pre-NPH hesitations do not reflect the same cognitive purposes that they would do in English. Chafe's [8] conclusion that "[the] fundamental reason for hesitating is that speech production is an act of creation" cannot be applied to German in terms of immediate, on-line processing as Goldman-Eisler [14] and Siegman [31] suggest is the case for hesitating in English. Instead, German pre-NPH hesitations may indicate cognitive processes that reflect longer-term planning than what is required of speakers of English.

5. Application

The experiments conducted for this pilot study represent a personal interest in ultimately ridding my non-native German speech of disfluencies, which most often occur as pre-nominal, pre-NPH hesitations and subsequent repair. The fact that I do repair is encouraging and indicative of respect for, if regrettably not a total command of, the German case and gender system. The root of the problem lies rather in a transfer of the typical English disfluency pattern: pre-NPH hesitations. Learners of German, it can be argued, may benefit from learning to hesitate as Germans do. Thus, studying native speaker disfluencies in speech can play a role in language acquisition, as Scanlan [25] has endeavored to investigate and which begs further consideration. Whether or not language students strive for native speaker-like fluency, which Jenkins [16] argues is not always the ideal, acquiring language-specific hesitation patterns may be a step in acquiring the accompanying cognitive processes which may, in turn, result in a greater understanding and command of the language and as well as control over spontaneous speech.

A further application of this research which offers truly exciting prospects is to first language attrition. The translation task data suggested that the participants attempted word-for-word translations of the English text, an approach to language production that is reminiscent of non-native speakers translating or transferring their L1 to the L2. It may also be a practice common to native speakers losing their L1 due to lack of use in favor of an L2. Consideration of hesitations and pauses may provide more evidence of whether or not hesitation patterns can be transferred and/or relearned or reacquired, and might also help to indicate degree of attrition. For recent work on language attrition, see Schmid, et al [27].

6. Acknowledgements

Contributing to the experiment design and pilot study data collection were students of my seminar, "Silence! An exercise in discourse analysis," offered at the Department of English Linguistics at Universität des Saarlandes, Saarbrücken, Germany. I would like to thank them collectively for their input, evaluation and personal anecdotes. I would also like to thank the two anonymous reviewers of the first submission of this text. I have done my best to improve this paper according to their suggestions, but I remain solely responsible for any shortcomings.

7. References

[1] Arnold, Fagnano & Tanenhaus. 2003. Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research*. 32:25-36

[2] Auer, P., & Uhmann, S. 1982. Aspekte der konversationellen Organisation von Bewertungen. *Deutsche Sprache*, 10:1-32.

[3] Batliner, Anton, A. Kiessling, S. Burger, & E. Noeth 1995. Filled Pauses in Spontaneous Speech. *Technical Report 88, Verbmobil Project.*

[4] Bock, J. K. 1986. Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12:575-586.

[5] Bock, K., & Levelt, W. J. M. 1994. Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics*, pp. 945-984. London: Academic Press.

[6] Brennan, S. E., & Schober, M. F. 2001. How listeners compensate for disfluencies in spontaneous speech, *Journal of Memory and Language*, 44:274-296.

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

[8] Chafe, W. 1985. Some reasons for hesitating. In Tannen, Deborah, and Muriel Saville-Troike (eds). *Perspectives on silence.* Norwood, NJ: Ablex.
[9] Clark, Herbert H. & Jean E. Fox Tree. 2002. Using uh and um in Spontaneous Dialog. *Cognition*, 84:73-111.
[10] Fox Tree, Jean E. (2002). Interpreting Pauses and Ums at Turn Exchanges. *Discourse Processes*, 24:37-55.
[11] Fox Tree, Jean E. & Josef C. Shrock. 1999. Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language*, 40:280-295.
[12] Fromkin, V. A. 1980. *Errors in linguistic performance: Slips of the tongue, ear, pen and hand*. New York: Academic Press.
[13] Garrett, M. F. 1975. The analysis of sentence production. In G. H. Bower (Ed.), *The psychology of learning and motivation (Vol.9)*, pp. 133-177. New York: Academic Press.
[14] Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. New York: Academic Press.
[15] Hansson, P. 1998. Pausing in spontaneous speech. www.ling.lu.se/persons/Petra/papers/Pauses98.html
[16] Jenkins, Jennifer. 2000. *The Phonology of English as an International Language*. Oxford: OUP.
[17] Langer, H. (1990). Syntactic normalization of spontaneous speech. *Proceedings of COLING 90*, pp. 180-183.
[18] Levelt, W. 1989. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
[19] Maclay, H., & Osgood, C. E. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15: 19–44.
[20] Makkai, Adam. 1980. *Periods of Mystery: Or, Syntax and the Semantic Pause*. Rice University Studies. Houston, TX. 66(2):125-41.
[21] Mayer, Jörg. 1997. *Intonation und Bedeutung*, http://elib.uni-stuttgart.de/opus/volltexte/1999/386/
[22] Mayer, Jörg. 1999. Prosodische Merkmale von Diskursrelationen. *Linguistische Berichte*. 177:65-86.
[23] Mukherjee, Joybrato. 2000. Speech Is Silver, but Silence Is Golden: Some Remarks on the Function(s) of Pauses. *Anglia: Zeitschrif für Englische Philologie* (Anglia). 118(4):571-84
[24] Oviatt, S. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9:19–35.
[25] Scanlan, Timothy. 1987. Improving Fluency in Spoken French through a Study of Native Pause Behavior. *Foreign Language Annals*. 20(4):345-352.
[26] Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. 1991. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, 60:362-367.
[27] Schmid, M., B. Köpke, M. Keijzer & L. Weilemar, eds. 2004. *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*. Amsterdam/ Philadelphia: John Benjamins
[28] Shattuck-Hufnagel, S., & Klatt, D. 1979. The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learningand Verbal Behavior*, 18:41-55.
[29] Shriberg, E. 1996. Disfluencies in Switchboard. *Proc. ICSLP '96*, Vol. Addendum, 11-14. Philadelphia, PA.
[30] Shriberg, Elizabeth. 2001. To 'errr' is Human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31:153–169.
[31] Siegman, Aron Wolfe. 1979. Cognition and hesitation in speech. *Of Speech and Time*. ed. by Aron Wolfe Siegman and Stanley Feldstein. Hillsdale, NJ: Lawrence Erlbaum Associates. p. 151-178.
[32] Stenstrom, Anna-Brita. 1986. A Study of Pauses as Demarcators in Discourse and Syntax. Aarts,-Jan (ed.); Meijs,-Willem (ed.). *Corpus Linguistics, II: New Studies in the Analysis and Exploitation of Computer Corpora*. Amsterdam : Rodopi. pp. 203-218.
[33] Swerts, M, A. Wichmann, & R.-J. Beun. 1996. Filled Pauses as Markers of Discourse Structure. *Proc. ICSLP'96*, Philadelphia, PA.
[34] Tseng, J. 2000. The treatment of V1 Parentheticals in the B8 Fragment. In J. Tseng (ed.), *Aspekte eines HPSG Fragments des Deutschen*. Arbeitspapiere des SFB 340, Bericht Nr. 156, UniversitätTübingen. pp. 56–83.
[35] Van-Winckel, Nance. 1982. The Role of Preconscious Thought in the Composing Process. A Journal of Composition Theory (JAC). 3(1-2): 102-115.
[36] Ward, Nigel. 2004. Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. *Speech Prosody 04*, pp. 325-328.
[37] Ward, Nigel. 2005. *Non-lexical conversational sounds in American English*. www.cs.utep.edu/nigel/pubs.html
[38] Wode, Henning. 1968. Pause und Pausenstellen im Deutschen. *Acta Linguistica Hafniensia: International Journal of General Linguistics* 11: 148-169.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 77-81.

The intra-word pause and disfluency in Dalabon

Janet Fletcher, Nicholas Evans*, & Belinda Ross***

* University of Melbourne, Australia
** University of Aarhuis, Denmark

Abstract

Earlier impressionistic analyses of Dalabon indicate that the grammatical word is often realized as either an accentual or an intonational phrase, followed by a pause. Unusually, it can also be interrupted by a silent pause, with each section being potentially (although not necessarily) realized as separate intonational phrases. Our analyses of pause duration and pause placement within grammatical words support these earlier impressions, although this use of the silent pause appears to be restricted to certain affix boundaries, and other phonological constraints relating to the following surrounding linguistic material. These interruptions also share certain characteristics of "normal" disfluencies however.

1. Introduction

There are relatively few experimental studies of disfluencies and repairs in typologically unusual languages. This paper examines silent pause placement duration and disfluency in Dalabon, a polysynthetic Australian language spoken in Arnhem Land in the Northern Territory of Australia (see Fig. 1). One interesting feature of Dalabon is that grammatical words can be interrupted by placing a silent pause after a pronominal prefix. Evans et al. [1] found that this is not a widespread feature of pausing in the language. It is also not at all clear what additional "functions" these pauses might have, if we can assume that speakers use pausing strategies as one means of structuring their spontaneous discourse. It is also not clear how these silent pauses differ from those that might more generally be associated with disfluency on the one hand, or the marking of larger discourse segments on the other. Figure 2 shows a typical example of this phenomenon. The prefix "kenh" is isolated from the rest of the following grammatical word by a long pause. The second (much shorter) pause also interrupts the fully inflected grammatical word.

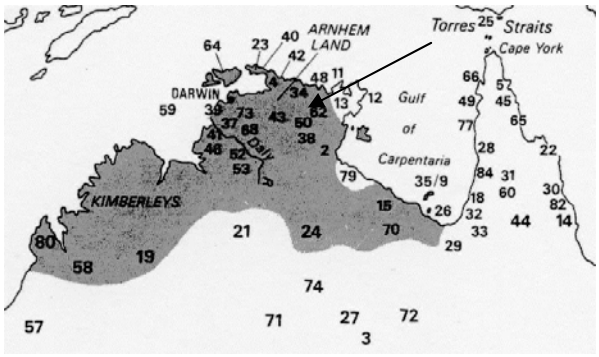


Figure 1: A map of Northern Australia showing where Dalabon is still spoken. The map also shows the location of other indigenous languages that are still spoken in Northern Australia

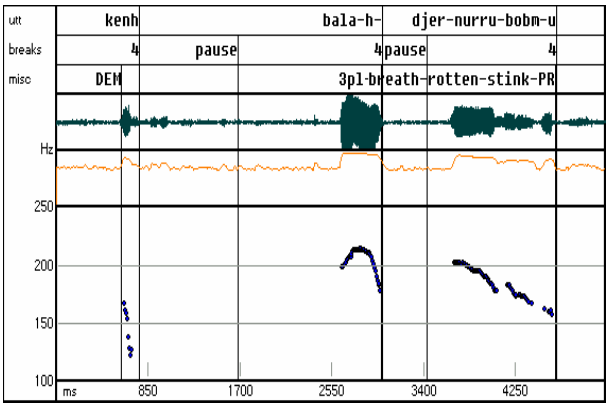


Figure 2: Speech waveform, fundamental frequency contour and rms amplitude trace for a stretch of speech illustrating an isolated prefix "kenh", a demonstrative.

Recent analyses of disfluencies (e.g. Shriberg, [2]) suggest that a distinction needs to be drawn between unfilled pauses that are part of the "editing" phase in the speech production process, and those that might be considered planning pauses. The temporary interruption may signal a following repair or correction. This does not appear to be the case in the above example. The first interruption in the grammatical word illustrated in Figure 2 is likely to be planned, in so far as the initial isolated unit reflects the properties of a full intonational phrase in Dalabon, with an intonational pitch accent and a final falling boundary tone. However, it is not clear whether the second pause (evident by the second silent stretch in the speech waveform) is the same kind of pause.

It appears that disfluency in Dalabon is rarely signaled by filled pauses like "um" or "uh" that you find in General Australian English or "eh" in Scottish English, for example. While it remains to be empirically tested, silent pauses seem to be the main cue to signal interruption to the speech production process in this language.

If we take into consideration, the three regions in a disfluency, (after Levelt, [3] & Shriberg, [2]), the Reparandum, the Editing phrase, and the Repair, it may well be that we need to examine the characteristics of the first and last regions with respect to disfluencies in Dalabon. Disfluency pauses appear to coincide with repetition, deletion or substitution of units, such as pronominal prefixes, which should otherwise be attached to a following verbal unit. When there is a true Repair, that is when the morpheme or morpheme sequence is either repeated or changed to a different unit, there is still a perceptibly clear silent pause. Initial observations also suggest that the intonation contour is slightly different at the edge of the Reparandum region, with a less clearly defined falling intonation pattern. Rather, the tune tends to be either slightly falling, or sustained mid-level.

Figure 3 below shows an example of an interrupted verbal word complex that shows the regions in disfluency. From the speech waveform, it is evident that there are two major silent pauses. From left to right, we see a Reparandum, which is then repaired, but this in turn is effectively both a Repair and a Reparandum, which by the third and final stretch of speech shown below, is a true Repair. It remains to be seen however, whether this is a typical pattern in spoken Dalabon narratives.

The aims of this study were to investigate intra-word pauses, and silent pauses in disfluency regions to see whether there was any significant difference between the duration of these pause-types, and the pauses that coincide with larger discourse segments. An additional aim was to see whether the right boundary of any Reparandum was also signaled by a different F0 level to those that usually signal the edge of complete grammatical units that coincide with full intonational phrase edges, or units consisting of detached prefixes that may or may not coincide with intonational phrase boundaries.

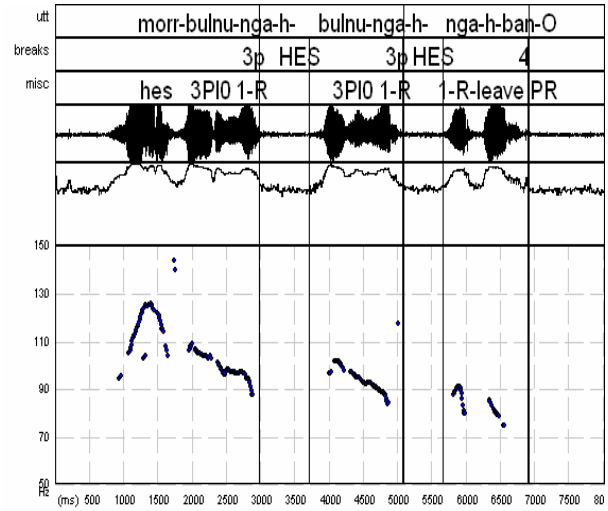


Figure 3. Speech waveform, F0 & Rms contour for a Dalabon utterance showing 2 unfilled pauses in Dalabon preceding 2 Repairs for the utterance: 'them, I, them, I, I leave them'

2. Method

The corpus in this study consisted of two narratives produced by a male and female speaker of Dalabon. Approximately 30 minutes of connected speech were analysed for the female speaker (MT), and 15 minutes for the male speaker (JC). The corpus was recorded by the second author in Northern Arnhem Land during a linguistic fieldtrip. The field tapes were transcribed in practical orthography (see Tables 1 & 2), and also glossed and translated by the second author.. The recordings were subsequently digitized at 22 KHz using esps/Waves+, running on a SUN work station in the Phonetics Laboratory, University of Melbourne. The signal files were then annotated using EMU (Cassidy & Harrington, [4]). Silent pauses were identified and annotated using the acoustic waveform and spectrogram as a guide. Any silent gap in the waveform of 200 ms or more was labeled as an unfilled pause. A conservative duration threshold was chosen because long stops in this language are often between 150-180 ms or more in duration (Evans et al [1]).

Break indices were also annotated. Earlier prosodic studies of Dalabon (e.g. Fletcher & Evans, [5]), as well as the current one, are located within the autosegmental-metrical (A-M) intonational framework (e.g. Pierrehumbert, [6] Ladd, [7]) among others. A version of ToBI (Tones and Break Indices) has also been devised for the language, whereby levels of prosodic constituency are also annotated, along with intonational targets (after Beckman & Ayers-Elam, [8]). Earlier work suggests that minimally three levels of constituency need to be acknowledged in Dalabon: a break index value of 1 indicates there is minimal juncture between adjacent words; a break index value of 3 indicates tonal juncture of some kind, i.e. a falling or rising intonation contour the end of a word, followed by a pitch restart on a following word. Provisionally, we will describe this as an accentual phrase. Finally, a break index value of 4 indicates a full intonational phrase. A break index (BI) 4 marks the highest degree of perceived juncture. The phonetic cue indicating a break 4 constituent is a phrase-final intonational movement, lengthening of a final syllable, and generally a silent pause. It was expected that detached prefixes that displayed the above characteristics were realized as well-formed prosodic constituents. In the case of disfluency regions however, a BI value of 4p was used because it was not at all clear whether the Reparandum region of disfluency was actually a full intonational phrase break. F0 values were extracted at these label points to get an indication of pitch height at the edge of Reparanda, versus full intonational phrase boundaries marked by a BI 4. The distribution of silent pauses, median pause duration was calculated. Instances of a grammatical word being interrupted by a pause were also noted, as were disfluencies.

Table 1. Consonant contrasts in Dalabon. The practical orthography used for each sound is included in parentheses.

		Place of Articulation					
		Peripheral		Apico-		Lamino-palatal	Glottal
		Bilabial	Velar	alveolar	retroflex		
Manner of Articulation	Short stop	(b)	(k)	(d)	(rd)	(dj)	(h)
	Long stop	(bb)	(kk)	(dd)	(rdd)	(djdj)	
	Nasal	(m)	(ng)	(n)	(rn)	(nj)	
	Lateral			(l)	(rl)		
	Rhotic			(rr)	(r)		
	Semi-vowel	(w)				(y)	

Table 2. Vowel contrasts in Dalabon

	Front	Central	Back
Close	i		u
Mid	e		o
Open		a	

minimal juncture between adjacent words; a break index value of 3 indicates tonal juncture of some kind, i.e. a falling or rising intonation contour the end of a word, followed by a pitch restart on a following word. Provisionally, we will describe this as an accentual phrase. Finally, a break index value of 4 indicates a full intonational phrase. A break index

(BI) 4 marks the highest degree of perceived juncture. The phonetic cue indicating a break 4 constituent is a phrase-final intonational movement, lengthening of a final syllable, and generally a silent pause. It was expected that detached prefixes that displayed the above characteristics were realized as well-formed prosodic constituents. In the case of disfluency regions however, a BI value of 1p was used because it was not at all clear whether the Reparandum region of disfluency was actually a full intonational phrase break. F0 values were extracted at these label points to get an indication of pitch height at the edge of Reparanda, versus full intonational phrase boundaries marked by a BI 4. The distribution of silent pauses, and mean pause duration were calculated. Instances of a grammatical word being interrupted by a pause were also noted, as were disfluency regions, where the editing phase was a silent pause. That is to say, we separated those kinds of grammatical word interruptions that involved the detachment of a prefix, from other kinds of interruptions that were more like a "conventional departure from fluency" (Shriberg, [2:160]).

3. Results

3.1. Pause duration and pause distribution

Figures 4-6 plot the durational distribution and length (ms) of all unfilled silent pauses, pauses that are part of the editing phase of a disfluency region, and pauses within grammatical words (after detached prefixes) for speakers MT and JC. Altogether, 648 silent pauses were measured across the corpus. As would be expected, the distribution of pauses is skewed to the left. The mean duration of silent pauses that are preceded by full intonational phrase boundaries, that coincide with fully inflect grammatical words is 941ms for speaker MT and 612 ms for speaker JC. For ease of exposition, we refer to these pauses as "standard" pauses. The pauses range in duration from 207 ms to 6160 ms for MT, and 200ms and 1874 ms for JC. The latter speaker pauses less often than speaker MT

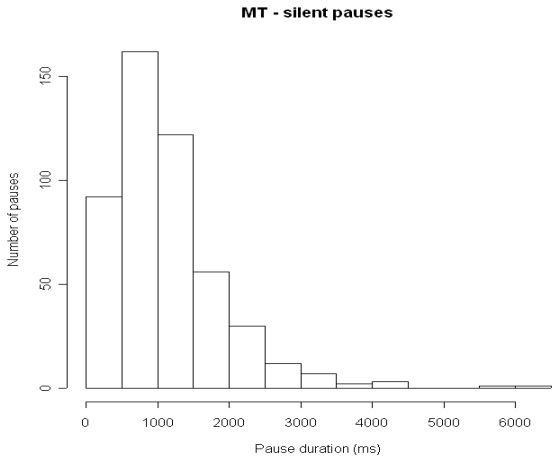


Figure 4a. Distribution and duration of intonational phrase final silent pauses produced by speaker MT.

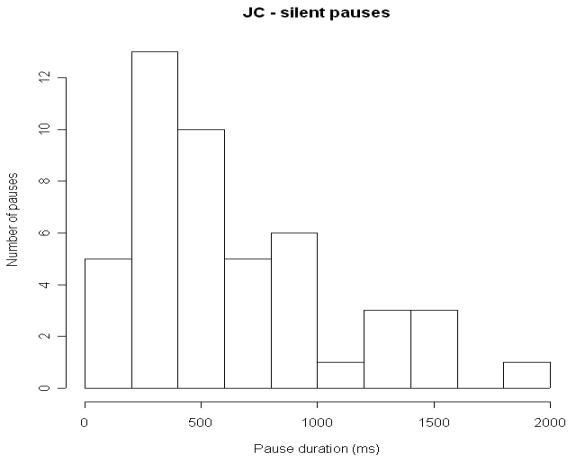


Figure 4b. Distribution and duration of intonational phrase final silent pauses produced by speaker JC.

The results of a one-way ANOVA are highly significant for speaker MT ($F=17.11$, $p<0.0001$) suggesting that there are clear differences in duration for the different pause types for this speaker. The differences are less significant for speaker JC ($F=3.42$, $p<0.05$), although a similar trend is apparent. Figures 4a and 4b, show that the proportion of hesitation silent pauses (i.e. pauses that constitute the "editing" phase of a Reparandum) are fewer in number than "standard" silent pauses, and tend to be somewhat shorter. There are 66 instances of silent pauses following a Reparandum and preceding a Repair for speaker MT, and only 18 for speaker JC. Comparing the pause durations plotted in Figures 5a & b, there are no significant differences between the mean duration of disfluency pauses (481ms) and within-word pauses, however ($t=1.03$, $p>0.05$) for speaker MT.

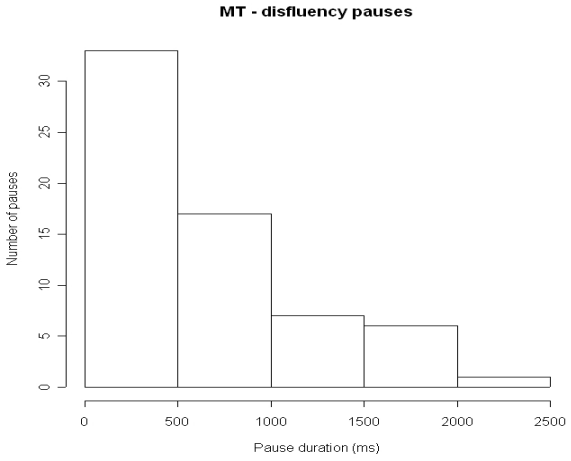


Figure 5a. Distribution and duration of "disfluency" silent pauses produced by speaker MT.

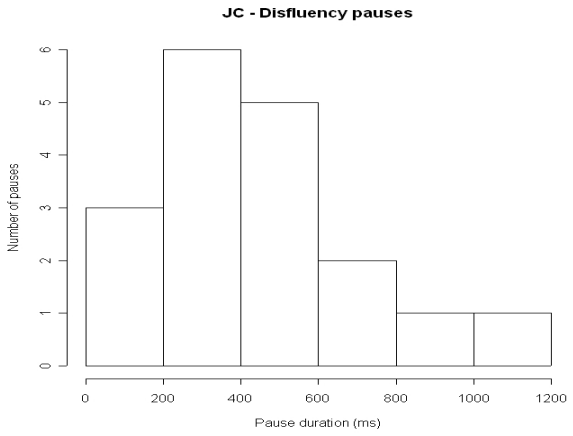


Figure 5b. Distribution and duration of "disfluency" silent pauses produced by speaker JC.

Figure 6 shows there are far fewer pauses within grammatical words, i.e. after detached prefixes, than standard pauses that occur after "complete" grammatical words (27 versus 488 instances for speaker MT). Speaker JC produced only 4 instances of detached prefixes followed by a pause. For MT, the difference in mean duration between the two types of pause is highly significant (t=5.05; p<0.0001), with intra-word pauses being more than half the length of standard pauses (440 ms versus 981ms). The four intra-word pauses produced by JC are also very short, being around 200-250 ms in length.

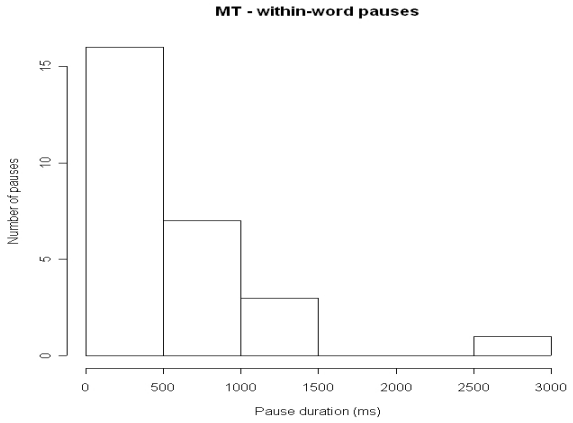


Figure 6. Distribution and duration of "within-word" silent pauses (not hesitation disfluencies) produced by speaker MT.

3.2 F0 analysis

Figures 7a & b plot the final measurable F0 value in the final voiced segment of the constituent preceding a) a "legal" within word pause i.e. at the right edge of a detached prefix b) a Reparandum, and c) a fully formed grammatical constituent and intonational phrase. Both a) and b) are labeled as 4p on the Break Index tier. The analysis of F0 values associated with the edge of BI 4 and 4p constituents shows that for speaker MT, (Figure 4a) there is a small difference between all three boundaries. Most BI4 prosodic constituents have a lower F0 value (116 Hz) at their right edge, than BI 4p constituents (138 Hz). This difference is weakly significant (w=5050; p<0.05). However there are no significant differences between F0 values at the edge of B14p constituents that are detached prefixes (the left most plot in Figure 6a), and F0 values at the edge of Reparanda (the middle plot).

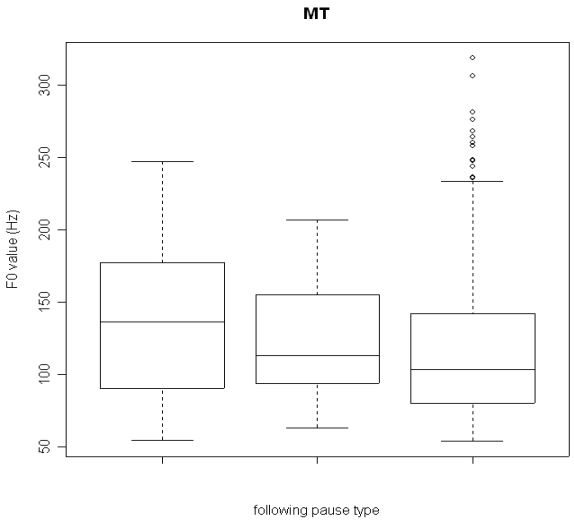


Figure 7a. F0 values extracted at the edge of verbal material preceding the three different pause types: within-word pauses (i.e. after detached prefixes), editing phase interruptions, and intonational phrase-final pauses for speaker MT.

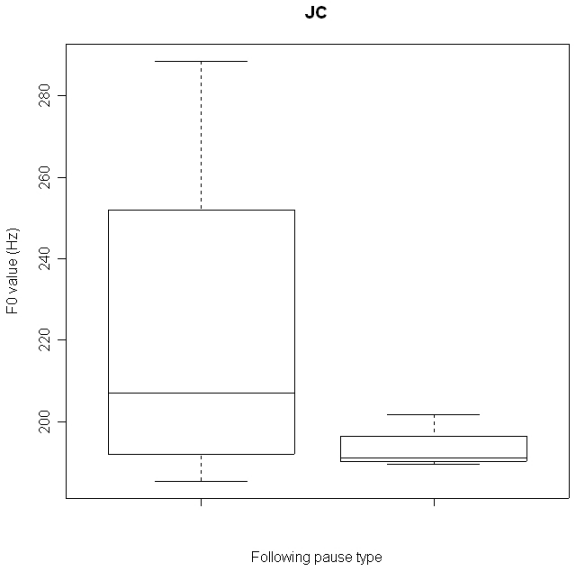


Figure 7b. F0 values extracted at the edge of verbal material preceding two different pause types: editing phase interruptions and intonational phrase-final pauses for speaker JC.

There are differences in the mean F0 values at BI4p and BI4 boundaries for Speaker JC, but these are not statistically significant (Figure 7b). However, speaker JC produced a large number of H% boundaries at BI4 edges which will have significantly influenced the F0 analysis shown in Figure 7b. At BI4p edges, a range of tune levels were observed, which is also clearly reflected here. Speaker MT, on the other hand produced relatively more falling tunes at the edge of BI4 constituents (i.e. L% boundary tones). Relatively low tune values were also apparent at BI4p edges but these were not as low as at BI4 edges. Presumably local pitch range effects like final lowering at the right edge of major discourse segments

will have also influenced the BI4 F0 values observed for both speakers.

4. Discussion

Our results suggest that there may not be significant phonetic differences between the phonetic cues at the right edges of Reparandum regions of disfluencies and detached prefixes. While there are clear pause duration differences at BI4 edges (i.e. "standard" pauses) and at interruption points in the discourse for both speakers, within this category, intra-word pauses (i.e. after detached prefixes) are of similar duration to editing phase pauses. Results of the F0 analysis are less clear cut. However, these values also reflect the different boundary tones used by speakers at BI4 constituent edges. Nevertheless, detached prefixes can exhibit similar boundary configurations to those observed at full intonational phrase edges, but these are also observed at the edges of Reparanda. A closer analysis of these boundaries will be undertaken to determine how many of the detached prefixes really do exhibit full prosodic characteristics of an intonational phrase.

It is perhaps no surprise that we have observed similar patterns between the two kinds of interruptions, particularly when we look at the conditions surrounding "detachment" of prefixes.The detached prefixes are the clearest example of a non-isomorphism between a grammatical word and a phonological word. The two relevant phonological requirements that must be met before pause can occur are that

(a) This must not split a foot, though since feet are constructed over morphemes in virtually all cases, this could also be phrased as a morphological requirement. Because feet have a bimoraic minimum, this effectively means that the prefix must be closed, if monosyllabic, and otherwise satisfy a disyllabic minimum

(b) The remainder of the word must have at least two syllables.

The detached prefixes in this case may have a fully formed intonational boundary. Hesitation pauses function similarly to traditional notions of the Editing phase of a Disfluency region. They occur in connection with repetition, deletion or substitution of units. Interestingly, many of the disfluency regions in this Dalabon corpus involve verbal words that include pronominal prefixes. In other words, the detached prefixes are often the elements that are repeated, or deleted, and are usually re-attached to the preceding word in the discourse. Editing phase interruptions differ from the intra-word pauses in this corpus, in that the unit preceding a pause is either repeated or changed to a different unit, whereas the detached prefix that precedes an intra-word pause is more likely (although not always) to constitute its own prosodic phrase (or BI4 constituent) with a fully formed intonational contour. Once again, a closer analysis of the tonal characteristics at the edge of these units needs to be undertaken.

Diachronically, there are reasons to regard the possibility of breaking a single verbal word into a number of "pause units" as a Dalabon innovation compared to related languages of the region. Comparison with other Gunwinyguan languages, spoken in the same region of Northern Arnhem Land, shows the situation in Bininj Gun-wok, rather than Dalabon, to be the norm – in other words, prefixes do not detach in these languages. Moreover, the emergence of sub-word pause units appears to be linked to a morphological innovation in Dalabon which has had important phonological consequences: the extension of a codal glottal-stop following pronominal prefixes to become the unmarked TAM value, signaling assertativity, rather than the marked type that it is in BGW, where it is confined to the much rarer 'immediate' aspect. We suggest speculatively that this is a historical evolution – what may have started out as a "normal" interruption or a disfluency many well have resulted in a process whereby the relatively long grammatical words of polysynthetic languages may well contain smaller phonological words with relevant prosodic characteristics associated with this level of prosodic constituency. The intra-word pause may well be an indicator of this change in progress in Dalabon.

Whilst preliminary in nature, these findings suggest that an examination of spoken discourse in languages other than the mainstream European or Asian languages may shed light on a number of phonetic and phonological issues, including the role of pause in the structuring of spoken communication

5. Acknowledgements

This research was supported by the Australian Research Council. Thanks to Debbie Loakes who assisted in the analysis of Speaker JC's narrative.

6. References

[8] Beckman, M. & Ayers-Elam, G.(1994). Guidelines for ToBI labelling http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html

[4] Cassidy, S. and Harrington, J. (2001). Multi-level annotation in the Emu speech database management system. *Speech Communication*, 33, 61-77.

[1] Evans, N., Fletcher, J. & Ross, B. (in review) Big words, small phrases: mismatches between pause units and the polysynthetic word in Dalabon. *Linguistics*

[5] Fletcher, J.& Evans, N. (2002). An acoustic intonational study of intonational prominence in two Australian languages. *Journal of the International Phonetic Association*, 32, 123-140.

[7] Ladd, D.R (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

[3] Levelt, W. (1983). Monitoring and self-repair in speech. Cognition 14, 41-104.

[6] Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. Unpublished MIT PhD thesis.

[2] Shriberg, E. (2001). To 'err' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetics Association*, 31, 153-169.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 83-88.

Repair-initiating particles and um-s in Estonian spontaneous speech

Tiit Hennoste

University of Helsinki, Finland & University of Tartu, Estonia

Abstract

Particles and um-s used in spontaneous Estonian speech as initiators of different types of repair are analysed. Our model and typology of repair based on conversation analysis is introduced. Three main types of repair and particles used to initiate those are described: prepositioned self-initiated self-repair, postpositioned self-initiated self-repair (addition, substitution, insertion and abandon), and other-initiated self-repair (reformulation, clarification and misunderstanding). In conclusion 6 groups of particles are brougth out by the role they play in the initiation of the repair sequence. Data come from Corpus of Spoken Estonian of the University of Tartu, which contains everyday and institutional speech, telephone and face-to-face conversations.

1. Introduction

This article is a part of a larger project, which has two aims: to analyze Estonian conversation and to build a Dialogue System which gives information in Estonian and follows norms and rules of human-human communication. One part of the project is find out how the different communication problems are solved in Estonian spontaneous conversation (see also e.g. [3], [8]). We are interested in markers in actual speech, which are used by humans to interpret speech and which are usable as markers of speech forming processes in Dialogue System.

Problem solving is a process, which has been defined and described differently and in different terms (structure schifts, hesitation, communication strategies, repair etc (see e.g. [1], [2], [13])). We have adopted model and term of Conversation Analysis (CA) as basis for our analysis (see e.g. [9], [15], [16]).

By CA repair organization is a process, which is used to solve different communication problems (grammatical mistakes, incorrect word-selection, changing of the sentence plan, misunderstanding, non-hearing etc).

First, repair is divided into four varietes according to different role of the participants:

1) self-initiated self-repair: repair is initiated and carried out by the speaker of the trouble source;

2) other-initiated self-repair: repair is initiated by the recipient and carried out by the speaker of the trouble source;

3) self-initiated other-repair: repair is initiated by the speaker of the trouble source and carried out by the recipient;

4) other-initiated other-repair: repair is initiated and carried out by the recipient.

The most used types in conversation are self-initiated self-repair and other-initiated self-repair. Other types are very rare in native dialogues (but much more used in e.g. L2 conversations).

Second, repair is divided into prepositioned repair (also stalling, hesitation signals etc in different models) and postpositioned repair (also reformulation, retrospective repair, structure schifts etc). The prepositioned repair is used when speaker needs more time than he/she has to solve his/her text-production problems. And so he/she postpones his/her next part of the text, using hesitation sequence. Postpositioned repair is used when some changes take place in the actual text.

The postpositioned repair could be divided by the positions where the repair can occur. There are two main positions. The first one is within the utterance containing the trouble source where self-initiated self-repair takes place. The second position is in the turn following the trouble source. This is a place where the other-initiated self-repair is mostly initiated.

The repair sequence has its beginning and end in speech. As repair is divagation from the "main line" of the conversation, it is natural, that the beginnig of the repair must be marked.

The most used repair initiators are pauses, um-s, particles, repetitions (*ja ja* 'and and'), drawling of the last sound of the word (*mina::* 'I::'), interruption of the word (*põhi-* = põhimõtteliselt 'in principle'), intonational breaks. Longer phrases, sentences and nonverbal fillers (gestures, gaze etc) are used very rarely.

Um-s are vocalizations (*er, um* etc), which are used as hesitation pause-fillers in different languages.

Particles (also *D-items, discourse particles/markers, pragmatic particles/markers, inserts* etc) are defined differently in different models. In our model they are items (words or word forms) which have no denotative/lexical meaning and no syntactic relations with other structures in dicourse/conversation (see [4]). Their use is defined by their pragmatical function. They may appear in their own in discourse (as *uh huh*) and may be attached prosodically to some larger structure. Some of them are homonyms of the other word classes (mainly adverbs or conjunctive words, e.g. *et* 'that' or *nagu* 'as', but also verb forms, e.g. *kule* 'hear').

In this article we will describe the particles and um-s, which are used to initiate repair and mark the beginning of the repair sequence in Estonian spontaneous speech. We will concentrate on three main types of repair: prepositioned and postpositioned self-initiated self-repair and other-initiated self-repair.

Our data come from Corpus of Spoken Estonian of the University of Tartu [6]. This corpus containes about 1 000 000 tokens. Subcorpus of 130 000 tokens is used in this study, both everyday and institutional speech, telephone and face-to-face conversations.

The CA transcription is used (see Transcription marks).

2. Prepositioned self-initiated self-repair (hesitation sequence)

The most used hesitation markers in our corpus are very short pauses (3200 items), drawlings (2400) and um-s. There are about 2600 um-s in analyzed corpus. About 75% of them work as hesitation signals in Estonian spontaneous speech. The others are used to initiate postpositioned self-initiated self-repair (see 4.3).

The most frequent um is *ee* in spoken Estonian (38% of all um-s; example 1), the other more frequent items are *õõ, ää* and *mm.*

(1) ***ee*** *(.) ja kui `kaua (.) mi- `mis ajal peab puh noh enne kaitsmist ütleme* ***ee*** *mul on viie`teistkümnes millal ma pean selle bakalaureusetöö `esitama.*[1]

'***EE*** *(.) and how many time (.) in what time I must bring let's say* ***EE*** *my day is the eleventh when I must bring my BA thesis*'[2]

There is one particle, which main function is to initiate prepositioned self-initiated self-repair in spoken Estonian: *noh* (untranslatable, example 2). This is also the most frequent particle in spoken Estonian (1056 tokens in our corpus).

(2) *C: ee tervist. ega ´teie ei oska kogemata öelda Tallinna ´busside aegu. (0.5)*
A: ja ´milliseid busse. (.)
C: ee (.) ***noh*** *millal lähevad ´viimased ´kiirbussid täna ´Tallinna. (.)*
A: kell=´öheksa on viimane ekspress.

'*C: ee hello. couldn't you tell me the departure times of the buses to Tallinn. (0.5)*
A: and which buses. (.)
C: ee (.) ***NOH*** *when the last express buses go to Tallinn today. (.)*
A: at nine o'clock is the last express.'

The other particles are used more rarely as hesitation signals. The more used are *nagu* 'as' (example 3); *tähendab/tändab* 'it means'; *jah* 'yes'; *see* 'this'; *ku(r)at* 'devil'; *ütleme* 'let's say' (example 4) etc.
The list of hesitational particles is open one, as there are lot of individual particles used by some or one person only.

(3) *C: mhmh (.) aga kuidas teil Itaaliaga on. kas n- (0.5) kas* ***nagu*** *on ´sama olukord.*

'And how is with Italy. Is it ***NAGU*** *(0.5) is it the same situation.'*

(4) *jaanuari´kuus nädalaks ´ajaks sis* ***ütleme*** *näiteks kui te võtaksite nüd aparta´mendi sis apartamendi ´hinnad ´kõiguvad kuskil* ***ütleme*** *nii .hh ´neljasaja:: neljasaja ´viie´kümnest ja sealt ´üles.*

'in january for a week then ***let's say*** *for example if you could take an apartment then the prices of the apartments are about* ***let's say*** *four hundred fifty and up.'*

Typical hesitation sequence (about 70%) consists of one hesitational unit (pause, drawling etc) in Estonian conversation. The other larger group is a combination of different means (25%), the most typical are:

- repetition of the conjunctive words + particle: *et et noh* 'that that NOH';
- two different particles: *nagu noh* 'as NOH', *noh ütleme* 'NOH let's say';
- pause + particle: (...) *noh* '(…) NOH';
- pause + um: (...) *ee* '(…) EE'.

The rest 5% are the repetitions of the same means (*noh noh* etc).

3. Postpositioned self-initiated self-repair (self-reformulation)

Self-reformulation is the repair, where speaker changes something in the his/her actual speech.

3.1. Model of self-reformulation

There are different models of the self-reformulation. Our model is based on the models of CA and of Enkvist & Björklund, who use Willem Levelt as basis (see [2], [13], [16]). Our model of the self-reformulation process is as follows:

1. The problem arises and the speaker decides to initiate reformulation.
2. He/she could begin immediately after the problem or with delay.
3. Typically some reformulation initiation marker is used at the beginning of the repair sequence (but it is possible to begin smoothly, without any marker).
4. Then the editing phase could follow sometimes, during which the speaker makes new construction. It is represented by hesitation sequence in the speech.
5. Then the reformulation (=new part of the text) follows.
6. Typically the repair sequence ends smoothly, but it is possible to mark it also explicitly by end marker.
7. There is a possibility to comment the process of the reformulation or the new construct by speaker him/herself or by hearer. Typically there is no any comments.

3.2. Typology of self-reformulation

There are different typologies of the self-reformulation. Our typology is based on CA (see e.g. [15]).

First, we divide reformulations into two groups. The basis is, whether the reformulation takes place in the same utterance (without intonational break) or not. There are two types of repair in the same utterance: substitution ja addition.

Substitution is a reformulation where some part of the utterance (typically word) or grammatical form of some part of the utterance is replaced with the new one (e.g. *vati* GENITIVE > *vatti* PARTITIVE in example 5). Addition is a reformulation, where a new item is added afterwards into utterance (*valget* in example 5). The most typical addition is adjectival attribute in Estonian.

(5) */---/ need olid siis iluduseks siis pandi* ***vati*** *(0.5) ee* ***vatti valget vatti*** *nagu oleks lumi sadand, (0.3)*

'/---/ those were as a decoration then was put the cotton [GENITIVE] (0.5) ee cotton [PARTITIVE] white cotton [PARTITIVE] like snow, (0.3)'

There are two additional reformulations which are used after intonational break: insertion and abandon.

Insertion is a reformulation, where speaker interrupts syntactical construction, adds a new one, finishes it and then finishes the interrupted utterance (example 6). Insertion itself is an information which is brought in after or before its "right place" in the text.

[1] C = caller; A= answerer/information officer in telephone conversation examples.
[2] The word-by-word translations are used in examples. The nontranslatable particles and um-s are written in upper case in translations (NOH).

(6) *ja siis mul tuli täna noh s ma mõtsin = et sööks midagi kui teatrist ära lähen = ja (.) siis mul tuli meelde et ma pole (1.0) isegi (.) (muud)* ***a pühapäeval jah sõin midagi*** *= aga (.) aga et ma pole sel nädalal* ***täna on neljapäev*** *et ma pole sel nädalal absoluutselt mitte ühtegi = m soolast asja söönd.*

'and then I remembered today that I want to eat something when I leave theater and (.) then I remembered that I haven't (1.0) even ***oh on Sunday yes I eat something*** *but (.) but that I haven't eaten this week* ***today is Thursday*** *that I haven't eaten absolutely nothing solty this week.'*

Abandon is a reformulation, when the utterance is interrupted and the new one is introduced (example 7).

(7) */---/ mm ma leian minu jaoks see siiski kontrolli kriteerium on (1.0) ma räägin kontrolli kriteerium on see.* ***ma olin tegelikult*** *kui me nüüd läheme tagasi Nikaraaguasse millest ma lootsin=et me tegelikult täna räägime.*

'/---/ mm I find it's a control criterion for me yet (1.0) I say it is a control criterion. ***I was really*** *if we now go back to Nicaragua what I hoped we will talk about today.'*

75% of self-reformulations are substitutions and 20% are additions in our corpus. The insertion and abandon are rare (about 5% together). Abandon is more used in radio intervievs and insertion in longer narratives. Of course, different combinations of the types are used in conversation.

3.3. Self-reformulation initiation markers

The different markers are used to initiate different self-reformulation types.

The most used initiator of the substitution is interruption of the last word before the substitution sequence (typically 1-2 syllables are pronounced: *arva-* (example 8).

The substitution-initiating particles are:

- *või / või siis / või seal / või parem / või ühesõnaga* ('or / or then / or there / or better / or in one word'; example 9);
- *ei / mitte* ('no'; example 10);
- *tähendab /tändab* ('it means'; example 11);
- *noh* (example 8);
- um-s.

The other hesitation signals are also used (pause, drawling; example 8).

(8) *no ma nimetaksin siis kohe nimodi käigupealt kahte assja, mida* ***ma: > (0.5) noh*** *mis on siis selles mõttes üllatus, (0.5) et ma* ***arva-*** *arvasin, et: ma lahkun: (0.5) pigem negatiivsete emotsjoonidega, (a)ga ma lahkusin väga positiivsetega. >*

'then I will mention two things at once, what ***I: (0.5) NOH*** *that are surprises in this sense (0.5) that I* ***thou-*** *thought that: I will leave (0.5) with negative emotions, but I leaved with very positive ones.'*

(9) */---/ ja:: ja=siis (...) mm kirik oli* ***või*** *tol korral oli väga külm ilm olnud (.) kirik ka väga külm.*

'/---/ and:: and=then (…) mm church was ***or*** *it was very cold weather this day and church also very cold.'*

(10) *.hh no sis ma olen nüd vahepeal jalutand tast kaugemale ku=ma lähen* ***ei*** *kui teil vaja on /---/*

'.hh then I have now walked far from him if I go ***no*** *if you need /---/'*

(11) *sest mina (.)* ***tähendab*** *ma tahtsin seda sulle rääkida=et (.) mul tuli täna teatris meelde täna ma ei ole ka loomulikult midagi söönd.*

'because I [long form of I] (.) ***it means*** *I [short form of I] wanted to tell you that (.) I remembered in theatre today that I haven't eaten nothing today too.'*

There are only some markers to initiate addition: interruption of the word, pause and um-s (example 12). No particles are used in our corpus.

(12) *veel võiks olla nii et ee (1.0) et ülejäänud* ***ra-*** *väheses rasvas /---/*

'furthermore it could be so that ee (1.0) that in the rest of ***grea-*** *few grease /---/ '*

The marker of abandon is intonational break. Sometimes hesitational markers are used at the beginning of the new utterance (pause (example 13), *noh* and um-s).

(13) *aga enne seda kui nad Väljaotsale läksid?* ***(.)*** *ma=tahan= veel=rääkida kuidas Lutsu (.) Palamuse kirikus ristiti (...)*

'and before they went to Väljaotsa? ***(.)*** *I want to tell you how Luts was baptised in Palamuse church'*

The marker of the insertion is intonational schift, and sometimes hesitational signals (pause, drawling and um-s) are used in addition.

Some particles and conjunctive words are used at the beginning of the insert sequence, but their role is not to initiate insertion but to show the semantical or/and pragmatical connection between insertion and main utterance. The most used particle is *no* (which is mainly the used to introduce a new (sub)topic in Estonian conversation, example 14; see [4]).

(14) *see Oskar Lutsu ema tädi* ***(.) õ no****=mälestuste raamatus nimetati seda ka Murumunaks see=sis andis /.../).*

'this aunt of mother of Oskar Luts ***(.) õ NO*** *in the memoirs she was called also Murumuna she gave /---/'*

4. Other-initiated self-repair (next-turn repair)

Other-initiated self-repair is initiated by hearer, who have found some problem in the speakers previous text. We call it next-turn repair, as 90% of those repairs are initiated immediately after the problematic turn in our corpus.

4.1. Typology of next-turn repair initiations

The typology of the other-initiated self-repair is also different in different problem solving models. We have divided those repairs into three groups: clarification, reformulation and misunderstanding (see also [3]).

Clarification is an initiation, by which the hearer repeats exactly or with some variation some utterance, phrase etc of the previous speaker to get confirmation that it was such (did you say that?).

Reformulation (also candidate understanding in CA) is an initiation, by which hearer gives his/her own interpretation (hypothesis, rewording, generalisation etc) to the speakers turn. His/her aim is to get confirmation, that his/her understanding is true (did you think that?)

Misunderstanding is an initiation, by which initiator reports that he/she did not hear or did not understand the previous information, or the information contradicted so much with his/her knowledge and beliefs that it must be checked.

There are two subtypes of misunderstanding:

1) the speaker only indicates that there was a problem;

2) he/she localises the problem more exactly.

About 50% of repair initiations are clarifications, 25% reformulations and 25% misunderstandings in our corpus.

4.2. Next-turn repair initiation markers

The main initiation means of the next turn repair are question phrases and sentences.

There are five types of questions in our dialogue act typology: wh-question, open yes/no question, closed yes/no question, question that offers answer and alternative question (see [8]).

Open and closed yes/no questions have similar grammatical form but they expect different answers. A closed question expects the answer yes or no (*Are you open in winter? – Yes.*) while an open question expects giving information (e.g. the question *Is there a bus that arrives to Tallinn after 8?* the client intends to learn the departure times of buses).

Closed yes/no questions and questions that offer answer are both questions that expect yes/no answer. Their difference lies in the presuppositions of the user. Asking a question that offers answer the speaker has an explicit opinion, hypothesis, and he/she is expecting a confirmation by the partner. No such presupposition exist in case of a closed yes/no question.

Clarification and reformulation expect yes/no answer (=closed yes/no question or question that offers answer), misunderstanding expects information (=open yes/no question, wh-question or alternative question).

There is a difference between question particles used in clarification and reformulation.

Clarification is mostly (55% of initiations) expressed by interrogative intonation only. The most frequent particle is *jah* ‘yes’ pronounced with interrogative intonation at the end of the utterance (20%; example 15).

(15) *A: ee=ütleme et=e kella ´kümnest=ee=h (.) ´neljani on ´vaba. (1.0)*
C: õõ siis paneks (2.0) kas kell ´kaks=või või pool ´kolm=või.
A: õõ=hh (.) pool=´kolm ***jah****?*
C: jah. (.)
A: ee ja kuidas ´nimi on.
C: Saabas. (.)
A: Saabas ***jah****.*
C: jah

’A: *ee=let’s say that=e from ten o’clock=ee=h (.) till four is free. (1.0)*
C: õõ then [we] would take (2.0) at two o’clock=or or a half past two or.
*A: õõ=hh (.) half past two=****yes****?*
C: yes. (.)
A: ee and what is name.
C: Saabas. (.)
A: Saabas ***yes****.*
C: yes ’

The second most used particle is *või/vä* (7%, example 16). The other particles are used rarely (*ühesõnaga* ‘in one word’, *et* ‘that’, *siis* ‘then’, *kas* ‘question particle of yes/no-question in Standard Estonian’).

(16) *A: ega ta üksi- kahekesi ei olnd, nad olid nelja-neljakkeisi.*
C: ***neljakesi = vä.***
A: mhmh? /---/

’A: she was not alone, there were four persons.
C: ***four-VÄ.***
A: yes.’

28% of reformulations are initiated using questions marked by interrogative intonation only. The main particle used here is *et* (22%) at the beginning of the utterance (example 17).

(17) *C: ee milliseid teil ´pakkuda on.*
A: $ e:i ´pakkuda ei ole meil ´ammu enam midagi. $
C: [aa hehe]
A: [ma=´mõtlesin et te] olete üks $ ´reisijatest. $
C: aa, ***et*** *teil on kõik ´välja*
A: oi loomu´likult.

’C: ee what do you offer.
A: $ no: we have nothing for a long time. $
C: [aa hehe]
A: [I thought that you] are one of the passangers.
C: oh, ***so*** *you have sold out all*
A: yes of course.’

The second most used particle is *jah* ‘yes’ at the end of the utterance pronounced with interrogative intonation (11%). The other particles are used less than 10%: *nii et* ‘so that’, *siis/sis* ‘then’, *kas, või/vä, tähendab* ‘it means’, *eks* ‘isn’t it’.

Misunderstanding is divided into two subgroups, according to the exactness of the localisation of the problem.

1. General initiations are used only to mark, that there was a problem in the preceding turn (example 18). Most of those initiations are formed by general question particles and question words only (*jah* ’really?’, *ah* ’what?’, *kuidas* ’what?’, *mis / mida / millega* etc ’what?’).[3] Sentences are used rarely (e.g. *mis sa ütlesid* ’what did you say?’).

(18) *K: räägi kuda siis on. (.)*
M: ***millega****.*
K: ülemusega sõit läks. (1.5)
M: no mis ta on, mis seal minna oli.
K: ***ah?***
M: mis seal minna oli.

’K: Tell now how it was. (.)
M: ***What****.*
K: Your trip with boss.
M: nothing interestong, normal.
K: ***What****.*
M: Normal.’

[3] *mis* is a prononun and a question word in Estonian, which is declinable.

2. More exact localisation of the problem is made by repetition of the problematic part of the turn (sentence, phrase, word) using question words that localize the problem (*kus*? 'where?', *kes*? 'who?' etc). This subtype is formulated almost all by wh-questions in our corpus.

The most frequent question word used in misunderstanding is *kuidas* (39%), which is usable with different intonation to indicate non-hearing or surprise ('what?'; example 18, 19) and to specify some element of the utterance ('how?'; example 20).

(18) *A: ´Estmar=´info, ´Leenu=kuuleb tere*
C: tere. (0.8) {Leenu.}
A: jah? (0.5)
C: rotilõks. (1.8)
*A: **kuidas**?*
C: rotilõks. (0.8)
A: jah, rotilõks
C: {´andke ´kõik.}
*A: **kuidas**?*
C: {´kõik kus ma saan ´osta.}

'A: ´Estmar=´info, ´Leenu=is hearing, hello
C: hello. (0.8) {Leenu.}
A: yes? (0.5)
C: rat trap. (1.8)
*A: **what**?*
C: rat trap. (0.8)
A: yes, rat trap
C: {give me all}
*A: **what**?*
C: {all [shops] where I could buy.}'

(19) *C: tere=ma palun (.) ´Teksako bensiini´jaama. (…)*
A: neli kaheksa kaks, üks kolm üks.
*C: oi ma ei ´kuule, **kuidas** palun.*
A: NELI KAHEKSA kaks, (.) null kolm null.

'C: hello=would you give me (.) ´Teksako oil station. (…)
A: four eight two, one three one.
*C: I don't hear, **what** please*
A: FOUR EIGHT two, (.) one three one.'

(20) *C: öelge (.) kus on Tartus ee ´Kaa sa´long. (1.5)*
*A: e=´**kuidas**=se sa´long oli*
C: ´Kaa. ´Kaa salong.

C: tell me (.) where Kaa salon is located in Tartu. (1.5)
*A: e=**how** this salon was*
C: ´Kaa. Kaa salon

5. Discussion

We can divide particles and um-s into two groups by the role they play in the initiation of the repair sequence in spontaneous Estonian speech.

The first group is particles and um-s used to initiate self-initiated-self-repair:

1) hesitation signals: um-s (*ee, õõ, aa, mm* etc), particles *noh, nagu, tähendab, ütleme, ku(r)at, jah* etc;

2) substitution initiators: um-s (*ee, õõ, aa, mm* etc), *noh, või* (*+siis/seal/parem* etc), *tähendab, ei/mitte* etc.

3) addition initiators: um-s (*ee, õõ, aa, mm* etc)

First, different particles are used differently in spoken Estonian.

um-s are universal self-initiated repair initiators and items which only function is to initiate self-repair in analyzed corpus. They are used in the hesitation sequense, and as initiators of the substitution and addition. They are also used at the beginning of the abandon and insertion sequences, but we have interpreted them as hesitation signals in those positions.

The most frequent repair particle is *noh*. It is unclear, whether *noh* is also universal self-repair initiator or not. There is no *noh* at the beginning of the addition and abandon sequences in analyzed corpus.

noh has also other functions in conversation. It could be a part of the editing phase of reformulation process, where it mostly signals that speaker has found the continuation (see 4.1). And *noh* is used at the beginning of the turn as particle, which signals the pragmatical connections between following and preceding utterances/turns (see [5]).

The other particles are used as particles only in some contexts, and some of them are also homonyms of the other word classes. E.g. *nagu* 'as' is also a softener and adverb; *tähendab/tändab* 'it means' is mostly an initiator of postpositioned repair (see 3.3), but it is also used as the initiator of the formulations, accounts, explanations etc; *jah* 'yes' is mainly response particle (answer to the yes/no question etc) and next-turn repair initiator (see 4.2); *see* 'this' is mainly pronoun and sometimes definite article, *ku(r)at* 'devil' is a swear-word (see also [4], [5], [12]).

The other particles are used typically in different functions at the same time. E.g. *ütleme* 'let's say' initiates example or proposal or works as hedge in Estonian dialogue, but it is almost always a part of hesitation sequence at the same time (see [7], [12]).

Second, there are differencies between self-repair varietes.

Particles and um-s were not used to initiate the abandon and insertion in our corpus. Of course, there are sometimes hesitation sequences at the beginning of the abandon and insertion. Particles and conjunctive words at the beginning of the insertion are used to show semantical or pragmatical connetion between insertion and main utterance (*no* etc).

There were only um-s and not particles used at the beginning of the addition in analyzed corpus.

Both particles and um-s are used to initiate hesitation and substitution. Some of the particles are used only at the beginning of the substitution (*või+siis/seal/parem* etc, *ei/mitte*), some are used also as hesitation signals (*noh, tähendab* 'it means'). It is possible to interpret the last ones as particles, that have two functions at the beginning of the substitution.

The second group is particles used by the hearer to initiate next turn repair:

1) particles and question words used to initiate clarification and reformulation: *jah* 'yes', *ek*s 'isn't it', *et* 'that', *või/vä, siis* 'then', *ühesõnaga* 'in one word', *tähendab* 'it means', *kas, nii et* 'so that';

2) particles and question words used to mark general misunderstanding: *jah* 'really?', *ah* 'what?', *kuidas* 'what?', *mis* etc 'what?';

3) wh-question words used to localize problem more exactly: *kus*? 'where?', *kes*? 'who?' etc.

There are two groups of words used in the inititation of the next-turn repair. The first is question words, which are used mostly in Standard Estonian and/or in "main line" questions (wh-question words, *kas, või/vä*).

kas is a question word of yes/no-question in Standard Estonian. *või/vä* is a yes/no-question word in spoken Estonian, developed from the conjunction *või* 'or' (see [14]).

The second group is question particles (*jah, eks, et, nii et, siis, ühesõnaga, tähendab, ah*).

et ('that') and *nii et* ('so that') are conjunctive words in Standard and spoken Estonian, and also particles ('so') in spoken Estonian. Their main particle function is to mark that the following utterance is a summary or conclusion from the previous text or that it is attributed to some other person than speaker (see [9]).

jah has several different functions. It's main function is response particle (answer to the yes/no question etc; see [4]). In this function it is used mostly separately or at the beginning of the utterance, rarely also in the middle of the utterance. As a question particle it is used at the end of the phrase or sentence in clarification and reformulation, and separately with interrogative intonation to initiate general misunderstanding. And as we have seen, it may be sometimes a hesitation signal.

siis 'then' is mainly an adverb and also particle with different functions in spontaneous Estonian (see [10]).

Eks is a tag question word.

Tähendab 'it means' is also an initiator of postpositioned repair and of the formulations, accounts, explanations etc.

There are some particles which are used in clarification and reformulation (*jah, et, siis, kas*). Some particles are used in reformulation (*eks, tähendab, nii et*), and some in clarification only (*või/vä, ühesõnaga*) in analyzed corpus. It is not clear, whether all those particles are usable in both repair or not.

There is a difference between the formulation of the repair initiating questions and "main line" questions in the dialogue (see [8]). Our analysis has shown that clarification and reformulation are formed by a question that offers answer in almost all cases (except one reformulation formed by wh-question in analyzed corpus), and majority of questions that offer answer (81%) are used for repair initiation.

The "main line" yes/no-questions are mostly closed or open yes/no-questions which are initiated mostly by particles *kas* (55-59%), *või/vä* (9-11%) or *kas+või/vä* (6-7%). The second important difference is that the "main line" questions are mostly formed as full sentences while repair initiations are mostly phrases (or single words). 79% of clarifications are phrases or words. Reformulation and misunderstanding is formed by a full sentence in half of cases.

6. Transcription marks

falling intonation	period.
fall not to low	comma,
raising intonation	question mark?
short interval (max 0.2 sec)	(.)
timed interval	(2.0)
nontimed longer interval	(…)
latching at end of utterance	word=
latching at beginning	=word
drawling	::
stress	` at the beginning of the stressed syllable
cut off	do-
inbreath	.hh
begin of overlap	[
end of overlap	]
loud sounds	UPPER CASE
{transcriber's inability to hear what was said}	
smile	hehe

7. References

[1] Dörnyei, Zoltan & Mary Lee Scott. 1997. Communication Strategies in a Second Language: Definitions and Taxonomies. *Language Learning*, vol. 47, pp. 173–210.

[2] Enkvist, Nils Erik & Martina Björklund. 1985. *Toward a taxonomy of structure schifts.* MS.

[3] Gerassimenko, Olga, Tiit Hennoste, Mare Koit & Andriela Rääbis. 2004. Other-initiated Self-repairs in Estonian Information Dialogues: Solving Communication Problems in Cooperation. *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, April 30–May 1, 2004. Ed by M. Strube, C. Sidner. Cambridge, pp. 39–42.

[4] Hennoste, Tiit. 2000. Sissejuhatus suulisesse eesti keelde IV. Suulise kõne erisõnavara 3. Partiklid [Introduction to Spoken Estonian IV. Particles]. *Akadeemia*, No. 8, pp. 1773–1806.

[5] Hennoste, Tiit. 2001. Sissejuhatus suulisesse eesti keelde IX. Lausung suulises kõnes 4 [Introduction to Spoken Estonian IX. Utterance in spoken Estonian 4]. *Akadeemia*, No. 1, pp. 179–206.

[6] Hennoste, Tiit. 2003. Suulise eesti keele uurimine: korpus [Studing Spoken Estonian: Corpus]. *Keel ja Kirjandus*, No. 7, pp. 481–500.

[7] Hennoste, Tiit. 2004. *Et*-komplementlause kesksete põhiverbide funktsioonid eestikeelses vestluses. [Main functions of complement verbs in Estonian converation]. *Keel ja Kirjandus*, No. 7, pp. 504–523

[8] Hennoste, Tiit, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson & Maret Valdisoo. 2005. Questions in Estonian Information Dialogues: Form and Functions. *Text, Speech and Dialogue. Proceedings of the 8th International Conference, TSD 2005*. ???+Berlin: Springer. In appear.

[9] Hutchby, Ian & Robin Wooffitt. 1998. *Conversation Analysis. Principles, Practices and Applications*. London: Polity Press.

[10] Jansons, Airi. 2002. Partikli *siis* funktsioonid suulises kõnes. [The functions of particle *siis* 'then' in Estonian conversation]. *Keel ja Kirjandus*, No. 9, pp. 612-629.

[11] Keevallik, Leelo. 2000. Keelendid *et* ja *nii et* vestluses [Tokens *et* and *nii et* in Estonian conversation]. *Keel ja Kirjandus*, No. 5, pp. 344–358.

[12] Keevallik, Leelo. 2003. *From Interaction to Grammar. Estonian Finite Verb Forms in Conversation.* Studia Uralica Upsaliensia 34. Uppsala: University of Uppsala.

[13] Levelt, Willem J. M. 1983. Monitoring and Self-repair in Speech. *Cognition*, vol. 14, pp. 41–104.

[14] Lindström, Liina. 2001. Grammaticalization of *või/vä* Questions in Estonian. *Papers in Estonian Cognitive Linguistics.* Ed. by Ilona Tragel. Tartu: University of Tartu, pp. 90–118.

[15] Schegloff, Emanuel. 1979. The Relevance of Repair to Syntax-for-conversation. *Syntax and Semantics, Volume 12: Discourse and Syntax.* Ed. by T. Givon. N.Y: Academic Press, pp. 261–288.

[16] Schegloff, Emanuel, Gail Jefferson & Harvey Sacks. 1977. The Preference for Self-correction in the Organization of Repair in Conversation. *Language*, vol. 52(2), pp. 361–382.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 89–92.

Repeats in spontaneous spoken French: the influence of the complexity of phrases

Sandrine Henry

Équipe DELIC, Université de Provence, Aix-en-Provence, France

Abstract

We here present the results of a descriptive study we conducted on 383 disfluent repeats from a corpus of spontaneous spoken French. We analyze noun phrases under construction and study whether there is a co-relation between the frequency of the repeats and the complexity feature of the phrases. We then focus on complex noun phrases in order to locate precisely the repeats. We also analyze how repeats affect structures such as [Preposition + Determiner + Noun] and what the constraints upon such structures are.

1. Introduction

Disfluency phenomena (such as repeats, word fragments, self-repairs, etc.) can be found in all spontaneous oral productions. Indeed, oral speech, as opposed to finalized writing which is a deferred production, is produced online and therefore has the specificity of retaining the traces of its elaboration. It is by fits and starts and later syntactic and/or lexical readjustments that oral spontaneous speech is elaborated. It is never delivered in a smooth fashion which could be compared to edited writing, that is to say a revised, corrected and perfect form! We here study repeats such as "malheureusement + c'est comme toujours on est obligé **de de** continuer **le : ++ le** voyage".

In the past years, researches on spoken French have thrived, in fields such as syntactic studies [3, 4], prosody [5], psycholinguistics [7, 11], computational aspects or human-computer dialogue [2]. Thus, a certain number of regular features have been identified in repeats: on the morpho-syntactic level, repeats mostly involve function words (9 repeats out of 10) which, most of the time, are monosyllables [9] as 41.5% are determiners, 26% pronouns and 13% prepositions [10]. These function words partake of the structuring of language and shape content words into syntactic units. Like Blanche-Benveniste [3], we have been able to check that repeats are subjected to syntactic constraints: they mainly appear at the beginning of phrases and their structure remains stable, that is to say the simple syntactic frame – without any lexical content – appears first, and the lexical filling comes second.

We have also established in collaboration with Campione & Véronis [9] that repeats present a variable degree of "disruptivity" according to the number of other disfluencies (silent and/or filled pauses) that are combined with them at the Interruption Point: when there is only one disfluency at the IP, it is the lengthening; when they are two disfluencies, the most frequent case is a combination of a lengthening and a silent pause.

We here focus on the realization of the repeat in NPs. We first give a distribution of the repeats according to the type of chunk involved. We will then analyze how the repeats are distributed according to the complexity feature of the phrase. Is it possible to co-relate the presence of a repeat with the complexity feature? We will also determine where the repeat is locating in the complex phrase: do repeats tend to appear more often in the head zone or in the expansion zone? Does the presence of a preposition in the noun chunk have an influence on the location of the repeat? Finally, among the prepositional noun chunks that have a [Preposition + (predeterminer) + Determiner + (modifier) + Noun + (modifier)] pattern, we will give an account of the most frequent types of repeats.

Our work is based on a corpus of spontaneous[1] speech of 17,000 words. It consists of Campione's corpus [5] (54 min, 8,500 words) to which we added, whilst preserving the original sampling of the corpus, 10 other extracts in order to obtain a corpus of 1h 47min. Most of the recordings are from the *CRFP* (French Reference Corpus)[2] and our corpus is composed of 20 speakers: 10 men and 10 women. The average length of the extracts is about 5 minutes, and the passages we selected are monologues where the speaker talks about his job, or evokes past events, etc. The speaker there answers questions from an investigator who does not intervene in the selected pieces.

2. Repeats and types of chunks

We have labelled disfluent repeats and found 383 occurrences in our corpus.

If, traditionally, the phrase, "a sequence of words composing a syntactic unit" [13], is considered to be the intermediate unit between the word and the sentence, we nevertheless remark that, in some cases, this unit can again be broken up into smaller units that are not words but chunks [1].

There are 4 types of chunks: noun (NC), verb (VC), adverb (AdvC) and adjective (AdjC) chunks. The distribution of repeats according to the type of chunk is as follows:

Table 1: Distribution of the repeats according to the type of chunks.

	NCs	VCs	AdvCs	AdjCs	Other	Σ
Repeats	263	91	8	5	16	383
%	68.7	23.8	2.1	1.3	4.2	100

We note that noun chunks obviously prevail (more than 2 repeats out of 3). This is linked to the strong involvement of determiners and pronouns in repeats. We have not found any example where the repeat would be on a content word only, such cases do exist, but they are not disfluent repeats.

Less than a fourth of the repeats occur in verb chunks and only a little over 3% in adverb and adjective chunks.

If we take a closer look at these results, we remark that approximately 1 repeat out of 3 (30.5%, 117/383) takes place in a chunk introduced by a preposition. When a chunk is introduced by a preposition, in 82% of cases it is a noun chunk, in 17% a verb chunk, and in only 1% of cases an adjective chunk.

Noun chunks, whether introduced or not by a preposition, are a privileged observation zone for repeats, and that explains why we propose a detailed study of these chunks.

[1] "Not based on a written piece, not learned by heart", Candéa [6].

[2] *Corpus de Référence du Français Parlé*, for more information, *cf.* [8].

3. Repeats in noun phrases

3.1. Definition of the complexity feature and results

We here adopt Clark & Wasow's definition of the complexity feature [7]: "NPs, however, range in complexity. *The mangy dog*, for example, is slightly more complex than *the dog* because of the added modifier, but *the dog down the street* and *the dog my neighbor owns* are much more complex because of the prepositional and clausal modifiers *after* the head noun. To simplify complexity, we divided NPs into *simple NPs*, which don't have anything after the head noun, and *complex NPs*, which do" (p.211).

Repeats can occur in phrases that present no expansion of the head noun. The noun phrase (NP) is then called "simple":

ex. 1: il faut être extrêmement vigilant car [**la : + la** loi]$_{NP}$ est euh ++ est précise là-dessus

When the NP is expanded, it is called "complex":

ex. 2: elle sort [**deux : deux** boudins qui étaient pleins de sciure]$_{NP}$

Among the 223 occurrences of repeats in NPs, 57% (126/223) take place in complex NPs, and 43% (97/223) in simple NPs. If these results indicate that repeats occur more frequently in complex NPs, only relative frequencies would however allow us to conclude that the complexity feature does influence the distribution of the repeats. We do not have such figures at the moment.

We have therefore restricted our analysis to "test words" which are the more frequent determiners in French, that is to say *le*, *la*, *les*, *un*, *une*, *des*. For each of them, we have identified, according to the complexity feature of the phrase (simple *vs* complex), the number of repeats *vs* the absence of repeats. When we divided the number of occurrences of repeats by the total number of occurrences (repeated + non repeated ones) for each determiner, we obtained the following relative frequencies:

Table 2: Relative frequencies of repeats of determiners *le, la, les, un, une, des* according to the complexity feature of the phrase.

	le	*la*	*les*	*un*	*une*	*des*
Simple NPs	5.4%	4.2%	5.2%	4.0%	2.2%	7.3%
Complex NPs	14.6%	7.4%	8.8%	6.9%	6.5%	11.8%

First at all, we can remark that for each of the six test words we retained for our study the relative frequencies of repeats are systematically higher in complex NPs than in simple NPs. The repeat rate is on average of 4.7% in simple NPs and reaches 9.3% in complex NPs.

In order to measure accurately the influence of the complexity feature on the presence of repeats, we have calculated the ratios of the relative frequencies of repeats in test words:

Table 3: Ratios of frequencies of repeats for each test word.

	le	*la*	*les*	*un*	*une*	*des*
Ratios	2.7	1.8	1.7	1.7	2.9	1.6

The table above shows that, according to the determiner involved in the repeat, the ratios change drastically, from 1.6 to 2.9. *Le* and *une* have the most important ratios (respectively 2.7 and 2.9); these elements are approximately 3 times more likely to be repeated in a complex NP than in a simple NP. The other determiners (*la*, *les*, *un*, *des*) have smaller ratios, between 1.6 and 1.8. The mean of the ratios amounts to 2.1 and it allows us to affirm that determiners are on average twice more likely to be repeated in complex NPs than in simple NPs.

Clark & Wasow [7] also observe a correlation between the complexity of the phrase and the frequency of the repeat for the determiners *the* and *a* located at the beginning of a noun phrase. So far, we have not taken into account the location of the determiner in the phrase and our results apply to determiners in the head chunk (containing the head noun) as well to those in the expansion chunk (containing the expansion of this head).

Why does the complexity of the phrase influence the occurrence of repeats? The most obvious and logical explanation is that the speaker has a lot more to plan in a complex phrase. Actually, he not only has to manage the structuring of the head chunk but also plan ahead the syntactic arrangement of the expansion chunk. The local constraint of a lexical search – the repeat as delaying tactics from the speaker – would be a lighter burden than the global structuring of the phrase. If it proves true, we should logically observe more repeats in the head zone than in the expansion zone. In order to check this hypothesis, we have located all the repeats in the phrases using our 6 test words.

3.2. Locating the repeat in complex noun phrases

In our corpus, determiners can be repeated:

- in the zone which contains the head noun:

  ex. 3: et en fait [(**les : les** personnes)$_{HEAD\ CHUNK}$ (d'un certain âge)$_{EXPANSION\ CHUNK}$]$_{NP}$ aiment toujours danser

- or in the zone which contains the expansion of the noun:

  ex. 4: il y a euh [(des fiches)$_{HEAD\ CHUNK}$ (sur **la la** faune)$_{EXPANSION\ CHUNK}$]$_{NP}$

As we expected, the determiners we selected as test words are mainly repeated in head chunks (83% of cases). The head chunk is therefore a privileged site for repeats of determiners and this proves that planning the whole of the noun phrase is a major constraint on the presence of repeats.

Moreover, the analysis of the complex noun phrases containing at least one repeat allows us to bring to light the following patterns:

- repeats in the head chunk:

  ex. 5: on a par exemple [(**ces + ces** fameux oeufs)$_{HEAD\ CHUNK}$ (en meurette)$_{EXPANSION\ CHUNK}$]$_{NP}$ je sais pas si vous savez

- repeats in the expansion chunk:

  ex. 6: alors là c'est c'est [(des soirées dansantes)$_{HEAD\ CHUNK}$ (**qui : qui** sont ouvertes à tout le monde)$_{EXPANSION\ CHUNK}$]$_{NP}$

- repeats in both the head chunk and the expansion chunk:

  ex. 7: j'ai eu très sincèrement l'impression que [(**ce : + ce** jour)$_{HEAD\ CHUNK}$ (**de : de** mon mariage)$_{EXPANSION\ CHUNK}$]$_{NP}$ a été le plus beau jour de ma vie

The following figure presents the distribution of complex noun phrases according to the location of the repeat:

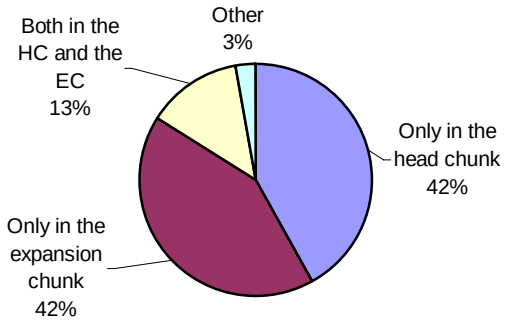


Figure 1: Distribution of complex NPs according to the location of the repeat.

Among the 112 noun phrases in which repeats occur, we notice that 13% of complex noun phrases contain repeats both in the head chunk and the expansion chunk. Repeating at the beginning of the phrase is therefore not enough to compensate for the difficulties the speaker encounters when planning the whole of the phrase. The constraint of a lexical search must not be neglected.

In addition to that, we can see that in 42% of cases repeats affect only the head chunk. The same proportion is to be found in the expansion chunk. This result seems to run counter to our previous conclusions on repeats of determiners. One would actually expect complex noun phrases to contain more repeats in the head than in the expansion zone, but it is not so. Why then?

A possible explanation would be as follows: contrary to the head chunk which, of course, contains the lexical head of the phrase and thus nearly systematically begins with a determiner (of any kind), the beginning of the expansion chunk can be composed of various elements, such as a preposition or a relative pronoun. The following figure shows the significant presence of prepositions in expansion chunks:

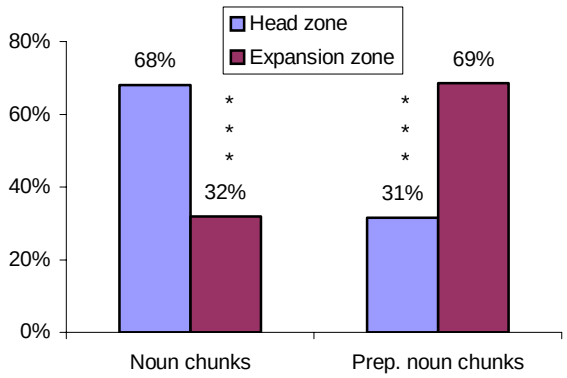


Figure 2: Comparison of the location of repeats according to the type of chunk.

When the complex noun phrase contains a determiner in the expansion chunk, the repeat tends to occur on the first element of the chunk rather than on the determiner. We are going to check this hypothesis.

4. Repeats and prepositional noun chunks

4.1. Patterns of repeats in prepositional noun chunks

Among the 96 repeats which affect prepositional noun chunks, we have kept only those which follow a [Preposition + (predeterminer) + Determiner + (modifier) + Noun + (modifier)]$_{PNC}$ pattern. We have found 59 occurrences. The study of the data allowed us to see six possible configurations, depending on whether the speaker goes back to the preposition or not:

- back to the preposition:

• the preposition only (47.5%):
ex. 8: on parle des journaux **dans dans** la plaine euh + euh bourguignonne

• the preposition and the determiner (23.7%):
ex. 9: là + **à ce : ++ beh à ce** lycée + j'ai eu des élèves absolument remarquables

• the whole of the prepositional noun chunk (1.7%):
ex. 10: et ensuite ++ euh on applique on euh l'émail **sur la pièce sur la pièce** en terre qui est déjà cuite

- later in the chunk:

• the determiner only (23.7%):
ex. 11: nous on les met dans **les : les** machines

• the determiner and the noun (1.7%):
ex. 12: aussi que je voulais dire + euh à propos des différentes terres + et de : justement que de **l'idée euh l'idée** reçue que se font les gens de la poterie

• the determiner and the “quantifier” of the noun (1.7%):
ex. 13: nous nous entendions bien avec **les deux : + les deux** Anglaises

The distribution of the repeats is as follows:

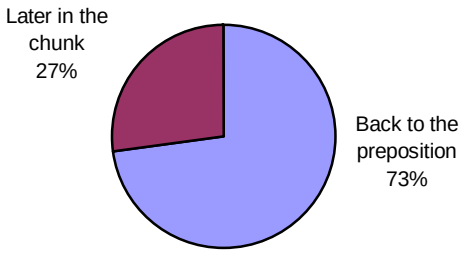


Figure 3: Frequency of the two major types of patterns of the repeats in prepositional noun chunks.

The figure above clearly indicates that, when the speaker begins a prepositional noun chunk, he tends to repeat the first element of the chunk (73%), that is to say the preposition or the unit composed of the preposition and the determiner. Cases when repeats appear later in the chunk are only a minority (27%).

Besides, in prepositional phrase expressions like *en dehors de*, *vis-à-vis de*, *dans la mesure de*, etc., the whole expression is rarely repeated:

ex. 14: et j'avoue qu'**avant de : avant de** me marier

We note that the repeat, most of the time, only affects the end of the fixed expression with *de* and not the whole of the expression[3]:

ex. 15: on essaie d'aider dans la mesure **de : + de** nos moyens actuels

ex. 16: et en fait moi je me suis aperçue au cours **de ces : de ces** soirées que beh il y a beaucoup de personnes d'un certain âge qui participent aux soirées dansantes

These expressions can not be broken down, and it is not possible to establish a relationship between the "head" and what follows[4]: ?*on essaie d'aider dans notre mesure*; that is the reason why we did not study these cases differently from the cases when the whole fixed expression is repeated.

4.2. Prepositional noun chunks as expansion chunks

When the prepositional noun chunk is the expansion chunk, the tendency to repeat the preposition only increases: 65% of cases *vs* 47,5%. Our hypothesis is thus confirmed: when complex noun phrases contain a determiner in the expansion chunk, the repeat tends to occur on the first element of the chunk rather than on the determiner.

5. Discussion

This study has permitted us to show that many different syntactic constraints bear an influence on repeats. We have been able to establish a co-relation between the complexity of the phrase and the frequency of the phenomenon. Indeed, the determiners we selected as test words are on average twice more likely to be repeated in complex noun phrases than in simple noun phrases. Furthermore, these elements are repeated mainly in head chunks (83% of cases). Our results are compatible with Clark & Wasow's findings on the English language. The "weight" of the phrase would be an additional constraint the speaker has to manage when he structures his speech.

As regards all the complex NPs in our corpus, we have counted as many repeats in the head zone as in the expansion zone. This result can be explained by the fact that the preposition – the juncture between the two chunks – is more often repeated than the determiner following it.

6. References

[1] Abney, Steven. 1991. Parsing By Chunks. In Robert Berwick, Steven Abney and Carol Tenny (Eds.), *Principle-Based Parsing.* Dordrecht: Kluwer Academic Publishers.

[2] Adda-Decker, Martine, *et al.* 2003. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. *Proceedings of DISS'03*, 5-8 September 2003, Göteborg University, Sweden, pp. 67-70.

[3] Blanche-Benveniste, Claire. 2003. La naissance des syntagmes dans les hésitations et répétitions du parler. In J.L. Araoui (Ed.), *Le sens et la mesure. Hommages à Benoît de Cornulier*. Paris: Editions Honoré Champion, pp. 40-55.

[4] Blanche-Benveniste, Claire. 1990. *Le français parlé. Études grammaticales.* Paris: CNRS Éditions.

[5] Campione, Estelle. 2001. *Étiquetage semi-automatique de la prosodie dans les corpus oraux: algorithmes et méthodologie.* Thèse d'état, Université de Provence, Aix-en-Provence, France.

[6] Candéa, Maria. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'« hésitation » en français oral spontané.* Thèse d'état, Université Paris III, Paris, France.

[7] Clark, Herbert H. & Thomas Wasow. 1998. Repeating Words in Spontaneous Speech. Cognitive *Psychology 37*, pp. 201-242.

[8] Équipe DELIC. 2004. Présentation du Corpus de Référence du Français Parlé. *Recherches Sur le Français Parlé 18*, Publications de l'Université de Provence, pp. 11-42.

[9] Henry, Sandrine, Estelle Campione & Jean Véronis. 2004. Répétitions et pauses (silencieuses et remplies) en français spontané. *Actes des XXV^èmes Journées d'Études sur la Parole*, 19-22 Avril 2004, Fès, Maroc, pp. 261-264.

[10] Henry, Sandrine. 2002. Étude des répétitions en français parlé spontané pour les technologies de la parole. *Actes de la 6^ème RECITAL*, 24-27 Juin 2002, Nancy, France, tome 1, pp. 467-476.

[11] Lickley, Robin. 1994. *Detecting disfluency in spontaneous speech.* Ph.D. thesis, University of Edinburgh.

[12] Martinie, Bruno. 1999. *Étude syntaxique des énoncés réparés en français parlé.* Thèse d'état, Université Paris X-Nanterre, Paris, France.

[13] Riegel, Martin, Jean-Christophe Pellat & René Rioul. 1999. *Grammaire méthodique du français.* Paris: Presses Universitaires de France (5^ème édition, 1^ère édition: 1994).

[3] We observe the same behaviour in adverbial expressions and complex determiners: "on fait beaucoup de : colonies : *beaucoup **de : + de*** choses comme ça", "on a quand même *un laps **de : de*** repos moi je suis du matin ma collègue est du soir". For more information on the behaviour of these determiners, see Claire Blanche-Benvensite's work, [4], pp. 109-111.

[4] *Cf.* Martinie [12], p.99.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 93-98.

Simulations of the types of disfluency produced in spontaneous utterances by fluent speakers, and the change in disfluency type seen as speakers who stutter get older

Peter Howell and Olatunji Akande**

* Department of Psychology, University College London

Abstract

The EXPLAN model is implemented on a graphic simulator. It is shown that it is able to produce speech in serial order and several types of fluency failure produced by fluent speakers and speakers who stutter. A way that EXPLAN accounts for longitudinal changes in the pattern of fluency failures shown by speakers who stutter is demonstrated.

1. Introduction

There are several ways in which loss of speech control can be accounted for. Three accounts have been discussed [15], one of which was proposed by the speech team in the Department of Psychology, University College London (UCL). This account, the EXPLAN model of fluency control [7, 9, 11], represents planning and execution as independent processes (the acronym for the model takes its name from these two processes). According to the model, fluency fails when a) planning is slow or b) execution rate is rapid. EXPLAN is a model of fluent speech control and how fluency can fail. The types of fluency failure observed in children who stutter (CWS) are similar to those seen in fluent children (though fluency failures occur more frequently in the CWS, which makes them useful for studying these events). Stuttering in adults who stutter involves events that are comparatively rare in fluent children and CWS. If the theory is right about its proposal as to how fluency failures arise in all children, it also needs to account for how and why fluency failures change in type when CWS become adults who stutter.

EXPLAN is assessed in this paper by simulating the way planning and execution processes interact according to the model. The performance of the model is examined to see if it can produce a) a fluent sequence of speech, and b) the different types of disfluency that arise when planning and execution are perturbed (slowed and speeded respectively). The paper reviews the history of EXPLAN in section two. Section three outlines the basic model that generates fluent speech according to the principles underlying the account (separate planning and execution components). In section four, the parameters representing planning and execution are perturbed to see whether incidence and type of disfluency are affected in the predicted manner. Finally, section five shows how the features seen in stuttering as it persists into adulthood, could be an adaptive change to the patterns of fluency failure seen in normal fluency development when the processes that led to these failures persist into adulthood.

2. EXPLAN theory

2.1. Planning

Typically, the elements in an utterance (the words, for example) vary in how complex they are to plan. Complexity can be reflected at different processing stages in utterance formulation or planning, including syntactic, lexical (word class and word frequency), phonological, prosodic and phonetic levels. Views that maintain that fluency failure is a result of planning complexity abound in the stuttering literature, and they usually emphasize the role of one linguistic level. For instance, Bernstein Ratner [3] emphasizes the syntactic level, a group at East Carolina consider word frequency is paramount, and Wingate [26] has promoted the view that the phonological level (and prosody in particular) is the primary source of the problem.

The planning component in EXPLAN allows complexity to be affected by any of these levels (though the evidence suggests that a pure syntactic deficit is not likely to be a determinant of stuttering [13, 14, 22, 23 for reviews]. To allow all the remaining levels to influence fluency failure, we have examined lexical forms (function and content words) separately. The reason for this is that lexical class correlates with the other factors that could specify complexity, so using lexical type integrates contributions from these several sources. For instance, content words are the only type of words stressed in English, so investigating lexical class is effectively also an examination of stress (and lexical class is easier to determine objectively than is stress).

Word frequency is conceived in a different way to stress in EXPLAN. Content words tend to be low frequency, so frequency correlates with lexical class. However, a case has been made that, in work on isolated words, word frequency can be dissociated from lexical class (there are low frequency function words) and that low frequency words are more likely to be stuttered than high frequency ones, irrespective of their lexical status. The implication to drop lexical status that is carried by this view is problematic because types of fluency failure depend on the type of word they occur on. Function words are repeated in their entirety or have pauses preceding them, content words, involve disfluency on their first part which can either be prolonged, repeated or have a break inserted between the initial (onset) and subsequent (rhyme) parts. It is hard to conceive how word frequency that operates independent of lexical class could account for why different types of disfluency are linked with different word types. Word frequency differences are idiosyncratic and also have an ephemeral property in an individual's speech. Though these properties make infrequent words unamenable to systematic study, there are some observations that can be made about the acquisition of words of different frequency: Function word usage will be reasonably stable once a mature syntactic system has been established. Content word usage will increase in frequency throughout life. Content word vocabulary in early life consists of relatively high frequency such words. Expansion of vocabulary at the start and throughout adulthood tends to involve adding low frequency content words. The latter aspect, in particular, that is specifically a feature of content words, has a critical role (as indicated in section four) in the current simulation for explaining the ontogenesis of the disorder into adulthood.

Content words tend to be phonetically and phonologically more complex than function words. The UCL team has used

phonological and phonetic measures as a way of quantifying different levels of difficulty within function and content word classes. Incidence of phonetic properties varies between these word classes and, more surprisingly, the incidence of phonetic properties varies across the age range 6 years to adulthood. This is shown in Figure 1 for words with difficult manners, words which are long and words which have contiguous consonants (shown separately for content and function words). Figure 1a (content words) shows significant increases over ages in use of each factor whilst there is no systematic increase for function words (Figure 1b).

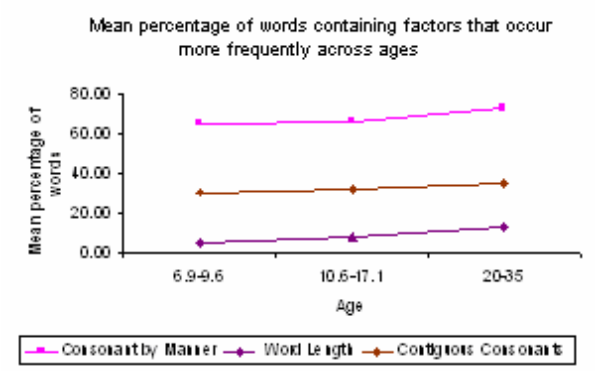


Figure 1a. Mean percentage of content words containing difficult manners, long words or contiguous consonants that occur more frequently in the speech of older speakers.

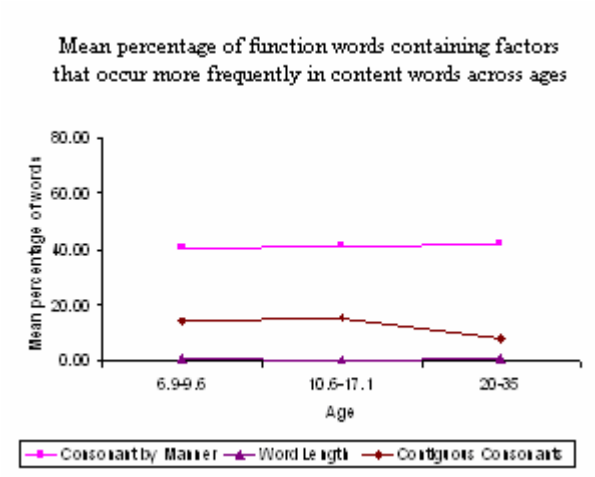


Figure 1b. Mean percentage of function words containing difficult manners, long words or contiguous consonants.

Comparison of percentage of words containing each of these factors also shows a big imbalance between usage of these phonetic properties between word classes (all factors occur less frequently in function words). The change over word type and ages for content words is probably a reflection that vocabulary increases with age and content words that are acquired later are more complex than those acquired early. Note that these variables also relate to word frequency as words acquired late are likely to be used infrequently. The phonetic and phonological properties correlate with stuttering rate for content words, but not for function words. This is shown in Figure 2 where phonetic difficulty is represented as the sum over four factors [18] (the three indicated earlier plus whether the word contained a dorsal consonant) for words marked as difficult (e.g. if a word contains a contiguous string of consonants, it scores a point).

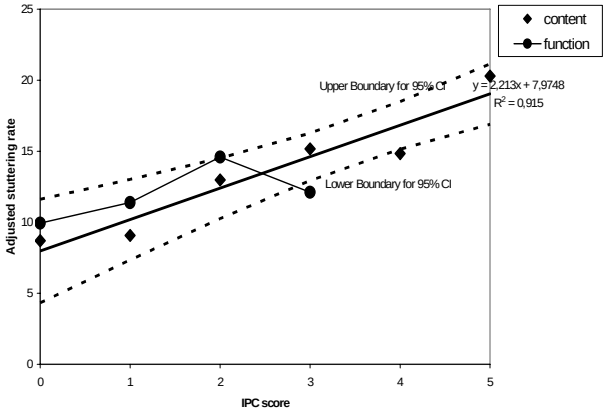


Figure 2. Adjusted stuttering rate (ordinate) versus number of times the four factors marked as difficult occurred (abscissa) for speakers aged over 18 years. The straight line is fitted to the content words and the upper and lower bounds around this line are indicated by the dashed line. The function word points are connected by a solid line.

There is a significant correlation for content, but not function words. The graph also shows that function words have a more limited range of phonetic difficulty. The lack of correlation with the difficulty measure for the function words underlines the importance of examining word types separately.

In summary, planning of function words appears to differ from the planning of content words: Content words show different form of fluency failure (disfluency on parts), this class includes attributes that relate to stuttering (stress, phonetic and phonological indices), across the members of the class, word frequency is low and this is also reflected at the phonetic level (for example, there is variation in usage of phonetic features between age groups for content words). Function words are more stable insofar as disfluencies on them involve complete words, this class does not carry stress in English nor do function words usually have complex phonetic properties and the incidence of the phonetic properties of words in this class does not vary across age groups. We assume that more difficult words take longer to plan and, if difficulty is measured phonetically, there are going to be influences specific to content words that vary with age (even into adulthood, as seen in Figure 1) that impact on fluency (Figure 2). Reference to function/content words can also be regarded as comment on stress and phonetic properties. Thus, if instead of lexical type, words were represented as bundles of phonetic features and divided into classes with high frequency of phonetic features, we would have come full circle insofar as these properties would define content and function types.

2.2. Execution

If timing constraints lead to stuttering and the speech execution system operates independent of planning, increasing execution rate should exacerbate fluency problems and decreasing execution rate should decrease the chance of fluency problems [16, 17]. As planning and execution are represented independently in EXPLAN, a speaker can be planning a different segment to that currently being executed. This necessitates examination of how planning and execution

interact which is the third component in the model (discussed next).

2.3. Interaction between planning and execution and disfluency types

Representing planning and execution processes separately allows speech to be planned in advance of the extract of the utterance currently being emitted. Though it is desirable to allow planning and execution to proceed on different timescales, some specification needs to be made about the point at which they interact.

If the plan for a later word in an utterance is not ready, fluency is likely to fail (though this is not inevitable, depending on the execution model). To flesh out this view, the contextual units within which utterances are planned and delivered need to be specified. We have used units that have been developed for other purposes in phonology. The contextual unit is the phonological, or prosodic, word, PW [24]. PW, as defined by Selkirk, have an obligatory content word that can be preceded and followed by various numbers of function words [See 10 for an alternative definition of PW]. Examples of PW are 'in the spring' (two function words precede the content word), 'I hit him' (one function word preceding and one function word following the content word). Speakers could start by planning 'in', and when complete can start its execution, while they plan 'the' and so on. As long as there is enough time to plan the next word, speech will proceed fluently.

Disfluencies occur when speakers are in the limiting situation and there is not sufficient time to plan the next word during the time allowed for the current word to be executed. The disfluencies that occur within PW can be conveniently divided into those which involve whole, and those which involve parts of, words. Whole word repetition has either been taken as signifying that speaker restarts an utterance because an error was made [19, 20] or stalls because a future word is not ready in time [5, 7, 9, 11, 21]. This class tends to involve function words which are more likely than not to start an utterance (more than 50% of utterances start with a function word in [2]) and contain the majority (more than 90%) of all types of disfluency including those on function words. Stalling explains why this happens on initial, not final, function words [25]. Stalling is also consistent both with the view that the speaker restarts at the boundary of a constituent that contains an error and that stalling occurs before a content word.

Disfluencies in content words are rarer. They often involve part of such words, in particular repetition or prolongation of the initial part (they are called advancings). Ambrose and Yairi [1] show that these make up less than 1% of disfluency for three age groups - fluent. Speakers who stutter have a higher rate (up to eight times). This type of disfluency occurs away from the constituent boundary [2] and occurs on words where phonetic complexity is high [6] and is seen mainly in older speakers who stutter [12].

3. Implementation representation of planning and execution and how they interact and lead to disfluency

The initial version of EXPLAN accounted for fluent production of a series of words and stalling and advancing fluency failure. Planning in this early version was represented as a line indicating the time taken (so the line for a content word was greater than that for a function word). Execution was represented similarly, though it has to occur with an offset relative to planning. The offset between the two indicated the interaction and the way this could lead to fluency failure.

Figure 3 shows the particular situation where planning of one segment is completed during the time the preceding segment is being executed, thus leading to fluent speech.

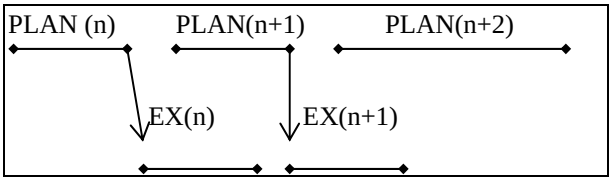


Figure 3. Diagrammatic representation of the temporal relationship between planning and execution for speech produced fluently. Time is along the abscissa. The epoch during which planning (PLAN) and execution (EX) occur are shown as bars in the top and middle rows respectively. Planning of adjacent words is shown in series for simplicity. Execution of word n commencing after its plan is complete and that there is sufficient time to plan the following word (n+1) while word (n) is being executed.

The case shown is at the limit where the next plan is just ready in time (when the speaker has pre-planned speech well in advance, that speech will also be fluent). Speakers need to gain extra time for planning when the subsequent segment is not ready. They can do this by pausing or repeating one or more prior segments (referred to as 'stallings'). If the segments that are planned rapidly are function words and the segments that take longer to plan are content words, the hesitation or repetition occurs around the function words. The function words are produced fluently during stalling and, for this reason, stalling is the least risky way of dealing with fluency failures and predominates in children's speech (whether they are fluent or stutter). This situation is depicted in Figure 4.

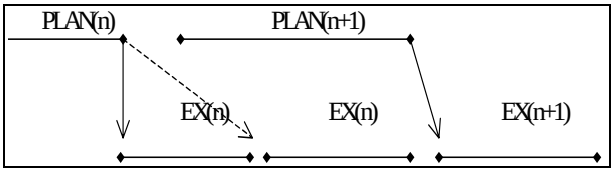


Figure 4. Representation of stalling as using the same conventions as in Figure 3. Execution of a prior word is complete before the plan for the following word is completed. In this case, after word n has been spoken the first time, it is repeated to allow more time to complete the plan of word (n+1).

Alternatively, the speaker can commence execution of the partly-prepared word [20]. If speakers do this, the plan may run out part way through the word (resulting in part-word disfluencies as shown in Figure 5).

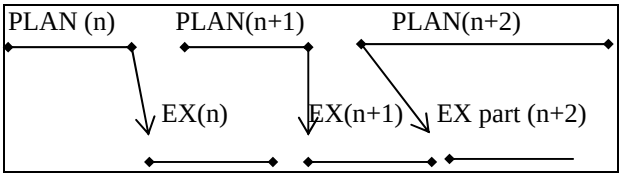


Figure 5. Representation of advancing as in Figure 3. After the first two words (n and n+1) have been completed, even though the plan of the next word is not ready, the speaker commences word (n+2) and the plan runs out resulting in part-word disfluency.

The part-word disfluencies that result would occur on segments that take a long time to plan, which would be content words in this case. The units within which this interaction occurs are PW (in particular those with initial function words).

In the next version of the model [8], planning was represented as an activation profile (such as that which would arise in a spreading activation account). Content words were assumed to have slower activation rates than function words. Selection of a word for execution was based on maximum activation as long as activation was above a minimum threshold. The approach taken was to show diagrammatically that there were occasions on which words were produced in correct sequence (fluent serial order), where preceding function word had higher activation than the content word that was next in sequence (which led to function word repetition) and when the content word had highest activation but was not at maximum resulting, if it was executed, in a part content word disfluency. In this model, planning differences were limited to gross differences between function and content words. The current version of EXPLAN includes more subtle planning rate differences and their effect on output when they are perturbed. In the first activation profile model, execution processes were not explicitly identified (they are complex and fill the interval represented by the difference between word initiation of two words which is affected by the point where a word was initiated, planning rate and threshold at which execution could be initiated).

4. Perturbation of parameters representing planning and execution

The range of disfluencies examined in the previous activation profile model was limited and the perturbations to the planning and execution processes were not examined systematically EXPLAN has now been implemented using a dynamic graphics package and speech output interface. A sequence of words is input (shown at left of each section in Figures 6-8 for 'in the morning' – a PW of the form FFC). Each red line represents buildup (rising line) and decay of a single word. The last point of the three red lines represents the activation level of each word at selected points in time. The threshold a word has to reach to be eligible for production is indicated. Figures 6, 7 and 8 show snapshots after 'in', 'the' and 'morning' have been produced (the way activations builds up dynamically will be demonstrated in the presentation).

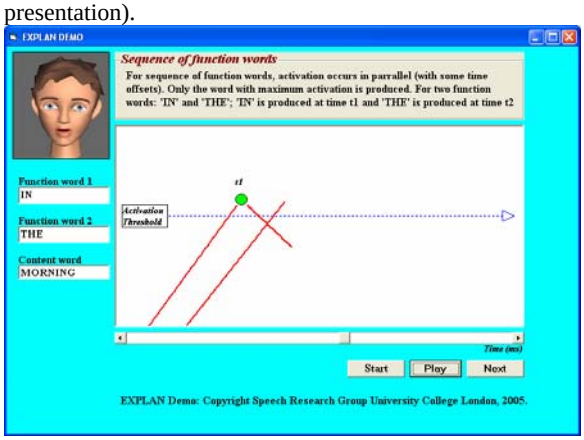


Figure 6. Activation rate (ordinate) over time (abscissa) for two function words followed by a content word. This still picture shows the situation after "in" has been executed and "the" is about to be produced.

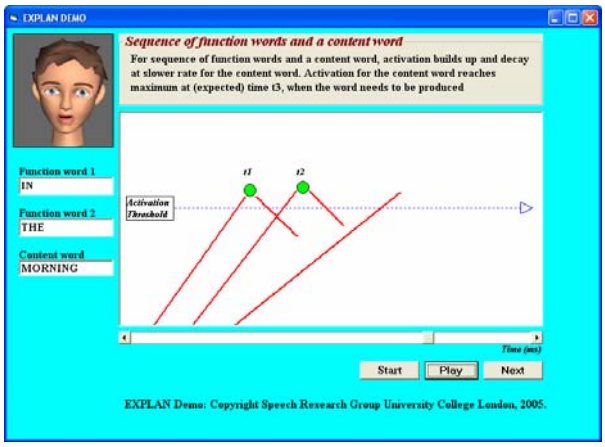


Figure 7. As Figure 6 showing the situation after "the" has been completed.

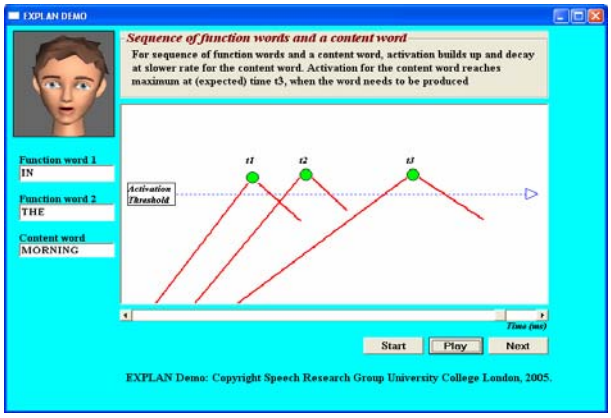


Figure 8. As Figure 6 showing the situation after "morning" has been completed.

Next the execution and planning inputs were perturbed in ways that should lead to the different types of disfluency. Execution rate changes were simulated by shortening the gap between words. As shown in Figure 9, this leads to the activation of the second function word being still above threshold after it has been produced once, while the content word is below activation threshold. Consequently, 'the' is repeated. After repetition of 'the', the content word is above activation threshold and can be produced.

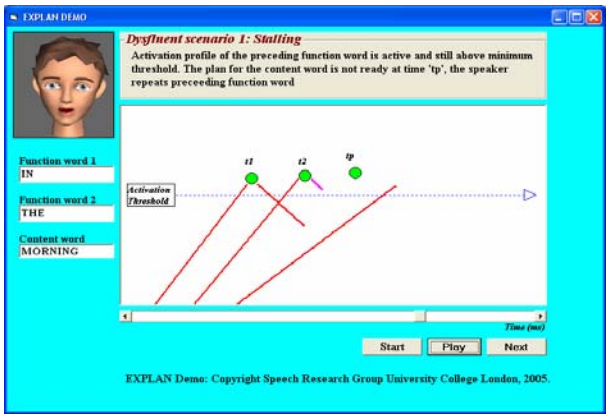


Figure 9. Diagram of the situation leading to function word repetition (see text for description).

Multiple repetitions of the same function word arise due to reactivation of the plan as proposed by Blackmer and Mitton [4]. The function word is reactivated however many times as are needed for the content word to reach its threshold for activation.

The detailed structure of phrase repetitions has not been described previously, so analysis was made of such events. The two main observations are: 1) Phrase repetitions occur most often on PW that start with two function words (as in the 'in the morning' example). This is predicted by EXPLAN (both words can be used for stalling). 2) PW of this form allow repetition of the first ('in in the morning') or second ('in the, the morning') word as well as phrase repetition ('in the, in the morning'). The PW of six CWS were examined for PW that started with two or more function words and which had repeated words. Repetition was at the phrase level for 9.7% of the sample (as in "in the, in the morning". Word repetition (90.3%) only involved the first word (as in "in, in the morning") (never the second). Repetition of the first word is consistent with Levelt's [20] main interruption rule, MIR (this rule states that speakers interrupt as soon as possible before difficulty is experienced). MIR is included in the simulator to produce phrase repetitions.

Planning rate changes were simulated by decreasing the slope of content word activations relative to function words. Decreasing the slope eventually results in the content word having a higher (but not full) activation than the function word which leads the speaker to produce an advancing type disfluency as shown in Figure 10. In this situation, only the plan of the first part of the content word is available to be produced.

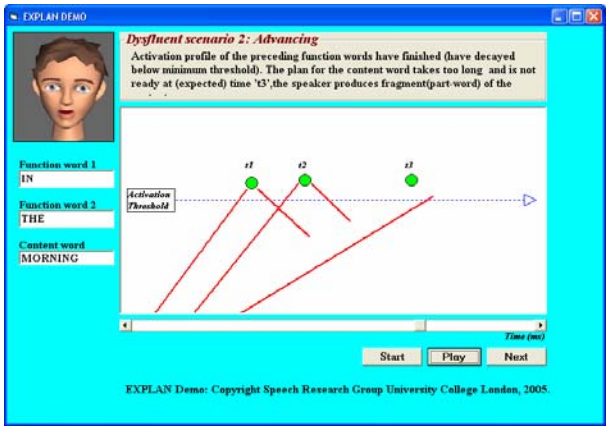


Figure 10. Diagram of the situation leading to part-content word disfluency (see text for description).

Variation in activation rates of word in the content word category occurs because of phonetic difficulty (Figure 2) which correlates with frequency with which these properties occur in content words over ages (Figure 1a). This is used in the simulations in section five to predict change in disfluency type for speakers who stutter over ages.

Recycling utterances already available, when applied to part of a plan, would lead to prolongation and part-word repetition – prolongation would arise when the onset consonant alone is available (which happens mainly on fricatives, laterals and nasals) and part-word repetitions when the plan is complete up to the onset-nucleus boundary or to onset plus nucleus, but typically not beyond that point in the syllable (i.e. not to the coda). The question that arises if this account is correct, is how speakers break out of the prolongation or repetition loop. This question did not arise in connection with whole word repetitions as planning of the subsequent content word continues and its execution takes over when it is completed. Here, however, the element being prolonged or repeated is also the element being planned. There are several ways that part of a word can be, for example, prolonged while at the same time its planning can continue. One way of achieving this is to allow the elementary constituents of a word that are available (e.g. up to the end of the onset consonant or the start of the coda) to be the elements that are reinitiated rather than the word forms. Planning of the rest of the word can then proceed if it is not complete.

5. Simulation of change from stalling to advancing with age

As noted, speakers who stutter change from producing stallings to advancing as they get older [12]. One explanation for why this occurs is that it arises because of the impact that changes in the frequency of usage of content words (function word usage remains constant once a child has learned the syntax of a language). Speakers who are inclined to produce a high incidence of stallings (CWS) use more content words as they get older. This dilutes the frequency of occurrence of all content words, making all words rarer and new words infrequent when they are acquired late, compared with when they are acquired early, in life. If the activation rate is determined in part by difficulty, it would tend to decrease for all words (this would apply to the words learned recently too). In turn this would tend to reduce the chances of content words reaching full activation and have the effect of increasing incidence of advancing-type disfluencies).

6. Summary and Conclusions

EXPLAN has been outlined and simulations of the main types of disfluency shown by fluent speakers and speakers who stutter have been presented. It has been shown how advancing type disfluencies could be a natural response to vocabulary changes in speakers prone to produce a high rate of disfluencies. As the emergence of this form of disfluency is based on processes that occur in fluent speakers as well as speakers who stutter, there would not appear to be structural central nervous system abnormalities that lead to the disorder.

7. Acknowledgement

This research was supported by Wellcome Trust grant 072639.

References

[1] Ambrose, N.G. & Yairi, E. (1999). Normative disfluency data for early childhood stuttering. *Journal of Speech, Language and Hearing Research, 42*, 895-909.

[2] Au-Yeung, J. Howell, P. & Pilgrim, L. (1998). Phonological words and stuttering on function words. *Journal of Speech, Language and Hearing Research, 41*, 1019-1030.

[3] Bernstein Ratner, N. (1997). Stuttering: A psycholinguistic perspective. In R. Curlee & G. Siegel (Eds.), *Nature and treatment of stuttering: New directions* (2nd edition). Needham, MA: Allyn & Bacon.

[4] Blackmer, E. R. & Mitton, J. L. (1991). Theories of monitoring and timing of repairs in spontaneous speech. *Cognition, 39*, 173-194.

[5] Clark, H. & Clark, E. (1977). *Psychology and Language: An Introduction to Psycholinguistics*. New York: Harcourt Brace.

[6] Dworzynski, K. & Howell, P. (2004). Predicting stuttering from phonetic complexity in German. *Journal of Fluency Disorders, 29*, 149-173.

[7] Howell, P. (2002). The EXPLAN theory of fluency control applied to the Treatment of Stuttering by Altered Feedback and Operant Procedures. In E. Fava (Ed.), *Current Issues in Linguistic Theory series: Pathology and therapy of speech disorders* (pp.95-118). Amsterdam: John Benjamins.

[8] Howell, P. (2003). Is a perceptual monitor needed to explain how speech errors are repaired? *Gothenburg papers in Theoretical Linguistics, 90*, 29-32.

[9] Howell, P. (2004a). Assessment of some contemporary theories of stuttering that apply to spontaneous speech. *Contemporary Issues in Communicative Sciences and Disorders, 39*, 122-139.

[10] Howell, P. (2004b). Comparison of two ways of defining phonological words for assessing stuttering pattern changes with age in Spanish speakers who stutter. *Journal of Multilingual Communication Disorders, 2*, 161-186.

[11] Howell, P. & Au-Yeung, J. (2002). The EXPLAN theory of fluency control and the diagnosis of stuttering. In E. Fava (Ed.), *Current Issues in Linguistic Theory series: Pathology and therapy of speech disorders* (pp.75-94). Amsterdam: John Benjamins.

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

[13]Howell, P. Davis, S. & Au-Yeung, J. (2003). Syntactic development in fluent children, children who stutter, and children who have English as an additional language. *Child Language Teaching and Therapy, 19*, 311-337.

[14] Howell, P. & Dworzynski, K. (in press). Planning and execution processes in speech control by fluent speakers and speakers who stutter. *Journal of Fluency Disorders*.

[15] Howell, P. Hayes, J. & Akande, O. (in press). Factors that determine the form and position of disfluencies in spontaneous utterances. Disfluencies in Spontaneous Speech.

[16] Howell, P. & Sackin, S. (2001). Function word repetitions emerge when speakers are operantly conditioned to reduce frequency of silent pauses. *Journal of Psycholinguistic Research, 30*, 457-474.

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

[18] Jakielski, K. J. (1998). *Motor organization in the acquisition of consonant clusters*. PhD thesis, University of Texas at Austin. Ann Arbor Michigan.

[19] Kolk, H.H.J. & Postma, A. (1997). Stuttering as a covert-repair phenomenon. In R.F. Curlee, & G. Siegel (Eds.), *Nature and treatment of stuttering: New directions*.(pp. 182-203). Boston: Allyn & Bacon.

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

[21] MacWhinney, B., & Osser, H. (1977). Verbal planning functions in children's speech. *Child Development, 48*, 978-985.

[22] Nippold, M. A. (1990). Concomitant speech and language disorders in stuttering children: A critique of the literature. *Journal of Speech and Hearing Disorders, 55*, 51-60.

[23] Nippold, M. A. (2001). Phonological disorders and stuttering in children: What is the frequency of co-occurrence? *Clinical Linguistics and Phonetics, 15*, 219-228.

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

[25] Strenstrom, A-B. & Svartvik, J. (1994). Imparsable speech: Repeats and other nonfluencies in spoken English. In N. Oostdijk & P. de Haan (Eds.), *Corpus-based research into language*. Atlanta, GA: Rodopi.

[26] Wingate, M. (1984). Stutter events and linguistic stress. *Journal of Fluency Disorders, 9*,295-300.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 99–102.

Factors that determine the form and position of disfluencies in spontaneous utterances

Peter Howell, Jennifer Hayes* , Ceri Savage*, Jane Ladd*, Nafisa Patel**

* Department of Psychology, University College London, London, England

Abstract

This presentation reviews work on types of disfluency in the spontaneous speech of fluent speakers and speakers who stutter. Examination is made of factors that determine where disfluencies are located. It is concluded that the phonological, or prosodic, word provides a good basis for explaining the distribution of different types of disfluency in spontaneous speech.

1. Introduction

In the previous paper [13], a model that simulated the types of disfluency that occur within phonological word (PW) units was presented. PWs were defined in that paper as having a content word as nucleus and function words as optional satellites that can precede and follow the nucleus. Thus 'I sprang up' is a PW with one function word preceding and one function word following the content word.

The issues examined in section two of this paper are the types of disfluency and the different taxonomies that have been applied to them. All the disfluencies seen in fluent speakers are also seen in speakers who stutter, though they occur more frequently. Thus, taxonomies developed for stuttered speech can validly be applied across a range of speakers. In section three previous evidence and new preliminary data are reviewed that support the view that PW units are appropriate for analysis of disfluencies in spontaneous utterances and better than other contextual units (syntactic ones in particular). In section four, alternative ways of defining PW are considered and whether they improve predictions about disfluencies in spontaneous speech over content- and function-based definitions of PW.

2. Selected taxonomies of disfluency types

Speakers produce a range of different types of disfluency in spontaneous speech and these disfluencies can be classified in various different ways. The main division is between errors proper (where a phone occurs in the wrong position for an intended word) and fluency failures that do not show signs of speech error according to the preceding definition. This definition of error would include word-selection errors (e.g. 'left' for 'right'). Speech errors per se are not the focus of this article (though they feature indirectly in one of the accounts discussed below). Speech errors, as defined here, have been reported to be rare even in the spontaneous speech of fluent speakers [8].

The common types of fluency failure involve hesitation, production of whole or parts of words (as in false starts). Two observations about these fluency failures are necessary: First, all speakers show them though they occur more frequently in the speech of some speakers than others. For instance, speakers who stutter show a higher incidence which makes their speech convenient for collecting a large sample. Also, as shown in the previous paper, the balance between different types of fluency failure that changes over development in speakers who stutter could arise as a self-organizing change as the language production system matures. Thus, there is no inherent difference between the forms of fluency shown by speakers who stutter and fluent speakers, only an imbalance in the frequency with which the different types of event occur.

.Second, it is commonly recognized that the various identifiably different types of fluency failure need to be sub-divided. Some of the widely identified types of fluency failure are repetitions involving parts of words, single whole words or phrases, filled and unfilled pauses, prolongation of phones in a word (usually the first one) and a break within a word. This list was used by Ambrose and Yairi [1] with stuttered speech who also included revision and interjections. The revisions are not included as they involve errors as defined earlier (the word or words selected are different to those intended). Many interjections are essentially filled pauses and when so are included in that category. The way the remaining types of fluency failure are sub-divided varies between research groups and is not theoretically neutral. Yairi and Ambrose [26] use an empirical criterion to divide the events into two classes. One class is stuttering-like disfluencies (SLDs) which are more prevalent in the speech of people who stutter than in fluent controls. This class includes part- and whole-word repetitions, prolongations and broken words. The second class is other disfluencies (OD) which consists of phrase repetitions and the classes (not considered here) of interjections and repetitions. Though the assignment of disfluency types to SLD and OD is empirical, it has theoretical implications insofar as it implies there are some categories of disfluency that are associated with stuttering and others that are not.

A second way of dividing up the disfluency events uses Levelt's [19] repair framework. Repairs occur when the speaker makes an error and corrects it. An example that includes a comprehensive list of repair events (many of these are optional in actual speech) would be "Turn left at the, no, turn right at the crossroads". This would be excluded from the disfluent events being included here as 'left' does not include the correct phones for the intended word ('right'). However, some repairs do not include an error as in the related example 'Turn, turn right at the crossroads'. This is assumed to reflect an underlying error that is detected and corrected before output and is referred to as a covert repair. Pauses (filled and unfilled), word and phrase repetitions would all signify covert repair processes. Kolk and his students [18] have developed the covert repair hypothesis (CRH) which takes these events as signs of repair to underlying errors. Fragments of words (as in prolongations, word breaks and part-word repetitions) might be parts of a word produced in error [9].

The final division of disfluencies is that according to the EXPLAN model [10, 12, 14] which basically produces a similar division to CRH (pauses, word and phrase repetitions in one class and prolongations, word breaks and part-word repetitions in the other) but accounts for them in a different way.

Ambrose and Yairi's taxonomy is empirical whereas CRH's and EXPLAN's categorizations also involve theoretical assumptions about underlying causes of disfluency. The principle difference between CRH and EXPLAN is that CRH considers disfluency events are reflections of underlying errors

whereas EXPLAN considers they are not, so they do not require an error-detection mechanism. According to EXPLAN, disfluencies should be divided into two classes because each class has a different role and the types of disfluency have their effect on different parts of utterances (in particular on different parts of PW).

The example used at the start of the article ('I sprang up') can be employed to explain the latter points about type and role of disfluency in more depth. 'Sprang' is the word most likely to be difficult to generate. The first class of disfluency involves the simpler function word preceding the content word 'sprang', which is usually either repeated or has a preceding pause. The role of pausing and word and phrase repetition involving the initial function word/s is to gain time to prepare the difficult content word. This class of disfluency has been referred to as stalling. The second class of disfluency involves the content word itself and usually involves disfluency on the first part of the word. This class of disfluency has been referred to as advancing. A way of contrasting the CRH and EXPLAN accounts is in terms of whether errors or timing are regarded as the paramount features that lead to fluency failure.

EXPLAN theory makes restrictive predictions about the positioning of each type of disfluency in PW which depends on the role attributed the respective type. The predictions and evidence about each type of disfluency are reviewed next.

3. Evidence on position and type of disfluency in spontaneous speech

3.1. Evidence for stalling

1. The view that word and phrase repetition and pausing represent an attempt to buy time (what is referred to here as 'stalling') has been proposed [5] and empirically tested [6, 20, 21] by several previous authors particularly those on child language.

2. Stalling at different position in an utterance. If word and phrase repetition serves a delaying role (as assumed to happen in stalling), these events should be located in early positions in utterances (in particular, prior to content word nuclei) and should be specific to function words. To examine whether position effects occur for function and content words, speech data from English and Spanish speakers who stutter were segmented into PWs. Both the English and Spanish data showed (1) Serial position functions for function words with higher disfluency percentage in earlier positions [2, 3]; (2) No such serial position effects occur for content words [2, 3].

3. Stalling prior to content words. As indicated under 2, stalling only works when it is done on function words that precede a content word. PW are ideal for these analyses as they can have function words prior to and after the single content word in such units. Thus, the function words that could serve a delaying role (those preceding the content word) and those that could not (those following the content word) can be coded unambiguously. The function words can then be examined to see whether it is only those that precede the content word that show the disfluencies we refer to as stallings. There are data form fluent speakers that support this prediction. Strenstrom and Svartvik [23] looked at repetition of subject ('he' in the phrase 'he hit him') and object pronouns ('him' in the phrase). There are two features to note about these examples: First, the units are a PW of the form FCF. Second, position relative to the content word is indicated by the different forms the words take. The prediction, if stalling applies, is that only the subject pronouns could serve a delaying role and so these are the only ones that should be repeated. This is exactly what Strenstrom and Svartvik found. Similar analyses have been made on English [2] and Spanish [3] using a wider range of PWs with essentially the same finding. Au-Yeung et al. [2] found significant effects for all the age groups they examined for English with higher rates with a high level of disfluency in pre-content function words and a low level on post-content function words. For Spanish [3], the difference in disfluency rate across these positions was significant for three age groups out of five.

4. Comparison of PW units for predicting position of stallings with other units. The serial and ordinal position effects indicated above may be mediated by units other than PW. Thus, syntactic units can be constructed from PW units, and then a syntactic unit will have a PW in initial position. This PW would lead to serial and ordinal position effects because they are in initial position in the syntactic unit, not because they are in initial position in a PW. Similar arguments would apply to units such as utterances [27]. To test between two alternatives (PW and utterances), utterances which contained two function words F1 (final in the utterance-initial PW) and F2 (initial in an utterance non-initial PW), […F1] .. [F2 …] …, were examined. F2 was stuttered more than F1 in spite of F1's earlier position in an utterance. This analysis shows that an utterance position effect cannot account for the effects of PW position.

5. Pause position in PW. Inserting a pause should also be a way of stalling. If so, pauses would be expected to be positioned at the start of a PW to play a delaying role. The majority (more than 50%) of pauses in stuttered speech occur at PW boundaries. One issue not examined to date is what happens to PWs that do not have initial function words (e.g. 'hit it'). Pauses could be used for delaying onsets when the content word is complex (in which case, such PW would have a higher incidence of pausing than other PW).

3.2. Evidence for Advancing

1. Words vary in difficulty and this should impact in different ways on words involved in stalling (function) and advancing (content) fluency failures. Stalling does not occur because the words themselves are difficult, but because an up-coming word is difficult. Advancings should occur on words that are difficult. Moreover, content words are not equally difficult and it should only be the difficult ones that attract advancing-type disfluency. To test these predictions, measures of word difficulty are required that can be applied to each class of word. Three metrics have been employed in work so far: The first was developed by Yairi's group and was an empirical measure that involved classifying words as to whether they contained a consonant string (CS), phonemes that were acquired late in language development (LEC) and whether the word was multisyllabic or not, MS [24]. They found little evidence of difficulty affecting stuttering rate for SLD and OD. Two facts are of note: First, they did not analyze the difficulty factors for word position and when that was done, effects were found particularly in older speakers who show more advancing-type disfluencies [15]. Second, as the OD and SLD classifications do not map directly onto stallings and advancings, their data may benefit from reanalysis using these categories.

The second measure, called the index of phonetic complexity or IPC for short, used an index based on babbled speech [17] that would have been thought to be particularly useful for younger speakers who stutter. As with the CS/LEC/MS metric, this was not useful with young speakers [25] but was with older speakers [7, 16]. It includes eight factors. The main drawback of this metric is that it is not word-position specific

which is an important factor to include as most stuttering occurs in initial position.

The third metric is one based on non-linear phonology. Non-linear means, in this context, classification based on hierarchical schemes that divide the syllable into onset, nucleus and coda (thus, it captures position-dependencies in stuttered speech. The scheme has six syllabic and two phonetic factors. It still needs to be fully investigated for stuttered speech, but has been employed with children who are developing fluently (see 3 below).

A feature of note about all three schemes is that they all include CS and LEC factors and these factors are consistently found to be related to disfluency. For the CS/LEC/MS and IPC schemes which have been extensively evaluated in connection with their impact on stuttering, effects of difficulty are found on content words in older speakers.

The focus on phonological and phonetic factors is not intended to suggest that the only 'difficulty' leading to advancing are at these levels. It is conceivable that lexical factors (frequency of occurrence, age of acquisition, name agreement, etc.),m syntactic, lexical and prosodic factors also make a word difficult and hence prone to mis-timing.

2. Does stalling occur before difficult content words? In unpublished work, Howell and Ladd looked at the difficulty of word following stalling disfluencies using the IPC metric and compared this with difficulty of words not following stalling disfluencies. They found that the content words that followed stalling were significantly higher in difficulty than the words from fluent contexts. From this it would appear that the difficulty of the up-coming word determines whether the function words are repeated or not. Howell and Patel performed a similar analysis for young fluent children using ANOPHS to score word difficulty. Once again, the difficulty of the up-coming word was found to determine whether the function words are repeated or not

3. Priming of function and content words. Priming can be used to increase the speed of online speech production [22]. According to EXPLAN priming of content and function words should yield different effects on fluency. Savage and Howell conduced an experiment in which intransitive picture descriptions were elicited after priming of either function or content words (e.g. 'He is' or 'swimming' respectively) and compared performance for 12 children who stutter and 12 fluent age and gender matched controls (mean age of both groups was six years). Both groups produced significantly fewer disfluencies in their target responses after content word primes than function word primes and produced significantly more silent pausing after function word primes than content word primes. The effect of priming was significantly greater for CWS than for the fluent children and the target responses of CWS after a function word prime contained content words of significantly longer duration than did those of fluent children. These results are consistent with the view that advancing the point in time when a content word has to be produced (function word priming) causes more fluency problems than priming content words, they are pre-prepared for processing particularly in CWS.

3.3. Does producing one type of disfluency prevent the other?

1. Though there are two classes of fluency failure, they are both reflections of the same underlying problem (tackling a difficult content word either by repeating the words before or on the content word itself). If stalling is effective, then the speaker should not produce an advancing after a stalling and advancings should only occur when there is no stalling. Stuttering in PW has been examined to see whether it occurs in the either-or manner predicted. For English [2] and Spanish [3], less than 3% of disfluencies occurred on both the initial function words and the content word. An estimate of what would be expected by chance can be made by assuming function and content word disfluencies are independent. A related t-test across speakers showed the disfluency rates below 3% were significantly lower than expected on the assumption that function and content word disfluencies are independent.

2. Although disfluency in a PW is one sort or the other (e.g. stalling or advancing), the relative incidence of stallings and advancings can vary within or between individuals. A related observation is that disfluency in young speakers occurs on function words, but on content words in older speakers (which suggests that stalling is more prevalent at young ages and advancing at older ages. Both English and Spanish, speakers had high disfluency rates on pre-content function words and low disfluency rates on the content word. Disfluency on the function words dropped off with age and, as it did so, disfluency on the content words increased (termed an exchange function). In [13], an account of this change based on word frequency changes was presented.

4. Comparison with other units that could potentially account for these phenomena

The work in section two supported PW as a unit for analysis of disfluencies. However, alternative definitions of PW are possible that would be correlated with content-based segmentations for English. For instance, stressed words or low frequency words could act as nuclei. Lexical status (content versus function) can be dissociated from stress in Spanish where function words as well as content words are stressed. Segmentation of the same speech material into PW based either on content/function or stressed/unstressed criteria were made for Spanish material. Examination of PW that were segmented differently according to the two schemes showed that both produced ordinal and serial position effects and an exchange function [11]. Thus both would seem appropriate bases for defining PW. In recent analyses, Howell and Hayes have shown the two forms of PW do, however, show a difference when pauses were examined. More pauses occurred prior to PW that use stressed words as nuclei.

5. Conclusions and future work

The main conclusion is that PW appear to be a good unit for predicting the nature and position of disfluencies in speech. This is in stark contrast with Bernstein Ratner's conclusion [4] that "we cannot actually tell whether ANY disfluency (stuttered or not) reflects problems with a specific *word*, the *clause* it is being embedded in, or a word three words *further down* in the utterance, etc.". There is also slight evidence (that based on pauses) that stressed words may be a better basis for specifying the nuclei of PW. Work on fluency failure in PW in fluent speakers lags behind that in speakers who stutter and this needs to be rectified.

6. Acknowledgement

This research was supported by Wellcome Trust grant 072639.

7. References

[1] Ambrose, N.G. & Yairi, E. (1999). Normative disfluency data for early childhood stuttering. *Journal of Speech, Language and Hearing Research, 42,* 895-909.

[2] Au-Yeung, J. Howell, P. & Pilgrim, L. (1998). Phonological words and stuttering on function words.

Journal of Speech, Language and Hearing Research, 41, 1019-1030.

[3] Au-Yeung, J. Vallejo Gomez, I. & Howell, P. (2003) Exchange of disfluency from function words to content words with age in Spanish speakers who stutter. *Journal of Speech, Language and Hearing Research, 46*, 754-765.

[4] Bernstein Ratner, N. (in press). Is phonetic complexity a useful construct in understanding stuttering? *Journal of Fluency Disorders*.

[5] Clark, H. & Clark, E. (1977). *Psychology and Language: An Introduction to Psycholinguistics.* New York: Harcourt Brace.

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

[7] Dworzynski, K. & Howell, P. (2004). Predicting stuttering from phonetic complexity in German. *Journal of Fluency Disorders, 29*, 149-173.

[8] Garnham, A.,Shillcock, R. C. Brown, G. D. A. Mill, A. I. D. & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics, 19*, 805-817.

[9] Hartsuiker, R. J. & Kolk, H. H. J. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology, 42*, 113-157.

[10] Howell, P. (2002). The EXPLAN theory of fluency control applied to the Treatment of Stuttering by Altered Feedback and Operant Procedures. In E. Fava (Ed.), *Current Issues in Linguistic Theory series: Pathology and therapy of speech disorders* (pp.95-118). Amsterdam: John Benjamins.

[11] Howell, P. (2003). Is a perceptual monitor needed to explain how speech errors are repaired? *Gothenburg papers in Theoretical Linguistics, 90*, 29-32.

[12] Howell, P. (2004a). Assessment of some contemporary theories of stuttering that apply to spontaneous speech. *Contemporary Issues in Communicative Sciences and Disorders, 39*, 122-139.

[13] Howell, P. & Akande, O. (in press). Simulations of the types of disfluency produced in spontaneous utterances by fluent speakers, and the change in disfluency type seen as speakers who stutter get older

[14] Howell, P. & Au-Yeung, J. (2002). The EXPLAN theory of fluency control and the diagnosis of stuttering. In E. Fava (Ed.), *Current Issues in Linguistic Theory series: Pathology and therapy of speech disorders* (pp.75-94). Amsterdam: John Benjamins.

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

[16] Howell, P. Au-Yeung, J, Yaruss, S. & Eldridge, K. (in press). Phonetic difficulty and stuttering in English.

[17] Jakielski, K. J. (1998). *Motor organization in the acquisition of consonant clusters*. PhD thesis, University of Texas at Austin. Ann Arbor Michigan.

[18] Kolk, H.H.J. & Postma, A. (1997). Stuttering as a covert-repair phenomenon. In R.F. Curlee, & G. Siegel (Eds.), *Nature and treatment of stuttering: New directions.*(pp. 182-203). Boston: Allyn & Bacon.

[19] Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41-104.

[20] MacWhinney, B., & Osser, H. (1977). Verbal planning functions in children's speech. *Child Development, 48*, 978-985.

[21 Rispoli, M. (2003). Changes in the nature of sentence production during the period of grammatical development. *Journal of Speech, Language and Hearing Research, 46*, 818-830.

[22] Smith, M. & Wheeldon, L. (2001). Syntactic priming in spoken sentence production - an online study. *Cognition, 78*, 123-164.

[23] Strenstrom, A-B. & Svartvik, J. (1994). Imparsable speech: Repeats and other nonfluencies in spoken English. In N. Oostdijk & P. de Haan (Eds.), *Corpus-based research into language*. Atlanta, GA: Rodopi.

[24] Throneburg, N.R. Yairi, E. & Paden, E.P. (1994). The relation between phonological difficulty and the occurrence of disfluencies in the early stage of stuttering. *Journal of Speech and Hearing Research, 37*, 504-509.

[25] Weiss, A. L., & Jakielski, K.J. (2001). Phonetic complexity measurement and prediction of children's disfluencies: A preliminary study. In B. Maassen, W. Hulstijn, R. Kent, H. F. M. Peters & P. H. M. M. van Lieshout (Eds.), *The 4th Nijmegen Conferences on Speech Motor Control and Stuttering.* Nijmegen: Uitgeverij Vantilt.

[26] Yairi, E. & Ambrose, N. G. (2004). *Early childhood stuttering*. Austin, TX: Pro-Ed.

[27] Yaruss, J.S. (1999). Utterance length, syntactic complexity, and childhood stuttering. Journal of Speech, Language and Hearing Research, 42, 329-344.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp 103-108.

Optional *that* indicates production difficulty: Evidence from disfluencies

T. Florian Jaeger

Stanford University, USA

Abstract

Optional word omission, such as that omission in complement and relative clauses, has been argued to be driven by production pressure (rather than by comprehension). One particularly strong production-driven hypothesis states that speakers insert words to buy time to alleviate production difficulties. I present evidence from the distribution of disfluencies in non-subject-extracted relative clauses arguing against this hypothesis. While word omission is driven by production difficulties, speakers may use that as a collateral signal to addressees, informing them of anticipated production difficulties. In that sense, word omission would be subject to audience design (i.e. catering to addressees' needs).

1. Introduction

Optional word omission, as optional that omission in complement clauses (1a) and non-subject-extracted relative clauses (1b), has been one of the testing grounds for audience design (i.e. the questions *when*, *how* and *to what extent* speakers react to addressees' needs [3]).

(1a) I believe (that) my brother stole your bike.
(1b) I mean everything (that) they spray out in the fields.

In earlier work with Thomas Wasow on non-subject extracted relative clauses (henceforth NSRCs) [17], we have shown that that omission correlates with the distribution of disfluencies. The current project compares two competing accounts of that correlation. This comparison pertains to the question of the extent to which audience design influences word omission. Do speakers mention that to *alleviate* production difficulties (as first suggested in [21])? Or, alternatively, does that serve as *collateral signal* [5] to the hearer that the speaker anticipates production difficulties? Before I describe these two hypotheses in more detail, I briefly summarize the relevant discussion in the literature.

Research on the influence of audience design on word omission has focused on the facilitation of comprehension, most specifically the avoidance of structural ambiguity [2, 25]. Consider (1a). If that is omitted, it creates temporary structural ambiguity (the NP my brother could also be an object argument of believe, as in "I believe my brother"). Indeed, reduced complement clauses take longer to read than their full-form counterparts [22] (I use the terms "full" and "reduced" to refer to complement and relative clauses with and without that, respectively). For this reason, using the full form has been argued to serve the facilitation of comprehension [2, 11-13, 25] (which would make word omission subject to audience design). However, evidence from several laboratory production studies [7, 14, 18, 21] and corpus studies [17, 21, 24] argues against such comprehension-facilitation accounts of word omission. Ferreira & Dell [7] found no evidence that speakers avoid reduced forms when this leads to structural ambiguity (not all reduced forms result in ambiguity). While this finding is a null effect and as such should be interpreted with caution, it has been replicated [18, 24].

Ferreira & Dell [7] propose that word omission is subject to availability-based sentence production: speakers use the reduced form when the material following the omitted word is readily available. This hypothesis has received a fair amount of support. For example, speakers are more likely to use full-form complement/relative clauses when the embedded subject is more complex (e.g. a full lexical NP rather than a pronoun) and therefore takes longer to plan [8, 16, 21, 24]. Elaborating on the idea of availability-based sentence production, Race & MacDonald [21] hypothesized that speakers insert that to *alleviate* production difficulties. Uttering a relativizer may give speakers more time to overcome production problems in the embedded clause. I dub this the Alleviation Hypothesis.

Alternatively, mentioning that (where it may be omitted) could be a *collateral signal* (as defined in [5]). More specifically, speakers may use that to signal that they anticipate production difficulties. I dub this the Signal Hypothesis. The Signal hypothesis is based on and informed by earlier research on collateral signals of production difficulty [6, 9, 10]. Consider the case of uh/um. Speakers intentionally use the fillers uh/um when they are likely to suspend speech [6], and addressees are sensitive to this signal [3, 6]. This means, even though the distribution of uh/um is driven by production difficulties, the use of uh/um is a case of *audience design*. The Signal Hypothesis applies this insight to optional word omission. Consistent with existing evidence, the Signal Hypothesis predicts that the distribution of that is driven by production pressures. But, unlike the Alleviation Hypothesis, the Signal Hypothesis attributes this correlation to audience design: mentioning that is a collateral signal to addressees.

The two hypotheses make different predictions about the distribution of disfluencies (as direct evidence of production difficulties) in full and reduced clauses. The Alleviation Hypothesis predicts that mentioning that at the beginning of a complement or relative clause reduces disfluency in the clause. The Signal Hypothesis makes the opposite prediction. To compare the two hypotheses, I conducted a large-scale corpus study of disfluencies in full and reduced NSRCs (future research will show whether the observations made here also apply to other word omission environments). While it is beyond the scope of this paper to show whether addressees are sensitive to relativizer presence (a prerequisite of the Signal Hypothesis), I examine whether addressees *could* interpret relativizers as signals of upcoming production difficulties.

Section 2 gives an overview of the database and methodology. Section 3 tests the Alleviation Hypothesis. Section 4 provides a preliminary test of the Signal Hypothesis. Section 5 incorporates the new findings into a model of relativizer omission. The implications of these studies are discussed in Section 6. Directions for future work are addressed in Section 7.

2. Method and data overview

The results presented here are part of ongoing work [15, 27] on all 4,400 NSRCs from Paraphrase version of the Treebank III Switchboard corpus [4]. The corpus consists of 650 parsed and part-of-speech-tagged telephone conversations on selected topics between two strangers. All conversations in the Switch-

board corpus are annotated for disfluencies [20]. For the current study, NSRCs with *wh*-relativizers (which, where, who, etc.) were excluded because tests revealed that omission of these relativizer was frequently unacceptable. Of the remaining 3,701 NSRCs, 1,601 (43.3%) were full NSRCs (i.e. with that); 2,100 were reduced NSRCs. All NSRCs and the information analyzed below were automatically extracted using Tgrep2 [23] and Perl scripts (available upon request).

2.1. Data overview

Of the 3,701 NSRCs in the dataset, 593 (16%) contained at least one disfluency (compared to 36% for of all types of clauses in the Switchboard corpus). About 1.3% were *part* of a larger disfluency, but this did not affect relativizer omission ($\chi^2 < 1$) and will therefore not be discussed further.

Counting separately each repetition of fillers, and restarts, etc., the NSRCs contained 793 disfluencies. 351 (44%) were fillers, predominantly uh/um (61%) and you know (32%):

(2a) … the nuclear [$_{NSRC}$ *that*, uh, they use] …
(2b) … things [$_{NSRC}$ *that*, you know, two people can do] …
(2c) … every time [$_{NSRC}$ I, uh, I spent money, I mean, cash] …

The remaining 442 disfluencies were restarts (including lexical, syntactic, and covert repairs in the sense of Levelt [19]) or complete suspensions. These repairs were on average 1.35 words long (STDEV = 0.96, ranging from 1 to 9 words).

(3a) … the way [$_{NSRC}$ our system, our court system works] …
(3b) … some aunts [$_{NSRC}$ *that*, uh, I, I, I do] …
(3c) … all the, [$_{NSRC}$ *that* we're doing here] …

The NSRCs contained on average 5.3 words (STDEV = 3.9; ranging from 2 to 42 words), resulting in an overall disfluency rate (counting fillers, restarts, and suspensions) of 0.04 disfluencies per word (i.e. every 25th word belongs to a disfluency; STDEV = 0.14, ranging from 0 to 3.5).

3. Testing the Alleviation Hypothesis

The purpose of the first study is to test the Alleviation Hypothesis. According to the Alleviation Hypothesis, uttering that buys times for the speaker to reduce production difficulties. The Alleviation Hypothesis does not state how often this additional time is sufficient to prevent a disfluency that otherwise would have surfaced (henceforth the success rate). For empirical evaluation, it is important to distinguish versions of the Alleviation Hypothesis based on the assumed success rate. The strongest Alleviation Hypothesis assumes a success rate of 100%. Hence full NSRCs should contain significantly less disfluency than reduced NSRCs. No direct predictions are made for fillers, but as suspensions/restarts are often preceded by fillers, full NSRCs should contain fewer fillers. Furthermore, relativizer presence may correlate with a higher likelihood of disfluencies immediately preceding the NSRC because higher workload during the planning of NSRCs may increase the need to buy time. Before I discuss weaker versions of the Alleviation Hypothesis, I test the strongest Alleviation Hypothesis.

I investigated the distribution of fillers, suspensions, and restarts immediately preceding and following the beginning of an NSRC (with or without a relativizer). For the analysis, all fillers form one group (separate tests for the two most frequent types of fillers you know and uh/um did not reveal significant differences with regard to relativizer omission). Suspensions and restarts form another group.

3.1. Results: Relativizers correlate with disfluencies

The results are summarized in Table 1. The first two rows list the percentage of full and reduced NSRCs that contain at least one disfluency either in the modified NP (i.e. preceding the NSRC) or within the NSRC itself. The last row gives the significance level of Fisher's Exact test. All results were significant. Full NSRCs are at least two times more likely to contain/to be preceded by disfluencies than reduced NSRCs.

Table 1: Percentage of NSRCs preceded by/containing disfluencies

	Fillers		Suspension/Restart	
	In NP	In NSRC	In NP	In NSRC
% of full NSRC	4.7%	8.8%	2.2%	17.5%
% of reduced NSRC	1.9%	4.2%	1.1%	7.9%
Fisher's Exact	$p < 0.001$	$p < 0.001$	$p < 0.02$	$p < 0.001$

3.2. Intermediate discussion

The observed correlations between disfluency and relativizer presence are predicted by production-driven accounts of word omission [7, 17, 21]. The positive correlation of relativizer presence with production difficulties preceding NSRCs is compatible with the Alleviation Hypothesis. However, the finding that relativizers correlate with a *higher* likelihood of disfluencies in NSRCs rejects the strongest Alleviation Hypothesis.

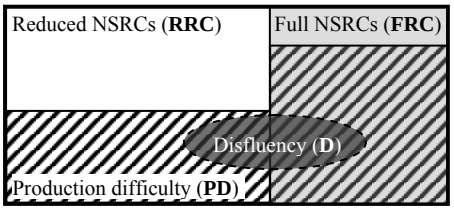
The results presented above are based on pooling all disfluencies in an NSRC. But maybe relativizers only help to alleviate production difficulties that occur at the beginning of NSRCs. In that case, pooling all disfluencies would unfairly bias against the Alleviation Hypothesis. To address this possibility, I conducted separate tests for the presence of suspensions/restarts in the first one to four words of NSRCs. The results, summarized in Table 2, still argue against the strongest Alleviation Hypothesis.

Table 2: Suspensions/restarts within the first *1..4* words of NSRC.

	Word 1	Word 1-2	Word 1-3	Word 1-4
% of full NSRC	5.5%	9.7%	10.9%	12.7%
% of reduced NSRC	2.7%	3.9%	4.9%	5.9%
Fisher's Exact	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Nevertheless, it could be that a weaker version of the Alleviation Hypothesis holds. Assuming a weak Alleviation Hypothesis, where the success rate of alleviation (due to relativizer insertion) is smaller than 100%, the results presented above may be due to an indirect association of that with more complex NSRC (e.g. because Alleviation isn't the *only* function of that). In that case, the higher likelihood of full NSRCs containing disfluencies could be a side effect. If full NSRCs are on average more complex, they will be associated with more production pressure, and they will more frequently result in production difficulties. To illustrate this idea, consider the Venn diagram in Figure 1 (the diagonal stripes indicate the subset of NSRCs that resulted in production difficulties).

Figure 1: Venn diagram of relativizer distribution, production difficulty, and disfluency in NSRCs (not drawn to scale)



The weak Alleviation Hypothesis predicts that relativizer presence alleviates production difficulties *after other factors are controlled for*. In other words, the proportion of alleviated production difficulties in full NSRCs (i.e. |[PD ∩ FRC] / D]| divided by |[PD ∩ FRC]|) should be higher than the proportion of alleviated production difficulties in reduced NSRCs (|[PD ∩ RRC] / D]| divided by |[PD ∩ RRC]|). These proportions cannot be directly calculated (since PD is not known), but one can try to control for as many factors as possible that contribute to making an NSRC a member of PD (i.e. factors that make an NSRC complex). Next, I describe the design of the statistical analysis that aims to do that.

3.3. Method and predictions

As in Section 3.1, two separate analyses were performed for fillers and suspensions/restarts. For the current study, a more sophisticated measure than relativizer presence was used as dependent variable: the normalized rate of disfluencies in the NSRC (i.e. the number of disfluent words per word). The number of words that were part of a disfluency was automatically extracted from the Switchboard Corpus. This number was divided by the overall number of words in that domain, yielding a normalized rate of disfluencies. This measure is finer-grained than disfluency presence (used in Section 3.1) and less correlated with NSRC length. Normalized disfluency rates proved a good measure of production difficulty; however, all results were confirmed for other measures of disfluency, cf. Section 3.5. Consider example (3a), repeated below. The NSRC contains six words, two of which are part of a disfluency (underlined). The NSRC's normalized disfluency rate in (3a) is therefore 2 / 6 = 0.33. The normalized disfluency rate preceding the NSRC in (3a) is 0 / 2 = 0.

(3a) … the way <u>our system</u>, our court system works …

Analyses of variance (ANOVAs) were conducted with subjects as random effects and relativizer as well as control factors as fixed effects. Including speaker effects in the ANOVAs accounts for the fact that speakers may have different base rates of disfluency. Table 3 summarizes the design of the two ANOVAs. The ANOVAs includes factors contributing to the complexity of NSRCs, and factors known to correlate directly with the presence of disfluencies as controls. Next, I briefly describe these controls.

Table 3: Design of ANOVAs

Dependent variable	Normalized disfluency rate in NSRC
Fixed effects [categorical]	Relativizer presence,
	Length of modified NP (in words)
	Is the NSRC subject a pronoun?
[continuous]	Rate of speech before NSRC
	Rate of speech within NSRC
	Length of RC (in words)
	Normalized rates of fillers in NP
	Normalized rate of suspension/restarts in NP
Random effects	Speaker

The rate of speech preceding and within NSRCs was included since rate of speech is known to correlate with disfluencies. The length of NSRCs (in words) was included since more complex NSRCs probably correlate with more disfluencies. Although the dependent variable already includes a control for NSRC length, NSRC length may correlate with a more than linear increase in disfluencies. Next, the normalized disfluency rates preceding the NSRC were included as a factor to control for production difficulties immediately preceding NSRCs (which may carry over into the NSRC). The NSRC subject's complexity was included because the availability-based sentence production hypothesis ([7], cf. Section 1) predicts that it should have an especially strong effect on relativizer presence. Finally, the grammatical function of the modified NP was included in case the NSRC's position within the matrix clause has an effect on disfluency rates. Other controls were tested but are not included in the final tests since they did not contribute to the model (Fs > 1).

If relativizers help to alleviate production difficulty, the normalized disfluency rate for NSRCs with a relativizer should be smaller than the rate for reduced NSRCs.

3.4. Results: Relativizers predict higher disfluency rates

Relativizer presence has a significant main effect on both the rate of fillers in the NSRC ($F1(1, 652) = 6.3$, $p < 0.02$) and the rate of suspensions/restarts in NSRCs ($F1(1, 989) = 6.1$, $p < 0.02$). But, contrary to the Alleviation Hypothesis, relativizer presence (i.e. full NSRCs) correlates with *higher* normalized disfluency rates even after the disfluency rates immediately preceding the NSRC are controlled for (cf. Table 4).

Table 4: Marginal means of the normalized disfluency rates

	Fillers	Suspensions/Restarts
Full NSRC	0.014	0.030
Reduced NSRC	0.005	0.019

NSRC length has a strong effect on both of the normalized disfluency rates (Fs > 100). This is due to an exponential increase in the average number of disfluencies for longer NSRCs. Furthermore, the speech rate in the NSRC significantly affects the rate of disfluencies (Fs > 15). The effect of the NP's grammatical function approached significance for the normalized rate of suspensions/restarts ($F1(3, 2863) = 2.5$, $p < 0.06$). All other factors failed to reach significance.

3.5. Discussion

The results argue against the Alleviation Hypothesis [21]. Full NSRCs are more likely to contain a suspension/restart (Section 3.1). This finding holds after controlling for other factors that could contribute to the amount of disfluency in NSRCs (i.e., the NSRC's complexity, the production pressure immediately preceding the NSRC, the rate of speech, and speaker effects). Thus even the weak Alleviation Hypothesis is hard to reconcile with the above results. Next, I discuss a couple possible objections to this conclusion.

First, it could be that normalized disfluency rates are not the best way to test the Alleviation Hypothesis. Maybe alleviation would show up in terms of a decrease in the absolute length of disfluencies or a decrease in the absolute number of disfluencies in the NSRC. Additional ANOVAs revealed that this was not the case. Regardless of which disfluency weight measure is chosen as the dependent variable (length, number, normalized disfluency rate), the effect of relativizer presence is either non-existent or in the opposite of the direction predicted by the Alleviation Hypothesis. Also, as shown in Section 3.2, it is not the case that relativizers only alleviate production difficulties immediately following them. This result was confirmed by an ANOVA with the factors from Table 3, but with the rate of suspensions/restarts in only *the first five words* as dependent variable. Contrary to the Alleviation Hypothesis, relativizers still were associated with a *higher* rate of suspensions/restarts ($F1(1, 893) = 11.1$, $p = 0.001$).

Interestingly, the results are consistent with the Signal Hypothesis. The presence of a positive correlation between dis-

fluency rates and relativizers after controlling for other factors suggests that speakers use relativizers as signals to the hearer.

4. Testing the Signal Hypothesis

The next question is whether addressees could potentially use relativizer presence as a signal that the speaker will be likely to run into production difficulties (as predicted by the Signal Hypothesis). In other words, given all information addressees have access to before they hear the beginning of an NSRC, does the presence of a relativizer provide additional information about the rate of disfluencies in the upcoming NSRC?

4.1. Method and predictions

Two separate ANOVAs were conducted for the normalized rates of fillers and suspension/restarts within NSRCs. In contrast to the analysis in Section 3, the factors included in these ANOVAs only contained information available to addressees prior to the beginning of the NSRC (cf. Table 5).

Table 5: Design of ANOVAs

Dependent variable	Normalized disfluency rate in NSRC
Fixed effects [categorical]	Relativizer presence,
	Grammatical function of modified NP
[continuous]	Rate of speech before NSRC
	Length of modified NP (in words)
	Normalized rates of fillers in NP
	Normalized rate of suspension/restarts in NP
Random effects	Speaker

Rate of speech was included as a factor since it usually correlates with disfluency rates. The length of the modified NP was included since producing complex NPs may consume more processing resources, which could cause a processing bottleneck when the speaker plans the NSRC. The grammatical function of the modified NP was included to account for effects that are due to the relative position of the NSRC within the matrix clause. Finally, to capture effects of experienced production difficulty at the point when the speaker is planning the NSRC, the normalized rates of fillers and suspensions/restarts immediately preceding the NSRC were also entered into the ANOVA.

The Signal Hypothesis predicts that relativizer presence exhibits a positive correlation with the rate of disfluencies, even after all information available to addressees prior to the relativizer is controlled for.

4.2. Results: Relativizers predict following disfluencies

Relativizer presence is a significant predictor of a higher rate of fillers ($F1(1, 604) = 12.2$, $p = 0.001$) and a higher rate of suspensions/restarts ($F1(1, 946) = 11.2$, $p = 0.001$).

Only one of the covariates had an effect. Fillers in the modified NP marginally predicted relativizer presence ($F1(1, 2942) = 2.8$, $p = 0.09$). Since fillers often occur in chains, this effect is expected and will not be discussed any further.

4.3. Discussion

The effect of relativizer presence on disfluency rates in NSRCs is predicted by the Signal Hypothesis. The effects are, however, rather subtle. Maybe relativizer presence only informs addressees of disfluencies early in NSRCs? Post-hoc tests were conducted on the rate of suspensions/restarts within the first two/the first five words of the NSRC. As expected, the effect of relativizer presence was stronger (both $Fs = 15.7$, $Ps < 0.001$). While the results are encouraging for the Signal Hypothesis, there are some issues that deserve discussion.

First of all, if relativizers are collateral signals, addressees should be sensitive to this information. While experiments show readers are sensitive to relativizer presence [11, 21], there is currently no evidence that addressees interpret relativizers as signals of production difficulties. Unfortunately, preliminary data searches only found ten interruptions by the addressee in the middle of an NSRC – too few to test whether interruption is more likely for full NSRCs (as predicted by the Signal Hypothesis). I leave this issue to future research.

Second, two potential confounds to the current study have to be addressed. First, it is well known that speakers lengthen words (including uh/um) when they are experiencing production difficulties [6, 9]. Maybe the difference in the relative rate of suspensions/restarts is solely driven by *lengthening* of relativizers (which, of course, can only occur in full NSRCs). Indeed, lengthening of relativizer that correlates positively (though only weakly) with the presence of suspensions/restarts at the beginning of an NSRC ($r = 0.09$, $p < 0.001$). A second potential confound is that the presence of a pause before the NSRC may signal to the hearer that disfluencies are to be expected. However, the existence of pauses is inversely correlated with relativizer presence (only 13% of all full NSRC are preceded by pauses compared to 20% of the reduced NSRC, $\chi^2 = 30.0$, $p < 0.001$). So, if anything, controlling for pauses should increase the correlation between relativizers and the disfluency rate in NSRCs.

To ascertain that the effect of relativizer presence holds after controlling for the existence of pauses before the NSRC as well as lengthening of that, I conducted an ANOVA including these two factors in addition to the factors mentioned above (I am indebted to Neal Snider for extracting information on pauses and the length of that from Switchboard). The duration of that was transformed by subtracting the mean duration from all cases with a relativizer (this was necessary to avoid colinearity between relativizer presence and relativizer duration). Cases without a relativizer were coded as having a duration of zero (i.e. these cases were treated just like cases with a relativizer of average length).

As expected, both absence of a pause ($F1(1, 2854) = 9.4$, $p < 0.001$) and lengthening ($F1(1, 2854) = 20.2$, $p < 0.001$) are significant predictors of following suspensions/restarts. The effect of relativizer presence was even stronger after controlling for these factors ($F1(1, 734) = 20.4$, $p < 0.001$).

In conclusion, the absence of a pause before an NSRC, lengthening of that, and the presence of that independently predict that the speaker will run into production difficulties at the beginning of the NSRC.

5. Modeling relativizer omission

The previous study shows that listeners could use the information provided by relativizer presence to predict that the speaker is (more) likely to produce disfluencies in the NSRC. The current study asks whether *speakers*' choice of full over reduced NSRCs is guided by anticipated difficulties and/or the presence of disfluencies preceding the NSRC. In other words, I develop a model of relativizer omission to see whether anticipated difficulty (here estimated by the presence of disfluencies) is a factor that drives the speakers' choice after other factors known to affect relativizer omission are controlled for.

5.1. Method

A Generalized Linear Model predicting relativizer omission was constructed using all factors known to account for a considerable amount of variation in relativizer omission (see A-H below). For more details on A-G, I refer to [8, 15, 16,

27]. H was included since ongoing work (together with Laura Staum) suggests that women use relativizers more frequently than men. Note that the inclusion of F and G assumes that speakers have *some* look-ahead into NSRCs. So does the inclusion of disfluencies in NSRCs as predictors (see below).

A Grammatical function of modified NP in matrix clause
B Determiner type of the modified NP
C Does the modified NP contain a uniqueness-requiring adjective or operator (e.g. the fastest person, …)?
D Does a light noun head the modified NP (e.g. way, …)?
E Does anything intervene between the modified NP's head noun and the beginning of the NSRC (e.g. my friend from New York that you met …)?
F Is the extracted element in the NSRC an adverbial NP?
G Is the NSRC subject a pronoun (e.g. the friend you met)?
H Gender of speaker

Earlier logistic regression models of word omission [15, 21, 24] did not model speaker effects. Such models rely on the assumption that each observation in the data set is independent of the other observations. However, datasets often contain several NSRCs from the same speakers, and speakers may have different base rates of relativizer omission (i.e. the assumption of the independence of observations is violated). Indeed, each speaker in the data set on average contributed about 10.8 NSRCs (STDEV = 9.2, ranging from 1 to 44).

I used logit Generalized Linear Mixed Models with A-H as fixed effects and normally distributed random intercepts to model speaker effects (as implemented in the R software library *glmmPQL* [26]; I am grateful to Joan Bresnan for pointing me to this solution). Logit models were used since the dependent variable (relativizer omission) is categorical.

The resulting mixed model (henceforth the standard model) predicts relativizer presence much more accurately (classification accuracy 75%) than the baseline model (a model that always predicts the more frequent event, here relativizer omission; classification accuracy 58%). The improvement was highly significant (change in -2log-likelihood = 455.0; $p < 0.0001$ based on a χ^2 with DF = 17). All factors contributed significantly to the standard model ($p < 0.01$ for factor C and H; all other factors $p < 0.0001$). Interestingly, the inclusion of speaker effects (STDEV of intercepts = 0.67) considerably improved classification accuracy (from 69% to 75%).

To test whether disfluency presence predicts relativizer presence after all of the above factors are controlled for, a separate model was fit for each of the four disfluency measures (cf. Table 6 below), by adding the disfluency measure to the standard model. Next, the goodness-of-fit of each of these four models was compared against the standard model's fit.

5.2. Results: Disfluencies predict relativizer presence

Table 6 summarizes the change in -2log-likelihood caused by adding any of the disfluency measures to the standard model as well as the significance level of that change based on a χ^2 with DF = 1 (adding a disfluency measure adds one free parameter to the model).

Table 6: Model improvement for each of the disfluency measures

	Fillers		Suspension/Restart	
	In NP	In NSRC	In NP	In NSRC
Coefficient in model	-0.02	0.89	-0.2	0.55
Change in -2log-LH	0	19.5	0.4	11.8
Significance level of χ^2	n.s.	$p < 0.001$	n.s.	$p < 0.001$

Information on fillers and suspension/restarts in the NSRC improves the model significantly. As expected, the coefficients for these two significant effects are positive (i.e. the more production difficulty a speaker anticipates, the more likely is a relativizer). Information on disfluencies preceding the NSRC does not improve the model.

5.3. Discussion

Under the reasonable assumption that speakers usually plan the beginning of NSRCs before they start pronouncing the relativizer, they may choose to insert a relativizer if they encounter production difficulties. The above results argue that speakers do this. Even after controlling for other factors known to favor relativizers, fillers and suspensions/restarts in NSRCs are significant predictors of relativizer presence.

Disfluency immediately before NSRCs does not seem to influence relativizer omission. This lack of an effect could be due to insufficient power. Note that there were fewer disfluent NPs (immediately preceding the NSRCs) than disfluent NSRCs. This asymmetry is not surprising since the NPs on average are much shorter (MEAN = 2.3 words, STDEV = 1.1, ranging from 1 to 8 words) than the NSRCs (MEAN = 5.3 words, STDEV = 3.9, ranging from 2 to 42 words). If, however, confirmed by other studies, the null effect of preceding disfluencies on relativizer omission would resemble Clark & Fox Tree's finding that uh and um contrast mainly in the length of the delays *following* them [6].

6. General discussion and conclusions

In earlier work with Thomas Wasow [17], we have presented preliminary evidence that relativizer presence is correlated with upcoming disfluencies. The studies presented here confirm that relativizers predict the presence of disfluencies in the NSRC *after controlling for other factors and speaker effects*. This finding is predicted by Ferreira & Dell's [7] availability-based hypothesis, according to which that is omitted when planning of the following material is finished. More generally, the results support the claim made in [7, 17, 21] that presence of that correlates with production difficulty.

Furthermore, the finding that relativizers are correlated with *more* disfluencies in the NSRC argues against the Alleviation Hypothesis. The available evidence suggests that relativizers do *not* help to alleviate production difficulties. On the contrary, the results support the Signal Hypothesis: relativizers are significant predictors of upcoming disfluencies even after properties of the modified NP as well as the disfluency rate before the NSRC are controlled for. In conclusion, while the current studies do not show that addressees *do* interpret optional relativizers as signals that the speaker is anticipating production difficulties, it shows that addressees *could* do so.

What are the consequences of these findings for the relation between audience design and optional word omission? The studies presented here show that optional that may be subject to audience design despite the fact that its distribution is governed by production pressures (in which case optional word omission would resemble uh/um, [6]). This is consistent with the observation that, although speakers do not seem to monitor their own speech for structural ambiguity, they use full-form clauses more frequently when addressees are present (compared to situations without addressees, [7]). This would be unexpected if optional that was uttered only to help speakers, but it is predicted by the Signal Hypothesis of optional word omission. If speakers use optional that to signal anticipated production difficulty, it makes sense that they use that more frequently in the presence of an addressee.

To end on a general note, the current finding argues that, for questions on the influence of audience design on morpho-syntactic variation (e.g., word omission, word order variation), it is misleading to focus exclusively on the avoidance of structural ambiguity. While current evidence suggests that speakers do not systematically avoid structural ambiguities (whether by prosodic phrase marking [1, 18], by insertion of disambiguating words [7, 24], or by choosing an unambiguous word order [1]), ambiguity avoidance is not the only way in which speakers can cater to addressees' needs: As shown by Clark & Fox Tree [6] for uh/um, speakers use collateral signals to keep addressees informed about the state of their production system.

7. Future work

More research on other word omission phenomena (e.g. complementizer omission, reduced subject-extracted relative clauses, omission of to after help) is necessary to see how general the finding presented here are.

With regard to the Signal Hypothesis, future research will decide whether addressees are sensitive to relativizer presence. Regardless of whether relativizers are signals or symptoms, a better understanding is needed of what kind of production difficulty that correlates with. For example, preliminary evidence suggests that relativizers also correlate with the presence of pauses immediately following them. Future research will show whether relativizers correlate specifically with suspensions of speech (as is the case for uh/um). Alternatively, speakers may use relativizers whenever they anticipate high workload. Ongoing work also investigates correlations of relativizers with the rate of speech before and within NSRCs.

8. Acknowledgements

I would like to express my heartfelt thanks to: Tom Wasow for continuous intellectual guidance; Herb Clark for discussions on the issue of audience design; Roger Levy and Joan Bresnan for statistical advice. Extraction of time information (that-length and pauses) would have been unfeasible without the incorporation of the Mississippi time-stamps into the Paraphrase Switchboard (by Shipra Dingare; LINK funding is gratefully acknowledged). I also owe thanks to Neal Snider, Dan Jurafsky, Laura Staum, Sasha Calhoun and Dave Orr for help with various aspects of this paper/project. None of the above necessarily agrees with the views presented here. This work has been funded in part by a Stanford Summer RAship.

9. References

[1] Arnold, J. E., Wasow, T., Asudeh, A., & Alrenga, P. 2004. Avoiding Attachment Ambiguities: the role of Constituent Ordering. *Journal of Memory and Language, 55*(1), 55-70.

[2] Bolinger, D. 1972. *That's that.* The Hague: Mouton.

[3] Brennan, S. E., & Williams, M. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34*, 383-393.

[4] Bresnan, J., Carletta, J., Crouch, R., Nissim, M., Steedman, M., Wasow, T., et al. 2002. *Paraphrase analysis for improved generation*: HRCR Edinburgh-CLSI Stanford.

[5] Clark, H. 2004. Pragmatics of language performance. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 365-382). Oxford: Blackwell.

[6] Clark, H. H., & Fox Tree, J. E. 2002. Using uh and um in spontaneous speaking. *Cognition, 84*, 73-111.

[7] Ferreira, V. S., & Dell, G. S. 2000. Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production. *Cognitive Psychology, 40*, 296-340.

[8] Fox, B. A., & Thompson, S. A. to appear. Relative Clauses in English conversation: Relativizers, Frequency and the notion of Construction. *Studies in Language*.

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

[10] Goffman, E. 1981. Radio talk. In E. Goffman (Ed.), *Forms of talk* (pp. 197-327). Philadelphia, PA: University of Pennsylvania Press.

[11] Hakes, D. T., Evans, J. S., & Brannon, L. L. 1976. Understanding sentences with relative clauses. *Memory & Cognition, 4*(3), 283-290.

[12] Hawkins, J. A. 2001. Why are categories adjacent? *Journal of Linguistics, 37*, 1-34.

[13] Hawkins, J. A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

[14] Jaeger, T. F., Fedorenko, E., & Gibson, E. 2005. Dissociation between Production and Comprehension Complexity. The 18th Annual CUNY Sentence Processing Conference, March 31st - April 2nd, 2005, Tuscon, AZ.

[15] Jaeger, T. F., Orr, D., & Wasow, T. 2005. *Comparing Frequency-based and Complexity-based Accounts (of Relativizer Omission)*, The 18th Annual CUNY Sentence Processing Conference, March 31st - April 2nd, 2005, Tuscon, AZ.

[16] Jaeger, T. F., & Wasow, T. 2005. *Processing as a Source of Accessibility Effects on Variation.* Paper presented at the Berkeley Linguistic Society.

[17] Jaeger, T. F., & Wasow, T. 2005. *Production Complexity Driven Variation: The case of relativizer distribution in non-subject-extracted relative clauses*, The 18th Annual CUNY Sentence Processing Conference, March 31st - April 2nd, 2005, Tuscon, AZ.

[18] Kraljic, T., & Brennan, S. E. 2005. Prosodic disambiguation of syntactic structure: For the speaker or for the hearer? *Cognitive Psychology, 50*, 194-231.

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

[20] Mateer, M., & Taylor, A. 1995. *Disfluency Annotation Stylebook for the Switchboard Corpus*. Unpublished manuscript, University of Pennsylvania.

[21] Race, D. S., & MacDonald, M. C. 2003. *The use of "that" in the production and comprehension of object relative clauses.* Paper presented at the 26th Annual Meeting of the Cognitive Science Society.

[22] Rayner, K., & Frazier, L. 1987. Parsing temporarily ambiguous complements. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 39A*(4), 657–673.

[23] Rohde, D. 2001. *Tgrep2 Manual*. Unpublished manuscript, Brain & Cognitive Science Department, MIT.

[24] Roland, D., Elman, J. L., & Ferreira, V. S. in press. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*.

[25] Temperley, D. 2003. Ambiguity avoidance in English relative clauses. *Language, 79*(3), 464-484.

[26] Venables, W. N., & Ripley, B. D. 2002. *Modern Applied Statistics with S* (Fourth edition ed.): Springer.

[27] Wasow, T., & Jaeger, T. F. 2005. *Lexical Variation in Relativizer Frequency*, Expecting the unexpected: Exceptions in Grammar Workshop at the 27th Annual Meeting of the German Linguistic Association.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 109-112.

Phrase-final rise-fall intonation and disfluency in Japanese: - A preliminary study -

Jumpei Kaneda

Graduate School of Human Science, Kobe University, Japan

Abstract

In Japanese conversations, rise-fall intonation with vowel lengthening often occurs on the final syllable of a phrase. This phrase-final rise-fall (PFRF) is a new type of intonation first reported in the 1960's. Researchers consider PFRF intonation a discourse marker which functions to sharpen the phrase boundary and retain the utterance turn, but other phrase-final intonation such as phrase-final lengthening (PFL) can have a similar pattern. PFLs are recognized as a type of disfluent speech with similar characteristics to PFRFs in terms of final-lengthening and having discourse functions. Also from reports about the spontaneity of speech, we assume that PFRFs would have a relation with disfluency, as well as with PFLs. To examine this assumption, this paper attempts to show the co-occurrence relation between PFRF and disfluency in the same utterance. The results show that PFRFs and PFLs have a relation to posterior disfluent units and suggest that both indicate speech planning strategies. Further, this paper speculates that a difference between PFRF and PFL is a difference in the purposes of speech planning: the latter represents ongoing linguistic editing while the former indicates adjusting the utterance according to the interlocutor's reaction. Disfluencies accordingly occur as effects from processes of speech planning.

1. Introduction

In Japanese spontaneous casual speech, it is often observed that the final syllable of an utterance non-final phrase (more precisely "intermediate phrase" [1]) is pronounced with a rise-fall intonation and the final vowel is more or less lengthened. This phenomenon has been called *shiriagari* "final-rising" intonation [2] or *shoko-cho* "rise-fall intonation" [3], but this paper calls this intonation "**phrase-final rise-fall** (later **PFRF**)". An example is shown in Figure 1 below.

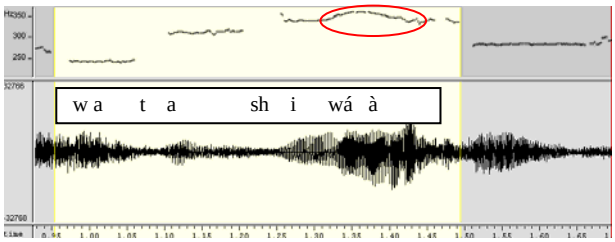


Figure 1: Pitch contour of *watashi-wa* "as for me" with a phrase-final rise-fall intonation (see the circled area)

Some researchers say a PFRF is used in order to attract the interlocutor's attention, and others say it has discourse functions (see Section 2.2). Although PFRF is not successfully explained yet, we can say that a PFRF shows at least a modality toward the interlocutor.

A PFRF always occurs on the end of a phrase, not a word. In Japanese, a head-final language with postpositions, *watashi-wa* is a single phrase formed by a noun *watashi* "me" and a topic marker *wa* as head of the phrase. A PFRF thus, in this case, occurs on the final syllable, the particle *wa*. Meanwhile, a PFRF seems to be a "postposition (particle) prominence" for the purpose of emphasizing the postposition, but we claim it is not. The reason is because a noun phrase with only a monosyllable postposition particle such as *wa*, *ga* and *o* containing a PFRF would indicate postposition prominence, but this is different from the case of a two-syllable postposition particle as illustrated below.

(1) Koobe-kara (ki-mashi-ta.)
(I came) from Kobe.
a. **ko**obe-**ka**ra
b. **ko**obe-karáà

Example (1) shows a contrast in intonation and meaning. Bold letters indicate that the syllable is pronounced with prominence (high pitch) and *koobe* has a lexical accent on the first syllable. (1)a is an example of postposition prominence, that emphasizes the meaning of the postposition particle when contrasting with other particles (not "to" but "from"). In this case the particle *kara* is accented on its first syllable independently from *Koobe*, but never on the second. This is phonologically different from (1)b, in which a PFRF appears on the last syllable. As for a difference in meaning, (1)a sounds emphasized on the particle while (1)b does not. In addition, a PFRF can appear on zero-particle NPs, adjectivals, adverbials, and VPs (not in the sentence-final position); this is another reason for differentiating PRFRs from postposition prominence. A PFRF can thus be said to be an intermediate phrase-level intonation phenomenon.

2. Background

2.1. History of PFRF

PFRF is a comparatively new phrase-level intonation phenomenon. According to Akinaga (1966) PFRFs existed in the 1960's at the latest [4], being used mostly by young women and children in the Kanto district [3]. Today the use of PFRF has spread throughout Japan as well as among male speakers. Teachers or lecturers also tend to use PFRFs since PFRFs are considered to be effective to help pupils understand when explaining things.

On the other hand, PFRFs are stigmatised by some older or conservative speakers of Japanese saying that it sounds rude, flippant or childish [2][3], and even among the younger generation, speakers who frequently use PFRFs are sometimes evaluated negatively [5].

2.2. Discourse functions of PFRF and their problem

Apart from sociolinguistic features mentioned in the previous section, some researchers affirm that PFRFs have pragmatic or discourse functions. In this section we see two functions: clarifying phrase boundaries (demarcating) and retaining the speaker's turn. As for the former, Sugito (1983) affirms PFRF's demarcating function [2]. The turn-keeping function is explained in Inoue (1997); PFRFs indicate continutity of the utterance and prevent the other speaker from interrupting [3].

The notion that a PFRF is a discourse marker is, however, incorrect. The above two functions are not inherent to PFRFs. With a normal (no rise-fall) phrase-final intonation, pauses in sentence-medial position can clarify the boundaries of phrases and keep the utterance turn. Postposition prominences (see Section 1) have these two functions, also. This implies that PFRFs originate from boundary (extra-)linguistic factors other than discourse. A possible extra-linguistic phenomenon is **disfluency**.

2.3. Disfluency and PFRFs

There is another phrase-final intonation, **phrase-final lengthening** without falling intonation (referred to as **PFL**). Figure 2 below is an example of PFL. The final vowel is lengthened and the pitch stays high.

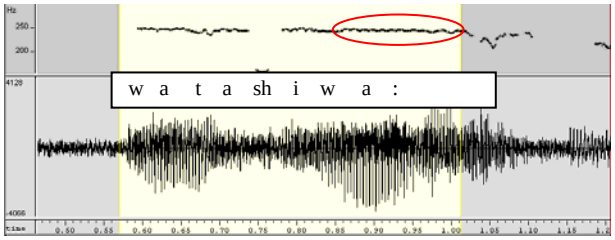


Figure 2: Pitch contour of "*watashi-wa* (as for me)" with a phrase-final lengthening (see the circled area).

A PFL intonation has two characteristics: one is a representation of disfluency, the other, a discourse function. It is thought that a PFL appears when the speaker hesitates to utter the following part of the utterance (cf. [6]). That is to say, a PFL occurs at the moment speech planning involving word selection, calculation, or recalling something is not completed. In this sense, phrases with a PFL intonation are similar to **fillers** such as *eeto* "let me see", *anoo* "uhm" and *nanka* "like." As for discourse functions, a PFL also has both demarcating and turn-keeping functions, mentioned in Section 2.2. Phrase boundaries are defined by a final vowel lengthening and the utterance turn does not move because the utterance is still incomplete.

A PFRF is similar to a PFL with regard to two points. First, both have some sort of final vowel lengthening. Second, the two discourse functions, demarcating and turn-keeping, are associated with both intonation patterns. The similarity implies that a PFRF intonation tends to appear in disfluent speech as well as does a PFL. For supporting this implication, the National Institute of Japanese Language shows that spontaneity of speech correlates with co-occurrence of disfluencies (e.g. fillers and fragments of word) and PFRF intonation [7]. In other words, in highly spontaneous speech such as casual conversations, disfluencies and PFRFs co-occur frequently. Although these data do not show that disfluency has a direct relation with PFRF, they at least suggest a possible relation between them.

2.4. Hypothesis

Based on the similarity of PFRF and PFL, and an expectation of a relation among spontaneity, disfluency and occurrence of PFRF shown in the previous section, this paper sets up a hypothesis that the PFRF intonation is a representation of disfluency. It is, however, difficult for this paper to fully verify this hypothesis, but we can test the expectation mentioned in the bottom of Section 2.3 that PFRFs have a relation with disfluency. If their relation is recognized, it will partially support our main hypothesis. This paper will thus test the relation between PFRF and disfluency by making use of the corpus of spontaneous dialogues as described below.

3. Method

3.1. Corpus

From a fifty-hour spontaneous dialogue corpus recorded during October and November 2004, utterances from 5 female informants between the ages 27 and 45 with a total amount of 15 hours were selected.

3.2. Method

The samples were utterances containing a pronominal expression *watashi* (sometimes pronounced with the first consonant dropped like *atashi*) "me" as topic or argument, but not as predicate. The samples were classified into 3 categories according to phrase-final intonation patterns and checked to see if there were co-occurrences of disfluency units such as otiose pauses and fillers in the same utterance. The co-occurrence check consisted of two parts, (i) presence of excessive pauses or fillers immediately before or after *watashi-(wa)* and (ii) co-occurrence of these disfluency factors in an anterior or posterior position to *watashi(-wa)* in the same utterance. Our hypothesis is that if the co-occurrence rate of disfluencies with PFRF intonation is significantly higher than that with normal intonation, and at the same time the co-occurrence with PFLs is also significantly higher than that with normal intonation, we can say that a PFRF is related to disfluency.

3.3. Reasons for collecting samples containing *watashi*

There are three reasons why utterances with *watashi* were chosen. First, *watashi* tends to appear in the topical position as it does in other languages (cf. [8]). If a representation of disfluency appears in the anterior position of an utterance, it may cause another disfluency representation in the rest of the utterance because of ongoing speech planning. We expect that more co-occurrences of disfluent elements like fillers will be observed if a PFRF is a representation of disfluency.

Second, when *watashi* appears in a zero-particle (bare) form or topic-marked form (followed by *wa*) it rarely has salience. Meanwhile, case-marked forms such as *watashi-ga* (subject), *watashi-o* (object) etc. are likely to be focused and pronounced with an emphatic prominence [9]. To avoid interference from focal prominence, we excluded case-marked forms from the target samples.

The third reason for adopting the utterances with *watashi* was because it is lexically unaccented unlike its masculine-only counterpart *boku*, which has an accent on the first syllable. *Watashi* is a general form indicating first person singular but since in casual speech it tends to be used by women, we chose only women for subjects.

3.4. Classification

The samples were classified by phrase-final intonation patterns on *watashi(-wa)* phrases into 3 categories, Normal, PFL and PFRF. Classifications were performed according to auditory perception by the author. PFLs had phrase-final syllable prolongation and no falling intonation, while PFRFs were those with rise-fall intonation on *watashi(-wa)*. The rest of the samples were labelled as Normal.

For disfluency elements, we focused on pauses and fillers. The criteria of pause as disfluency were (i) perceptible silent section inside an utterance and (ii) it sounded longer than the expected length. Judgments of pause were done according to auditory impressions of phrase-final intonation patterns. Fillers were units with no linguistic content (in a traditional sense) like *eeto*, *anoo*, *nanka*. Interjections or discourse makers like *un* "yes", *e* (surprise), *a* (finding), *hee* "I see" etc. were not treated as fillers.

4. Results

4.1. Basic data

Before testing the relation between phrase-final intonations and disfluency, we analyze non-disfluent types of phrase-final intonation. Distribution of each phrase-final intonation pattern is shown in Table 1 below.

Table 1: Distribution of phrase-final intonation

Type	Normal	PFL	PFRF	Total
watashi	143	22	8	173
watashi-wa	71	23	19	113
Total	214	45	27	286

These data show that PFRFs are fewest among the 3 categories (9.4%). Focusing on the rate of *watashi-wa*, categories PFL and PFRF are significantly higher than Normal (PFL $p<0.05$, PFRF $p<0.001$), though PFL and PFRF are not significantly different. Relations between the presence of particle *wa* and phrase-final intonation will be discussed later in this section.

Also we observed co-occurrences of interjectional phrase-final particles like *ne* after *watashi(-wa)* which appear very frequently in colloquial Japanese. Table 2 shows the distribution of particles among intonation patterns.

Table 2: Distribution of phrase-final particle among intonation patterns

PF-particle	Normal	PFL	PFRF	Total
None	207	37	18	262
Ne	6	6	8	20
Others	1	2	1	4
Total	214	45	27	286

Comparing the categories None and *Ne*, PFL and PFRF are significantly different from Normal category (PFL $p<0.005$, PFRF $p<0.001$), i.e., the particle *ne* is likely to appear with either PFL or PFRF intonation patterns. As is seen in Table 1, there were 8 cases in which non wa-marked *watashi* was pronounced with a PFRF, while 6 of them were accompanied by phrase-final particle *ne*. This means *watashi* is less likely to appear without a phrase-final particle (2 out of 8) when pronounced with a PFRF intonation. Adding *wa*-marked watashi-*wa* with a PFRF (19 samples), only 2 samples of 27 (7.4%) appeared without postpositional morpheme (*wa, ne* and *sa* in Other category). For PFLs, 17 out of 45 samples were of *watashi* without any particle. This fact may be a hint for finding differences between PFL and PFRF. Further discussion is continued in Section 5.

4.2. Disfluency-related data

In this section, we report the main analyses, and the possibility of a relation between phrase-final intonation patterns with occurrences of disfluency units such as otiose pauses and fillers defined in Section 3.4. First, we examined the presence of idling pauses or fillers immediately before or after *watashi-(wa)*. Table 3 shows the distribution of co-occurrences of fillers **right before** *watashi(-wa)* and Table 4 for pauses. Samples in which *watashi(-wa)* comes first in the **utterance** were excluded (41 out of 286).

Table 3: Co-occurrences of FILLERS just BEFORE *watashi(-wa)*

Fillers	Normal	PFL	PFRF	Total
-	152	23	13	188
+	30	16	11	57
total	182	39	24	245

Table 4: Co-occurrences of PAUSES just BEFORE *watashi(-wa)*

Pauses	Normal	PFL	PFRF	Total
-	130	31	20	181
+	52	8	4	64
Total	182	39	24	245

With regard to the co-occurrence rate of fillers (Table 3), categories PFL and PFRF are significantly higher than Normal (both $p<0.001$); with regard to pauses (Table 4), no significant difference was observed. Next, co-occurrences of Fillers **immediately after** *watashi-wa* are shown in Table 5, those of Pauses in Table 6. 17 utterances in which *watashi(-wa)* came after the predicate were out of the subjects.

Table 5: Co-occurrences of FILLERS just AFTER *watashi(-wa)*

Fillers	Normal	PFL	PFRF	Total
-	164	26	16	206
+	35	17	11	63
Total	199	43	27	269

Table 6: Co-occurrences of PAUSES just AFTER *watashi(-wa)*

Pauses	Normal	PFL	PFRF	Total
-	181	28	14	223
+	18	15	13	46
Total	199	43	27	269

Unlike the immediately anterior position (Table 3, 4), for the co-occurrence rate of both Fillers and Pauses, PFL and PFRF were significantly higher than the category Normal (Fillers: PFL $p<0.005$, PFRF $p<0.01$; Pauses: both $p<0.001$). In sum, in the situation of co-occurrence with PFLs and PFRFs, Fillers are more strongly related than Pauses, and disfluency units are more likely to co-occur after *watashi(-wa)* than before.

Finally, we examined co-occurrences of disfluent units (Pauses and Fillers) in an anterior or posterior position to *watashi(-wa)* in the same utterance. Table 7 shows the results for anterior positions, Table 8 for posterior positions. Table 7 excludes samples with the utterance-initial *watashi(-wa)*, and Table 8 those with *watashi(-wa)* after the predicate.

Table 7: Co-occurrences of disfluent units FORMER than *watashi(-wa)*

Fillers	Normal	PFL	PFRF	Total
-	108	18	9	135
+	74	21	15	110
Total	182	39	24	245

Table 8: Co-occurrences of disfluent units LATTER than *watashi(-wa)*

Pauses	Normal	PFL	PFRF	Total
-	128	12	7	147
+	71	31	20	122
Total	199	43	27	269

The co-occurrence rate of disfluent units **anterior** to *watashi(-wa)* (Table 7) was significantly higher in the PFRF category than in Normal ($p<0.05$), but PFL was not significantly different from Normal, while for **posterior** positions (Table 8) the rate of disfluency units which co-occurred is significantly higher in both PFL and PFRF than Normal (both $p<0.001$). Similar to the immediately prior and post positions (Tables 3 to 6), disfluent units posterior to *watashi(-wa)* are significantly more likely to co-occur with PFLs and PFRFs than those of anterior.

In conclusion, PFLs and PFRFs were observed to have a co-occurrence relation with posterior disfluent units rather than

anterior. In this sense, we can say PFRF intonation has a relation with disfluency similar to that with PFLs. This implies that PFRFs represent disfluencies in a way similar to PFLs.

5. Discussion

In Section 4, PFRFs as well as PFLs have been shown to relate to disfluecy. We will review similarities of PFRF and PFL. As mentioned above, both intonation patterns function to retain the turn in discourse. In addition, when talking to the speaker her/himself, i.e. soliloquies, both PFLs and PFRFs seldom occur. The filler *anoo*, which was mentioned to be similar to PFL intonation in section 2.3, is also unsuitable for soliloquies [10]. Sadanobu & Takubo (1995) explained that *anoo* is "used when the speaker makes a linguistic editing in the mind" and "functions to maintain the speaker-interlocutor interface (1995:79)." PFRFs and PFLs are similar to *anoo*, and in this sense they indicate speech planning in process.

There remains, however, another problem: what is the difference between PFRF and PFL? In Section 4.1, a difference between PFRF and PFL is that the former is likely to co-occur with the particle *wa* or *ne*, not in bare form. When a particle appears after a noun, they can close as one phrase, while bare nouns without postposition are not always independent phrases. From this notion, PFRFs can appear only when the phrase is formed while PFLs do not have such a constraint; they can appear on an unclosed noun phrase. As evidence, our corpus has one sample in which bare *watashi* is pronounced with its final vowel lengthened and after that the particle *wa* appears with a PFRF, like *watashi-i-wáà*. The phrase is not closed yet at the moment of lengthening of the third syllable *shi* and on the fourth syllable *wa* the phrase is closed and can be pronounced with a PFRF intonation. In this sense a PFL is not exactly a phrase final intonation.

Another difference between PFRF and PFL is reported by Ichikawa (2005). Analysis of the relation between phrase-final intonation and the interlocutor's reactions shows that PFRFs cause the interlocutor to nod or do an *aizuchi* (supportive response) at a higher rate than do PFLs [11]. This means a PFRF can give the interlocutor a chance to react more than a PFL does. We saw that PFRFs and PFLs resemble the filler *anoo*. Sadanobu & Takubo (1995) classified the linguistic function of *anoo* into two operations: "searching for the name" and "considering an appropriate expression"(ibid) [10]. The former is a purely linguistic operation while the latter is a pragmatic one. Applying this idea to phrase-final intonation, I speculate that PFLs are representations of ongoing linguistic editing, such as word selection, remembering things or calculating, while PFRFs, like the above, appear when the speaker monitors her/himself to adjust the utterance according to the interlocutor's reaction or the discourse situation. These operations require a considerable amount of work for the brain. When the working memory or buffer runs short, disfluencies will occur, fillers and otiose pauses will interrupt linguistic information, and the speech speed therefore will decrease. Also, the speaker's attention to the interlocutor's reactions will retard the tempo of speech with presence of PFRF intonation.

6. Conclusion and ending remarks

This paper has shown the relation between a PFRF intonation and disfluency and tried to elucidate the difference between the two phrase-final intonation types, PFRF and PFL. The study has problems mainly in its method; judgments of prosodic features such as pitch form and length of pause were performed by the auditory impression of a single listener. In future studies, we will utilize more objective methods such as quantitative analysis and multiperson judgments. In addition, utterances of male speakers will also be examined.

7. Acknowledgements

This work is sponsored by (i) Grants-in-Aid for Scientific Research A 16202006 from the Ministry of Education, Culture, Sports, Science and Technology, and by (ii) Strategic Information and Communicative R&D Promotion Programme 041307003 from Ministry of Internal Affairs and Communications.

8. References

[1] Pierrehumbert, Janet & Mary Beckman. 1988. *Japanese Tone Structure*. Cambridge, Massachusetts: MIT Press.

[2] Sugito, Miyoko. 1983. Nihongo-no Akusento-to Intoneshon (Accents and Intonations in Japanese). *Kotoba-to Onsei (Language and Speech)*. Agency for Cultural Affairs.

[3] Inoue, Fumuo. 1997. Intoneshon no Shakaisei (Sociality of Intonation). Sugito, Kunihiro, Hirose & Kono(Eds.) *Nihongo Onsei 2 Akusento, Intoneshon, Rizumu to Pozu (Japanese Speech 2, Accent, Intonation, Rhythm and Pause)*, pp. 143–168. Tokyo: Sanseido.

[4] Akinaga, Kazue. 1966. Nihongo-no Hatsuon –Intoneshon nado (Pronunciation of Japanese – Intonation and other features). *Koza Nihongo Kyoiku (Seminar in Japanese Language Education)*, vol. 2, pp. 48-60. Waseda University Laboratory of Language Education.

[5] Hara (Sasaki), Kaori. 1993. The Acoustic Feature and Sensory Impression of What Is Called "High-Rising" Intonation in Japanese. *Studies in Language and Culture*, vol. 11, pp 61-71. Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies

[6] Nakagawa, Akiko & Toshiyuki Sadanobu. 2003. The Contrastive Study of "Disfluency" between Japanese and Chinese: Preliminary Research. Proceedings of *The 1st JST/CREST International Workshop on Expressive Speech Processing*. 21-22, February 2003. In Kobe University, Japan.

[7] National Institute of Japanese Language, the. 1999-2003. *The Corpus of Spontaneous Japanese, Preliminary Analyses II*. http://www2.kokken.go.jp/~csj/public/j6_2.html

[8] Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Cole (ed.) Radical Pragmatics, pp. 223-255. New York: Academic Press.

[9] Kaneda, Jumpei. 2005. Unmarkedness of Zero-Particle NPs. *KLS 25*, pp. 315–325. Kansai Linguistic Society.

[10] Sadanobu, Toshiyuki & Yukinori Takubo. 1995. The Monitoring Devices of Mental Operations in Discourse: A Case of "Eeto" and "Ano(o)." *Gengo Kenkyu (Studies in Linguistics)* vol. 108, pp. 74-93. The Linguistic Society of Japan.

[11] Ichikawa, Akira. 2005. Multimodal Dialogue Corpus and Analysis of Speaker's Nod. Report of *Realization of Advanced Spoken Language Information Processing from Prosodic Features*. http://www.gavo.t.u-tokyo.ac.jp/tokutei_pub/houkoku/corpus/ichikawa.pdf.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 113-116.

Evaluation of vowel hiatus in prosodic boundaries of Japanese

Shigeyoshi Kitazawa

Shizuoka University, Hamamatsu, Japan

Abstract

We investigated V-V hiatus through J-ToBI labeling and listening to whole phrases to estimate degree of discontinuity and, if possible, to determine the exact boundary between two phrases. Appropriate boundaries were found in most cases as the maximum perceptual score. Using electroglottography (EGG) of the open quotients OQ, pitch mark and spectrogram, the acoustic phonological feature of these V-V hiatus was found as phrase-initial glottalization and phrase-final nasalization observable in EGG and spectrogram, as well as phrase-final lengthening and phrase-initial shortening of the morae. A small dip was observable at the boundary of V-V hiatus showing glottalization. The test materials are taken from the "Japanese MULTEXT", consisting of a particle - vowel (36), adjective - vowel (5), and word - word (4).

1. Introduction

In normal fluent speech, phrase as well as word boundaries become obscure because of fluency and become difficult to segment. This is the salient problem of speech recognition and speech synthesis. Marks such as juncture, punctuation, focus, and prominence in a stream of speech sound are crucial for effective use of prosodic corpus. Resolution of such hiatus plays an important role in listening comprehension. In Japanese, there are very few studies about hiatus, but Kawahara states that preceding vowels spread into following syllables [1].

This paper presents results of a study concerning the boundary between morphological units, i.e., words and phrases in a Japanese sentence. Here we investigate the phrasal boundary in an utterance comprising a transition between a final mora of a preceding accentual phrase and an initial mora of the succeeding accentual phrase consisting of the same two vowels, i.e., vowel-vowel hiatus.

J-ToBI, a prosody annotation scheme, defines the phrase structure vaguely as BI label with 5 different degrees as perceived disjuncture [2]. We tried to measure this ambiguous disjuncture quantitatively through a series of perceptual experiments. Results were also investigated using EGG analyzed data (open quotient), F0, speech waveform, and spectrogram. These observed disjunctures matched with discontinuities of articulatory measurements.

2. Vowel-Vowel Hiatus in Japanese

Since Japanese almost exclusively consists of open syllables ending with a vowel (with the exception of some syllables ending in ～ん), if the following phrase begins with an initial vowel, a vowel-vowel (V-V) hiatus arises, the same vowel continues without pause. This vowel sequence is very common in Japanese:

body of a phrase	vowel	|	vowel	body of a phrase

2.1. Morphonology of vowel-vowel hiatus

Possible Japanese vowel-vowel hiatus consists of the following structure:

front phrase		rear phrase
noun + particle	→	predicate
morpheme + adverb	→	adjective
a part of compound word	→	a part of compound word

Example hiatus was taken from our corpus (3.1). The most frequent occurrence is with particles, and the next most frequent occurrence is with adjectives.

The most common phrasal unit is a morpheme (e.g. a noun) + a particle which bears an accent to compose a particle (joshi) | vowel initial phrase.

a morpheme (e.g. a noun) + a particle (joshi) | a vowel initial phrase
ga | *aru*, *wa* | *ame*, *sika* | *arimaseN*, *ni* | *iQte*, *te* | *ekizo*, *to* | *omou*, *wo* | *osiete*, *no* | *otaku*

The second type is an adjective (fukushi) | vowel initial phrase.

an adjectives (fukushi) | vowel initial phrase
mada | *atarasii*, *iQtai* | *itu*, *mosi* | *ikite*, *seQkaku* | *utouto*, *kitiNto* | *okonau*

The third less frequent type is a compound word (word | word), such as, *komugi* | *iro*, *takusii* | *ichidai*.

compound word
komugi | *iro*, *takusii* | *ichidai*

Similar phenomena are observable in the TIMIT.

examples in TIMIT
She | *is thinner than I am* (sx5: /iy | ih/). *Combine all the* | *ingredients in a large bowl* (sx118: /iy | ix/). *Where were you while we were* | *away?* (sx9: /axr | ax/)

2.2. Phonological realization of Japanese hiatus

There are a number of possible factors that help perception of Japanese V-V hiatus.

2.2.1. Phrase-initial glottalization

Glottalization of word-initial vowel is a common phenomenon of world languages [3]. It is more strongly pronounced if the word has a stress or accent at the beginning of the word.

2.2.2. Phrase-final nasalization

Voiced velar consonant is nasalized at the non-word-initial position in Tokyo Japanese. This nasalization contrasts with the following word-initial vowel that should not be nasalized. This sort of hiatus resolution occurs very often since noun phrases consisting of a noun + a particle *ga* are very common in Japanese, and such phrases can be followed by a predicate *aru* for example, composing a *a*|*a* hiatus.

2.2.3. Lengthening and shortening

Phrase-initial syllable or mora is shortened, while phrase-final syllable or mora is lengthened. This mora timing is a built in rhythm of Japanese as well as other languages. Duration of the concatenated vowel might be segmented with a built in

timer of the perception mechanism. The mechanism will help the human hearing to resolve the hiatus.

2.2.4. Morphonological constraints

Part of speech plays some role in realization of the hiatus. Vowel sequence at the phrase boundary often occurs in the environments stated in 2.1. Such constraints help to resolve the hiatus.

3. Prosody data base

Phonetic prosodic labeling is performed on voice data collected for Japanese prosody database.

3.1. Japanese MULTEXT prosody corpus [4]

The Japanese version of MULTEXT (multi-language prosody corpus) is created by the specification of EUROM1 [5]. It aims at recording same-content of speech consisting of 40 small paragraphs, then the extraction of prosody parameter, and the prosody notation of five languages.

Speakers are native speakers of the Tokyo dialect. A text is given for a reading and to evoke a simulated spontaneous utterance. Speech was recorded with apparatus based on the specifications of EUROM1, in an anechoic chamber, using a B&K 1/2 capacitor microphone, a DAT recorder (SONY PCM2300). In addition, electroglottograph is recorded with an EGG (KAY (Co.) 4338) from which F0 and open quotient are extracted.

3.2. Phonetic and prosodic labeling

Phoneme segmentation by hand-eye is good, but still is difficult to segment when the same two vowels connect. Those cases were conventionally marked at the mid point to achieve equality of morae duration [6].

J-ToBI labeling is applied for prosodic annotation according to the manual [2]. Although, the X-JToBI [7] extended the J-ToBI in spontaneities of speech, e.g. descriptions of fillers and disfluencies, it does not describe V-V hiatus. J-ToBI is sufficient for our prepared speech.

4. Method of hiatus analysis

The prosodic boundary of phrases was segmented with reference to the waveform (speech and EGG) and the spectrogram of wide-band and narrow-band, and then evaluated by listening to the separated accentual phrases.

4.1. Perceptual analysis of phrase

The hiatus we treat is a V-V boundary between adjacent accentual phrases in Japanese. Samples were taken from the Japanese MULTEXT prosodic corpus spoken by a female speaker fhk. The examined phrases consist of 45 phrases producing hiatus of /a/|/a/, /i/|/i/, /u/|/u/, /e/|/e/, /o/|/o/. There is no gap or transition between these two vowels.

4.1.1. Preparation of speech materials

In order to investigate deviations of V-V segment boundary, the following short speech waveforms are prepared. Referring to the hand labeled boundary as a fixed point, a front phrase and a rear phrase are separated and excised for speech materials in a perceptual experiment. The excising points are moved forward and backward from the fixed point with a step width of one vocal cord vibration period up to 5 periods (vertical lines in Figure 1 like pitch marks). As a result, it amounted to 11 speech sounds for each side, to a total of 22 speech sounds per hiatus.

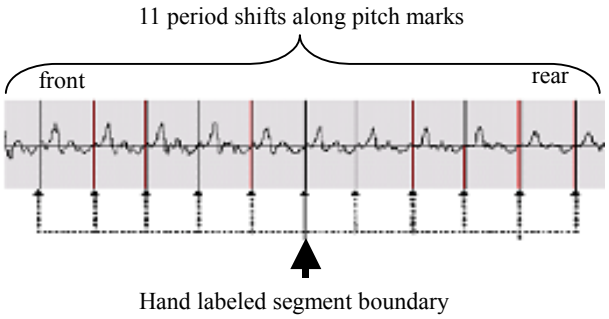


Figure 1: Pitch marks at zero-crossings with trimming for evaluation of perceptual sharpness of either cut. Trimming is done from the right side for front parts of the shift, and from the left side for rear parts

4.1.2. Phrase listening [6]

Speech sounds are presented in random order for each subject. Subjects were asked to judge the naturalness or the sharpness of each phrase sound, paying special attention to the ending and beginning. Responses were scored on a scale from 5 to 0, with 5 points awarded for natural clear-cut speech, and 0 for utterances appearing completely unnatural or contaminated with the adjacent component. Each answer is scored from +2, +1, 0, -1, -2 accordingly. Subjects' answers are summed and averaged for individual speech materials. The listeners participating in the perceptual experiments were 6 male students and 2 female students.

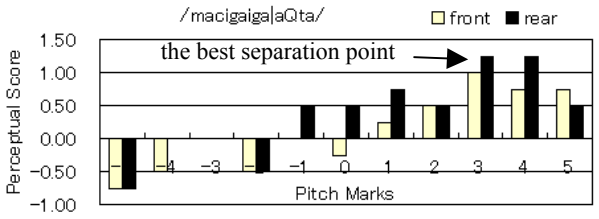


Figure 2: Phrase listening result for */macigaiga\aQta/* "There was some mistake." Perceptual score shows sharpness of either cut and separation of the front (white bar) and the back (black bar) phrases consisting with an */a|a/* hiatus.

4.2. Electroglottography waveform analysis

Electroglottography waveforms were analyzed for the open quotient (as shown in the bottom tier of the Figure 4) and the fundamental frequency (the middle tier in Figure 4) was computed from each glottal cycle using the KAY CSL tool [8]. The open quotient is related to voice quality, i.e., over 50% is harsh voice, 50% is modal voice, and 20-30% is breathy voice. The quotient changes smoothly along time, but abrupt change can be an evidence of glottalization. The fundamental frequency extracted from EGG is an instantaneous F0, i.e., an inverse of the pitch period, drops simultaneously with glottalization as well. This F0 differ from the conventional F0's that are smoothed by a filter.

5. Analysis results of vowel-vowel hiatus

Hiatus resolution is possible based on glottalization in most cases and nasalization in some cases. And the resultant lengthening of phrase-final vowel and shortening phrase-initial vowel is common.

5.1. Hiatus observed in phrasal tone

An intermediate phrase is composed as a chunking of several (only rarely more than three) accentual phrases in Japanese. An intermediate phrase boundary is often marked by a pause or *pseudo-pause*. Also, F0 declining characteristic of the intermediate phrase, however, is known as *catathesis*. We sometimes call this a *pitch-reset*.

Figure 3 shows an accent phrase *iQtai* "in the world" is emphasized, while *icunji* "when?" is focused and emphasized. A pitch-reset is observable in the succeeding phrase, and the pitch-range is extended. Perceptual results show that there is a clear hiatus at +2 period backward with scores 1.0 for the front phrase and 1.25 for the rear phrase. The F0 has began to increase already at the end of the former phrase /i/. A dip in open quotient indicates glottalization. Lengthening at the phrase ending as well as shortening at the phrase start up was not observed because the rear phrase was emphasized.

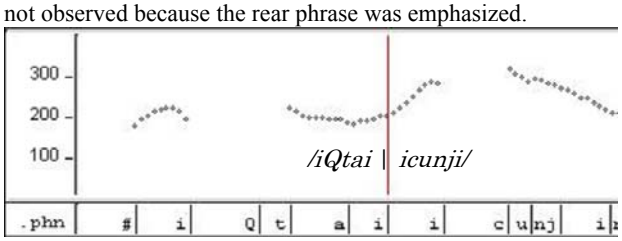


Figure 3: An example "Pitch Reset" from */iQtai | icunji/* "When in the world…?".

5.2. Phrase-initial glottalization

In most cases, the phrase-initial vowel is stressed on its phonation by glottalization. This is also true in cases where the preceding phrase-ending vowel is the same as the following phrase-initial vowel. An example is shown in Figure 4 below, showing that the EGG open quotient goes up once to 65% then decreases down to 56% and then goes up again to 61%. The most appropriate point to separate the phrases is before this bottom (8.907s.). Simultaneously, the F0, which is computed from the EGG period, showed lowering: 249 Hz to 195 Hz then 239 Hz movements.

Similar phenomena were observed in other cases in different degrees of prominence. Among 45 hiatuses analyzed, 17 items showed clear glottalization with dip in F0 and open quotient, and 23 items showed weak glottalization accompanied with other features, with the remaining 5 items showing phrase-final nasalization. If the phrase-initial word is emphasized, the boundary showed prominent glottalization, while a particle *ga* | *a*-initial phrase and a particle *wo* | *o*-initial phrase depict rather vague features of glottalization if the following phrase is not emphasized. In those cases, the acoustic features are too small in magnitude to detect visually. Although syntactic condition may help to resolve hiatus, probably human perception of glottalization must be far more sensitive than present signal processing.

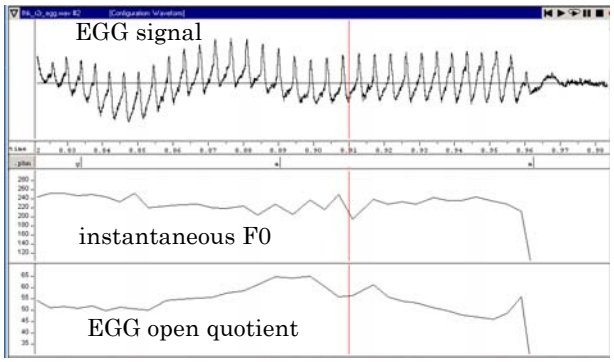


Figure 4: A boundary of a */a|a/* hiatus in */macigaiga | aQta/* "There was some mistake." Perceptual scores were shown in Figure 2.

5.3. Phrase-final nasalization

Switching from nasalization to non-nasalization may sign a perceptual cue, and indicates a segmental boundary by spectrogram texture as relatively lower high frequency energy for nasalized speech. The nasalization contrast is observable even in a continuation of the same vowel. A phrase-final particle *ga* has to be nasalized in Tokyo Japanese. Around the central part of Figure 5., a long vowel /*a*/ in the context of /*ga | aru*/ is shown with a vertical bar that is the best separation point between two phrases. The left part is nasalized (albeit weakly) and the right part is not nasalized. Although the speaker fhk, as a younger generation in the drift of phonological change, tends to pronounce these *ga* with non-nasals or at least weakly nasalized, spectrographic contrast is not clear.

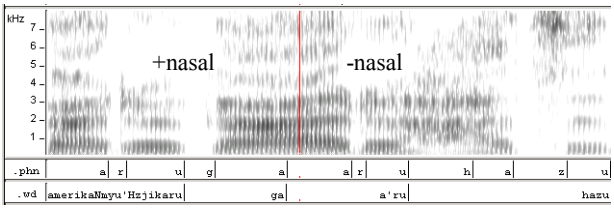


Figure 5: A contrast of *nasalized+/-* at the phrase boundary of */myuHzjkaruga-aruhazu/* "There must be a musical program".

Nasalization is helpful to detect hiatus since the preceding phrase-final nasalization ends and the following phrase starts without nasalization. This +/-nasal contrast as well as glottalization resolves many hiatus (in our example, +/-nasal contrast alone was 4 cases and 12 cases accompanied with glottalization as well as nasal contrast).

5.4. Segmental duration analysis

A phrase ending mora is lengthened to indicate the end of a phrase, while a phrase initiating mora is shortened in order to catch up with the isosyllabic mora timing. A V-V sequence of the same vowels has duration of about two morae, however, the boundary is usually searched in the right half region.

Statistics of our 45 hiatus, the ratio of duration of vowel segment in the preceding phrase-final position to the following phrase-initial position was 1.7 in average. In cases the following word is emphasized, the preceding vowel is not lengthened, but the following vowel keeps normal duration, then the ratio becomes as low as 0.76 for example.

Figure 7 plots 45 hiatuses examined along their measured durations of individual vowels in pairs of the frontal and the following. In reference to the diagonal line, which shows the equal duration, the figure suggests that the former (phrase-ending) vowel is usually longer than the following (phrase-initial) vowel. In a few cases, the following vowel is a little longer than the former vowel, and the former vowel is usually shortened.

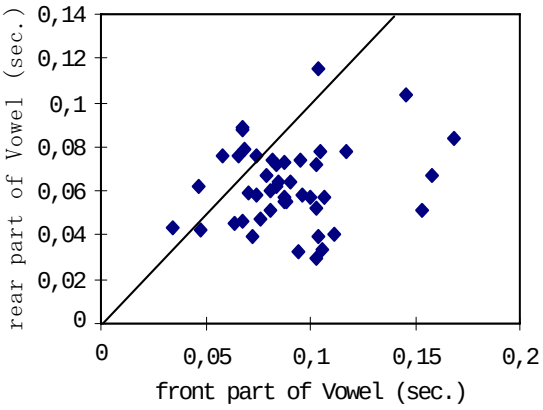


Figure 6: Durations of vowels in hiatus. Diagonal line, being the equal duration of connected vowels, suggests phrase-final lengthening and phrase-initial shortening.

6. Conclusion

J-ToBI labeled phrase boundaries are examined through perceptual evaluation of disjuncture, i.e. tidiness or flawless perfection. We investigated V-V hiatus by listening to whole phrases. The best perceptual score was obtained in most cases as the maximum perceptual score of a single peak.

A phrase-final particle | a vowel, the most common pattern of V-V hiatus, was found to have the following acoustic features: (1) *ga* | *a, nji* | *i, no* | *o*: these show +/-nasal contrast in the spectrographic pattern, since *ga* is normally nasalized while the following *a* is not nasalized. (2) *wa* | *a, sjika* | *a, te* | *e, to* | *o*: phrase initial vowel is glottalized. This glottalization is observable in F0 drop and a dip in EGG open quotient. (3) In *wo* | *o*:, another frequent pattern, glottalization is not so distinct since EGG open quotient is not stable, but spectral change is also useful.

A phrase-final adjective | a vowel and a word ending a vowel | a word beginning a vowel are cases characterized with stronger glottalization than the above-mentioned cases.

Phrase-initial glottalization observable in EGG open quotient, F0 or period of each cycle, and phrase ending nasalization are all important in resolving the hiatus phenomena.

Duration of vowels, constitutes hiatus, depend on mutual emphasis of the phrases adjacent, however, usually the former is longer than the follower.

The findings in this paper indicate that some small abrupt discontinuity in vocal source generator is sharply sensed by our auditory system to effectively segment phrases, words, and phonemes. Accordingly, speech synthesizers may need to take much more care in their smoothness and discontinuity of the artificial vocal source generator so as to cause effective phonological prosodic signs as well as to prevent unnecessary signs to confuse listeners.

7. Acknowledgements

This research is based on the domain research specific (B) subject number 12132204.

8. References

[1] Kawahara, Shigeto, 2003. On certain type of hiatus resolution in Japanese, *Phonological Studies*, 6, 11-20, ed. Phonological Society of Japan, Tokyo: Kaitakusha.

[2] Venditti, Jennifer J., 2002. The J-ToBI model of Japanese intonation. In S. - A. Jun (ed.) *Prosodic Typology and Transcription: A Unified Approach*. Oxford: Oxford University Press.

[3] Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M., 1996. Glottalization of word-initial vowels as a function of prosodic structure, Journal of Phonetics, 24, 423-444.

[4] Kitazawa Shigeyoshi, Kitamura Tatsuya, Mochiduki Kazuya, and Itoh Toshihiko, 2001. Preliminary Study of Japanese MULTEXT: a Prosodic Corpus. International Conference on Speech Processing, Taejon, Korea, 825-828.

[5] Campione, E., & Veronis, J., 1998. A multilingual prosodic database. 5th International Conference on Spoken Language Processing (ICSLP'98), Sidney, 3163-3166.

[6] Kitazawa Shigeyoshi, Kiriyama Shinya, Itoh Toshihiko, and Yukinori Toyama, 2004. Perceptual Inspection of V-V Juncture in Japanese, SP2004, 349-352.

[7] Maekawa, K., Kikuchi, H., and Igarashi, Y., 2001. "X-JToBI: An Intonation Labeling Scheme for Spontaneous Japanese", Technical Report of IEICE, SP 2001-106, 25-30. (in Japanese)

[8] Instruction Manual Electroglottograph (EGG) Model 4338, Kay Elemetrics Corp., Lincoln Park, NJ 07035-1488 USA (April 1995).

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 117-121.

Important and New Features with Analysis for Disfluency Interruption Point (IP) Detection in Spontaneous Mandarin Speech

Che-Kuang Lin, Shu-Chuan Tseng, Lin-Shan Lee*

Graduate Institute of Communication Engineering
National Taiwan University, Taipei, Taiwan
Institute of Linguistics, Academia Sinica, Taipei, Taiwan*

Abstract

This paper presents a whole set of new features, some duration-related and some pitch-related, to be used in disfluency interruption point (IP) detection for spontaneous Mandarin speech, considering the special linguistic characteristics of Mandarin Chinese. Decision tree is incorporated into the maximum entropy model to perform the IP detection. By examining performance degradation when each specific feature was missing from the whole set, the most important features for IP detection for each disfluency type were analyzed in detail. The experiments were conducted on the Mandarin Conversational Dialogue Corpus (MCDC) developed by the Institute of Linguistics of Academia Sinica in Taiwan.

1. Introduction

Most speech recognition systems can successfully process well-formed and well-spoken utterances. However, for ill-formed utterances frequently appearing in spontaneous conversation, properly modeling the ill-formness is a very important but still very difficult problem. One of the primary sources of ill-formness is the presence of disfluencies. Accurate identification of various types of disfluencies and properly utilizing such messages can not only improve the recognition performance, but provide structural information about the utterances.

The structure of disfluencies is usually considered to be decomposed into three regions: the reparandum, an optional editing term, and the resumption. The disfluency interruption point is the right edge of the reparandum. The purpose of the research presented in this paper is to identify useful and important features in automatic detection of such disfluency interruption point (IP) in spontaneous Mandarin speech, and analyze how these features are helpful. The disfluencies considered here in this paper include the following four categories:

(1) Direct repetitions: the speaker repeats words in a way that can not be justified by grammatical rules. Many other cases of repetitions in Mandarin Chinese are perfectly legal syntactic constructions for emphasis purposes and so on, which should be excluded from this study.
(2) Partial repetitions: only part of a word (including compound words) is repeated.
(3) Overt repairs: the speaker modifies expressed words within utterance.
(4) Abandoned utterances: a speaker abandons an utterance and starts over.

Consider the following example (overt repairs):

shi4 jin4kou3 EN chu1kou3 ma1?
is import [discourse particle] export [interrogative particle]
*Do you import * uhn export products?*

In this example, "uhn" is a filled pause and "export" is meant to correct "import", which is an overt repair. Here '*' denotes the right edge of the reparandum region, or the interruption point (IP) to be detected and analyzed here.

Consider another example (direct repetition):

yin1wei4 yin1wei4 ta1 you3 jian4shen1 zhong1xin1
because because it has fitness center
*Because * because it has a fitness center.*

Here the speaker repeats the word "because" to restart the sentence.

It has been suggested much earlier [2] that there exists a certain acoustic "edit signal" serving as a cue indicating that fluent speech had been interrupted. Although it may be difficult to find a single cue for such purposes, several prior studies had indicated that combinations of more cues can be used to identify disfluencies with reasonable success [3,9,10].

In this paper, a whole set of acoustic-prosodic features were considered and disfluency interruption point (IP) detection was tested and analyzed. Two types of features were considered here, duration-related and pitch-related. Detailed analysis regarding which features are the most important for which types of disfluencies and possible reasons are then discussed. Below the corpus used in this research is first introduced in section 2, while the acoustic-prosodic features investigated are summarized in section 3. Section 4 briefly describes the two approaches used for the disfluency interruption point (IP) detection. Section 5 finally presents the experimental results and relevant analysis.

2. Corpus Used in the Research

The corpus used in this research was taken from the Mandarin Conversational Dialogue Corpus (MCDC) [13], collected from 2000 to 2001 by the Institute of Linguistics of Academia Sinica in Taipei, Taiwan. Several studies have been conducted on MCDC to analyze various phenomena of spontaneous conversational Mandarin speech [13,14,15]. This corpus includes 30 digitized conversational dialogues with a total length of 27 hours. 8 dialogues out of the 30, with a total length of 8 hrs, produced by nine female and seven male speakers, were annotated by adopting a taxonomy scheme of four groups of spontaneous speech phenomena: disfluency, sociolinguistic phenomena, particular vocalization, and unintelligible or non-speech sounds. Disfluencies here include breaks, word fragment, overt repairs, direct repetitions, abandoned utterances, discourse particles, and markers. In this paper, we only deal with direct repetitions, partial repetitions, overt repairs and abandoned utterances. The 8 hrs of annotated dialogues as mentioned above were used in this research. Due to the mono-syllabic structure of Chinese language, i.e., in Mandarin Chinese every character has its own meaning and is pronounced as a monosyllable, while a word is composed of one to several characters (or syllables), every syllable

boundary is considered as a possible interruption point (IP) candidate in this research. Table 1 summarizes the data used in the following experiments. As can be found, 96.3% and 96.4% of the syllable boundaries are non-IPs. The total number of IPs is limited in the annotated corpus, which makes the studies and analysis slightly difficult.

Table 1: The summary of experiment data.

	train	test
Data length	7.1hr	1.1hr
Number of non-IPs	92189	14231
Number of IPs	3569	536
Chance of non-IPs	96.3%	96.4%

3. Prosodic Features

We tried to define a whole set of acoustic-prosodic features for each IP candidate, or each syllable boundary, and use them to detect the IPs. Many prosodic features have been proposed and proved useful for such purposes [5,11], and it has been found [4] that it is important to identify better features. Because this research is focused on IP detection, we tried to identify some IP specific features. Moreover, considering the special feature of Mandarin Chinese, including the mono-syllabic structure as mentioned above and the tonal language nature, some acoustic phenomena for Mandarin spontaneous speech may be quite different from those in English. Such consideration were reflected here by constructing a new set of features.

3.1. Pitch-related Features

Pitch information is typically less robust and more difficult to use [11], partly due to the variabilities of pitch values across speakers and speaking contexts, partly due to serious pitch tracking errors. A pitch contour stylization method has thus been used and smoothing out the "micro-intonation" and tracking errors has been found helpful for English [5,11]. For a tonal language such as Mandarin Chinese, however, such "micro-intonation" apparently carries tone or lexical information, and thus should not be removed, although some approaches of pitch contour smoothing are still certainly needed. Syllable-wise pitch contour smoothing and Principal Component Analysis (PCA) have both been shown to be helpful in identifying key characteristics in the pitch contours and performing the tone recognition in Mandarin Chinese [8,12].

We used PCA for syllable-wise pitch contour smoothing, instead of piece-wise linear stylization. For each syllable, the pitch contour was decimated or interpolated to become a vector with fixed dimension. PCA was then performed on such training vectors. By choosing the principal components with the largest eigenvalues, we projected the fixed dimension vectors onto the subspace spanned by the principal components to obtain the smoothed version of the pitch contours. Various pitch-related features were then extracted from these smoothed pitch contours, such as the pitch reset for boundaries being considered, and so on. Several syllable-wise pitch-related features found useful in tone recognition [8] were also used here, such as the average value of normalized pitch within the syllable, the average of absolute value of pitch variation within the syllable, the maximum difference of normalized pitch within the syllable and so on, all evaluated for the syllable before and after the boundary being considered.

3.2. Duration-related Features

Duration features such as pause and phone duration features have been used to describe prosodic continuity and preboundary lengthening [5,11]. By carefully examining the characteristics of IPs in our corpus, we hypothesized that deviation from normal speaking rhythmic structure is an important cue to disfluency IP detection. For example, relatively sudden, sharp, discontinuous changes in speaking rate were consistently observed across IPs. We also hypothesized that certain ways of integration of pause and syllable duration fluctuation are important characteristics of the rhythmic structure of speech. Considering these observations, we derived the following set of duration-related features to try to detect IPs.

We first computed the average and standard deviation of syllable duration over several syllables before and after the boundary being considered. Then we calculated the ratio of the former to the latter. The possible ranges for evaluating the above statistics included one, two, three syllables as well as extending to the nearest pauses on both sides. Other groups of duration-related features were generated by jointly considering the pause duration and the duration parameters of the syllables before or after the pause. The product of these two different duration parameters represented some integration of the two types of information. Alternatively, normalizing the syllable duration parameters by the duration of a nearby pause being considered may emphasize the fluctuations of these syllable duration parameters. Finally, a total of 38 such duration-related features were considered.

4. IP Detection

We use the acoustic-prosodic features mentioned above to detect IP events given all syllable boundaries after the first pass recognition giving all the recognized syllables. Two analytical approaches were used in the IP detection, the decision tree and the maximum entropy model. The task here is simplified as a two-class classification problem. For each syllable boundary, a decision of "non-IP" vs. "IP" is made. Due to the very limited number of IPs of different disfluency types in the available annotated corpus, they were grouped together as a single class of "IP", and all other boundaries then belong to the class of "non-IP". Because IPs are relatively rare events, the approach of ensemble sampling previously proposed [6] was used on the training data to equate the prior probabilities for the two different classes. This made the approaches more sensitive to any inherent prosodic features that can distinguish the classes. In the first set of experiments, we used decision trees [5] to learn from the data, to identify the useful features, decide how to use them and finally use them to detect the IPs. The decision was made based on the posterior probabilities for the leave nodes where the syllable boundary being considered went to. In the second approach of the maximum entropy model [1], on the other hand, each feature is expressed by a binary feature function, and the model tried to find the appropriate parameters for each feature function with the constraint that the expected values of the various feature functions match the empirical averages in the training data. Some improved approaches were then developed trying to integrate the nice properties of the two different approaches, in which the threshold values for the various features learned from the decision tree were used as the quantization thresholds in defining the feature functions used in the maximum entropy model. This approach turned out to be better than the original maximum entropy model. The

results reported below for maximum entropy model were obtained by this integrated approach.

5. Experiment Results & Further Analysis

5.1. IP Detection Results

The IP detection results in terms of recall and precision rates using decision tree and maximum entropy model are listed in Table 2. While decision tree achieves moderate and balanced recall and precision rates, maximum entropy model trades degraded recall for significantly better precision. As far as performance of a recognition system is concerned, a false alarm is usually more harmful than an omission. In other words, for the purposes here it is preferred to achieve as high precision as possible while having a high enough recall rate. As a result, maximum entropy model may be more appropriate for the purpose here.

Table 2: Recall and precision rates for IP detection with decision tree and maximum entropy model.

	Recall	Precision
Decision Tree	73.15	73.03
Maximum Entropy	56.38	81.95

5.2. Duration- and Pitch- Related Features for Different Disfluency Types

To get a further insight into the characteristics of various disfluency categories and the IP detection process, we tried to find the relation between the features used and the IP detection performance. A partial feature selection analysis was performed upon the full feature set mentioned earlier. In this approach, we excluded each single feature from the full set and then perform the complete IP detection process in each small experiment, to find out how much the IP detection performance was degraded due to the missing of this single feature. Here the performance is in terms of recall rate only. Because we grouped all the four types of disfluencies together into a single class, precision for each disfluency type was not obtainable, while recall was. Only the results of maximum entropy model were discussed here due to space limitation, although almost the same trend was found in the decision tree approach.

First, to see how pitch-related and duration-related features contribute to the IP detection of different types of disfluencies, we compared the performance degradation for the four disfluency types being considered with respect to the two feature categories. In Figure 1, we show the most serious performance degradation caused by removing one single feature from the two categories of either pitch-related or duration-related features. We find that for overt repair and partial repetition, pitch-related features play relatively more important role for IP detection, and this is especially apparent for overt repair. This is in good consistency with the earlier findings [16] that overt repairs are produced partly because the correction of the delivered information is required, and partly because the speaker changes his/her language planning. It is often true that when overt repairs are produced within utterances, the F0 level of the onset of the resumption part is approximately reset to that of the onset of the reparandum. In other words, the resumption part should fit seamlessly into the original utterance after removing the problematic items. Then the cleaned utterance should look like a natural utterance that obeys the normal F0 declination.

In addition, intonation units have been defined and analyzed in Mandarin conversation [17], which are unique characteristics in spoken language different from syntactic units. They are also found to be highly related to the language planning process. Moreover, it has been observed [16] that almost all reparandum parts are themselves intonation units. The behavior of overt repair is just similar to that of a new intonation unit with respect to the preceding one. All these imply that overt repairs have a lot to do with the intonation units and thus pitch-related features. All these are consistent with the results here, i.e., the cues carried by pitch-related features provide important information for overt repair detection.

On the other hand, we also find that for direct repetition IP detection, the duration-related features are more important, and for abandoned utterances IP detection, both pitch-related and duration-related features have equally important impact.

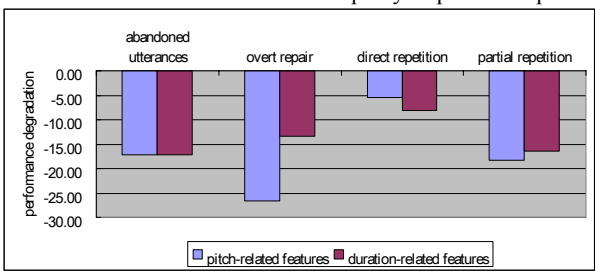


Figure 1: Performance degradation (recall degradation) for the four disfluency types with respect to the two feature categories.

5.3. Pitch-related Features for IP Detection of Different Disfluency Types

The 13 features found to be the most important in IP detection for the four different types of disfluencies are represented by symbols (a) to (m) with their definitions as listed in Table 3, where the upper and lower halves are for pitch-related and duration-related features respectively.

In Table 4, for each of the four categories of disfluencies, we list the symbols for the two pitch-related features causing the most serious recall rate degradation, or the two most important pitch-related features, together with the associated recall rate degradation, in the columns labeled as "pitch-related". We can see that the average pitch value within a syllable, used in features represented by symbols (b) and (d) in Table 3, appears to be very important in three out of the four types of disfluencies, regardless of different smoothing methods used. This suggests that the level of pitch is a very good cue for disfluency IP detection, probably due to the tone information carried and the intonation unit property as mentioned earlier. In particular, the absence of this feature degrades the performance very severely on partial repetitions and abandoned utterances. Direct repetition, on the other hand, is much less influenced. Moreover, the difference of maximum and minimum pitch values within a syllable, used in features represented by (e) and (f) in Table 3, is beneficial to IP detection of direct repetitions and partial repetitions.

It has been found [16] that as far as Mandarin Chinese is concerned, the overt repairs, direct repetitions, and partial repetitions tend to be shorter. The main reason is probably that in Mandarin Chinese there is no inflection and the word order can vary to a great extent, speakers can re-initiate at the morphological boundary immediately after some inappropriateness is sensed. Moreover, it was also found that simple direct repetition repeating only one syllable usually dominates [16]. With many of such mono-syllable repeats, the pair of (partially) repeated and re-initiated syllables very often exhibit highly similar pitch contours. With the tone information inside these contours, pitch level (features (b) and (d)) and range (features (e) and (f)) can thus capture the evidence of short direct repetition and partial repetition.

Another important pitch-related feature in Table 4 is the difference of pitch value across boundaries (used in the feature represented by (a)). This feature somehow conveys to what degree the speaker resets the pitch at this boundary. The reset of pitch is often the evidence of starting a new intonation unit, which is probably also the beginning of a new planning unit. This may be the reason why this feature is very important in the detection of abandoned utterances and overt repair IPs.

Table 3: The definitions of features used in Table 4. Note that for certain parameter z evaluated for each syllable boundary, Δ(z) is the difference of the parameter z for two neighboring syllable boundaries.

	Feature ID	Definition
Pitch-related features	(a)	Δ(difference of pitch slope across boundary)
	(b)	Δ(average pitch value within a syllable), with pitch value obtained from raw f0 value
	(c)	averaged absolute value of pitch slope within a syllable, with pitch value obtained from linear approximation
	(d)	Δ(average pitch within a syllable), with pitch value obtained from PCA
	(e)	Δ(difference of maximum and minimum pitch value within a syllable), with pitch value obtained from raw f0 value
	(f)	Δ(difference of maximum and minimum pitch value within a syllable), with pitch value obtained from linear approximation
Duration-related features	(g)	Δ(ratio of the duration for the syllable before the boundary to the pause duration at the boundary)
	(h)	ratio of the duration for the syllable after the boundary to the pause duration at the boundary
	(i)	product of the duration for the syllable after the boundary with the pause duration at the boundary
	(j)	Δ(product of the duration for the syllable after the boundary with the pause duration at the boundary)
	(k)	Δ(ratio of the duration for the syllable after the boundary to the pause duration at the boundary)
	(l)	syllable duration parameter ratio across the boundary, with the duration parameter being the average over 3 neighboring syllables
	(m)	standard deviation of (product of the duration for the syllable before the boundary with the pause duration at the boundary)

Table 4: The recall rate degradation when excluding an pitch-related/duration-related feature for different types of disfluencies. (with definitions of features listed in Table 3).

Disfluency Types	Most Important Features (recall degradation)		Second Important Features (recall degradation)	
	pitch-related	duration-related	pitch-related	duration-related
abandoned utterances	(a) (-17.25)	(g) (-17.25)	(b) (-14.97)	(h) (-14.97)
overt repairs	(c) (-26.67)	(i) (-13.33)	(a) (-20.00)	(j) (-13.33)
direct repetition	(d) (-5.40)	(k) (-8.10)	(e) (-5.40)	(l) (-8.10)
partial repetition	(b) (-18.21)	(h) (-16.33)	(f) (-18.21)	(m) (-16.33)

5.4. Duration-related Features for IP Detection of Different Disfluency Types

Table 4 also showed similar analysis with respect to duration-related features, in which we list the two most important duration-related features, together with the associated recall rate degradation, for the four types of disfluencies, in the columns labeled as "duration-related". Although duration-related features are beneficial to direct repetition detection as mentioned above, they also help indicate IP of other types of disfluencies. First, jointly considering both the syllable duration and pause duration was shown to be useful across all kinds of disfluencies. Combining through ratio of syllable duration to pause duration (represented by (g), (h) and (k) in Table 3) is relevant to IP detection of abandoned utterances, direct and partial repetitions, while overt repairs and partial repetition benefit from the product of them (represented by (i), (j) and (m) in Table 3). The ratios may have normalized the syllable duration with respect to the breathing tempo of the speaker, if any, which was revealed by the pause duration fluctuation. The results showed that such features are actually useful.

Moreover, a specific feature for direct repetition is the character duration ratio across boundary (represented by (l)), implying how the speaking rate was fluctuating. This showed that direct repetitions usually cause significant speaking rate deviation, and this is consistent with the observation obtained before [16], in which it was concluded that the repeated words in the resumption are shorter than those in the reparandum part, because the direct repetition itself often provides no new information. Partial repetitions also exhibit similar properties to those of direct repetition. The contribution of standard deviation (represented by (m)) to partial repetition may thus be also due to the duration fluctuation related to partial repetition. Although the effect of standard deviation (feature represented by (m)) on direct repetition is not shown on Table 4, it indeed stands right behind (being the third important, not shown in the table), which supports the above argument.

6. Discussion

A whole set of features to be used for disfluency IP detection is developed, tested and analyzed. The most important features for each disfluency types were identified and discussed considering the linguistic characteristics of the disfluencies. The false alarms obtained in the detection output remains to be a major problem for further applications. One possible approach toward this direction is probably to adopt some kind of disfluency type-specific rule-based methods to further discriminate the false alarms.

7. References

[1] Berger, A. L., Della Pietra, S. A., & Della Pietra, V. J. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39-72.

[2] Hindle, D. 1983. Deterministic Parsing of Syntactic Nonfluencies. *Proc. ACL'83*, pp.123-128.

[3] Lickley, R. J. 1996. Juncture Cues to Disfluency. *Proc. ICSLP'96.*

[4] Liu, Y., et al. 2005. Structural Metadata Research in the EARS Program. *Invited paper. Proc.ICASSP'05.*

[5] Liu, Y., Shriberg, E., & Stolcke, A. 2003. Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources. *Proc. Eurospeech'03*, pp. 957-960.

[6] Liu, Y., Shriberg, E., Stolcke, A., & Harper, M. 2004. Using Machine Learning to Cope with Imbalanced

Classes in Natural Speech: Evidence From Sentence Boundary and Disfluency Detection. *Proc. of ICSLP'04.*

[7] Lin, C.-K. & Lee, L.-S. Improved Spontaneous Mandarin Speech Recognition by Disfluency Interruption Point (IP) Detection Using Prosodic Features. *Proc. Eurospeech'05.* (to appear).

[8] Lin, W.-Y. & Lee, L.-S. Improved Tone Recognition for Fluent Mandarin Speech Based on New Inter-syllabic Features and Robust Pitch Extraction, *Proc. ASRU'03*.

[9] Nakatani, C. & Hirschberg, J. 1994. A Corpus-based Study of Repair Cues in Spontaneous Speech. *JASA*, pp.1603-1616, 1994.

[10] Shriberg, E. 1999. Phonetic Consequences of Speech Disfluency. *Proc. ICPhS'99*, pp. 619-622.

[11] Shriberg, E., et al. 2000. Prosody-based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication*, pp. 127-154, 2000.

[12] Tian, J. & Nurminen, J. 2004. On Analysis of Eigenpitch in Mandarin Chinese. *Proc. ISCSLP'04*.

[13] Tseng, S.-C. 2004. Processing Spoken Mandarin Corpora. *Traitement automatique des langues*. Special Issue: Spoken Corpus Processing. 45(2): 89-108.

[14] Tseng, S.-C. 2003. Repairs and Repetitions in Spontaneous Mandarin. In *Proceedings of Workshop on Disfluency in Spontaneous Speech (DISS 03)*. Ed. Robert Eklund. Gothenburg Papers in Theoretical Linguistics 90. pp. 71-74. University of Gothenburg.

[15] Tseng, S.-C. 2005. Syllable Contractions in a Mandarin Conversational Dialogue Corpus. *International Journal of Corpus Linguistics*. 10(1): 63-83.

[16] Tseng, S.-C. Repairs in Mandarin Conversation. *Journal of Chinese Linguistics.* (to appear).

[17] Tao, H.-Y. 1996. *Units in Mandarin Conversation*. Prosody, Discourse, and Grammar. John Benjamins Publishing Company.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 123-126.

Influence of manipulation of short silent pause duration on speech fluency

Tobias Lövgren & Jan van Doorn

Umeå University, Umeå, Sweden

Abstract

Ordinary speech contains disfluencies in the form of hesitations and repairs. When listeners make global judgements on speech fluency they are influenced by the frequency and nature of the individual disfluencies contained in the speech. The aim of this study was to investigate a single dimension, pause duration, in the perception of speech fluency. The method involved simulation of pause duration within naturally fluent speech by manipulating existing acoustic silences in the speech. Four conditions were created: one for the natural speech and three with step wise increases in acoustic silence durations (average x2, x4 and x7.5 respectively). In a forced choice task listeners were asked to judge the speech samples as fluent or non fluent. The results showed that the percentage of judgements of disfluency increased as the pause durations increased, and that the difference between the unmanipulated speech condition and the two conditions with the longest pause durations were statistically significant. The results were interpreted to indicate that the individual dimension of pause duration has an independent influence on the judgement of fluency in ordinary speech.

1. Introduction

Speech fluency is generally characterised by the presence or absence of disruptions to speech flow [8]. Studies of naturally disfluent speech have identified the types of disfluency that can be found in normal speech [7] and have reported their frequency of occurrence [12]. Those speech disruptions which are generally considered to influence speech fluency include word or part word repetition, phrase repetition or revision, prolongation of sounds, filled pauses (uh, um etc) and silent pauses [7].

Various methodologies have been used to study the influence of individual factors on speech fluency. In particular, digital simulation of speech features opens possibilities for investigation into how factors that have been identified as features of disfluent speech affect listener perception. It enables independent manipulation of factors that could not otherwise be controlled in natural speech. Recent studies have taken advantage of accessible high quality speech technology to simulate disfluent speech. For instance, Amir and Yairi [1] manipulated vowel durations and between-word pauses in part word and full word repetitions in order to investigate features that distinguish stuttered speech from normally disfluent speech.

While the presence of pauses has been identified as a contributor to the perception of disfluent speech, only a subset of all pauses found in natural speech influence the perception of speech fluency. Those pauses which perceptually disrupt the smooth flow of speech, sometimes referred to as hesitation pauses [11] frequently occur at non syntactic boundaries, and have been referred to by Duez [4, 5] as within-constituent pauses i.e. pauses occurring between strongly connected elements. Hesitation pauses can be signaled by a filler such as 'um' (filled pauses), a period of silence (silent pauses) or prepausal lengthening [13].

For silent pauses a duration of about 200 ms silence has been considered as a threshold before the silence is consistently perceived as a pause [17]. Thus, in investigations on pausing it is common to eliminate periods of silence less than 200-300 ms [9]. However, it should be noted that shorter silent periods have also been found to be associated with a perceived pause. For instance intervals as short as 60 ms could be associated with perceived pauses [2], while it has also been reported that in certain acoustic environments (such as prepausal lengthening) a pause could be perceived in the absence of any silent period [6, 13]

Martin & Strange [10] reported that listeners found hesitation pauses (in the form of silent pauses) to be less salient than syntactic pauses, in that they were frequently disregarded in tasks where listeners were specifically asked to identify pauses. However, Duez found that such pauses were not entirely disregarded by listeners because the identification rate for silent within-constituent pauses was higher when the pause duration was longer [5].

Studies on speech fluency have looked at some individual contributing factors to the perception of fluency, but have not specifically looked at pause characteristics. On the other hand studies on pauses have looked at the influence of pause characteristics on the perception of the pauses themselves, but not on the perception of fluency. The purpose of this study was to investigate the link between duration of silent hesitation pauses and the perception of speech fluency.

Specifically, the aim of this study was to investigate the influence of artificial manipulation of the duration of silent intervals that occurred at non syntactic boundaries in ordinary speech on the global perception of speech fluency. The study manipulated duration of existing acoustic speech silences in natural speech. Naïve listeners judged samples of original and manipulated speech as fluent or disfluent in a forced choice task.

2. Method

2.1. Source speech material

The speech material for this experiment was required to have duration of sufficient length for listeners to obtain a global impression of fluency (15s according to Dalton & Hardcastle [3]), and needed to have disfluencies only in the form of natural silent pauses imbedded in it. The planned method involved manipulating the duration of natural silent pauses within the speech rather than the insertion of pauses in order to minimise disruption of other prosodic features in the vicinity of the pause.

Material from professional newsreaders met the specific criteria for the purposes of this study. It has been established that professional newsreading can be considered highly fluent speech, where natural pauses are used at syntactic boundaries for linguistic emphasis [14]. Acoustic silences at non syntactic

boundaries e.g. those that formed occlusions in unvoiced word or syllable initial plosives were present in the material to enable simulation of pauses within syntactical units. It was also possible to select material that did not contain other types of disfluency.

Speech material from 10 news programs from Swedish national television was recorded using VHS video recording (hi-fi audio quality with a dynamic range 20 Hz-20 kHz) and reviewed for selection of suitable excerpts. The criteria for selection were that the speech must come from presenters who were in the recording studio and excluded material that could be offensive to the listeners in the experiment (e.g. reports of violent crime).

From this material four speech excerpts were selected. They were each 15-20 s long, read by four different professional newsreaders (two male, two female), and each containing three or four clear silent periods at occlusions of word or syllable initial unvoiced plosives or natural non syntactic silent pauses. The silent periods were spaced at a frequency within the range found for normal disfluencies in ordinary speech [12]. A description of the original speech material, giving number of pauses and distribution of their duration for each sample is shown in Table 1.

Table 1: Description of the original speech excerpts selected for silent pause manipulation.

Sample	# words	# silent pauses	Duration of silent pauses
1	37	4	27,31,51,110 ms
2	42	3	41,53,53 ms
3	43	4	40,46,48,102 ms
4	36	3	44,73,73 ms

2.2. Manipulation of pauses

The audio signals from the VHS video recordings of the selected speech excerpts were recorded directly into the speech analysis program Praat version 4.2. Praat was selected as the program with the most suitable function for duration manipulation.

The boundaries of each acoustic silence were carefully identified within Praat, and in the first instance each silence was stretched x2.5, x5, and x10 respectively to produce three conditions with step wise increase in pause duration. Using the stretch function in Praat allowed the exact acoustic environment of the silence to be preserved, and thus avoided any acoustic disjunction that could be perceived as artificial by listeners. A multiplicative increase in silence duration was chosen in an attempt to maintain the relativity of the pause lengths, and thus improve the naturalness of the speech samples that contained simulated silences.

The simulated material was then reviewed by two independent listeners who were asked to comment on the naturalness of each sample. For the speech samples that were judged as unnatural, those acoustic silences that were perceived as artificial were reduced in absolute duration, yet still maintained step-wise increases from one condition to the next. The revised material was re-judged for naturalness by a further two independent listeners. This process was repeated until there was consensus that the speech did not sound artificial. This process produced a spread of multiplicative factors for each pause-lengthened condition, but for each individual silent pause there remained a step-wise increase from one condition to the next. The actual range of pause durations and corresponding multiplicative factors for each condition can be seen in Table 2. The speech material (in Swedish, with translation to English) showing the location of each manipulated pause, along with the actual pause durations in each condition can be seen in Appendix 1.

Table 2: Description of manipulated pauses

Condition	Mean & range of pause duration (ms)	Mean & range of % increase
C1	57 (27 – 110)	-
C2	117 (65 – 237)	214 (120 - 298)
C3	212 (98 – 479)	394 (149 - 633)
C4	403 (184 – 975)	744 (300 - 1263)

2.3. Listening speech material

The eventual listening material was prepared from the 16 samples (4 samples x 4 silent pause conditions per sample). Specific sequences of these samples were prepared for presentation to the listeners. A Latin squares construction was selected so that any one listener would hear four utterances: one utterance from each speaker in one of the four conditions. Thus 16 different listening sequences were prepared, each consisting of four excerpts. Each sequence was placed on a single track on a listening CD, with a 3 second gap between each of the four samples to allow time for listeners to record their judgments.

2.4. Listeners

Thirty two listeners were used, 14 males and 18 females aged between 20 and 32 years. All were students of medical and allied health programs at Umeå University in Sweden. The listeners had no previous experience in assessing speech, had not studied phonetics, and had no personal experience of stuttering in close friends or family.

2.5. Procedure

Each listener completed a simple questionnaire to establish basic listener information (age, gender, university education). They were then given a verbal description of the concept of speech fluency, and instructions regarding the required task - a forced choice judgement on whether they considered each speech sample to be fluent or not. They were then presented with two practice examples to familiarise them with the nature of the task prior to the commencement of the task. Each person listened to the four samples on a specific track on the CD, using a portable CD player and a set of AKG K130 headphones. All listeners listened from the same portable CD player, using the same headphones, and in the same listening environment. The listeners listened once to each sample and wrote down their spontaneous judgement (fluent/non fluent) at the end of each sample. At the completion of the task the listeners were then asked if they had a close association (family or friends) with stuttering. No listeners reported such association.

2.6. Data analysis

For any one speech sample there were two listener judgements giving eight judgements per pause condition. The total number of non-fluent judgements was counted across speech samples for each condition. A non parametric Kruskall-Wallis test was used to look for overall statistical significance of the pause condition on the number of disfluent judgements. In the case of overall statistical significance from the Kruskall-Wallis test, post hoc testing between individual pause conditions was then planned using Mann-Whitney U tests with Bonferroni correction.

3. Results

3.1. Disfluency judgements

The percentage of judgements of disfluent speech for each condition can be seen in Table 3. There is a steady increase in the percentage of disfluent judgements from the condition C1 (unmanipulated) to Condition C4 (longest durations).

Table 3: Percentage of judgements of disfluent speech for each pause duration condition.

Condition	Mean duration of silent periods	% of disfluent judgements
C1	57 ms	12.5
C2	117 ms	34.4
C3	212 ms	71.9
C4	403 ms	93.8

A Kruskall-Wallis test showed that there was an overall significance effect of pause duration condition $H(3) = 12.23$, $p<0.01$. Post hoc testing with Mann-Whitney U tests using BonFerroni correction for multiple testing showed that the differences were significant between Condition C1 (no manipulation) and Conditions C3 and C4 respectively. There was also significant difference in the number of judgments of disfluency between Conditions C2 and C4

The percentage of samples that were judged disfluent within each condition as a function of the mean duration of the silent interval in each condition can be seen in Figure 1.

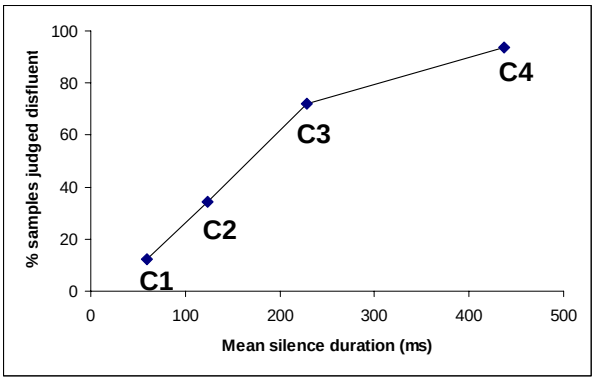


Figure 1: Percentage of speech samples judged disfluent vs. mean duration of silent interval for each of four pause conditions.

The statistically significant increase in the percentage of samples judged to be disfluent in Condition C3 (71.9%) compared with Condition C1 (12.5%) corresponds to an increase in the mean silence duration from 57 to 212 ms.

4. Discussion

In a forced choice judgement of global speech fluency the percentage of disfluent judgements increased when silent pause duration was artificially increased in increments from the original speech condition C1 (with pause durations from 27–110 ms). This increase reached significance for conditions C3 (pause durations 98–479 ms) and C4 (184– 975 ms).

Interpretation of these results must be made within the experimental limitations of the study. Using a design of forced-choice categorical judgements by naïve listeners presents difficulties for those samples where listeners are equivocal about categorising in a dichotomous task. For fluency judgements listeners have been shown to be able to use a continuum scale [1]. However, for naïve listeners the use of a forced choice of presence or absence of a global speech feature is a simple, effective and reliable task in experimental conditions. The design has been able to detect a clear distinction between each condition, and demonstrate a steady increase in the percentage of disfluent judgements from Conditions C1 to C4.

The changes in fluency judgement that occurred in the present study can be attributed to the artificial elongation of existing pauses in the naturally fluent speech of professional newsreaders. It would have been easier to generalise the result if spontaneous speech had been used. However, in order to achieve the aim of isolating the influence of pause duration on speech fluency, it was essential to use material that did not contain any other disfluencies besides silent pauses. The use of newsreading material rather than spontaneous speech needs to be considered in the interpretation of the results. The material was clearly recognizable as excerpts from mass media news, and spoken by well known media personalities. It is thus likely that listeners expected high levels of fluency from these professional readers, and would have judged any hesitations more harshly than they would have if it were another reading situation, and probably also for hesitations in spontaneous speech.

Using artificial manipulation means that it is not possible to be certain that the fluency judgements were not influenced by an artifact of such manipulation. There are two possible sources of artifact - either that the actual process of electronic manipulation of the pauses themselves, or that those manipulations introduced an unnatural distribution of pause durations. Experimentally both those factors were addressed. First, the increase in pause duration was carried out using a stretching function which avoided introducing any electronic boundaries into the acoustic signal. Secondly a multiplicative increase was employed to preserve the relativity of the distribution of pause durations in the samples. Finally, the preparation of the eventual listening material involved naturalness judgements by listeners who were independent of those who participated in the fluency judgements.

An interesting speculation can be made from these results when they are considered in conjunction with findings by Duez [5] for identification rate of within-constituent pauses. The present study found that listeners perceived speech to be disfluent when there were pause durations of 98–479 ms, while Duez found only low rates of pause identification for the same types of pauses (within-constituent) that had durations in the range 250-400 ms. It appears that the global judgement of speech disfluency has occurred for pause durations that were in the range where identification rates of the pauses themselves were only low.

It is thus feasible that this experiment has shown that pause durations required for speech to be judged disfluent are shorter than those when the listening task is to identify the presence of a pause. Such an interpretation would be consistent with the notion that listeners are able to judge global features of speech better than they can perceive individual features for multifactorial features of speech such as voice and speech quality, and speech nasality. Within the speech pathology literature there are numerous reports that show that the reliability of perceptual judgements for individual factors that contribute to a specific speech disorder is found to be inferior to judgements of more global features.

In conclusion, the findings of this experiment demonstrate a relationship between the single factor of pause duration and judgement of global speech fluency when newsreading speech is artificially manipulated. The range of pause duration for which there is a significant increase in the number of disfluency judgements should be further investigated to make a direct comparison with the durations for which such pauses themselves can be identified.

5. Acknowledgements

Technical assistance from Anders Asplund and Thierry Deschamps is gratefully acknowledged. The authors would also like to thank Eva Strangert for her constructive suggestions. Finally, thanks must go to the listeners who took the time to participate in this study.

6. References

[1] Amir, O. & Yairi, E. 2000. The effect of temporal manipulation on the perception of disfluencies as normal or stuttering. *Journal of Communication Disorders*, 35, 63-82.

[2] Campione, E. & Véronis, J. 2002. A large-scale multilingual study of silent pause duration. *SP-2002*, 199-202

[3] Dalton, P. & Hardcastle, W. J. 1977. Disorders of fluency. New York: Elsevier.

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

[5] Duez, D. 1985. Perception of silent pauses in continuous speech. *Language and Speech*, 28, 377-389.

[6] Duez, D. 1993. Acoustic correlates of subjective pauses. Journal of Psycholinguistic Research, 22, 21-39.

[7] Guitar, B. 1998. *Stuttering: An integrated approach to its Nature and Treatment.* 2nd edition. Baltimore: Williams and Wilkins.

[8] Hegde, M. N. 1978. Fluency and fluency disorders: their definition, measurement and modification. *Journal of Fluency Disorders*, 3, 51-71.

[9] Kirsner, K., Dunn, J., Hird, K., Parkin, T. & Clark, C. 2002. Time for a pause. *Proceedings of the 9th Australian International Conference on Speech Science and Technology*. Melbourne, December 2-5 2002. Australian Speech Science and Technology Association Inc.

[10] Martin, J. G. & Strange, W. 1968. The perception of hesitation in spontaneous speech. *Perception & Psychophysics*, 3, 427-438.

[11] Ruder, K. F. & Jensen, P. J. Fluent and hesitation pauses as a function of syntactic complexity. *Journal of Speech and Hearing Research*, 15, 49-60.

[12] Searl, J. P., Gabel, R. M. & Fulks J. S. 2002. Speech disfluency in centenarians. *Journal of Fluency Disorders*, 35, 383-392.

[13] Strangert, E. 1990. Perceive pauses, silent intervals, and syntactic boundaries. PHONUM 1, 35-38. University of Umeå: Department of Phonetics.

[14] Strangert, E. 1993. Speaking style and pausing. PHONUM 2, 121-137. University of Umeå: Department of Phonetics.

[15] Susca, M. & Healy, E. C. 2001a. Listeners' perceptions along a fluency-disfluency continuum: a phenomenonological analysis. *Journal of Fluency Disorders*, 27, 135-161.

[16] Susca, M. & Harley, E. C. 2001b. Perceptions of simulated stuttering and fluency. Journal of Speech, Language and Hearing Research, 44, 61-72.

[17] Zellner, B. 1994. Pauses and the temporal structure of speech. In E. Keller (Ed.) *Fundamentals of speech synthesis and speech recognition*. Pp 41-62.

7. Appendix

Excerpt 1.

En upptäckt vid Karolinska Institutet kan få stor betydelse för patienter ***med /27, 65, 121, 247/ typ****-två-diabetes. Ett forskarlag* ***har /51, 120, 197, 349/ kartlagt*** *den insulinproducerande betacellen och har* ***upp /110, 132, 164, 330/ täckt*** *hur* ***insulinproduktionen /31, 67, 98, 184/ kan*** *regleras genom att man tar bort en del av cellen*

English translation.

A discovery at the Karolinska Institute could have great significance for patients with ***[P1]*** *type two diabetes. A research team has* ***[P2]*** *investigated insulin producing beta cells and has dis-* ***[P3]*** *covered how insulin production* ***[P4]*** *can be regulated by removing some of the cells.*

Excerpt 2.

Och först ska vi berätta att den kraftigaste jordbävningen på fem år nu på morgonen skakade stora delar av Taiwan. Jordbävningen mätte sju komma ***noll /53, 105, 184, 340/ på*** *richterskalan vilket är* ***förhållandevis /41, 93, 180, 339/ kraftigt****, men epicentrum låg under havet, elva mil öster om Taiwans* ***öst /53, 114, 237, 450/ kust****.*

English translation.

And first we will report that the strongest earthquake for five years this morning shook extensive regions of Taiwan. The earthquake measured seven point zero ***[P1]*** *on the Richter scale which is relatively* ***[P2]*** *powerful, but the epicentre lay under the ocean eleven thousand kilometres east of Taiwan's east* ***[P3]*** *coast.*

Excerpt 3.

Den svenska säkerhetspolisen ägnar allt mer resurser åt att spana på grupper av islamistiska extremister som man hävdar ***finns /102, 237, 479, 975/ i*** *Sverige. Säpo tror* ***att /48, 112, 241, 444/ grupperna*** *vistas i Sverige för att* ***planera /46, 82, 174, 254/ nya*** *terrordåd och för att rekrytera* ***nya /40, 119, 253, 505/ terrorister*** *bland unga svenska muslimer.*

English translation.

The Swedish security police are allocating further resources to investigate groups of Islamic extremists that are claimed to exist ***[P1]*** *in Sweden. SÄPO believes that the* ***[P2]*** *groups are staying in Sweden to plan* ***[P3]*** *new terror attacks and to recruit new* ***[P4]*** *terrorists from among young Swedish Moslems.*

Excerpt 4.

Brittiska forskare har utvecklat en metod för att förvara vaccin så att det inte behöver ***vara /73, 139, 226, 416/ kallt****. Det här kan* ***leda /73, 169, 287, 559/ till*** *att hanteringen av vaccin blir billigare och att fler barn i* ***u-länderna /44, 82, 133, 251/ kan*** *vaccineras.*

English translation.

British researchers have developed a method of storing vaccine so that it does not need to be ***[P1]*** *cold. This can lead* ***[P2]*** *to handling being cheaper and more children in underdeveloped countries* ***[P3]*** *can be vaccinated.*

.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 127–131.

Disfluency markers and their facial and gestural correlates. Preliminary observations on a dialogue in French

Elgar-Paul Magro

Paris III – Sorbonne Nouvelle University, Paris, France

Abstract

The aim of this article is to try to establish any observable regularities between the vocal and the visual expression of disfluency markers in a French spontaneous dialogue. The data show different configurations for different types of disfluency markers. Thus "euh"s are typically accompanied by mutual eye contact and no gesture; interrupted eye contact takes place less frequently, on occasions where speech planning is more seriously impaired (syntactical disruption and combination of "euh" with other disfluency markers). False starts seem to be typically accompanied by gesture production whereas eye contact can be maintained if the speaker relies or not on the listener to resolve the speech production problem. The article takes up the idea that disfluency markers can be classified along a continuum throughout the speech formulation process, going from the most discreet to the most prominent. It suggests that the more prominent the disfluency, the more likely is the visual channel to play a role (interrupted eye contact and gesture production).

1. Introduction

Few researchers so far have considered hesitation phenomena from a multimodal perspective. When they do, they either focus on facial expressions, particularly on eye contact, or else they focus on hand movement (or gesture), rarely on both.

This pilot study mainly tries to address the following two questions:

a) which regularities can be observed between the vocal and visual channels during disfluencies in spontaneous speech?

b) what light do these regularities throw upon the speech formulation process itself?

At this stage, it is important to specify and define the variables of the visual mode that have been take into account for the purpose of this study and to define them for the sake of clarity. They amount to two: eye contact and gesture. Eye contact is what normally takes place between the speaker and the listener during a conversation (Kendon [6], Bouvet & Morel [1]. Although inter-individual differences have been observed, the speaker typically looks less often in the direction of the listener than the latter looks at him. By gesture, one must understand here "the movements of the hands and arms that we see when people talk" and that "are closely synchronized with the flow of speech" (McNeill [9] : 1, 11).

2. Literature review

Published research (Exline & Winters [5], Kendon [6], Brossard [2]) states that, during hesitation, the speaker typically looks away from the listener. The main reason put forward to account for this behaviour is that, during hesitation, the speech formulation process requires extra cognitive efforts on the speaker's part; thus the speaker prefers to look away from the listener so as to avoid being distracted by his gaze. In this article, the term "interrupted eye contact" will be used when the speaker looks away from the listener during disfluent speech; the term "mutual eye contact" will be used to design instances when the speaker keeps looking at the listener despite disfluency markers in his speech.

Research taking into account hesitation and gesture points out differing views according to whether the disfluencies arise from a speaker expressing himself in his mother tongue or else in a foreign language. As regards native speakers, McNeill [9] and Seydeffinipur & Kita [12] stress the absence of gesture during hesitant speech. In the case of non-native speakers, various authors (e.g. Proceedings of ORAGE 1998 [7], 2001 [3]) point out the abundance of gestures accompanying the speech formulation process.

On the basis of the above assumptions, we set out this study with the following starting hypothesis, or expectation: namely, that the disfluencies of the corpus under observation would be mainly accompanied by interrupted eye contact and rather limited gesture production.

3. Corpus

The corpus used for the purpose of this study is a spontaneous dialogue between two female French friends, Anne and Soline, aged 25, both coming from the region around Paris and having pursued tertiary education. What is meant by "spontaneous" here is speech that is not linked to any prior written text and not recorded in a speech laboratory – in fact, the recording took place at the house of a common friend of the two speakers.

The general theme of discussion was not imposed and the conversation went from reminiscing holidays to discussing boyfriends. The excerpt which was analyzed is a 90-second narrative by Soline, particularly rich in disfluencies as the speaker encounters difficulties, amongst others, in remembering the proper names of the places she visited in Italy with her parents some years ago.

An audio and video recording (on MiniDisk and digital cameras respectively) was first carried out. As illustrated in Figure 1 below, the two speakers were seated at an angle, at an approximate distance of half a metre from each other. Each video camera was positioned on either side of the speakers outside the interactive space of the dialogue, at an approximate distance of one and a half metres and at an angle of 45° from the speakers. The cameras were neither totally visible, nor totally hidden, but both speakers stated at the end of the recording that they had not been upset by the presence of the cameras.

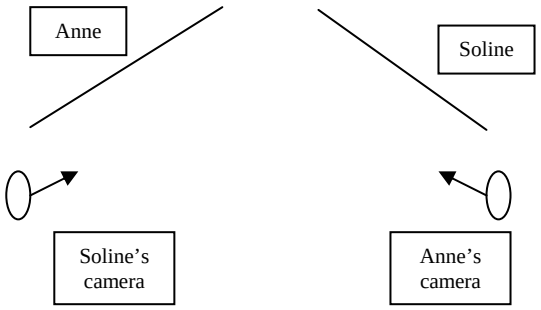


Figure 1: Diagram showing the recording set-up for the corpus.

At a later stage, a synchronized editing of the video was obtained on computer format (AVI file). This made it possible for Soundforge to be used to transcribe and observe minutely the eye, hand and arm movements of the speakers (25 images per second). The transcription of gestures was based on the identification of their movement phases : preparation, stroke and retraction (cf. McNeill [9], Bouvet & Morel [1]). On the other hand, a rigorous observation of the eye and head movements of Soline (the main speaker) over a long sequence of the recording made it possible to pinpoint the direction of Soline's gaze, that is whether it was oriented towards Anne (the listener) or elsewhere. Although a small error margin is not to be totally excluded, the observation work carried out as well as the observer's intuitive experience of speaker and conversation participant made it possible to infer whether a disfluency was accompanied by mutual or interrupted eye contact.

The examples illustrated in the plates below show the end result of the edited video: Soline, the main speaker in this excerpt, and the one whose disfluencies have been analyzed, is on the right frame of each photo; on the left frame is Anne, the listener. As can be observed in these illustrations, each frame not only shows the head, trunk, hands and knees of one particular speaker but also includes the hands of the other speaker. This detail, which initially came across as an undesirable result of the recording, was actually beneficial for the transcription of the gestures produced, as the hands of each speaker could be observed from two different angles (one on each frame).

4. Results

The data show that not all disfluencies are accompanied by interrupted eye contact and lack of gesture. On the contrary, a detailed analysis suggests that different disfluency markers tend to display different types of facial and gestural behaviours.

For the purpose of this study, only two types of disfluency markers were taken into account, namely "euh"s (the French equivalent of the English "uh"s and "um"s) as well as false starts. One of the main reasons behind this choice lies in the high frequency with which they occur in the dialogue excerpt analyzed in this paper. Besides, it is generally admitted that "euh"s are considered to be the most typical markers of hesitation in French, whereas false starts are easily identifiable and enable to consider larger stretches of disfluent speech than "euh"s do.

4.1. The case of "euh"s

4.1.1. General remarks

Soline's speech, whose disfluencies have been analyzed, contains fifteen occurrences of disfluent "euh"s. Four of these occurred at overlaps in her speech (at moments when Anne, the listener, gives some sort of backchannel signals). Since it is not clear whether these four "euh"s are a result of a real disfluency in Soline's speech or rather a way to assert her wish to keep her turn, they have not been taken into account for this study.

Out of the eleven remaining "euh"s, only one was accompanied by gesture production, which seems to comfort McNeill's [9] claim. On the other hand, only three out of eleven were accompanied by interrupted eye contact, whereas during the remaining eight the speaker keeps looking at the listener. This seems quite surprising in the light of the common point of view adopted by the doxa.

Let us first give an example of each category of "euh", one with interrupted and another with mutual eye contact. The numbers within brackets indicate the duration of silent pauses in milliseconds, the colons shows the amount of lengthening (each colon marks an additional lengthening of 200ms).

E.g. (1): *c'est en (550)* ***euh:::*** *(600) ben près de Florence c'est quoi la région?*

(it's in (550) euh (600) well next to Florence what's the region?)

Soline cannot remember the name of the region around Florence in Italy and after producing an 800ms euh, she decides to ask Anne for help. The eye contact is interrupted at the beginning of the utterance and is resumed on the word "Florence". Soline's averted gaze does not fix one particular point but is very mobile. Cf. Plate 1 below.

Plate 1: Illustration of Example (1). Disfluent "euh" - interrupted eye contact and absence of gesture: "*c'est en (550)* ***euh:::*** *(600) ben près de Florence c'est quoi la région?*".

E.g. (2) *et* ***euh:::*** *donc* ***euh:*** *eux ils ont l'habitude de louer une maison* ***euh:::*** *là-bas*

(and euh so euh they have the habit of renting a house euh over there)

Soline is trying to explain the geographical context for the voyage she is narrating. During these three "euh"s of varying length (350 to 800ms) which she produces within the same utterance, she does not look away from Anne. Cf Plate 2 below.

Plate 2: Illustration of Example (2). Disfluent "euh » - mutual eye contact and absence of gesture. *"et **euh:::** donc euh: eux ils ont l'habitude de louer une maison euh::: là-bas*

4.1.2. "Euh"s and their distribution

At this stage we tried to find ways to possibly account for the distribution of these two categories of "euh"s (with mutual or interrupted eye contact). So we tried to look at the impact that the following three factors may have on their distribution: namely (a) the duration of the "euh"s, (b) their combination with a silent pause and (c) the degree of syntactical disruption they bring about.

4.1.2.1 The duration of "euh"s and eye contact

Out of the three "euh"s accompanied by interrupted eye contact, one is lengthened (350ms) and two are very long (750, 800ms). On the other hand, the duration of the eight remaining "euh"s accompanied by mutual eye contact varies from brief (200ms) to very long (800ms). Although one would expect duration of "euh"s to be of some significant influence on the behaviour of the speaker's gaze direction, the small numbers of "euh"s under observation here do not make it possible to make any statement in this sense.

4.1.2.2 The combination of "euh"s with silent pauses and eye contact

Since no correlation could be found between eye contact and the combination of "euh"s with vowel lengthening, we tried to see if the data suggested any possible correlation between eye contact and the combination of "euh"s with other disfluency markers.

The three "euh"s during which the speaker looks away from the listener are immediately followed by a silent pause. On the contrary, none of the eight other "euh"s during which the speaker does not turn away her look from the listener is immediately followed by a silent pause. This suggests that the combination of an "euh" with a silent pause might have a direct relation with eye contact during speech.

4.1.2.3 The level of syntactical disruption and eye contact

Another interesting observation to point out is that two of the three "euh"s that are accompanied by interrupted eye contact are also accompanied by an interrupted syntactical structure. They are the same two "euh"s that are immediately followed by a silent pause (cf. 4.1.2.2. above). In other words, the speaker does not finish off the syntactical structure that precedes the "euh" and finds the need to start afresh (either a repetition or a new start).

On the contrary, this syntactical disruption does not happen in any of the eight "euh"s that are accompanied by mutual eye contact. In all of these instances, if the "euh"s were to be deleted from the speaker's discourse, the resulting syntax would not bear the traces of any disfluency.

This might suggest a possible correlation between the extent of syntactical disruption followed by an "euh" and the interruption or otherwise of eye contact.

4.2. The case of false starts

The discourse of Soline in the dialogue excerpt under observation contains seven false starts, that is seven syntactical structures that are left abandoned in the favour of new ones in the immediate linguistic context. Six of these are not isolated but appear in sequences. These figures do not include, for obvious reasons, any incomplete utterances located at turn-taking frontiers, in which case they are clearly interrupted by the listener and not by the speaker herself.

4.2.1. Gesture

False starts present a major difference compared to "euh"s in the visual channel. Contrary to "euh"s, they are predominantly accompanied by gestures (six out of seven). Moreover, when false starts appear one after the other in a sequence, as is twice the case in our corpus, gesture seems to illustrate and throw light on the formulation process that is under way.

Let us have a closer look at a sequence of three false starts in our corpus.

E.g. (3): Soline is trying to give a list of the places she and her parents visited in Tuscany. After mentioning Florence and Siena, she takes up the same syntactical structure to add a third place name, but apparently she does not manage to recall it immediately. This gives rise to the following sequence of structures, of which (c), (d), (e) (in italics) are false starts:

(a) on a fait Flo^{rence}
(b) on a fait ^{Sienne} (1300)
(c) on a $^{fait ::}$
(d) parce qu'on é^{tait} dans le: dans la ré$^{gion:}$
(e) on avait lou$^{é::}$ une:
(f) à Grosse^{to} on était (600)

(we visited Florence / we visited Siena / *we visited / cos we were in the region / we had rented a* / Grosseto that's where we were)

It is interesting to note that no gesture accompanies the first of the three successive false starts: (c). Presumably, the speaker at this stage is not yet conscious of the amplitude which her disfluency is going to cover.

The second false start (i.e. (d)) is accompanied by a gesture which clearly shows the cognitive effort of word-searching going on. The fingers of the speaker's right hand, oriented at an angle towards the listener, rub against the thumb six times. This gesture accompanies the whole of the utterance. Incidentally, this utterance introduces a change of strategy : following the unproductive attempt in (c) to find the desired place name, the speaker moves out of the narrative per se and shifts to an explicative mode (cf. the connector "parce que" and the change in tenses from narrative (passé composé) to extranarrative ones (imparfait, plus-que-parfait). Plates 3A and 3B try to illustrate this finger-rubbing gesture.

Plate 3A: Illustration of Example (3d). False start – mutual eye contact and word-searching gesture (repeated finger-rubbing). *"**parce** qu'on é^tait dans le: dans la ré^gion:"*

Plate 3B: Illustration of Example (3d). False start – mutual eye contact and word-searching gesture (repeated finger-rubbing). *"parce qu'**on** é^tait dans le: dans la ré^gion:"*

Since the word-searching in (d) above was not fruitful, the speaker gives a new twist to her speech planning strategy and prefers to start afresh by evoking a concrete referent this time, namely the lodging they had rented (later on in the dialogue, we learn it was a house). The gesture that is produced now is no longer a gesture of word-searching, but one which shows the grasp of an object. In fact, the speaker raises her right hand, then palm facing downwards she lowers it whilst cupping it in the form of the claws of a bird of prey. Plausibly enough, the object being grasped is the house she has in mind or possibly the place whose name she is after. The gesture of prehension is illustrated in Plate 4 below.

Plate 4: Illustration of Example (3e). False start – mutual eye contact and gesture of prehension (grasping). *"on avait lou^é:: **une**:"*.

This strategy pays off as before the speaker finishes her utterance in (e), she has already found the name she was looking for, namely the town of Grosseto. Interestingly enough, this final syntactical structure which puts an end to the series of false starts analyzed above is also accompanied by a gesture. This time it is an abstract pointing gesture, whereby the forefinger of the speaker's right hand points out to an imaginary point within the conversation space. The wider space circumscribed during the previous false start thanks to the gesture of prehension, is now clear-cut, precise: the proper name has at last been identified. Plate 5 illustrates this pointing gesture.

Plate 5: Illustration of Example (3f). End of false start series – mutual eye contact and pointing gesture. "*à Grosse^to on était*".

4.2.2. Eye contact

As far as eye contact is concerned, no clear-cut tendency arises in the case of false starts. In three out of seven occurrences the eye contact is mutual, in the remaining four it is interrupted. It is not very clear to us yet at this stage which are the main factors that may account for this distribution. One hypothesis is that it may significantly depend on whether the speaker decides or not to rely on the listener or not to overcome the source of his disfluency.

It may be of interest to note that when false starts appear in a cluster (which is often the case in our corpus), the first false start of the series behaves differently to the following false starts. Thus in the series analyzed in 4.2.1. above, the first false start is accompanied by interrupted eye contact, following which the speaker looks back at the listener till the final completed utterance arises. This possibly bears a link with the speaker's change of narrative strategy pinpointed above. In the second series (equally made up of three false starts) it is the other way round. On the other hand, the only occurrence of a false start in isolation is accompanied by interrupted eye contact. Thus at a certain point in all the false starts observed, there is interrupted eye contact.

5. Discussion

It may be argued that disfluency in speech takes place along a sort of continuum throughout the speech formulation process going from the more discreet to the more prominent markers. The results observed above suggest that the more prominent the disfluency in the vocal channel, the more prominent it is likely to be in the visual channel.

As a matter of fact, only the more prominent "euh"s seem to be accompanied by an interrupted eye contact; that is, those "euhs" which are combined to a silent pause and/or the need for a syntactical restart. Gestures do not intervene at this point. As regards false starts, they cover longer stretches of disfluency than "euh"s, often contain other disfluency markers (vowel lengthening, internal repetition as in (3d) above or even combination with an "euh" as in (1)), and their incompletion clearly brings about a higher level of syntactical disruption than "euh"s. Which means they come off as more prominent disfluency markers than "euh"s. Interestingly, one can note that this is where gestures take place.

Thus one may argue that the more the speech formulation seems to be seriously impaired, the more this will be marked on both the vocal and the visual level.

It seems interesting at this stage to mention that different researchers have shown interest in the role of gesture within the speech production process, proposing different speech production models which place gesture at either the conceptualization level or the formulation level (e.g. de Ruiter [4], Kita [7] and Krauss, Chen & Gottesman [8]). Our preliminary findings do not offer much support to either of the two major models but we understand that, at a later stage, our research may contribute to this debate, which is still open.

6. Conclusion

The aim of this study was to present the preliminary results obtained upon close observation of 1.5 minutes of a conversation between two young female French friends.

This paper has argued that the generally acclaimed views on facial and gestural behaviour during disfluencies (interrupted eye contact and lack of gesture) give only a partial picture of what really goes on. The initial working hypothesis thus needs to be nuanced.

The data show different configurations for different types of disfluency markers. Thus eye contact shows a tendency to be mutual during "euh"s, except where speech planning seems to be more seriously impaired (combined disfluency markers and syntactical disruptions); the data suggest that false starts may prefer interrupted eye contact although no statement can be made in favour of this. Gestures seem to be typically absent with "euh"s whereas they typically accompany false starts, illustrating thereby the progress in the speech planning process.

Finally, the paper suggests that the more the speech formulation is impaired alongside the disfluency continuum, the more markers arise in the visual mode (going from interruption of eye contact to the production of gestures).

It is clear that no general conclusions can be drawn out of such a small corpus. Indeed, the results obtained have to be validated by using much vaster material, ideally constituted of data that is greater both in size and in nature. A further analysis of longer stretches of the same dialogue is currently under way.

Future analysis must also take into account the rates and configurations of eye contact and gesture production in fluent stretches of speech. Such a comparison of fluent and disfluent speech will enable to establish how typical of disfluent speech are the trends of facial and gestural behaviour that have been observed. The observation of the timing relation (e.g. Seyfeddinipur & Kita [12]) between gesture, gaze and speech during disfluencies may also be of interest in future studies.

7. Acknowledgements

This study is part of an ongoing doctoral research supervised by Prof. Mary-Annick Morel [1], [10]. Our gratitude also goes to Prof. Laurent Danon-Boileau [10] and Dr. Danielle Bouvet [1] for their inspiring comments on certain aspects of the research presented in this paper.

8. References

[1] Bouvet, Danielle & Mary-Annick Morel. 2002. *Le ballet et la musique de la parole.* Paris-Gap: Ophrys.

[2] Brossard, Alain. 1992. *La psychologie du regard*. Neuchâtel: Delachaux et Niestlé.

[3] Cavé, Christian, Isabelle Gauïtella & Serge Santi (éds.). 2001. *Oralité et gestualité. Interactions et comportements multimodaux dans la communication*, Actes du colloque Orage 2001, 18-22 juin 2001, Aix-en-Provence, France. Paris: L'Harmattan.

[4] de Ruiter, Jan Peter. 2000. The production of gesture and speech. In McNeill, David (ed.). *Language and gesture*. Cambridge : Cambridge University Press, pp. 284-311.

[5] Exline, Ralph V. & Lewis C. Winters. 1966. Affective relations and mutual glances in dyads. In Tomkins, Silvan S. & C. Izard (eds.). *Affect, cognition and personality*. London : Tavistock.

[6] Kendon, Adam. 1967. Some functions of gaze direction in social interaction, *Acta Psychologica*, vol. 26, pp.1-47.

[7] Kita, Sotaro. 2000. How Representational Gestures Help Speaking. In McNeill, David (ed.). *Language and gesture*. Cambridge : Cambridge University Press, pp. 162-185.

[8] Krauss, Robert M., Yihsiu Chen & Rebecca F. Gottesman. 2000. Lexical gestures and lexical access: a process model. In McNeill, David (ed.). *Language and gesture*. Cambridge : Cambridge University Press, pp. 261-283.

[9] McNeill, David. 1992. *Hands and mind. What gestures reveal about thought*. Chicago and London: The University of Chicago Press.

[10] Morel, Mary-Annick & Laurent Danon-Boileau. 1998 (rééd. 2001). *La grammaire de l'intonation. L'exemple du français oral.* Paris – Gap : Ophrys.

[11] Santi, Serge, Isabelle Gauïtella, Christian Cavé & Gabrielle Konopczynski (éds.). 1998. *Oralité et gestualité. Communication multimodale, interaction*, Actes du colloque Orage 1998, 9-11 décembre 1998, Besançon, France. Paris: L'Harmattan.

[12] Seyfeddinipur, Mandana & Sotaro Kita. 2001. Gesture as an Indicator of Early Error Detection in Self-Monitoring of Speech. *Proc. DISS'01*, 29-31 August 2001, Edinburgh, Scotland, pp. 29-32.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 133–138.

Disfluency & Behaviour in Dialogue: Evidence from Eye-Gaze

Hannele Nicholson[1], Ellen Gurman Bard[1], Robin Lickley[2],
Anne H. Anderson[3], Catriona Havard[3] & Yiya Chen[1]

[1] University of Edinburgh, Edinburgh, Scotland
[2] Queen Margaret University College, Edinburgh, Scotland
[3] University of Glasgow, Glasgow, Scotland

Abstract

Previous research on disfluency types has focused on their distinct cognitive causes, prosodic patterns, or effects on the listener [9, 12, 17, 21]. This paper seeks to add to this taxonomy by providing a psycholinguistic account of the dialogue and gaze behaviour speakers engage in when they make certain types of disfluency. Dialogues came from a version of the Map Task, [2, 4], in which 36 normal adult speakers each participated in six dialogues across which feedback modality and time-pressure were counter-balanced. In this paper, we ask whether disfluency, both generally and type-specifically, was associated with speaker attention to the listener. We show that certain disfluency types can be linked to particular dialogue goals, depending on whether the speaker had attended to listener feedback. The results shed light on the general cognitive causes of disfluency and suggest that it will be possible to predict the types of disfluency which will accompany particular behaviours.

1. Introduction

Types of disfluency distinguished by their form are also distinguishable by other characteristics. Repetition disfluencies are the most common in spontaneous speech [21]. In a pioneering paper, Maclay & Osgood showed that repetitions precede content words more often than function words [22]. Repetitions have been linked to strategic signalling commitment to both listener and utterance [10, 12]. The prosodic cues for repetitions are linked to certain strategies in dialogue [25]. Savova showed, however, that the prosodic cues to repetitions differ from the cues to a substitution, providing support for the notion that disfluency types have distinct sources in the cognitive processes underlying the production of speech in dialogue [26].

It is already clear that disfluencies of different types cause different processing problems for the listener. While repetitions cause less disruption than false starts [a kind of deletion disfluency] for a word recognition task, [13], repetitions are more difficult for trained transcribers to detect than false starts of the same length [20].

Disfluency has been linked to cognitive causes by Levelt [17], who proposes that some disfluencies occur for covert cognitive reasons while other disfluencies are overt corrections. Lickley found that disfluency types vary systematically across turn types whereby turns that involve planning typically involve more self-corrections than utterances which are responses to queries [18]. Replies to queries, on the other hand, tend to involve more filled pauses (ums, uhs) and repetitions in order to buy time [18]. Thus, it seems that certain types of disfluencies have already been linked to certain dialogue behaviours.

More recently, psycholinguistic studies of a speaker's eye-gaze at a visual array have revealed that speakers look at objects involved in the process of speech perception and production. [15, 28]. Speakers who made a speech error when performing a simple object naming task had spent just as long gazing at the object as they did when they named it fluently. Apparently, then, disfluency did not result from either long or hasty examination of the object to be named. Disfluency does not appear to be a measure of perceptual problems per se.

Instead, disfluency is related to the cognitive burdens of production [5]. We will use disfluency to discover whether there is a cognitive cost involved in taking up information needed to pursue a dialogue task. We will then show that this cost is put to good use: the locations of disfluencies reveal that they are appropriate responses to the information that speakers have garnered.

The information in question underpins what is thought to be a crucial task in dialogue: each participant must maintain a model of her interlocutors' knowledge so as to adjust to their mutual knowledge both what she says and how she says it. Most views of dialogue now assume that speakers will take some interest in indications both of the listener's knowledge about the domain under discussion and of the listener's satisfaction with the communication just made. Clark and Krych [9], for example, propose that speakers monitor listeners' faces for all manner of feedback, much as they track listeners' utterances. Horton and Gerrig [16] acknowledge the costs of this operation, suggesting that complete uptake and application of listener information could prove to be taxing in some cases, so that utterances will be less perfectly designed for the audience as the cognitive burden increases.

To determine whether garnering cues to listener knowledge is indeed costly to production, we use a variant of the map task [2, 7]. As in the original task, players have before them versions of a cartoon map representing a novel imaginary location. The Instruction Giver communicates to the Instruction Follower a route pre-printed on the Giver's map. The present experiment manipulates time-pressure and the modality or modalities in which a distant confederate delivers pre-scripted feedback to the speaker's instructions. Verbal feedback affirms comprehension of some instructions and declares general incomprehension of others. Visual feedback, in the form of a simulated listener-eyetrack projected onto the map, may correctly go to the named map landmark or wrongly advance to another. Where both modalities are used, their feedback may be concordant or discordant across modalities. Scripted and simulated responses are used to control the conditions under which speakers are operating. Genuine speaker eye-gaze is tracked.

We use eyetracks, rather than sight of the speaker's direction of gaze, to represent listener feedback for two reasons. First, simulated gaze is much easier to control than genuine gaze on the part of the confederate. Second, though facial expressions and direction of gaze have real value, tasks with a visual component produce remarkably little inter-interlocutor gaze [[1,3,11]]. To allow simultaneous performance of the task and uptake of listener information, the

listener's 'eyetrack' was superimposed on the map (See Figures 1 and 2).

The present paper will examine two kinds of disfluency diistinguished by previous research, repetitions and deletions. In the current definition, a repetition is produced when the speaker repeats verbatim one or more words with no additions, deletions, or re-ordering, as in (1)

(1) Now you want to ***go*** **go** just past the tree

Repetitions are thus a single faulty attempt at communicating the same message in the same form. In contrast, a deletion has occurred whent the speaker interrupts an utterance without restarting or substituting syntactically similar elements, as in (2)

(2) A MOVE 36 You need to be just under…
A MOVE 37 Do you have a White Mountain?

Thus, deletions abandon one communicative act in favour of another.

In this setting, there seem to be two distinguishable predictions. Clark and Krych [9] predict good uptake of all visual cues to listener knowledge and suitable application of the information. Horton and Gerrig [16] predict that the more complex the input, the more difficult will be both uptake of cues and the production of suitable speech. Thus there should in principal be an increase in dsfluency if speakers observe negative visual feedback ('follower gaze' at wrong landmarks) and if there ar conflicts between verbal and visual feedback.

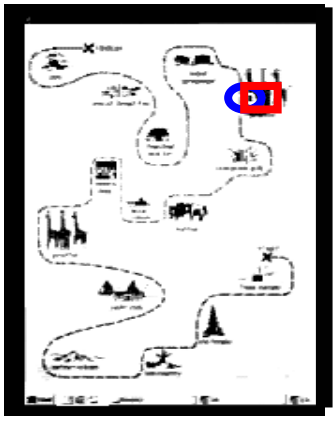
1.1. Task and procedure

All the materials come from an experiment which used conversations between subject Instruction Givers and a confederate Instruction Follower. Each subject was greeted individually with the confederate. Each subject was naïve to the status of the confederate and during post-experimental debriefing, none reported any suspicions. Both subject and confederate were told that whoever took the role of Instruction Giver should guide the Instruction Follower, from a marked start-point to buried treasure. Subject and confederate then 'negotiated' that the subject would be Giver and the two were taken to separate rooms. The Giver was seated 60 cm from a flat screen monitor displaying the map. Labelled landmarks and map designs were adapted from the HCRC Map Task Corpus [2]. Eye tracking movements were recorded using a non-invasive Senso-Motor Instruments remote eye-tracking device placed on a table below the monitor. Eye movements were captured with Iview version 2 software. The tracker was re-calibrated at the beginning of each trial. Speech was recorded in mono using Asden HS35s headphone- microphone combination headsets. Video signals from the eye tracker and the participant monitor were combined and recorded in Mpeg with Broadway Pro version 4.0 software.

Feedback from the confederate took two forms. Visual feedback consisted of a simulated eyetrack, a small red square advancing from landmark to landmark once each landmark was named, and showing saccades of random length and direction. The visual feedback was under the control of the experimenter, who advanced the feedback square to its next programmed position when the Giver first mentioned a new route-critical a landmark. When feedback was scheduled to be wrong, the square moved to a landmark that had not been named. When feedback was to be correct, the feedback square advanced to the landmark just named. Similarly, verbal feedback came from the confederate subject who read pre-scripted responses. Just as with the visual feedback, the confederate provided verbal feedback when the speaker uttered the first mention of the landmark in question. Figures 1 and 2 illustrate possible events.

Instruction Follower: 'Yes, got it.'

Figure 1. Discordant feedback. Circle = Giver's gaze; Square = Follower's feedback (wrong location).



Instruction Follower: 'Okay, that's fine'

Figure 2. Concordant feedback. Circle = Giver's gaze; Square = Follower's feedback (correct location).

1.2. Experimental Design

The experiment crossed feedback modality (3), single modality group (2), and time-pressure (2). In the *No Feedback* conditions, subjects saw only the map. In the *Single-Modality* condition, subjects in the Verbal Group got verbal feedback only, while those in the Visual Group had only visual feedback. Finally, in the *Dual-Modality* condition, all subjects received both visual and verbal feedback. The two modalities might be discordant or concordant. *Concordan*t feedback consisted on average of 8 instances of positive verbal and correct visual feedback, and 6 instances of negative verbal and wrong visual feedback per map. In each map, *discordant* feedback included roughly 3 instances of negative verbal and correct visual feedback, and 6 instances of positive verbal and wrong visual feedback. This design is portrayed in Table 1. In half of the trials, speakers under *time-pressure* had three minutes to complete the task; in *untimed* dialogues there was no time limit.

Table 1. The relationship between the Experimental_Groups and the various Feedback Modalities.

Experiment	Feedback Modalities		
	None	Single	Dual
Verbal Group	None	**Verbal**	Verbal + Visual
Visual Group	None	**Visual**	Verbal + Visual

Thirty-six subjects with normal uncorrected vision were recruited from the Glasgow University community. All were paid for their time. All encountered all 6 conditions. Six

different basic maps were used, counter-balanced across conditions over the whole design. Subjects were eliminated if any single map trial failed to meet criteria for feedback or capture quality. The feedback criterion demanded that the experimenter advance the feedback square between the introduction of the pertinent landmark and the onset of the following instruction in all cases where the feedback was scheduled to be errant and in 70% where the square's movement was scheduled to be correct. The capture criterion demanded that at least 80% of the eye-tracking data was intact. Subjects were also eliminated if on debriefing they revealed any suspicions about the nature of the interlocutor.

2. Results

2.1. Baseline effects: Words

Since the opportunities for disfluency increase with increasing amount of speech, it is important to note effects of the experiment's design on word counts. Word counts for whole and part-words show less speech with time-pressure (425 words/trial on average) than without (579): ($F_1(1,34) = 24.38$, $p < .001$). Visual Group Single-Modality trials (459 words) were shorter than the corresponding Dual-Modality trials (590 words) with no corresponding change for Verbal subjects (Feedback Modality x Group: ($F_1(2,68) = 8.65$ $p < .001$; Bonferroni: $t = -6.4$, $p < .001$). Since Dual-Modality Conditions do not differ between groups (Verbal: 616, Visual: 590), we can use this condition to examine the relationships between disfluency and gaze or dialogue events.

We also examined speech rate across the experimental conditions. To calculate speech rate we divided the Giver words per map by the total Giver speaking time for the map (the summed durations of all conversational moves less the summed durations of both simple and filled pauses). Time-pressure had no significant effect on speech rate. The interaction between Feedback Modality and Group ($F_1(2,68) = 4.87$, $p < .02$) presented in Table 2, is due only to a difference between the No-Feedback (.34) and Dual-Modality (.30) conditions for the Verbal Group (Bonferroni $p = .004$). Again Dual Modality conditions are alike.

Table 2. Speech rate (Words/Total speaking time) means from Feedback Modality x Group interaction

Experiment	Feedback Modalities		
	None	Single	Dual
Verbal Group	.340	.303	.304
Visual Group	.344	.343	.340

2.2. Baseline effects: Gaze

In order to test for the relationship between disfluency and Giver gaze, it was necessary to determine whether all conditions in which a Giver might gaze at a feedback square actually did succeed in directing the Giver's attention to the square. To check for overlap of gaze between Giver and 'Follower', the video record of feedback and Giver Gaze were analyzed frame by frame for the landmark at which each was directed. When Follower Gaze and Giver Gaze were on the same landmark, the Giver was considered to be looking at the feedback square. Here we report the number of feedback episodes [task sub-portions containing in feedback] in which *any* frame contained an instance of gaze at the feedback square].

Givers did not make use of all their opportunities by any means (Figure 3). Nor did they use their opportunities equally (Visual feedback x Verbal feedback: $F_1(1,34) = 7.70$, $p < .01$). Strangely enough, Givers used fewest opportunities in an important concordant condition, the one in which the Follower was clearly lost: the Follower square was hovering over a wrong landmark while the Follower was simultaneously providing negative verbal feedback (verbal- vis-: .366). These attracted less gaze than another concordant condition – when the Follower needed no help because she was in the right place and said so (verbal+ vis+: .511). Similarly Givers looked less when the Follower was lost but claimed not to be (verbal+ vis-: .448) than when she was correct but claimed to be lost (verbal- vis+:.591) (Bonferroni *t*-tests at .008). A simple description says that speakers are most likely to track listeners, the listener's location falls under their own gaze, which is occupied by the things they are describing. Apparently, spekaers prefer not to go off-route to learn the whereabouts of an errant follower.

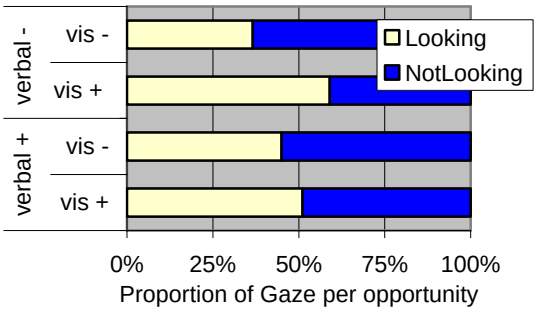


FIGURE. 3 Proportion of feedback episodes attracting speaker gaze to feedback square: Effects of combinations of visual and verbal feedback in dual channel conditions

2.3. Disfluencies Overall

The first author labeled disfluencies according to the system devised by Lickley [19] as repetitions, insertions, substitutions or deletions. She used Entropic/Xwaves software to listen to, view and label disfluent regions of speech. Spectrograms were analyzed whenever necessary. Each word within a disfluent utterance was labeled as belonging to the reparandum, the interregnum, or the repair. A reparandum involves speech that is either overwritten, expunged or retraced in the repair [19]. Repairs typically 'replace' the error in the reparandum. Since deletions are typically abandoned utterances, they have no repair [19, 27].

Because disfluencies are more common in longer utterances [6, 10, 25] we divided the number of disfluencies in a monologue by its total number of words, yielding disfluency rate as a dependent variable.

Disfluency rates were submitted to a by-subjects ANOVA for Group (2) (Verbal vs. Visual), Time-pressure (2) (timed vs. untimed) and Feedback Modality (3) (none, Single-Modality, Dual-Modality). The baseline No-Feedback conditions differed between Verbal and Visual groups (Group * Modality: $F_2(2,68) = 5.21$, $p < .01$; Bonferroni, $t = 2.94$, $p < .02$). This difference can be explained by a single subject in the Verbal Group who was an outlier in terms of disfluency. Because of this subject, there was no effect of Feedback Modality within the Verbal Group, while the Visual Group showed the expected increase in rate of disfluency between No Feedback and Single- (Bonferroni $t = -4.12$, $p = .001$) or Dual-Modality conditions (Bonferroni $t = -5.77$, $p < .001$). Since Single and Dual Modality conditions did not differ, we can proceed to examine only the Dual Modality conditions in the expectation that conflicting feedback (only found in Dual Modality) *per se* is not an overall cause of disfluency.

2.4. Disfluency Types: Repetitions v Deletions

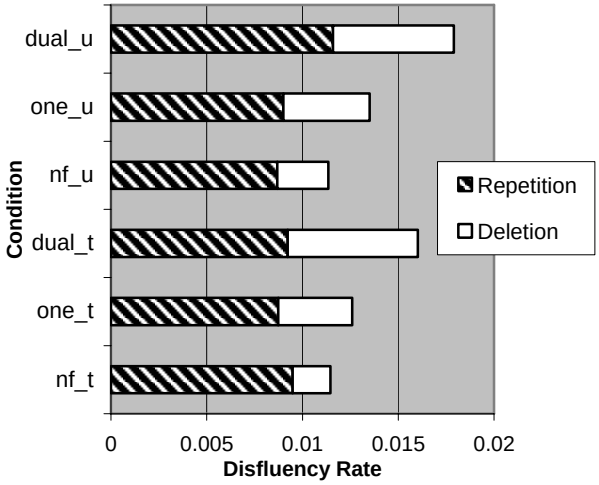


Figure 4. Rates of disfluency by type and experimental condition for the Verbal and Visual Groups combined. nf = no feedback, one = Single-Modality feedback, dual = Dual modality feedback; t = timed, u = untimed.

An initial investigation of deletions and repetitions begins to separate them. Figure 4 displays their distributions across experimental conditions. Independent analyses were done for each type of disfluency; that is one analysis within deletions only and one within repetitions only.

As found in [23], only deletion rate showed any significant effect of feedback: Deletion rate rose significantly with each additional feedback modality (No Feedback .002, Single-Modality .004, Dual-Modality .007; $F_1(2,68) = 21.00$, $p < .001$; all Bonferroni t-values < .01). There were no effects of time-pressure on deletion rate and no significant interactions.

For repetitions on the other hand, an interaction between Time-pressure and Group ($F(1,34) = 6.27$, $p < .02$) revealed that subjects were more disfluent in the untimed condition (.012) of the Verbal Group than they were anywhere else in either the Verbal or the Visual Group, timed or untimed, though the internal comparisons were not significant.

3.5 Disfluency & Eye-Gaze

Within the Dual-Modality condition, the experimental design contrasted positive and negative feedback in the two modalities. However, the modalities are concordant or discordant only if the Giver actually takes up both visual and verbal feedback. The tendency for more speech in conditions with verbal feedback suggests that subjects were attending to what the confederate Follower said. Eye-tracking enabled us to tell when the Giver had actually looked at the Follower's visual feedback. As Figure 3 made plain, Givers do not take up the same proportion of concordant and discordant feedback. They gazed most at one kind of discordant feedback (negative verbal + correct visual) and least at a concordant condition (negative + wrong visual feedback).

To look for disfluency in truly vs potentially concordant and discordant situations, we examined disfluency per feedback opportunites in concordant and discordant situations contrasting those in which Givers did or did not look at Follower feedback. In fact, Givers who attended to discordant feedback from the Follower encountered subsequent fluency problems. The number of disfluencies per feedback opportunity was greatest following a discordant feedback episode in which the Giver had actually gazed at the Follower feedback square (.333), a significantly higher rate than following a concordant feedback episode which had drawn the Giver's attention (.205) (Bonferroni $t = -3.51$, $p = .001$ within by-subjects Group (2) x Giver attention (looking v not looking) x Concordance of modalities (concordant v discordant: $F_1(1,34) = 7.24$, $p = .01$). None of the other pairwise comparisons was significant.

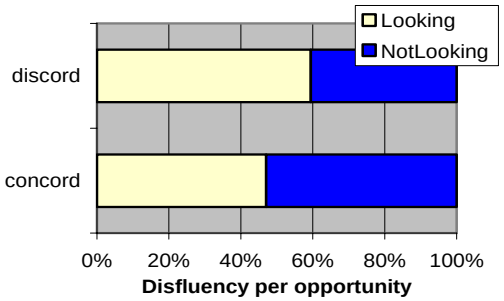


Figure 5. Rate of repair disfluencies per concordant or discordant feedback opportunity with respect to whether the Giver was either looking or not looking at the Follower. The difference is significant when the Giver looked at the Follower.

3.6 Disfluency Type, Gaze & Motivation

So far we have seen that speakers' gaze behaviour is not randomly distributed. It follows certainly problems (a Follower on-route who claims not to be) and ignores others (a Follower off-route who claims to be on-route). We have also seen that on those occasions when an instruction Giver actually takes in enough information to see what is amiss, he or she is more likely to speak disfluently. The question we ask here is whether these disfluencies are part of well formed communicative processes. If the information taken in by examination of listener feedback is properly processed by the speaker, what s/he says disfluently will be something appropriate to the situation. To determine whether this is really the case, it was necessary to classify utterances by their goal or motivation. To do this, the first author examined all 564 repetitions and 280 deletions occurring in the Dual Modality feedback condition.

The first stage of this process was to identify an interval for analysis. All dialogues were coded according to the HCRC Conversational-Game-Move coding scheme [8]. In this system, each turn is decomposable into conversational Moves, or sub-units of the dialogue. For example, a speaker might 'Instruct' by giving directions or 'Align' when noting that the Follower has gone astray. Analyses began with the Move that carried the disfluency. The coder searched backwards from the Interruption Point of the disfluency to the most recent Giver Move introducing a new landmark. The start time was considered to be the Giver's first mention of a new landmark while the end time was the Interruption point of the disfluency or for deletions, the end of the repair.

The second stage was to identify Giver gaze behaviours within these intervals. The gaze record of the speaker for this time-span was then checked and disfluency was coded as 'Looking' if there were any overlaps of Giver and Follower Gaze from the introduction of the landmark to the end of the disfluency. All others were coded 'Not Looking'.

Third, each disfluency was classified by Motivation, the content of the repair. Repetitions necessarily occur within the same dialogue Move, while deletions are almost always a single abandoned Move, so that the repair effectively lies in the next Move. Motivations were classified under two major

goals: either the speaker was 'confirming' that the Follower was at a correct or incorrect landmark or the speaker was 'reformulating' by adding, elaborating, or correcting information being transmitted. Examples of goal and disfluency combinations are given in Table 3 below.

Table 3. Examples of disfluencies by goal and type. For repetitions, both reparandum and repair appear in bold text. For deletions, just the reparandum appears in bold text since the repair is effectively non-existent.

Disfluency Type	Dialogue Goal	
	Confirmation	Reformulation
Repetition	'**That's, That's** just fine	'Eh you travel directly ehm sort of **north…north** and east'
Deletion	'So loop around the waterfall **over**….Yeah, there'	'Um **can you si-**…it's to the left of that'

Since appropriate confirmation of position should depend on the Giver actually determining where the Follower was, we would expect confirmations to accompany gaze at the follower. Since the arrival of the Follower at the goal or her movement off route should complete the execution of a series of instructions, all the Giver need do is cease instructing and declare the Follower to be right or wrong. Accordingly, deletion disfluencies are appropriate: in this view they mark a sequence of instructing, checking, and, finally, abandoning any ongoing instruction for a new a phase in the dialogue.

Our second goal category, reformulation, can also repair communication problems but by elaborating the material serving the current goal. Typically [14], speakers have to look away from their interlocutors when formulating complex material. Also on the grounds of complexity, we might expect not looking and reformulating to accompany repetition disfluencies [10].

Analyses of Giver's Gaze (2: looking vs. not looking), Motivation (2: confirmation vs. reformulation), Disfluency Type (2: repetition vs. deletion) and Time-pressure (2: timed vs. untimed) showed part of this pattern.

We predicted that reformulations would attract repetition disfluencies and confirmations would attract deletions. As Figure 6 illustrates, numerically repetitions (confirmation = 0.083; reformulation = 0.403) and deletions (confirmation = 0.245; reformulation = 0.186) worked as predicted ($F_1(1,34) = 59.60$, $p < .001$). The predicted effect of Motivation, however, was significant only for repetitions (F_1 (1,34) = 124.17, p < .001).

We predicted that looking at the feedback square would yield confirmations and not looking would accompany reformulations. In fact, only when Givers did not gaze at the Follower's square was the prediction met: there was a higher rate of reformulations than confirmations (Gaze x Motivation: *F(1,34) = 9.27, p < .01*, Bonferroni *t* at $p = .008$.).

Since we have an association between reformulations and repetitions, and one just reported between reformulations and not looking at the interlocutor, we tested for the effects within repetitions and deletions separately. Though the Giver tended not to look at the Follower square during repetition disfluencies, the trend is weak because it appears to hold only in the Verbal Group (Disfluency Type x Gaze: *F(1,34) = 3.59, p = .067;* Gaze x Motivation x Experiment: *F(1,34) = 8.62, p < .006;* Bonferroni at $p = .001$). For deletion disfluencies, the effect of gaze depends on motivation: deletions classified as confirmations were, as we predicted, more common when the Giver took the opportunity to look at the Follower (Bonferroni at $p = .008$), whereas deletions classed as reformulations showed an insignificant tendency to be more common when the Giver was not looking at the Follower (Motivation x Gaze: *F(1,34) = 8.61, p < .01).* Thus, there were associations between disfluency type and motivation type and between disfluency-motivation combination and gaze.

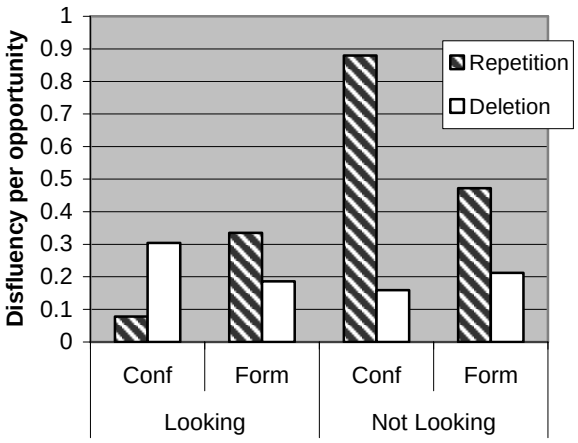


Figure 6. Rates of Repetitions and Deletions per opportunity with respect to Behaviour type, either confirmation (Conf) or reformulation (Form) and Gaze. The difference is significant for Repetititions but not for Deletions.

3. Discussion and Conclusions

Although the visual feedback provided the Giver with the Follower's exact location at any point during the interaction, this information had a cost. The Giver tended to gaze away from the Follower's location. Gaze aversion during difficulty is a common phenomenon found in conversational analysis and gaze studies [14, 15], and we find that gaze itself makes for production difficulty: speakers are more disfluent if they look at the follower feedback. Furthermore, Givers tended not to look at concordant negative feedback which clearly indicated trouble, though they did look at discordant feedback when the Follower was easily found – on the landmark being described.

When a Giver noticed this discordance, disfluency often occurred as result, presumably because the speaker was burdened with resolving the conflicting verbal and visual signals and in a sense handling the Follower's confusion. Disfluency, it seems, tend to co-occur first with uptake of the speaker's whereabouts and misalignment in dialogue, as predicted in [24]

If speakers are committed to tracking and accommodating listeners' knowledge [9, 10], and if repetitions indicate commitment to listener and message, Givers should visually attend to their Followers whilst making a repair: a committed speaker might be expected to assist a Follower who is clearly in difficulty by looking at the Follower's feedback and tailoring any following utterances to them. Instead, repetitions tended to associate with reformulation and thus by reformulation to gaze aversion during critical need. Looking at the follower instead accompanied deletions, as the Giver abandoned a Move in order to confirm or deny the listener's progress. Thus, it seems deletions, or false starts were associated with attending to the Follower but not with commitment to the utterance.

The present paper has added a psycholinguistic and dialogue perspective to the taxonomy of disfluency. We found that speakers are disfluent in different ways depending

upon the dialogue task in which they are currently engaged. The nature of listener feedback and the Giver's uptake of information about the listener both had effects.

4. Acknowledgements

The authors thank Yiya Chen and Catriona Havard for running the experiment and administering the gaze and transcription coding. This work was supported by EPSRC Research Grant GR/R59038/01 to Ellen Gurman Bard and GR/R59021/01 to Anne H. Anderson.

5. References

[1] Anderson, Anne H., Ellen Gurman Bard, Cathy Sotillo, Alison Newlands, & Gwyneth Doherty-Sneddon. 1997. Limited Visual Control of the Intelligibility of Speech in Face-to-Face Dialogue. *Perception and Psychophysics*, 59 (4), pp. 580-592.

[2] Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Gwyneth Doherty, Simon Garrod, Steve Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Cathy Sotillo, Henry S. Thompson, & Regina Weinert, 1991. The HCRC Map Task Corpus. *Language and Speech***,** vol. 34, pp. 352–366.

[3] Argyle, Michael & R Ingham. 1972. Gaze, mutual gaze and proximity. *Semiotica*, 6. pp. 289-304.

[4] Bard, Ellen Gurman, Anne H. Anderson, Marisa Flecha-Garcia, David Kenicer, Jim Mullin, Hannele Nicholson, Lucy Smallwood & Yiya Chen, 2003. Controlling Structure and Attention in Dialogue: The Interlocutor vs. the Clock. *Proceedings of ESCOP, 2003*, Granada, Spain.

[5] Bard, Ellen Gurman, Robin J. Lickley, & Matthew P. Aylett. 2001. Is Disfluency just Difficulty? *Proceedings of DiSS'01*, Edinburgh.

[6] Bard, Ellen Gurman, Anne H. Anderson, Cathy Sotillo, Matthew Aylett, Gwyneth Doherty-Sneddon & Alison Newlands. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, vol. 42, pp. 1–22.

[7] Brown, Gillian, Anne H. Anderson, George Yule, Richard Shillcock, 1983. *Teaching Talk*. Cambridge: Cambridge University Press.

[8] Carletta, Jean, Amy Isard, Steve Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, & Anne H. Anderson, 1997. The reliability of dialogue structure coding scheme. *Computational Linguistics*, vol. 23, pp. 13–31.

[9] Clark, Herbert H. & Meredyth A Krych. 2004. Speaking while monitoring addresses for understanding. *Journal of Memory and Language.* vol. 50, Issue 1, pp. 62-81.

[10] Clark, Herbert H. & Thomas Wasow, 1998. Repeating words in Spontaneous Speech. *Cognitive Psychology*, vol. 37, pp. 201–242.

[11] Exline, Ralph V., P. Jones, & K. Maciorowski. 1977. *Race, affiliation-conflict theory and mutual vision attention during conversation.* Paper presented at the meeting of the American Psychological Association.

[12] Fox Tree, Jean & Clark, Herbert H.. 1997. Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition. 62. pp. 151-167*

[13] Fox Tree, Jean. 1995. The effects of false-starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory & Language.* 34. pp. 709-738.

[14] Glenberg, Arthur M, Jennifer L. Schroeder & David A. Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory and Cognition.* Vol. 26, (4). pp. 651-658

[15] Griffin, Zenzi M., 2005. The Eyes are right when the Mouth is Wrong. *Psychological Science*. Vol 15, number 12, pp. 814-821

[16] Horton, William S. & Richard J. Gerrig. 2005. The impact of memory demands on audience design during language production. *Cognition*, vol. 96. pp. 127-142.

[17] Levelt, Willem J.M., 1989. Monitoring and self-repair in speech, *Cognition*, vol. 14, pp. 14–104.

[18] Lickley, Robin J. 2001. Dialogue Moves and Disfluency Rates. *Proceedings of DiSS '01*, *ISCA Tutorial and Workshop*, University of Edinburgh, Scotland, UK, pp. 93-96.

[19] Lickley, Robin J. 1998. HCRC Disfluency Coding Manual *HCRC Technical Report* 100. `http://www.ling.ed.ac.uk/~robin/maptask/disfluency-coding.html`

[20] Lickely, Robin J. 1995. Missing Disfluencies. *Proceedings of ICPhS*, Stockholm, vol. 4. pp. 192-195.

[21] Lickley, Robin J. 1994. *Detecting Disfluency in Spontaneous Speech.* PhD. Thesis, University of Edinburgh.

[22] Maclay, Howard & Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15, pp. 19-44.

[23] Nicholson, Hannele, Ellen Gurman Bard, Robin Lickley, Anne H. Anderson, Jim Mullin, David Kenicer & Lucy Smallwood, 2003. The Intentionality of Disfluency: Findings from Feedback and Timing. *Proc. Of DiSS'03, Gothenburg Papers in Theoretical Linguistics 89. pp.15-18*

[24] Pickering, Martin & Simon Garrod, 2004, Towards a mechanistic theory of dialogue: The interactive alignment model. *Behavioral & Brain Sciences.* 27 (2), pp. 169-190.

[25] Plauché, Madelaine & Elizabeth Shriberg, 1999. Data-Driven Subclassification of Disfluent Repetitions Based on Prosodic Features. *Proceedings of the International Congress of Phonetic Sciences*, vol. 2, pp. 1513–1516, San Francisco.

[26] Savova, Guergana & Joan Bachenko. 2002. Prosodic features of four types of disfluencies. *Proceedings of DiSS'03.* Gothenburg University, Sweden. pp. 91-94.

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

[28] Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, Julie Sedivy. 2000. Integration of visual and linguistic information in spoken language comprehension. *Science.* 268, pp. 1632-1634.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 139–144.

Lexical bias re-re-visited. Some further data on its possible cause.

*Sieb Nooteboom**

*Utrecht Institute of Linguistics OTS, Utrecht University.

Abstract

This paper describes an experiment eliciting spoonerisms by using the so-called SLIP technique. The purpose of the experiment was to provide a further test of the hypothesis that self-monitoring of inner speech is a major source of lexical bias ([1; 10; 11; 14]. This is a follow-up on an earlier experiment in which subjects were explicitly prompted after each response to make a correction in case of a speech error. In the current experiment both the prompt and the extra time for correction were left out, and there was no strong time pressure for the subject in giving his response. It is shown that under these conditions many primed-for spoonerisms are replaced by other, mostly lexical, errors. These 'replacing' or 'secondary' errors are more frequent in the condition priming for nonword-nonword errors than in the condition priming for word-word errors. Response times obtained for replacing errors are considerably and significantly longer than response times for overtly interrupted errors, and also longer than response times for the primed-for spoonerisms. This suggests that a time-consuming operation follows the primed-for spoonerisms in inner speech, and replaces those with other speech errors, often to preserve lexicality of the error.

1. Introduction

Lexical bias is the phenomenon that phonological speech errors tend to create more real words than nonwords, other things being equal. For quite some time there have been two competing explanations for lexical bias. One explanation, the so-called feedback explanation, is that it results from immediate reverberation of neural activation between the phoneme and the word form level in the mental production of speech [e.g. 2; 16]. Another explanation, proposed by those who reject the existence of immediate feedback between different levels of speech production, is that lexical bias results from self-monitoring of inner speech, nonwords being rejected and repaired more often than real words before they are uttered [10; 11; 12; 14]. Of course, these explanations do not logically exclude each other. Recently it was argued on the basis of careful experimenting that under certain conditions lexical bias has two sources, both immediate feedback and self-monitoring of inner speech [7]. The current paper focuses on self-monitoring. It provides experimental evidence for a hidden self-monitoring of inner speech by which primed-for nonword errors are either early interrupted immediately after pronunciation has begun, or replaced by real words in a time-consuming operation before pronunciation has started. The issue whether or not the same predicted data could also be explained by immediate feedback will come back later in this introduction, and also in the discussion section.

Recently it was found [12; 14] in an experiment eliciting spoonerisms with the so-called SLIP technique [1] that when nonword-nonword spoonerisms are primed for, there are significantly more early interruptions than when word-word spoonerisms are primed for. This was interpreted as evidence that in inner speech nonlexical errors are more frequently detected and rejected than lexical errors. If so, this would support a self-monitoring account of lexical bias in phonological speech errors.

In that experiment (to be called Exp03 from now on) subjects were explicitly prompted after each response to correct themselves if they detected a speech error. Hundred ms after the offset of the to-be-spoken word pair, a visible prompt to speak the last word pair seen aloud was presented (during 900 ms) followed by a blank screen (during 110 ms). After this prompt another visible prompt was presented (also during 900 ms) and also followed by a blank screen (during 110 ms), meant to elicit a correction in case of error. This procedure was meant to provoke corrections of complete spoonerisms (or other speech errors). In this respect the technique was not successful: Very few corrections of complete spoonerisms were made. Possibly, however, the instruction to correct any detected speech error, combined with the time-pressure in the experiment caused the subjects to pay special attention to speech errors in their inner speech, and made them reject nonlexical errors in inner speech more easily than lexical errors, employing a quick and dirty criterion of lexicality.

It was decided to run another experiment (Exp05) with the SLIP technique, this time without the prompt and the extra time for making corrections, but also with less time pressure. In most other respects the experiment was similar to Exp03. In this new experiment there was no signal before which the response had to be given other than the next word pair to be read silently. The first two word pairs presented were never followed by the prompt to speak the last word pair seen aloud. This meant that subjects soon detected they could relax during these first two word pairs. It was thought, in line with a suggestion by Hartsuiker et al. [7], that the absence of a time limit would decrease the number of early interruptions and increase the time-consuming contribution of self-monitoring inner speech to lexical bias. The relevant question here is how this contribution of self-monitoring would surface, if not in the number of interrupted nonword-nonword spoonerisms. In Exp03 and many similar experiments described in the literature, there were many cases where the primed-for speech error was not made, but instead another speech error was made showing the same exchange of initial consonants, as when the stimulus BAD GOOF does not turn into GAD BOOF, but into GAS BOOK instead. From the self-monitoring account of lexical bias, one may predict that (a) in the condition priming for nonword-nonword errors such replacing errors are far more frequent than in the condition priming for word-word errors, and (b) that such replacing errors are much more often lexical than nonlexical. Within this view, there are two successive errors being made in inner speech. The first error made is the one which was primed for (GAD BOOF), which then is rejected and replaced either by the correct target, or by another error (like GAS BOOK). If this view is valid, one predicts that (a) response times for errors like GAS BOOK are longer than response times for interrupted errors like G..BAD GOOF, and (b) response times for errors like GAS BOOK are longer than response times for

errors like GAD BOOF, because the first error derived from two successive operations, and the latter from only one.

The reader may note that models incorporating immediate feedback of activation between sounds and words as in [2; 16] would probably also predict that phonological errors like GAD BOOF might be replaced in inner speech by real word errors like GAS BOOK, because activation would not only reverberate between the fitting sounds and the correct targets BAD GOOF but also with these similar words, especially when the priming for the consonant exchange is strong. In the current view, however, there is supposed to be some kind of trade-off relation between early interruptions like G..BAD GOOF (that cannot be explained by immediate feedback and clearly result from self-monitoring inner speech), and errors like GAS BOOK, the former being made under time pressure, the latter being made instead of early interruptions like G..BAD GOOF, when subjects are more at ease and have more time. An important prediction from this view, as argued above, is that response times will be significantly longer for secondary errors like GAS BOOK than both for errors like G..BAD GOOF and for predicted, primed-for, exchanges like GAD BOOF. It is currently not clear what specific predictions could be derived with respect to these response times from models exhibiting immediate feedback. The reader may note, however, that if there is immediate feedback between sounds and word forms, this feedback is always there, and potentially affects response times of all reponses, also all correct and fluent responses. This is different from the effect of detecting and repairing speech errors in inner speech on response times. This effect should be only there when in inner speech a speech error has been made. This issue will come back in the discussion.

2. Method

The method used was basically the same as the one applied by Baars et al. [1]: Subjects were to read silently Dutch equivalents of word pairs like DOVE BALL, DEER BACK, DARK BONE, BARN DOOR, presented one word pair at the time, until a prompt told them to speak aloud the last word pair seen. However, there was no white noise applied to the ears of the subjects as in [1] and [7]. The reason white noise was not applied is that this would very likely make self-repairs of completed speech errors in overt speech rather scarce. I needed these errors, though, to support my claim that there are two classes of overt self-repairs, viz. self-repairs of errors in inner speech (G..BAD GOOF) and self-repairs in reaction to overt speech (GAD BOOF...BAD GOOF). In [14] it was demonstrated that two such classes can be separated on the basis of the distribution of offset-to-repair intervals. This issue will not return in this paper.

2.1. Stimulus material

Priming word pairs consisted of pairs of Dutch CVC words with a visual word length of 3 or 4 characters, visually presented in clear black capital print on a computer screen, in a white horizontally oriented rectangle against a greyish green background and intended to be read silently. In total there were 36 test word pairs, 18 potentially leading to word-word and 18 potentially giving nonword-nonword spoonerisms. The latter were derived from the first by changing only the final consonants (cf. [2]). Each word pair was either preceded by 3, 4, or 5 priming word pairs, chosen to prime a spoonerism, as in the sequence *give book, go back, get boot* preceding the test stimuli *bad goof*, or by 3, 4 or 5 non-priming word pairs, providing a base-line condition. In this experiment the priming word pairs were not preceded by additional non-priming word pairs, as was the case in Exp03 as an attempt to hide the purpose of the experiment from the subjects. Note also that the minimum number of precursor word pairs whether priming (preceding the test stimuli) or not (preceding the base-line stimuli) was 3, so that clever subjects could soon discover that they could relax during the first two precursor word pairs. The priming word pairs all had the reverse initial consonants as compared to the test word pair, and the last priming word pair always also had the same vowels as the test word pair. There were 2 stimulus lists, being complementary in the sense that the 18 word pairs that were primed for spoonerisms in the one list were identical to the 18 word pairs providing the base-line condition in the other list, and vice versa. In this experiment there were no fillers other than the base-line stimuli that were identical to the test stimuli in the other stimulus list.

The initial consonants of priming word pairs and test word pairs were chosen from the set /f, s, , v, z, b, d, p, t, k/. Each set of 18 word pairs was divided in 3 groups of 6 stimuli with equal phonetic distance between initial consonants, viz. 1 , 2 or 3 distinctive features. For example, /f/ versus /s/ differ in 1 feature, /f/ vs. /p/ differ in 2 features, and /f/ vs. /z/ differ in 3 features. After each test and each base-line stimulus word pair the subject saw on the screen a prompt consisting of 5 question marks: “?????” (cf. [2]). In addition to the set of test and base-line stimuli described so far there was a set of 7 stimuli with a variable number, on the average 4, of non-priming preceding word pairs to be used as practice for the subjects, and of course also followed by a prompt to speak.

2.2. Subjects

There were 102 subjects, virtually all being staff members and students of Utrecht University, all with standard Dutch as their mother tongue and with no self-reported or known history of speech or hearing pathology.

2.3. Procedure

Each subject was tested individually in a sound proof booth. The timing of visual presentation on a computer screen was computer-controlled. The order in which test and base-line stimuli, along with their priming or non-priming preceding word pairs were presented was randomized and different for each odd-numbered subject. The order for each even-numbered subject was basically the same as the one for the immediately preceding odd-numbered subject, except that base-line and test stimuli were interchanged. Each (non-) priming word pair and each “?????”-prompt was visible during 900 ms and was followed by 100 ms with a blank screen. The subject was instructed, on seeing the "?????" prompt to speak aloud the last word pair presented before this prompt. Fifty subjects were, after the practice word pairs, presented with list 1 immediately followed by list 2, the 50 other subjects were presented with list 2 immediately followed by list 1. This meant, of course, that each subject was presented with the same to-be-spoken word pair twice, once as a test stimulus and once as a base-line stimulus. The hope was that in this way more speech errors might be elicited than otherwise would be the case, and that there would be no significant difference in the data between the two stimulus lists. The advantage would also be that each subject would more or less serve as his or her own control, which is important because in this type of experiment subjects behave very differently. All speech of each subject was recorded, and digitally stored on one of two tracks of a DAT. On the other

track of the DAT a tone of 1000 Hz and 50 ms duration was recorded with each test or base-line stimulus, starting at the onset of the visual presentation of the "?????" stimulus. These signals were helpful for orientation in the visual oscillographic analysis of the speech signals, and indispensable in measuring response times.

2.4. Collecting the data

Reactions to all test and stimulus presentations were transcribed either in orthography, or, where necessary, in phonetic transcription by the present author using a computer program for the visual oscillographic display and auditory playback of audio signals. Response times for all correct and incorrect responses, to both base-line and test stimuli were measured by hand in the two-channel oscillographic display from the onset of the 50 ms tone (coinciding with the onset of the presentation of the visual "?????" prompt) to the onset of the spoken response. The onset of the spoken response was in most cases defined as the first visible increase in energy that could be attributed to the spoken response. However, the voice lead in responses beginning with a voiced stop was ignored because in Dutch duration of the voice lead appears to be highly variable and unsystematic both between and within subjects (cf. [17]), within the current experiment showing a range from 0 to roughly 130 ms. In those cases where (interrupted or completed) responses were followed by a (correct or incorrect) self-repair, the duration of the offset-to-repair interval (the interval between first and second response) was measured.

3. Results

3.1. Analysis of spoonerisms

The current design with less time pressure and no urge to correct speech errors led to only half the number of speech errors per test stimulus found in Exp03. The previous design provided 56 (3.1%) completed spoonerisms and 371 (21%) speech errors in total as responses to 1800 test stimulus tokens (36 test stimuli x 50 subjects), the current design led to 56 (1.5%) completed spoonerisms and 317 (8.6%) speech errors in total as responses to 3672 test stimulus tokens (36 test stimuli x 112 subjects). The average response time for fluent and correct responses to test stimuli was 527 ms in the previous experiment and 489 ms in the current experiment, suggesting that the subjects in the current experiment were less plagued by conflicting production patterns probably because priming for spoonerisms was less effective. As the main difference between the two designs was the presence or absence of a prompt and extra time for correction, it seems that the explicit need to correct in Exp03 provided extra mental stress and led to relatively many speech errors.

The first issue in analyzing the results was if the rather unorthodox decision to present each subject two times with each stimulus word pair, once in the first stimulus list presented and once in the second stimulus list presented, albeit in different contexts (priming for a spoonerism or not), led to different patterns for the first and second presentation. This was not the case. There was no significant difference in the pattern of speech errors between the first and second presentation, neither was there a significant different in the pattern of response times between the first and second presentation. Therefore it was decided to analyze the results of the two presentations together.

Given the rather low number of speech errors, one could doubt whether the test conditions as compared to the base-line condition were effective enough. The relevant data are given in Fig. 1.

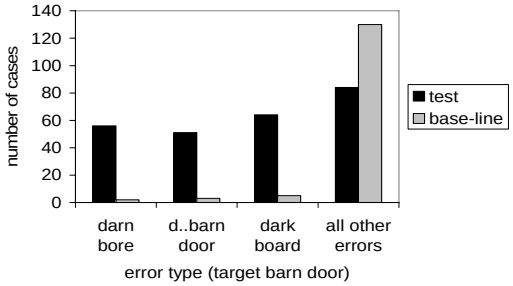


Fig. 1. Number of speech errors of different types separately for the test conditions and the base-line condition (df=3; $chi^2 = 140$; $p<0.001$)

Obviously the relative effectiveness of priming was good enough. In Exp03 there was a significant lexical bias, word-word complete spoonerisms being much more frequent than nonword-nonword spoonerisms. This was at least partly compensated by early interrupted spoonerisms being much more frequent in the condition priming for nonword-nonwords spoonerisms than in the condition priming for word-word spoonerisms. The data for both experiments are given in Fig. 2.

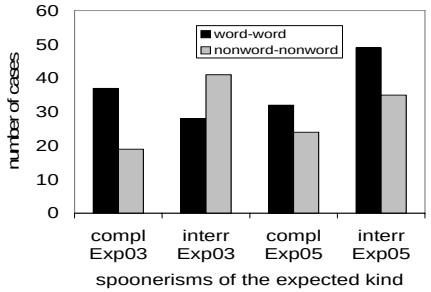


Fig. 2. Completed and interrupted spoonerisms of the primed-for kind in Exp03 and the current experiment. In Exp03 the distributions of completed and interrupted errors differed significantly between the test conditions, in Exp05 they do not.

Against expectations, in the current experiment there is no significant lexical bias, at least not in the spoonerisms that are fully identical with the primed-for spoonerisms, and no complementary distribution of interrupted errors. Also, it was expected that in Exp05 there would be relatively less interrupted exchanges than in Exp03, but there are more. However, these do not show the interaction with word-word versus nonword-nonword priming that was found in Exp03. It is also noteworthy that, if we take completed and interrupted responses together, in Exp '05 there seem to be relatively few responses to stimuli priming for nonword-nonword. There are only 44 such responses whereas there are 63 responses to stimuli priming for word-word errors. In Exp03 this was 58 as compared to 65. This suggests that somehow responses to stimuli priming for nonword-nonword errors got lost in Exp05.

In Fig. 2 complete spoonerisms were considered to be only those spoonerisms that are fully identical with the primed-for spoonerisms, because it was thought that other exchanges of initial consonants were not controlled for lexicality. As mentioned in the introduction, in most such experiments described in the literature, in order to make up for low numbers of errors, to begin with Baars et al. [1], and recently in Hartsuiker et al. [7], complete spoonerisms include other full and partial exchanges of the two initial consonants. In Fig.

1 it became clear that the relative frequency of such other exchanges (the DARK BOARD responses to BARN DOOR) is controlled by the priming versus base-line conditions. Might it be the case that the missing responses in the nonword-nonword condition are hiding in these "other exchanges" that were removed from further analysis? Would there perhaps be significantly more lexical "other exchanges" in the nonword-nonword condition than in the word-word priming condition?

If in inner speech nonword errors are indeed more frequently replaced by other, possibly lexical, errors than real-word errors, according to the current argument, there should be more replacing exchange errors in the condition priming for nonword-nonword errors than in the condition priming for word-word errors. As the argument is about elicited speech errors in inner speech, the analysis checking its validity should be limited to those responses where one be can be reasonably sure that the attempt to elicit a consonant exchange was initially (that is in inner speech) successful. To that end we assume that all responses not being expected spoonerisms that start with the initial consonant of the second stimulus word fall in that category. This would include DARK BOARD and DARK DOOR for BARN DOOR, but not BARK DOG for BARN DOOR. Note that these cases are counted irrespective of the source of the replacing words. Also clear intrusions from earlier parts of the experiment are counted. What is relevant to the present argument is not the source of the replacing word (for the possible source of these secondary errors, see the discussion), but rather whether the number of replacing words is controlled by the priming condition. As it happens, in Exp05 such cases number 22 in the word-word and 50 in the nonword-nonword condition. Now we count the number of these cases where at least one word is replaced by a real word, and the number of these cases where at least one word is replaced by a nonword. This leads to the data in Fig. 3. These data include 3 cases, 2 in the word-word and 1 in the nonword-nonword condition, where 1 word was replaced by a real word and the other by a nonword. The comparable data for Exp03 are also given.

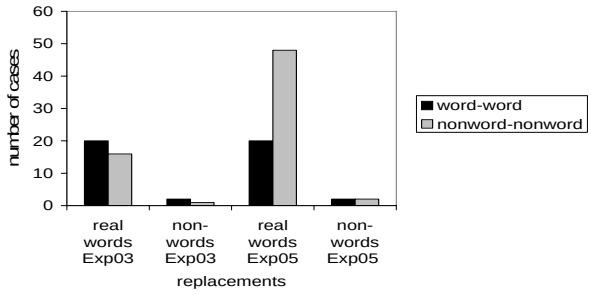


Fig. 3. Numbers of cases where a primed-for spoonerism is turned into another speech error by replacing 1 or 2 (non)words by other real or nonwords.

It seems that we have found here our missing responses in the nonword-nonword condition of Exp05. Whereas in Exp03 responses were significantly more often interrupted in the nonword-nonword than in the word-word condition, in Exp05 elicited speech errors are significantly more often replaced by other, lexical, speech errors in the nonword-nonword than in the word-word condition. If we only look at the number of replacements in the current experiment, 52 for the nonword-nonword condition and 20 for the word-word condition, this difference is highly significant on a simple sign test ($p<0.0001$), whereas in Exp03 the difference, if anything, goes the other way. If we look only at the numbers of replacing real words, and forget about the very low numbers of nonword replacements, the distributions are significantly different for the two experiments (df= 1; $chi^2=6.797$; $p<0.01$).

The strategy of subjects to interrupt and repair nonword-nonword spoonerisms more often than word-word spoonerisms in reaction to detecting such errors in inner speech, found in Exp03, seems to be replaced in the current experiment by a strategy to replace nonword speech errors in inner speech more often than real word errors by real words before any response is given. The combined data of the two experiments suggest that there is a trade-off between early interruption and replacement by real words of nonword errors. Possibly, this trade-off is controlled by the difference in the degree of time pressure in the task of the subjects.

3.2. Supporting evidence from response times

So far, the current analysis works from the assumption that if under the conditions of priming for spoonerisms, another (partial) exchange error than the primed-for spoonerism is found, this is the result from two successive processes, the first one creating the primed-for spoonerism in inner speech, the second rejecting this spoonerism or one of its words, replacing it by another error, before pronunciation is started.

This is different from what happens when the primed-for spoonerism is interrupted and overtly repaired after its overt production has started, as in G..BAD GOOF. In the latter case the repair takes place openly, so it cannot consume part of the response time before any overt speech act takes place (note also that the very fact that in cases like G..BAD GOOF speaking the erroneous form is initiated might indicate that speech production is started too hastily, before the self-monitoring of inner speech has had a chance to detect and repair the error).

It is also different from the situation where the primed-for spoonerism is actually made, and not replaced by another speech error, because here also only one of the two processes takes place before the response is given. If the current reasoning is valid, one thus expects that response times of errors like GAS BOOK for BAD GOOF are longer than response times for errors like G..BAD GOOF, or GAD BOOF for BAD GOOF, because the first case involves two consecutive error-producing processes and the last two cases only one. The relevant data are given in Fig. 4.

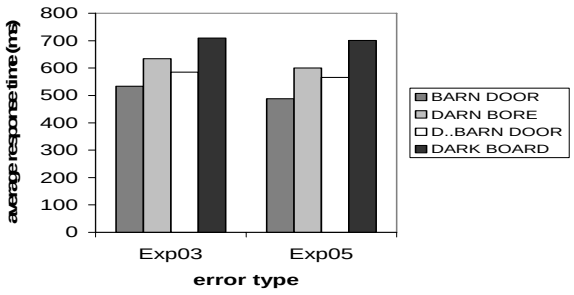


Fig. 4. Average response times for four types of responses, viz. fluent and correct responses (BARN DOOR), spoonerisms that are fully identical to the primed-for spoonerisms (DARN BORE), early interrupted spoonerisms (D..BARN DOOR), and full or partial exchanges that deviate from the primed-for spoonerisms (DARK BOARD). Data separately for Exp03 and Exp05.

Separately for each experiment response times were submitted to a univariate analysis of variance with type (BARN DOOR vs DARN BORE vs D..BARN DOOR vs DARK BOARD in response to BARN DOOR) and priming condition (word-word versus nonword-nonword) as fixed

factors. There was no significant effect of priming condition, nor a significant interaction. There was however, in both experiments a significant main effect of error type ($p<0.0001$ for both experiments). Of course this main effect is mainly due to the fact that correct and fluent responses are faster than erroneous responses (Obviously, within the category of correct and fluent responses, there hide a number of cases where the primed-for spoonerism was made in inner speech and then corrected before pronunciation. In these cases, response times are potentially as least as long as for the replacing speech errors. But as speech errors are rare, these cases are not many, and do not contribute much to the average response time for correct and fluent responses). Apparently, making an error costs time. The main prediction is about differences in response times between erroneous responses. In Exp05 a Tukey test showed that correct and fluent responses (BARN DOOR) differed significantly from all other response types, predicted spoonerisms (DARN BORE) did not differ from interrupted spoonerisms, but did differ significantly from replacing errors (DARK BOARD), which had longer response times. In fact, replacing errors had significantly longer response times than all other error types. The pattern was very similar for Exp03, except that correct responses did not differ significantly from interrupted responses, and replacing errors did not differ significantly from predicted spoonerisms, but did differ significantly from both correct responses and interrupted spoonerisms.

These data suggest that the replacing speech errors (DARK BOARD) result from a time-consuming (on the average some 100 ms in Exp05 and some 80 ms in Exp03) self-monitoring operation in inner speech, during which the primed-for spoonerism is rejected and replaced with another speech error that is nearly always lexical. Rejection of the primed-for speech errors in inner speech, preceding the hidden replacement, obviously employs a criterion of lexicality in Exp05 but not in Exp03, whereas overt early interruption of the primed-for speech errors, as we have seen, employs a criterion of lexicality in Exp03 but not in Exp05. Note also that interrupted speech errors have relatively short response times as compared to the other error types, as if pronunciation was started too hastily, making interruption necessary for self-repair.

4. Discussion

The main findings of the present experiment are to some extent unexpected. In experiments employing the SLIP technique, lexical bias in phonological speech errors has been demonstrated to be a rather robust phenomenon, in most experiments leading to more full exchanges of initial consonants in word pairs when lexical spoonerism are primed for than when nonlexical spoonerisms are primed for (e.g. [1; 3; 7; 8; 14], but see [2]). It has also been shown that lexicality of the first error word is the main determinant of lexical bias [8]. Lexical bias has also been demonstrated in spontaneous speech errors ([4; 14] but see [5; 6]). Although the pattern of complete spoonerisms in the current experiment basically corresponds to the common pattern, the difference was not significant. Also, where in Exp.'03 the common pattern of lexical bias in completed spoonerisms was mirrored by a greater number of interrupted spoonerisms when nonlexical than when lexical spoonerisms were primed for, this pattern was completely absent from the current data.

The data of the current experiment, particularly when compared with the data of Exp03, strongly suggest that the strategies of the subjects in experiments with the SLIP technique are very sensitive to differences in design and task. Simply by removing the visible prompt and extra time for correction after each response, subjects reacted in general faster and made only half the number of speech errors they made in Exp03. Apparently, they were more at ease. Although subjects made more, not less interrupted spoonerisms, they obviously did not employ a criterion of lexicality in overtly interrupting speech errors, as they seemed to do in Exp03. The criterion of lexicality was definitely there, though. In the current, obviously more relaxed conditions, nonlexical spoonerisms were much more often rejected and "repaired" in inner speech than lexical spoonerisms, where "repaired" here refers to cases where the outcome is a new speech error, not identical to the target. A lexical bias in producing these secondary errors is overwhelmingly present, both in the sense that such "repairs" more often occur in the condition priming for nonlexical than in the condition priming for lexical spoonerisms, and in the sense that these secondary speech errors are virtually always lexical themselves. The assumption that these secondary errors are made only after the primed for spoonerism has been rejected is strongly supported by the considerable and significant difference in response times between interrupted spoonerisms and secondary speech errors.

The rejection of nonword-nonword spoonerisms in inner speech became observable in Exp03 in the distribution of early interrupted spoonerisms. In Exp05 the rejection of nonword-nonword spoonerisms in inner speech is observable in the number of primed-for spoonerisms that are replaced with alternative speech errors. This difference is also reflected in the much greater effect of error type on response times in Exp05 than in Exp03. Primed-for nonword-nonword spoonerisms that in Exp03 were interrupted under the time-pressure resulting from the prompt to correct, were under the more relaxed conditions of Exp05 replaced with alternative lexical errors. This finding provides further evidence for self-monitoring being the main cause of lexical bias.

Unavoidably, the question should be asked whether the same data could also be explained by immediate feedback of activation between phoneme level and word form level in speech production ([2, 4, 16]). Obviously, such feedback could in principle generate errors like GAS BOOK for BAD GOOF and DARK BOARD for BARN DOOR, as there would be reverberation between the active phonemes GA.. BOO or DAR...BO..(after the phoneme exchange has been made), and these words. That errors like GAS BOOK and DARK BOARD are more numerous in the condition priming for nonword-nonword spoonerisms than in the condition priming for word-word spoonerisms, might be explained in a feedback account by the presence or absence of competition with the elicited speech errors: In the nonword-nonword condition there is no such competition because nonwords are not represented in the lexicon. However, there are two arguments why the current data reflect self-monitoring rather than feedback. One argument is the trade-off between an effect of priming condition (nonword-nonword vs word-word) on early interruptions in Exp03 and on the number of secondary lexical errors in Exp05. This trade-off demonstrates that the strategies of the subjects are highly variable and influenced by the precise task structure. Such variability one rather expects from semi-conscious self-monitoring that is controlled by focus and level of attention than from immediate feedback of activation within the mental production of speech, that is supposed to be automatic and more or less indifferent to attentional control.

The other somewhat related argument is from the distribution of response times. Unfortunately, the models in

[2; 16] were not set up to predict response times. However, the distribution of response times shown in Fig. 4 strongly suggests that in these experiments response times are mainly a function of whether or not a speech error has been made in inner speech, and whether or not this speech error has been rejected and replaced by another speech error before speech is initiated. In the majority of cases, where responses are fluent and correct, response times remain much shorter and show no or hardly any effect of priming condition. The differences in Fig. 4 between fluent and correct responses on the one hand and secondary speech errors like DARK BOARD on the other are in the order of 200 ms. This seems to reflect the working of a repair strategy that only becomes operative when an error has been detected in inner speech. Of course, this does not exclude that there is immediate feedback of activation in speech production, nor that the current data are affected by such feedback. Note, however, that a potential effect of immediate feedback on response times would not be limited to cases where a speech error had been made in inner speech. Immediate feedback is supposed to be automatic and always present. Indeed, earlier a small but significant effect of priming condition on response times of correct and fluent responses was found, that could possibly be attributed to automatic feedback between sound level and word form level ([12]). But the current data supply a link between the detection and repair of speech errors in inner speech on the one hand, and differences in response times that are much greater than the differences discussed in [12] on the other. These findings can easily be accounted for by assuming that self-monitoring of inner speech for speech errors employs a criterion of lexicality, and that the choice of a repair strategy is strongly influenced by the task structure.

One point remains, however, bringing up the issue of feedback again. If subjects so frequently replace the words and especially the nonwords of an elicited speech error with other words, where do these other words come from? Many (but far from all) of these words used in secondary speech errors are words intruding from earlier parts in the experiment. Supposedly these are still relatively active (cf. [9]). In those cases, which are many, where these words share phonemes with the correct target and/or the elicited error in inner speech, possibly these phonemes contribute to provide extra activation, and, particularly in the absence of syntactic and semantic constraints, the words concerned may then "fire" and become rapidly available for pronunciation. However, in order for phonemes to contribute to the activation of intruding word forms, there must be some kind of feedback between phonemes and word forms. This would of course easily be accommodated by models like in [2; 16] incorporating immediate feedback. There is another way, however. Roelofs [15] has suggested that there may be feedback via the inner perceptual loop employed by self-monitoring. This would make it possible that the phonemes of rejected words in inner speech contribute to the selection of other, replacing, words. Such a mechanism would make a major contribution to lexical bias in speech errors.

Acknowledgement
Thanks are due to Theo Veenker for his help in setting up the experiment, to 112 subjects who suffered a bit from being forced to make speech errors, to Hugo Quené for his help in analyzing and interpreting the data, and to Hugo Quené and Anne Cutler for critical comments on an earlier version.

5. References

[1] Baars, Bernard B.J., Michael T. Motley & Donald G. MacKay. 1975. Output editing for lexical status in artificially elicited slips of the tongue. Journal of Verbal learning and Verbal Behavior 15, 382-391

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

[3] Dell, G.S., 1990. Effects of frequency and vocabulary type on phonological speech errors. Language and Cognitive Processes, 5, 313-349.

[4] Dell, G.S., Reich, P.A., 1981. Stages in sentence production: An analysis of speech error data. Journal of Verbal learning and Verbal Behavior 20, 611-629.

[5] Del Viso, S., J. M. Igoa, J.E. Garcia-Albea, 1991. On the autonomy of phonological encoding: evidence from slips of the tongue in Spanish. Journal of Psycholinguistic Research 20, 161-185.

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

[7] Hartsuiker, Rob, Martin Corley & Heike Martensen,2005. The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related Beply to Baars, Motley, and MacKay 1975. Journal of Memory and Language 52, 58–70.

[8] Humphreys, Karin. 2002. Lexical bias in speech errors. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

[9] Kolk, H.H.J., 1995. A time-based approach to agrammatic production. Brain and Language 50, 282-303.

[10] Levelt, Willem J.M.. 1989. Speaking. From intention to articulation. Cambridge Massachusetts: The MIT Press.

[11] Levelt, Willem J.W., Ardi Roelofs, Antje S. Meyer, 1999. A theory of lexical access in speech production. Behavioral and brain sciences 22, 1-75.

[12] Nooteboom, Sieb G., 2003. Self-monitoring is the main cause of lexical bias in phonological speech errors. In: Eklund, R. (Ed.), Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop. 5–8 September 2003, Göteborg University, Sweden. Gothenburg Papers in Theoretical Linguistics 89, ISSN 0349–1021, pp. 25–28.

[13] Nooteboom, Sieb G., 2005. Listening to one-self: Monitoring speech production. in Hartsuiker, R, Bastiaanse, Y, Postma, A., Wijnen, F. (Eds.), Phonological encoding and monitoring in normal and pathological speech, Hove: Psychology Press.

[14] Nooteboom, Sieb G., in press. Lexical bias re-visited. Detecting, rejecting and repairing speech errors in inner speech. To appear in Speech Communication.

[15] Roelofs, Ardi. 2005. Planning, comprehending, and self-monitoring. In: Hartsuiker, R, Bastiaanse, Y, Postma, A., Wijnen, F. (Eds.), Phonological encoding and monitoring in normal and pathological speech, Hove: Psychology Press, pp. 42-63.

[16] Stemberger, Joseph P., 1985. An interactive activation model of language production. In: Ellis, A.W., (Ed.), Progress in the psychology of language Vol. 1, pp 153-186. London: Erlbaum.

[17] Van Alphen, P.M., 2004. Perceptual relevance of prevoicing in Dutch. Unpublkished doctoral thesis, Radboud university.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 145-149.

The re-adjustment of word-fragments in spontaneous spoken French

Berthille Pallaud

Parole et Langage, Université de Provence, Aix-en-Provence, France

Abstract

A study of word-fragments in spoken French has been undertaken for a few years on the basis of non directive talks corpora recorded and transcribed according to GARS'conventions (DELIC currently). These disfluencies are often analyzed within the framework of disfluent repetitions. The observations made on these two types of disfluencies led us to distinguish them. The aim of our study is to describe on the one hand insertions which take place in relation to the word interruptions and their re-adjustment, and on the other hand, to specify the types and localizations of retracing which follow these interruptions. Two kinds of incidental clauses were observed at the time of the readjustments which follow these disturbances. Some, (the more numerous) are syntactically linked to the fragment or with its retracing, others are not. Moreover, the word-fragments which will be modified are the only one to be dependent on the type of localization. For the others, this localization does not make it possible to predict the category of interruption (complemented or unfinished). Our results on word-fragments, confirm however that in contemporary French, the retracing at the head of the nominal or verbal group which contains the disfluency remains the simplest example (at the same time the most frequent, [5]. Nevertheless, a third of the retracing either does not go back to the beginning of the Group, or exceeds it.

1. Introduction

If the fluidity of an oral statement is measured by the rhythmic regularity in its production, it is clear that the statement is not fluent but disfluent [14, 12]. All speakers produce oral speech with a certain variability in the flow, pauses, whether they be silent or not [6], and in the lengthening of linguistic elements, some of which were studied at the point where the statement was interrupted, either in the middle of a word, such as in the case of a word-fragment, or at the boundary of words, such as syntagm interruptions which were or were not followed by word repetitions [15]. These interruptions are also characteristic of oral disfluencies. This study, using the grid paradigm [2], in fact has highlighted the progression of the statement by successive syntactic steps. These "halts" in the production of the text were due not only to silent or filled pauses, known as hesitation pauses [6], but also to word repetitions and fragments.

These last two types of stumbling, that is to say disfluencies, have rarely been distinguished (2,5). So, we thought it was important that they be characterized separately, and the big difference in frequency suggested that there were two kinds of phenomena. The results already obtained confirmed this hypothesis; in fact, they affected neither the same syntactic places nor the same grammatical categories [10, 11].

The aim of this study on word-fragments was to specify their syntactic position and the place in the statement where the subject returned to when it happened. Shriberg & Stockle [14] showed that by repeating words (where there is always a return to the statement) the speaker tended to go back to the beginning of the word with which he had problems formulating. This question regarding word-fragments, will only be considered from the position (of the word and the return) and morpho-syntactic and syntactic aspects.

2. Corpus and methodology

The study on spoken French throughout France, headed by Claire Blanche-Benveniste, was carried out in 1998 and 1999 on 20 corpora, the results of which were collected, with the exception of one, by the GARS. The entire survey was followed and digitized - the sound and transcription - by the DELIC team. In accordance with GARS, when transcribing conventions, which enjoin an orthographical transcription of the oral statements, the word- fragments are noted using a hyphen attached to the fragment of the word, which is then detected automatically.

All but two recordings were done in private, using non-directive interviews and only two speakers. Only two corpora were recorded in public, which consisted of improvised talks in front of a group of 40 people. All the speakers were adult.

These corpora were not labelled. The analysis was carried out exclusively on the statements of the interviewed speaker. The 441 extracted pieces of data from these corpora were then labelled (not automatically) and entered on a spreadsheet (using Contextes software by J Véronis and Excel). Our study was based on a medium flow of 200 words/min, with the total duration of the group lasting 7 hours 51 min. The average length of time of the GARS corpora was 3,080 words, thus an average duration of 16 min with the extreme values of these durations being 1,307 and 4,931 words. Hence, we found in this corpus subset, the average frequency of the apparition of word-fragments in a statement was 1/57sec - from 1/23 sec. to 1/8 min.

3. Results

3.1. Types of word-fragments

Three types of word-fragments were distinguished [10] in relation to the syntactic place occupied by what followed the word-fragment. When the following element occupied the same syntactic place, this insistence could result in completing the word-fragment (completed fragments) or replacing it (modified fragment). If the element, which followed the fragment, belonged to another syntactic place, the fragment of the word was left unfinished (unfinished fragments). A little more than half of the word-fragments were completed; modified and unfinished fragments constituted the remainder equally:

Ex. 1 completed fragment: EDF, 20,5 qui ont la possibilité de **r-** de **remonter** euh autour d'un axe

Ex. 2 modified fragment: EDF, 19,5 5 en revenant à nos **ber- Bassins** Versants Intermédiaires qu'on a qu'on par- qu'on *a* parlés précédemment euh

Ex. 3 unfinished fragment: EDF, 15,2 2 toutes les usines nous appellent nous **di-** et nous donnent par forme de fax ou d'e-mail euh

The completed and modified fragments were the only cases where the speaker resumed or "repaired" the flow of his/her speech by going back to the interrupted part of the statement. The unfinished fragments were those where the speaker simply continued the statement without any reparation or retracing.

The proportions in these three categories showed that these involuntary truncations appeared a lot more frequently as hesitation markers when developing the text than as the sign of an error to be corrected. Our observations confirmed the studies of Schegloff et al. [13] concerning auto-corrections (completed fragments in our study) and Cappeau's findings on syntagm fragments [4].

Contrary to the disfluent repetitions [11], these hesitations were mainly related to the nominal or verbal lexicon (70%) and much less to the functional word category. In the same way, the hesitation was more frequently a repetition than a word-fragment before the verb. The functional-words were significantly less frequently modified and more often left unfinished.

When the fragment was repeated, it was, in 82% of the cases, completed (instead of 59%) and barely modified or left unfinished. This last, very different result (kh2 = 29.85; d.d.l.= 2; p<.001) of what was observed on the simple fragment justified an in-depth study of the retraces and readjustments after the interruption of the word as the repetition of the fragment seemed to allow the speaker to complete more frequently the started word, and less frequently to modify it or leave it unfinished.

3.2. Analyses of re-adjustments after the interruption of words

Two types of phenomena could be defined: firstly, linguistic **insertions** of elements, and secondly, **retracing** of the statement (whether it be by syntactic insistence or by the continuation of the statement on the syntagmatic axis). The term "retracing" was only employed if an insistence occurred in the same syntactic place (thus the case of the completed or modified word-fragments).

3.2.1. Insertions

The structure of the disfluencies [5, 14) which differentiated the three phases in stumbling revealed two possible spaces of insertion following the truncation: the space which preceded the retracing of the statement and that which began with the retracing of the statement.

The three phases of this word-fragment structure are as follows:

* **the reparandum** (RM): indicates the word-fragment The interruption point (IP): establishes the final boundary of *the reparandum*: (it is identified by the hyphen in our transcriptions).
* The **interregnum** (IM, *the hyatus* in [5]): indicates the moment between the final boundary of *the reparandum* and the initial boundary of *repair*.
* The **repair** or **reparans** (RR): represents the repaired, repeated or modified part of the *reparandum*;

Thus, it is a 3-phase structure: **interruption, latency and the retracing of the statement:**

C24Bnanc, 24,2 ça **dép-** (Reparandum RM, IP)
ouais [Space IM]
ça *euh* dépend (Reparans, space RR**)**
+ moi au début j'avais demandé à aller en au Népal

Two kinds of insertions were distinguished:

- Firstly, interpolated clauses which were inserted into the statement without being syntactically linked (pauses voiced or not, enunciator and parenthetic clauses).
- Secondly, interpolated clauses syntactically linked to the interrupted statement and its retracing: the repeated word-fragments (IM space only), repetitions, added, removed or replaced items (RR space).

3.2.1.1 The first space ([IM space] "Interregnum") was only filled in 13 % of the cases.

Enunciator insertions (voiced pauses or not, *enfin, ben, bon*, etc), parenthetic items and repetitions of the word-fragment were observed:

Ex 4 EDF, 14,4 c'est qu'au premier choc euh **prétro-** [*euh*] **pétrolier** euh

Ex 5 C6bBesan 2,8, et j'ai **regard -** [*et j'ai en fait choisi ça par hasard*] j'ai **regardé** ce qu'il fallait avoir

Ex 6 CäBelfo, 3,2, euh et **j'ai-** [*j'ai*] **j'aimais** pas quoi

The moment, which followed the interruption of word, was thus a potential space for enunciator or parenthetic insertions. However, in our corpora, it was only "used" by the speaker in one out of 10 cases. Besides, if we had taken out repetitions of word-fragments, which was rare, we would never have observed sequences of successive truncations as one can find in the case of stammering [21, 1, 7]

3.2.1.2. The second space ([RR space], "Reparans") began with the retracing of the statement which re-established the continuity in its development where it had initially been interrupted. We found the same insertions as described in IM space, although they were fewer:

(parenthetic clause) EDF, 18,5 pour euh mettre les **ni-** [les *comment* les **niveaux**] euh d'eau corrects
(silent pause) EDF, 13,2 mais la co-génération restera **u-[une + une] solution**
(repetition) C24bNan, 2,1 c'était un peu euh un peu **com- comment *comment*** vivre au quotidien

In this second space (RR), it was assumed that it could be possible to find the same insertions In fact, they accounted for only 3%. We found in particular, insertions which would modify the interrupted statement by this truncation, which was not therefore, the moment of correction. Thanks to these "back modifications", this space comprised of, in total, two times more insertions than IM space, that is to say, 22% to 13%. There were three types of modulating statement insertions: term additions, suppressions, and replacements.

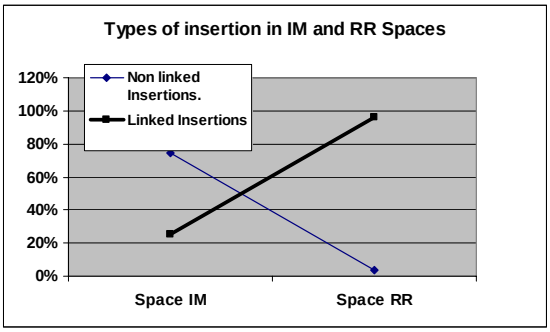


Figure 1 Insertion types linked or not syntactically to the interrupted statement in IM and RR spaces

Ex 7 Interpolated clauses which modulate the statement
*Retracted item C24bNan, 20, 1 il y **en** av- il y avait deux filles de dix-huit ans
*Added item C7cBorde, 12, 11 le judoka commence à être pas mal et euh s'il a s'il s'en- **continue à** s'entraîner
*Replaced item EDF 2,41 qui est un peu **plus** mo-un peu **moins** modulable

In total, a third of the word-fragments were followed by various insertions, in IM and RR spaces, and this in fact occurred mainly in the completed word-fragment category. As our observations on the repeated fragments suggested, these insertions, encompassing especially the completed fragments, seemed to play a role in the lexical research by facilitating it. These two spaces (Fig.1) were not composed of the same types of insertion. The space located just after the truncation seemed to be where the non-syntactically linked interpolated clauses were located, whereas they were practically absent when the retracing started. However, in the RR space, the majority of the interpolated clauses adjusted the retracing of the statement.

3.2.2. Retracing

Retracing after interruption only occurred where fragments were completed or modified, these being the only word-fragments followed by an insistence on the same syntactic place (i.e. a retracing). The unfinished fragments were precisely identified because what followed the fragment did not belong to the same syntactic place. The re-adjustment in this case was not a retrace but a continuation of the statement. The proportions of these three word-fragment categories validated the conclusions of Levelt [8] on the phenomena of an interruption in statements where the speakers retrace more than they continue their statement after an interruption: 78% to 22%. We were interested in the position of the interrupted word and the retracing which followed.

It was not possible to describe the exact place of retracing after an interruption without specifying beforehand, **the place of this interruption,** which might or might not have taken place at the beginning of the nominal or verbal group: [1]

At the beginning: C24aNanc 1, 3 euh la mygale **s' – s'arrime** avec ses ses crochets sur sa sa proie

Not at the beginning: C7dBord 5, 5 et et après j'ai **vou- voulu** changer

Clark and Wasow [5] stated that retracing after the interruption of a constituent, whether it be a group or a syntagm, more frequently occurred at the beginning of this constituent. If this was the same for the word-fragment and the latter was already at the beginning of the group, then the place where retracing would begin would not have the same signification as it would if the fragment was not already there. Indeed, if the word-fragment was not at the beginning of group but the retracing went back to this point, this would reveal a linguistic constraint which would be impossible to show as the fragment would already be at the beginning of the group.

3.2.2.1. Localization of the word-fragment

Out of the 436 word interruptions where it was possible to determine if they took place at the beginning of the group or not, more than two thirds of them did not (72%): in fact, there were no differences between the nominal and verbal groups. Hence, a large majority of these disfluencies (in French the lexicon is seldom found at the beginning of a group, but is preceded by a determinant or a pronoun) occurred later in the nominal or verbal group.

3.2.2.2. Localization of the retracing which followed a word-fragment which was not at the beginning of a group.

The mass result on the development of the word-fragment was that when this fragment was completed, it was always (without any exception in our corpora) through at least a minimal retracing of the fragment. There was never a simple completion of the fragment by the missing fragment (of the type: **un li vre**). This was another "qualitative" feature which made it possible to distinguish a disfluent statement from a statement produced by a person who stammers [16].

In addition, we noted that the localization of the disturbance did not make it possible to predict if there would be a retracing or a simple continuation of the statement, and in fact, wherever the fragment was localized, we found a similar proportion (one out of five) of unfinished fragments (without retracing). It was the same for the fragments which would be completed. Only those fragments which would be modified were significantly more numerous as the disturbance was not located at the beginning of the group ($kh2 = 10,46$; $p<.01$; d.d.l.=2):

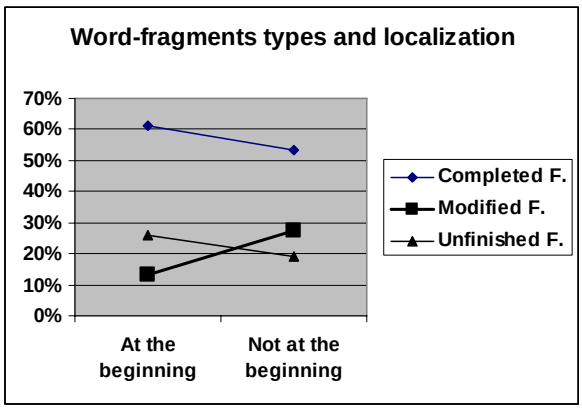


Figure 2 Localization of the word-fragments (whether or not at the beginning of a group in relation to the word-fragment types)

Moreover, the types of observed retracing could be regrouped into three categories:
* "**Minimal** " retracing which did not go back to the beginning of the group: CorpusEDF, 12,6 la note est est salée hein la note **énergé- énergétique**
* "**Beginning of the Group** "retracing which went back to the beginning of the nominal or verbal group: C5cBelfo, 2,2 alors ce monsieur s'est approché et il a dit ***vous* li- *vous* lisez** l'Est Républicain donc vous êtes de l'Est

[1] We refer here to the hierarchical concepts employed by Blanche-Benveniste [2}

* "**More** " retracing which started well before the group: Corpus EDF, 8,38 mais ***l'eau*** **est de vingt mè-l'entrant est de vingt metres cubes**
This type of disfluency did not make any difference to the more general results shown by [5]: retracing generally took place at the beginning of the nominal or verbal group (71%) although 29% did not actually conform to this model. Some (19%) did not retrace back to the beginning of the group and others (10%), on the contrary, exceeded this limit. This could however, have been a characteristic of our speaker. It would obviously be necessary to check these results on a much larger corpora comprising of more speakers.

3.2.2.3. Localization of the retracing which followed a word-fragment located at the beginning of the group.
In our corpora, only 28% of the fragments were found at the beginning of the group. The completed and modified fragments, followed by retracing, in fact accounted for 92% and this in turn was followed by a readjustment located at the place of the word-fragment; that is to say at the beginning of the group. As there was no opposition, retracing took place before the statement (retracing More), although 6% exceeded this limit (*[2*) and retraced back to earlier elements (*[1*):

*C6cBesan, 6, euh dans la mise en pratique *[1* je *vois* pas *[2* **c** - *[1* je *verrais* pas *[2* **comment** faire correspondre les choses
*Tropr102, 3,1, *[1* je crois que*[2* **C** + *[1* je crois que*[2* **c'est** la famille des N. mais j'en suis pas sûr

4. Conclusion

Repetitions and word-fragments described in the "standard" spoken French corpora, showed that these phenomena were very frequent and took an active part in the development of the statement through successive insistence on the syntactic places. Half of the involuntary truncations in the statements could be qualified as hesitations. If the function of the fragments, which would be followed by a modification, appeared clearly to be an error, that of unfinished word-fragments was less clear as the speaker did not confirm nor annul his production by retracing, and only this would reveal if it was an error or a hesitation.

Repetitions and interruptions of words are two distinct phenomena which do not affect the same syntactic places. We repeat more functional words than lexical words while we stop in the middle of a lexical item rather than a functional word. As regards the stammerers, Zellner [16, p482] noted that on the contrary, at the time of a hesitation, and not of a stammering, "whatever the rate of disfluencies, the monosyllabic functional words – which enables the speech to be structured – tend to be subjected to more accidents than the other words". However, we never observed, as was seen in 40% of the disfluencies produced by stammerers, "syllables gradually produced in several utterances" in our corpora, nor did we observe the completion of word-fragments without retracing to the beginning of the word. From time to time, although rarely, (once every 33 minutes), a word-fragment was repeated before the word was actually completed (this happened in 80% of the cases).

We studied the re-adjustments concerning where the speaker continued after having stopped in the middle of a word from two points of view: firstly, the types of insertions (30% of the cases) and secondly, the retracing of the statement (75% of the cases). Two types of insertions were distinguished: those which were non-syntactically linked to the statement and those which were. The first were observed, above all, immediately after the interruption, and sometimes, although rarely, when retracing started. Syntactically linked insertions were used when retracing started. These modulations consisted primarily of word additions or replacements while retracted elements were infrequent.

In the large majority of cases, word interruptions occurring in the lexicon, did not appear at the beginning of the nominal or verbal group. Retracing could occur on the interrupted word, go back to the beginning of the two groups or even exceed this limit. If the most numerous cases were those (as noted by Clark and Wasow, [5]) where the retracing of the statement went back to the beginning, a third of the retracing did not obey this schema. Nearly 20% was in fact "minimal" and did not go back to the beginning of the group whereas the others, on the contrary, exceeded the nominal and verbal group. The number of speakers was insufficient, but it seemed that the retracing (6 to 10 %) which exceeded the limits of the group where the interruption took place, did not depend on the localization of the latter.

5. References

[1] Bensalah, Y. 1997. *Pour une linguistique du bégaiement*. Paris, L'Harmattan

[2] Blanche-Benveniste C. 1997. *Approches de la langue parlée en français*. Paris, Edition Ophrys.

[3] Candéa M. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Étude sur un corpus de récits en classe de français*. Thèse d'État, Université Paris III (Sorbonne Nouvelle).

[4] Cappeau P. 1998. Quelques mots sur quelques bribes liées au genre. In: Bilger M, Van den Eynde & Gadet F, (Eds) *Analyse linguitique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste.* Peeters. Leuven, Paris, pp. 301-311.

[5] Clark et Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology, 37*, pp 201-242

[6] Duez D. 2001. Signification des hésitations dans la production et la perception de la parole spontanée. *Revue Parole, 17-18-19*, pp 113-138.

[7] Van Hout A. 2002. *Les bégaiements. Histoire, psychologie, évaluation, variétés, traitements.* Paris, Masson, 2ème édition. (1997)

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

[9] Pallaud B. 2002 a. Les amorces de mots comme faits autonymiques en langage oral. *Recherches Sur le Français Parlé, 17,* pp. 79-102.

[10] Pallaud B. 2003 a. Achoppements dans les énoncés de français oral et sujets syntaxiques. In Merle J.M. (Ed.). *Le Sujet*. Paris : Éditions Ophrys, Faits de Langue, pp. 91-104.

[11] Pallaud B. & Henry S. 2004. Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. In *Le poids des mots. Actes des 7èmes Journées Internationales d'Analyse statistique des Données Textuelles*. Louvain-la-Neuve, 10-12 mars 2004. Louvain, PUL, vol 2, pp.848-858.

[12] Pasdeloup V. 1992. A Prosodic Model for French Text-to-Speech Synthesis. A Psycholinguistic Approach. In BAILLY. Gérard; BENOIT, Christian; SAWALLIS,

Thomas R. (eds). *Talking Machines. Theories. Models, And Designs*, pp. 335-348

[13] Schegloff E., Jefferson H. and Sachs H.. 1977. The preference for self-correction in the organization of repair in conversation.. *Language*, *53*, *2*, pp. 351-382.

[14] Shriberg E.& Stolcke A..1998. How Far Do Speakers Back Up In Repairs? A Quantitative Model. *Proc. Intl. Conf. on Spoken Language Processing*. vol. 5, pp. 2183-2186, Sydney, Australia.

[15] Shriberg E. (1999). Phonetic Consequences of Speech Disfluency. Symposium on The Phonetics of Spontaneous Speech (S. Greenberg and P. Keating, organizers). *Proc. International Congress of Phonetic Sciences*. vol. 1, pp. 619-622, San Francisco.

[16] Zellner, B. (1992). Le bé- bégayage et euh … l'hésitation en français spontané. Actes des 19 èmes Journées d'Études sur la Parole, J.E.P. Bruxelles, pp. 481-487.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 151–155.

Disfluencies as a window on cognitive processing.

An analysis of silent pauses in simultaneous interpreting

Myriam Piccaluga[*], *Jean-Luc Nespoulous*[**], *Bernard Harmegnies*[*]

[*] Université de Mons-Hainaut, Mons, Belgique
[**] Université de Toulouse-Le Mirail, France

Abstract

The paper focuses on silent pauses observed in the productions of subjects involved in simultaneous interpreting tasks. Four bilingual subjects with various degrees of expertise in interpreting and various degrees of mastery of the languages involved (French and Spanish) have been recorded while interpreting utterances of French and Spanish talks. The source discourses had been perturbated by changes both in speech rates (by time compression) and in auditory quality (by addition of a parasiting noise). On the basis of acoustical analyzes performed on the subjects' productions, statistical analyzes focus both on the number and on the duration of the observed pauses. This double approach enables investigations of the kind of cognitive disturbances caused by the independent variables and allows further speculation on the semiology of the pauses durations.

1. Introduction

It is well known that a vocal signal used by a speaker in order to communicate is not produced at a constant rate but, on the contrary, involves many solutions of continuity. These disfluencies may take many forms (repetition, self-correction, vocalic lengthening, etc.). In this study we will focus on phenomena of that kind that are characterized by a break in the phonic flow: in other words, we will focus on *silent pauses* (as opposed to 'ums', hesitation phenomena and other 'filled pauses').

Authors who have investigated this domain generally agree that these breaks in the flow of speech take various forms. Some pauses result from simple physiological necessity (breathing) or articulatory requirements (linked to the realization of certain phonemes: the silence that precedes the release of a plosive, for example); some have a semantic function, and contribute to the discourse strategy of the speaker (delimiting sense groups, for example); still others are linked neither to articulatory mechanisms nor to segmentation in sense units, are not part of the communicative intention of the speaker, but stem from difficulties in speech production which can arise at any point in the speech production cycle (while planning the utterance or accessing the lexicon or even during phonetic implementation, etc.).

These interruptions, whatever their form and cause, are therefore inherent in all ordinary speech production. However, in situations where speech production is combined with another task, breaks in the flow of speech may be linked to imperatives other than those which usually lie behind speech production.

Simultaneous interpreting is an especially interesting task, when seen from this point of view, as it combines simple production of the target language with the preliminary tasks (which appear to an unsuspecting onlooker to be performed simultaneously) of understanding the source language and translating from source to target. Thus, for the interpreter, the production of speech signal is only the last link in a chain of complex cognitive processes.

During the interpreting process, various factors can affect the pauses, or cause them to appear when, under ordinary conditions, they would not be observed. Silences may be linked to difficulties in comprehending the initial message, and to various processes, such as searching for an equivalent translation of the source term in the target language, and indeed difficulty in expressing the concepts in the target language. In other words, it would be reasonable to think that the proliferation and/or lengthening of these pauses is linked to difficulties of various kinds in performing the complex task of interpreting.

This may seem obvious, but it must be observed that few of the researchers who have taken an interest in the process of interpreting have carried out systematic or sustained studies of pauses. Some studies of pauses are carried out on the basis of purely subjective analysis [2]. On the other hand, when a study is based on acoustic analysis, a summarizing approach is usually taken, designed to measure global indices such as, for example, the ratio of the total duration of pauses to total phonation time [1,6,8,9], and extensive studies like those carried out on other speech situations [3,4] are sorely lacking in this field.

In this article we shall employ a dual analytical approach (centered both on pauses frequencies and duration) which will aim to target accurately the characteristics of pauses which appear in the speech of subjects involved in an interpretation task, with the goal of shedding light on the cognitive functions which are active while the subject is interpreting.

2. Method

2.1. Subjects

There were four subjects. All are from the Barcelona area, where they had been living for several decades. They were all female, had a good command of French and Spanish, and spoke both languages regularly. However, for all four subjects, Spanish was the dominant language. Aside from these similarities, the subjects differed in their interpreting expertise and in their command of their languages.

Two of the subjects ('*int1*' and '*int2*') were professional interpreters, with less than 5 years' and 20 years' professional experience respectively. One subject ('*stud*') was in her third year of translation and interpreting studies. The other ('*biling*') had no experience of interpreting. Moreover, their command of French (and therefore their degree of bilingualism) was variable: the two subjects with the most balanced bilingualism were *int2* (very early contact with French, which she used very frequently with family) and *biling* (a teacher of French as a foreign language in a Spanish university); the two other subjects use French only for professional purposes, in interpretation situations.

2.2. Source corpora and linguistic combinations

Each subject was asked to perform six interpreting tasks, each consisting of interpreting a conference speech which was originally given on the floor of the European Parliament and then re-recorded in a laboratory by expert native speakers. The speeches were on issues of general policy and did not involve specialized vocabulary. Each subject had to interpret three speeches in each of the two combinations (French to Spanish and Spanish to French).

2.3. Disruptions introduced into source corpora

Disruptions were introduced into each text in the laboratory. These were: firstly, a local alteration in the speed of the source speech (increasing the rate through a reduction of 80%, 70% or 60% of the total duration of speech without modifying the characteristics of F_0); and secondly, local addition of noise interference (0 dB, 3 dB or 6 dB with respect to the average level of the source discourse).

Each source speech was divided into one-minute portions, without the subject's knowledge; portions 2, 4 and 6 were affected by disruptions, while the other portions were unaffected. Noise interference and temporal compression were applied to each disrupted portion. The levels of independent variables *compression* and *noise interference* were organized according to a Latin square design, which was designed to balance out the associations between levels of the independent variables, and to neutralize the possible semantic effect of the content of the speech.

2.4. Acoustic processing

The subjects performed their interpreting tasks at the Faculty of Translation and Interpreting, Barcelona Autonomous University. The results were analyzed acoustically in the phonetics laboratory at the University of Mons-Hainaut. All acoustic analysis was carried out using Multispeech software from Kay Elemetrics. Pauses were detected using a pitch analysis algorithm, with manual corrections based on interactive listening and examination, at relevant points, of narrow-band spectrograms. The data collected were converted into Excel spreadsheets. For the purposes of the present study, we will only include pauses which could legitimately be considered [4, 5, 7, 10] as neither physiological nor articulatory in nature; that is, pauses longer than 200 ms.

2.5. Statistical analysis

The quantitative analysis that we performed aimed to test the hypothetical effect of our independent variables ('IV') and of their interactions. In each case, we investigated the effect of the IVs *noise interference*, temporal *compression* and linguistic *combination* as well as *subject*. Note that the IV *subject* is considered to be fixed, in the same way as the other variables. This is due to the fact that the selection of subjects was not done randomly; on the contrary, it was the result of close study with the aim of identifying individuals who were prototypical of the groups they represented. This inferential analysis therefore is not intended to lead to any form of generalization to the universe of the subjects, but is designed to explore and shed light on the validity, for future research, of the individual characteristics of these subjects which led us to choose them for this study.

We shall rely on variance analysis to analyze the interval data. However, in so far as the distributions of pause duration are highly dissymmetrical, we will apply an inverse hyperbolic tangent transformation, with a view to giving the dependent variable the Gaussian features required by variance analysis. For numerical analysis of data, we rely on a fully saturated hierarchical loglinear model, which enables us to process nominal data in a conceptual framework which is close to that of variance analysis.

3. Results

3.1. Number of pauses

As table 1 shows, two IVs appear to have a significant effect[1]: *subject* and *combination*. Conversely, neither of the two IVs related to disruption appear to have significant effects.

Of the interactions, only two are clearly significant: the interaction of *subject* and *combination* on the one hand, and the interaction of *noise interference* and *compression* on the other.

Table 1: results of loglinear analysis applied to all pauses longer than 200 ms.

Source of Variation	Degrees of Freedom	Chi-square	Signif.
subject	3	54.927	0.000
combination	1	17.878	0.000
noise interference	2	3.518	0.172
compression	2	0.026	0.987
subject*combination	3	13.802	0.003
subject*noise interference	6	10.967	0.089
combination*noise interference	2	0.507	0.776
subject*compression	6	2.251	0.895
combination*compression	2	0.747	0.689
noise interference*compression	4	13.866	0.008

A descriptive study of the differences shows that the significance of the *subject* variable can be principally attributed to the difference between *int2* (who was remarkable for the large number of pauses she made) and the three other subjects, who recorded lower numbers of pauses and for each of whom the number of pauses was approximately equal.

The significant difference attributable to the effect of the *combination* variable is linked to the fact that the number of pauses was generally lower in the French to Spanish combination than in Spanish to French. In other words, the subjects paused less often when they were speaking the dominant language.

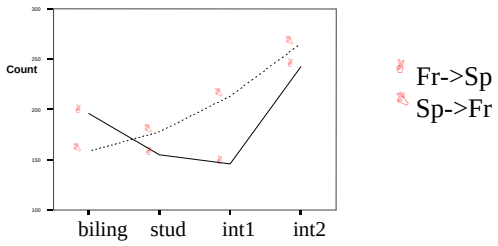


Figure 1: number of pauses by *subject* and by *combination*

Analysis of the dual interaction *subject*combination* (see fig.1) shows that the difference in numbers of pauses between the two combinations varies according to subject. This number is particularly low in *int2*, particularly high in *int1*, and

[1] Due to lack of space and a desire for clarity, we will confine ourselves, in this study, to the analysis of the main effects of these independent variables and to the dual interactions between them; analysis of higher-level interactions did not offer substantial additional information.

intermediate in the two other subjects. It is notable that *biling* is the only subject who paused more often in Spanish than in French.

The study of the interaction between *noise interference* and *compression* shows that, generally, the number of pauses increases as a function of the amount of noise interference, but this increase varies considerably as a function of the rate of compression: there is a very clear increase when the level of compression is high or medium.

3.2. Duration of pauses

Variance analysis shows a significant effect for the *subject*, *combination* and *noise interference* IVs. Conversely, no significant effect was observed when considering the effect of (temporal) *compression* in isolation. A significant effect can be observed for the interactions *subject*combination*, *compression*combination* and *compression*noise interference*.

Table 2: results of variance analysis applied to all pauses longer than 200 ms in duration.

Source of Variation	Degrees of Freedom	F	Signif.
subject	3	9.735	0.000
combination	1	16.819	0.000
noise interference	2	8.896	0.000
compression	2	1.557	0.211
subject*combination	3	9.967	0.000
subject*noise interference	6	0.124	0.993
combination*noise interference	2	2.816	0.060
subject*compression	6	1.235	0.285
combination*compression	2	5.292	0.005
noise interference*compression	4	4.020	0.003

In other words, the VI temporal *compression* may well have no effect on the duration of pauses when considered in isolation, but its effects are felt when its interaction with *linguistic combination* and *noise interference* is considered.

It can be observed that in the group as a whole, the mean duration of pauses is 768 ms. From this point of view the differences between subjects are great, as the mean pause duration for *int2* (620 ms) differs by more than 300 ms from that observed for *stud*. The lowest durations were observed in *biling* and *int2*, and the greatest in *stud* and *int1* (see fig. 2)

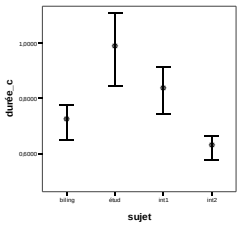


Figure 2: mean and standard deviation of duration of pauses according to *subject* (left to right: *biling, stud, int1* and *int2*)

In the Spanish to French combination, the subjects tended generally to produce shorter pauses (670 ms on average) than in the French to Spanish combination(865 ms on average), as shown in fig. 3a.

Under the influence of an increase in noise interference, the average pause duration tended to increase. Fig. 3b shows that from *noise interference* level 1 to level 2, a reduced increase of the order of 30 milliseconds can be observed. Conversely, between levels 2 and 3, the difference was of the order of 200 ms (706 ms compared to 902 ms on average).

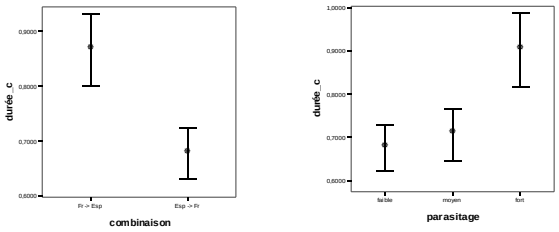


Figure 3: a (left) :mean and standard deviation of duration of pauses according to linguistic *combination* (left to right: *Fr->Sp* and *Sp->Fr*) and b (right) to *noise interference* (right: *low, medium, high*)

Analysis of the dual interaction *subject*combination* (fig. 5) confirms the previously noted difference in profile of the four subjects. However, this shows that, from subject to subject, the difference in pause duration between the two combinations manifests itself in very different ways. Thus, the mean durations for *int2* are approximately equal (the difference being around 30 ms). *Biling* shows greater differences but these is still only of the order of 122 ms on average. Conversely, *int1* and *stud* show much greater differences between combinations (of the order of 200 ms and 500 ms respectively).

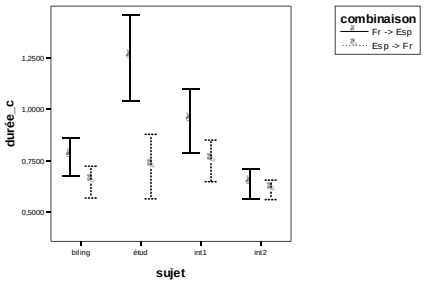


Figure 5: mean and standard deviation of pause duration according to the interaction *subject*combination* (left to right: subjects *biling, stud, int1* and *int2*)

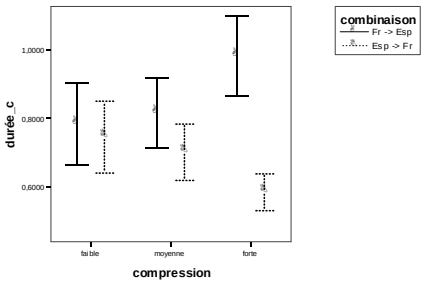


Figure 6: mean and standard deviation of pause duration according to the interaction *compression*combination* (left to right: *low, medium* and *high* compression)

Under the joint effect of *compression* and *combination*, average pause duration varies in opposed directions. Thus, a consistent increase in duration as a function of an increase in compression rates in the French to Spanish combination can be observed (784 ms, 816 ms and 982 ms respectively). Conversely, in the Spanish to French combination, a consistent decrease is observed (746 ms, 702 ms and 585 ms).

In any case, analysis shows that, whatever rate of compression is applied, the duration corresponding to the lowest level of noise interference is lower than the duration

corresponding to the highest level of noise interference. However, regular gradation is only observable under high levels of compression.

4. Conclusion

The two types of analysis that we carried out (on number and on duration of pauses) showed in each case that dependent variables are sensitive to the IVs used in our analysis. However, the combination of approaches (number and duration of pauses) seems to be particularly relevant, in so far as it enables richer observations to be made. Thus it can be observed that the subject who had the greatest linguistic and interpreting expertise (*int2*) was manifestly superior in terms of the number of pauses. But analysis of duration shows that this subject also produced the shortest pauses. It can also be seen that if the three other subjects could not be differentiated by how many pauses they made, differ one from another in the durations of their pauses. In contrast, it was observed that *stud*, who produced the lowest number of pauses, recorded the longest average pause duration.

We also saw that pauses were more numerous in the Spanish to French combination than in French to Spanish; these were of longer duration in the French to Spanish combination than in the Spanish to French combination. These observed tendencies are, however, sensitive to the *subject* effect: the three subjects having some degree of contact with the world of professional interpreting had higher scores when speaking French than when speaking Spanish, but the opposite was recorded in the subject with no interpreting experience (*biling*). The number of pauses is revealed to be sensitive to the level of interpreting expertise, while the pause duration seems to be sensitive to linguistic expertise: the subjects recorded different pause durations in each of the two combinations, but it was also noted that for *biling* and *int2* these differences were only minor, while for the other two subjects the differences in duration were greater.

The relevance of joining these approaches together is confirmed by the emergence of an effect that noise interference has on duration, while this IV appeared to have no effect on the number of pauses.

We observed, moreover, that the variable *compression* seems to have no direct effect on the number or duration of pauses. However, it would be false to assume that pauses are completely insensitive to variations in compression; no direct effect was observed, but several quite complex interaction effects were observed, which reveal only a general tendency for compression, in association with other variables, to have an indirect effect on the quantity of silence, particularly if the level of compression is high.

Data collected in this exploratory framework do not enable us to make inferences of any guaranteed validity. However, it may be observed that these data aid in gaining a better comprehension of the cognitive processes at work in this domain. The subject with the greatest linguistic and interpreting expertise notably made a large number of short pauses, which illustrates a markedly functional division of the source speech into regular chunks, which are then reproduced in the target language, separated by short pauses; conversely, the student subject, whose expertise both in language and interpreting was lacking, made fewer pauses but these were rather long, which did not accord with the optimal strategy displayed by the more experienced subject, and which probably illustrates disorganization in the production of the target language linked to various breakdowns in the interpretative process.

The general tendency for subjects to pause more often and for shorter periods when the speech is delivered in their dominant language could also be interpreted in terms of their having more effective chunking strategies for their stronger language; this could also be linked to difficulties in production of the target language, which is a hypothesis that could possibly partly explain the greater inter-combination contrast in subjects with less knowledge of French.

The effect of noise interference is clearly linked to comprehension phenomena, and it is useful to note that this effect mostly manifests itself in the appearance of pauses that are longer as a function of the level of noise interference.

Temporal compression effects were not observed; this could be explained either by the fact that an increase in source discourse speed had no consequences, or by the fact that, by choosing a threshold criterion of 200 ms, we prevented ourselves from observing pauses whose duration was shortened when the subject needed to speak more quickly. If this is the case, then these effects are linked to low-level production phenomena. Conversely, the contribution of the variable of compression to interaction effects (for high levels of compression) suggests difficulties of another kind, linked to comprehension and/or translation problems for such a fast source speech.

These observations could lead to further speculation about modeling the interpretative task, as they are rooted in analysis of low-level phenomena, but are aimed at understanding high-level phenomena.

From the point of view of basic research, the study of pauses in the task of interpreting is relevant in so far as it helps us work towards a better understanding of the cognitive processes in the interpreter. Moreover, from the point of view of applied research, this study is justified in that it is directly linked not just to the intelligibility of the speech produced by the interpreter, but also to the perception by the speaker of the interpreter's confidence, based on discontinuities in his/her speech; therefore, research of this kind contributes to the study of interpreting quality.

5. References

[1] Barik, Henry C. 1973. Simultaneous interpretation: Temporal and quantitative data, *Language and Speech*, 16, 237-270.

[2] Boschian Schiavon, Vesta. 1983. Velocità di parola e interpretazione simultanea, Università degli studi di Trieste: Monografie, N°20, Trieste.

[3] Campione, Estelle & Véronis, Jean. 2002. A Large-Scale Multilingual Study of Silent Pause Duration. In B. Bel & I. Marlien (Eds.), *Proceedings of the Speech Prosody 2002 conference*, 199-202.

[4] Campione, Estelle & Véronis, Jean. 2004. Pauses et hésitations en français spontané, *Actes des 25èmes Journées d'Etudes sur la PArole (JEP 2004*), 109-112.

[5] Candéa, Maria. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané.* Thèse de doctorat, Université Paris III.

[6] Gerver, David. 1969. The effects of source language presentation rate on the performance of simultaneous conference interpreters, In E. Foulke (Ed.), *Proceedings of the 2nd Louisville conference on rate and/or frequency controlled speech*, University of Louisville, 162-184.

[7] Goldman-Eisler, Frieda. 1968. *Psycholinguistics: experiments in Spontaneaous Speech*, Academic Press, London and New York.

[8] Lee, Tae -Hyung. 1999. Speech proportion and accuracy in simultaneous interpretation from English into Korean, *Meta*, 44, 2, 261-267.
[9] Yagi, Sane M. 2000. Studying style in simultaneous interpretation, *Meta*, 45, 3, 520-547.
[10] Zellner, Brigitte.1994. Pauses and the Temporal Structure of Speech, In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition*, 41-62.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 157-162.

Disfluency in speech input to infants? The interaction of mother and child to create error-free speech input for language acquisition

Melanie Soderstrom[*] *& James L. Morgan*[*]

[*] Brown University, Providence, USA

Abstract

One characteristic of infant-directed speech is that it is highly fluent compared with adult-directed speech. However, the speech that infants hear still contains disfluencies. Such disfluencies might potentially cause problems for infants during language development. We first analyzed samples of spontaneous speech in the presence of infants (both adult- and infant-directed) and found that under ideal circumstances the speech infants hear is highly fluent. Under less than ideal circumstances infants hear much more highly disfluent speech - however this disfluent speech is almost entirely adult-directed. While grammatically ill-formed, the prosodic structure of these disfluencies might signal their ill-formedness to the infants. In a preference experiment, 10 month olds listened longer to infant-directed speech samples containing prosodic disfluencies than to equated samples without disfluency. However, this effect was found in only one of two counterbalancing groups. Using adult ratings of low-pass versions of these speech samples, we found that infants' preferences were correlated with the adults' perception of the relative disfluency of the samples. A follow-up experiment using adult-directed disfluencies found that while the 10 month olds showed no differences in their listening preferences, older infants preferred to listen to the fluent speech. These results suggest that younger and older infants attend differently to infant and adult-directed speech, and that older infants may be able to differentiate grammatical adult-directed input from input distorted by disfluency. We discuss implications of these findings for language acquisition.

1. Introduction

The extent to which the disfluent character of speech might pose a problem for the learner has been a subject of great interest in the study of language acquisition. Chomsky [1] famously asserted that the language input to the child included many "interrupted fragments, false starts, lapses, slurring, and other phenomena that can only be understood as distortions of the underlying pattern". While typical adult-adult speech might well fit this description, studies of child-directed speech have found it to be highly fluent. For example, Newport, Gleitman & Gleitman [8] found only 1 disfluent child-directed utterance out of 1500, and only 4% of utterances were untranscribable due to mumbling or slurring. Even the adult directed speech in this study was fairly fluent - 5% disfluent and 9% untranscribable. This finding, which has been supported by numerous studies of child-directed speech, suggests that the input to the child might not be subject to the distortions that characterize normal adult-directed speech.

Nevertheless, the input to the language learner may not be unswervingly well-formed. For one thing, a recent estimate of the total language input to an infant found that only 15% of the speech heard by the infant is directed to that infant - an additional 30% was directed at an older sibling [12]. Therefore, if the ambient adult speech is processed by infants, between 55 and 85% of the language input in the earlier stages of grammatical development might be of the type described by Chomsky.

Even if infants do not process a significant number of explicitly ungrammatical utterances due to disfluency, the ill-formed prosodic structure of disfluent utterances may be a source of misinformation about the grammatical structure of the language. A growing body of literature suggests that infants as young as 2 months are highly sensitive to the prosodic structure of utterances [e.g. 5, 6] and by 6 months can use prosodic information to organize fluent speech [e.g. 7, 9, 10] into grammatically-relevant units, even before lexical information (including word boundaries) is available. This "prosodic bootstrapping" theory depends on the *prosodic* well-formedness of the input to the language learner. Studies such as [8] examined only major errors in grammatical structure of the utterance, what they referred to as "true garble", but not disfluencies in the phonological or acoustical structure of the utterance. Such prosodic disfluencies might themselves be a critical source of error in this initial process of grammatical development. On the other hand, if infants can detect these prosodic disfluencies, they may provide cues to the infant about the reliability of the utterances as a source of information about the grammar.

The current study asks three questions about the possible effect of disfluency on early stages of language acquisition:

(1) Does infant-directed speech contain prosodic disfluencies?
(2) Are infants able to differentiate between fluent prosodic breaks and prosodic disfluencies in infant- and adult-directed speech?
(3) Are there developmental differences in infants' sensitivities?

In the second and third parts of this paper, we will provide some analyses of the presence and properties of disfluencies in the speech input of infants in different environments. In the fourth and fifth parts, we will examine behaviorally adults' and infants' sensitivity to the prosodic properties of disfluencies in speech.

2. Corpus analysis I: Best case scenario

2.1. The corpus

Maternal speech to two infants was collected every 1-3 weeks in a natural home setting while the infant was 6-10 months old. As part of a larger study [11], both mothers were asked to record about an hour per week of speech, and to make recordings at least 30 minutes long, however some recordings were shorter than this. In total, approximately 8.5 hours were collected during this period for the first mother (MOT1), and 14 hours for the second mother (MOT2). These recordings were then transcribed using the CHAT transcription system [4], for a total of 9067 utterances for MOT1 and 10604 utterances for MOT2. The majority of these utterances were infant-directed. However, for MOT1, there were also a

significant number of adult- and child-directed utterances. These are reported separately.

Each utterance was then coded for a variety of syntactic and prosodic features. In particular, major utterance-internal prosodic breaks (i.e., anything transcribed with a comma, #, [/], etc.) were coded qualitatively as follows:

(1) F: Any fluent, prosodic break at the conjunction of two well-formed grammatical units.
(2) D: Any disfluent prosodic break caused by restart, reformulation, speech error, etc.
(3) O: Any prosodic break not caused by restart, reformulation, etc., that was otherwise odd or ill-formed (primarily pauses for thought). In this first analysis this category included prosodic breaks which were odd either because they were prosodically ill-formed, or because they occurred at a grammatically inappropriate location.

2.2. Prosodic disfluencies in speech samples

Our samples of speech to two young infants found relatively few examples of utterance-internal prosodic disfluencies compared with utterance-internal fluent prosodic breaks (Table 1). These data suggest that only about 5-10% of prosodic breaks are disfluent in infant-directed speech (Formula: # D + O breaks / total # prosodic breaks).

Table 1: Fluent, disfluent, odd, and total prosodic breaks in infant-directed speech (per utterance). Bottom Row: Disfluent breaks per total breaks).

	MOT1	MOT2	MOT1-IDonly
F/utt	.390	.219	.330
D/utt	.013	.008	.007
O/utt	.013	.014	.007
Tot/utt	.416	.241	.344
(D+ O)/tot	.063	.091	.041

These data present a relatively good picture for prosody as a source of information for infants about the structure of their language. However, these data provide in some sense a "best case scenario" picture of the speech input environment of infants. These recordings were taken at home, in relative quiet (although for some parts of the MOT1 recordings, the father and siblings were present and provided a noisier input environment). Not all of an infant's speech input will be obtained in such ideal conditions. For example, INF1 (MOT1's infant) was in a busy daycare during the day, and was exposed to a variety of children and caregivers. While we do not have recordings of this environment, we do have a recording by another mother in the larger study, MOT4, during a wait in a busy airport terminal. This transcript allows us to examine a more complex input environment.

3. Corpus Analysis II: Worst Case Scenario

3.1. Transcript

This transcript was obtained while the infant was 2.5 months old. Mother, father and infant were waiting in an airport terminal for their flight to be called. The mother chose this time to make one of her recordings. During this time a variety of other people in the terminal interacted with the mother, so a smaller percentage of the utterances in this recording were directed at the infant. This environment was likely to generate a much larger number of disfluencies not only because many of the utterances were adult-directed, but also because of the high ambient noise level, and because the mother was engaged in an animated discussion with strangers.

Due to the greater number of disfluencies overall, we distinguished in this analysis between prosodically and syntactically odd breaks. A prosodic break was only considered an O if it was *prosodically* odd. If it sounded fluent but was grammatically inappropriate, it was classified as a C. Utterances were then separately analyzed for the grammaticality of the word groupings caused by both fluent and disfluent prosodic breaks.

(1) C: Any fluent-sounding prosodic break, regardless of the grammatical context.
(2) R: Any disfluent prosodic break caused by restart, reformulation, speech error, etc.
(3) O: Any prosodic break not caused by restart, reformulation, etc., that was prosodically odd or ill-formed (primarily pauses for thought).

3.2. Prosodic disfluencies in a complex environment

Clearly, this mother produced both a larger rate of prosodic breaks per utterance overall, and also a much larger proportion of disfluent breaks (Table 2). However, the overall rate of breaks she produced in infant-directed speech was comparable to that of the other two mothers, and if anything, the rate of infant-directed disfluent breaks was smaller.

Table 2: Absolute numbers and prosodic breaks per utterance for MOT4

	# breaks	breaks/utt	# breaks-ID	breaks/utt-ID
C	273	.503	122	.396
R	42	.077	2	.006
O	54	.099	1	.003

Proportionally, this mother's infant-directed disfluencies accounted for only 2% of the total infant-directed prosodic breaks in the speech samples. However, looking at the total sample, this proportion is much higher, up to 26%.

Table 3 gives the percentages of phrases and clauses that are grammatically ill-formed when we consider utterances, all prosodically bound word sequences (i.e. word sequences to either side of a C, R, or O), only phrases and clauses bound on at least one side by an R or O break, and finally phrases and clauses bounded only by well-formed prosodic breaks. Isolated single words were excluded from this and the following analysis, because they are by defnition well-formed grammatical units.

Table 3: Percentages of syntactically ill-formed units by prosodic context for MOT4

	Utterances	All pros units	R's	O's	C only
Phrases	.054	.228	.778	.375	.028
Clauses	.157	.151	.467	.640	.080

Clearly, the mother in this transcript is producing a large number of syntactically ill-formed prosodic units due to disfluency. However, the vast majority of prosodically well-formed word sequences (the "C only" column) are also syntactically well-formed (97% of phrases and 92% of clauses). These data suggest that the ability to detect prosodic disfluency might be of great benefit to infants in determining which word sequences constitute reliable input for language acquisition.

The picture looks even better if we examine the relative number of grammatically well-formed and ill-formed phrases/clauses preceding fluent and disfluent prosodic breaks, including disfluent utterance endings (which were not counted in Tables 2 and 3). Ninety-nine percent of word sequences

preceding fluent prosodic breaks are grammatically well-formed. Of the 6 ungrammatical units preceding a fluent-sounding prosodic break, 4 are directly following a disfluency. Of the remaining two, one was formed by the insertion of ”you know” in the middle of a phrase. The other was preceded by an ”um”, indicating disfluency, but simply did not sound disfluent to the ear.

Table 4: Number of grammatical and ungrammatical phrases preceding fluent and disfluent prosodic boundaries.

	Fluent (C)	Disfluent (R)	Odd (O)	Disfluent Endings
Gramm.	474	4	8	7
Ungramm.	6	29	27	24

Based on our transcribers’ judgments, MOT4’s adult-directed speech in this sample, though highly disfluent, also contains highly reliable cues to grammaticality. The following section examines whether these prosodic cues to disfluency are detectable by naive adult listeners, and more importantly, by infants.

4. Differentiating fluent and disfluent speech: Infant-directed speech

The data from the preceding section suggest that in some circumstances, infants might be exposed to a large amount of disfluency in speech. However, the ungrammatical word groupings cause by this disfluency appeared to be reliably marked prosodically. We examined both adults’ and infants’ ability to detect prosodic cues to disfluency.

4.1. Stimuli

Disfluent utterances directed at the infant or older siblings were culled from the transcripts of MOT1. Utterances were discarded if they contained background noise, were not clear enough to be fully transcribed, or if they were highly repetitive due to restarting. Due to these selection limitations and the small number of disfluencies overall, we were not able to be particular about our disfluent utterance types. These included a variety of utterance types, including questions and declaratives, and utterance lengths. They covered a range of disfluency types, including interruptions, repetitions, restarts and reformulations.

Because these utterances were created in a spontaneous speech environment, it was not possible to obtain paired fluent samples which matched the disfluent samples on variables such as length, intensity, number of syllables, prosodic structure etc. A pilot attempt to directly use the speech samples from the recordings as fluent controls was deemed too variable. Therefore, in order to control for these other factors, both fluent and disfluent utterances used in the experiment were created in the laboratory, and produced by the first author. Disfluent utterances were produced by listening to, and mimicking as closely as possible, the disfluent speech of the mother. Fluent utterances were created by transforming the disfluencies into fluent prosodic breaks.

For example, the fluent utterance (2) was created from the disfluent utterance fragment (1):

(1) Should we start +//. Oh, a big yawn from you.

(2) Should we start? Oh, a big yawn from you.

Utterances were grouped into passages of 6 utterances each. Eight disfluent passages and eight matched fluent passages were produced. Each passage was about 20 seconds in length. While mimicked disfluencies are probably not prosodically identical to real disfluencies, we felt that this method was the best compromise between a controlled experiment and using stimuli that are most representative of the kinds of disfluencies that infants are likely to hear. It is likely, however, that this design underestimated the prosodic differences between fluent and disfluent speech - this will be pursued further below.

Table 5 provides information about the properties of these infant-directed disfluent utterances and their fluent matched controls.

Table 5: Average properties of infant-directed disfluent and fluent utterances

	Length (ms)	Syllables	Repeated words	Prosodic boundaries
Fluent	2438	12.0	0.313	1.23
Disfluent	2459	12.1	0.354	1.61

There were no significant differences between length, number of syllables, or number of repeated words. The disfluent passages contained significantly more utterance-internal prosodic boundaries than the fluent passages ($t(47)=4.08$, $p < .001$). Along with prosodic disruptions, the disfluent passages contained 1 um/uh and 18 part-words.

4.2. Adult Ratings

In order to measure how disfluent our “disfluent” samples were, we presented the samples to 8 adult raters, who were asked to judge how “fluent” or “disfluent” they sounded. This served two purposes - to establish that our disfluent samples did indeed contain prosodic cues to their disfluency, and to ascertain that our “fluent” controls actually sounded more fluent.

Samples were low-pass filtered at 400 Hz (with 100 Hz smoothing) so that the raters heard only the overall prosodic character of the speech samples, and not the high-order phonological or lexical information. Raters were asked to rate each individual utterance from the speech samples, from 1 (highly fluent) to 4 (highly disfluent). All 8 raters scored the disfluent utterances on average as more disfluent than the paired fluent utterances. This difference was highly significant by two-tailed t-test ($t(7)=7.688$, $p < .001$), suggesting that the disfluent and fluent utterances were discriminable based on prosodic characteristics. However, the difference in rating was not very large - an average of 2.63 for disfluent utterances, and 2.40 for fluent utterances. In order to try to pull out this difference, a second set of ratings was obtained with a larger (1-7) scale, but the means were similarly close (4.8 versus 4.1 on the wider scale), so the original ratings were used in subsequent analyses.

This result not only validates our stimuli, but also provides evidence that disfluencies in infant- and child-directed, while rare, are detectable based on their prosodic characteristics alone - at least by adult listeners. We next examined whether infants are sensitive to these prosodic cues.

4.3. Participants

Thirty-two 10 month old infants participated in the experiment. They ranged in age from 309 days to 342 days. There were fifteen males and seventeen females. An additional 2 infants were tested but not included in the experiment due to fussiness. All participants were normally developing infants with normal hearing from Providence, RI, USA, and had parents and caregivers who were native speakers of American English.

4.4. Design

If infants are able to differentiate fluent and disfluent speech samples, they are likely to exhibit differences in their listening preferences between the two passage types. Because it is often difficult in behavioral research to predict whether infants will show a preference for the ill-formed stimulus (a novelty effect) or the well-formed stimulus (a familiarity effect), a two-tailed statistical test is employed to determine whether infants show a difference in their preferences, regardless of the direction of that difference.

Infants were divided into two counterbalancing groups. Prior to the test phase there was a pre-test phase to familiarize the infants with the procedure. This consisted of two trials of speech stimuli, one of which was a 10 second repetition of a fluent, prosodically well-formed phrase; the other was a 10 second repetition of a prosodically ill-formed phrase-like word sequence. The stimuli came from a separate study [9], and were chosen to minimize the impact of the pre-test phase on infants' preferences during the test phase. Each group then heard all 8 test passages. Group 1 heard the disfluent versions of passages 1-4 and the fluent versions of passages 5-8, while group 2 heard the fluent versions of passages 1-4 and the disfluent versions of passage 5-8. The passages were presented once each in a random trial order.

4.5. Procedure

Infants were tested using the Headturn Preference Procedure. In this procedure, a flashing light is paired with a sound stimulus. When the infant looks toward the flashing light, the trial is initiated and the sound stimulus begins to play. When the infant looks away for at least 2 consecutive seconds, the trial ends. Infants' preference for the difference stimulus types is measured by their total orientation time toward the flashing light on each trial. For further details on this method, see [2].

4.6. Results

Across the two groups, infants preferred the disfluent passages, but this difference was non-significant. However, an examination of the two groups found that while group 1 showed no difference in their preferences, group 2 significantly preferred the disfluent passages (8.9 s mean looking time) to the fluent passages (7.4 s) ($t(15) = 2.174$, $p < .05$, two-tailed). This result is suggestive that infants do differentiate fluent and disfluent speech. But given the lack of statistical significance across the two groups, they must be interpreted with caution. However, recall that the way the stimuli were created may have minimized the prosodic differences across the sample types. If so, it is not surprising that infants would show only a weak difference in their preferences. Furthermore, because of the stringent criteria applied to the samples taken from the transcripts, there was some difficulty finding enough disfluent utterances to use. The best utterances were selected from a larger pool of possible samples. Therefore, the later-chosen samples were potentially less disfluent overall, and these were the samples that group 1, which showed no differences, heard as disfluent.

Because only one group of infants had a listening preference, we wished to confirm that the infants' preference was based on the fluent/disfluent distinction we were investigating. We therefore performed a correlational analysis between the infant listening preferences and the adult ratings. Because the infant listening preferences were obtained for the sample passages and the adult ratings on the individual utterances, we took an average of each rater's scores across the utterances in the sample passages. These 16 averaged scores were used in a Pearson correlation with the infant listening preferences. There was a weak but significant correlation ($r = .441$, $p < .05$, one-tailed) overall. Looking by group, Group 2's listening preferences were significantly correlated with the ratings ($r = .749$, $p = .016$, one-tailed), while Group 1's listening preferences were not correlated ($r = .031$, $p > .25$).

5. Differentiating fluent and disfluent speech: Adult-directed speech

The infant behavioral results with infant-directed speech were equivocal. While there was a significant difference in listening preferences for one of the two groups, the effect was not significant across the two groups. Might the infants show a more reliable listening preference with adult-directed stimuli? On the one hand, the disfluencies in the adult-directed speech may be more salient as well as simply more numerous. On the other hand, because these utterances are adult-directed, the speech is more rapid, and the prosodic characteristics of the speech are less salient. Furthermore, infants show greater attention to infant-directed than adult-directed speech overall [2]. Therefore, the disfluencies in adult-directed speech may be more difficult to detect. If infants show a reliable difference in their preferences for the disfluent and fluent versions of these adult-directed utterances, it will add support to our claim that infants differentiate disfluent from fluent speech, particularly in the speech mode for which it would be most beneficial because of the greater presence of disfluencies - i.e., adult-directed speech. Furthermore, if infants show a more reliable difference with the adult-directed samples, this would be a unique case where the properties of adult-directed speech might actually be more beneficial to infants than those of infant-directed speech.

5.1. Stimuli

Forty-two disfluent utterances were culled from the same transcript used in the "worst case scenario" analysis. Criteria for selection, and the process for creating the final stimuli were similar to those for infant-directed speech. However, the constraint on background noise in the sample was relaxed as long as the utterance was fully intelligible. We also included utterances with more repetition in order to maximize the disfluency in the samples.

5.2. Adult Ratings

We used the same procedure as with the infant-directed stimuli to obtain ratings from 8 adult listeners, using the 7 point scale. As with the infant-directed stimuli, the raters were able to discriminate the fluent and disfluent samples with a high degree of accuracy ($t(7)=7.27$, $p < .001$). The average rating for disfluent utterances was 4.52 and for fluent utterances 3.26. Because we had a larger selection of utterances, we then chose the utterances with the largest differences as the stimuli for the behavioral study. Five utterances were selected for each of four fluent and disfluent passage pairs. All utterances selected had a difference of at least 1 point between their average fluent and average disfluent rating. Thus, in this study, the difference between the fluent and disfluent versions was more extreme in terms of adults' perceptions. Also, in order to better control for the small differences in length between the fluent and disfluent versions, the versions were equated for length by slowing down the shorter of the two and speeding up the longer of the two. These doctored stimuli were then rated a second time to verify that they were still highly discriminable by fluency.

Table 6 provides information about the properties of these adult-directed disfluent utterances and their fluent matched controls.

Table 6: Average properties of adult-directed disfluent and fluent utterances (length is pre-adjustment)

	Length (ms)	Syllables	Repeated Words	Prosodic boundaries
Fluent	3565	17.6	0.85	1.85
Disfluent	3601	17.25	0.95	2.8

As with the infant-directed utterances, there were no significant differences between starting length, number of syllables, or number of repeated words. The disfluent passages contained significantly more utterance-internal prosodic boundaries than the fluent passages ($t(19)=4.50$, $p < .001$). Along with prosodic disruptions, the four disfluent passages contained a total of 5 um/uhs and 9 part-words.

5.3. Participants

Twenty-three 10 month old infants participated in the experiment. They ranged in age from 309 days to 339 days. There were eleven males and twelve females. Fourteen older infants, ranging from 20 to 24 months, also participated, in an ongoing experiment. There were six males and eight females in this age group. An additional 4 infants at this age were tested but not included in the experiment due to fussiness. All participants were normally developing infants with normal hearing from Providence, RI, USA, and had parents and caregivers who were native speakers of American English.

5.4. Design & Procedure

The design and procedure were identical to the first experiment for the 10 month olds, except that the fluent and disfluent version of each passage were both presented to each infant. So each infant heard all four disfluent passages, and the corresponding four fluent passages. For the 22 month olds only, a variant of the testing procedure was used in which an image on two T.V. screens are paired with the sound stimuli instead of flashing lights. This was done to maintain the interest of the older infants throughout the testing session.

5.5. Results

The 10 month olds showed no listening preferences for either the disfluent or fluent stimuli. Mean looking time to the disfluent passages was 7.3 s and to the fluent passages 7.1 s ($p > .5$). Only twelve out of twenty-three infants preferred the disfluent passages. This finding suggests that despite their greater disfluency overall, ten month olds are not sensitive to the prosodic cues to disfluency in adult-directed speech.

By contrast, our preliminary results with the fourteen 22 month olds suggests that they prefer the fluent speech stimuli (8.3 s) to the disfluent stimuli (6.9 s). However, with this small number of participants to date, the difference is not significant.

5.6. Follow-up studies

We are currently gathering more data to determine whether this difference in the older infants with the adult-directed stimuli is in fact significant, indicating that infants at this age differentiate disfluent from fluent utterances. If so, a second experiment using low-pass filtered speech will help to determine whether the older infants are using prosodic as well as grammatical cues to disfluency. Addtionally, it will interesting to run a comparable study with this older age range using infant-directed stimuli, to determine whether there is in fact a difference between the younger infants and older infants in how well they detect disfluency in adult- versus infant-directed speech.

6. Discussion

The current study examined two contexts for disfluency in maternal speech: One, a large, infant-directed corpus, in which speech samples were collected in the home environment in relative quiet, and the other a smaller sample collected in the chaotic environment of an airport waiting room, in which the mother was interacting with strangers as well as her own infant. This latter sample contained a much higher number of disfluencies than the more infant-directed corpus. However, ungrammatical word sequences were reliably marked by prosodic disfluency.

We next examined the extent to which these prosodic cues are detectable by adult raters and infants in a behavioral test. Disfluent and fluent utterances were highly discriminable by adult raters, especially the adult-directed speech stimuli. However, the picture with the infants is more complex.

In an initial experiment with infant-directed speech samples, one out of two groups of infants showed a significant difference in their preferences for the fluent and disfluent samples, but the other group did not. In the group showing a preference, the preference was correlated with the disfluency scores given to the samples by adult raters. Overall, these results are suggestive that infants are able to differentiate fluent and disfluent speech in infant-directed speech, but clearly further research is needed.

We next examined whether infants are able to differentiate fluent and disfluent utterances in adult-directed speech. Despite the more salient disfluencies in these samples - at least, according to adult ears - the ten month olds did not show any differences in their listening preferences. By contrast, an ongoing study with an older age group, 22 month olds, found a preference for the fluent passages.

Overall, these findings suggest that infants are indeed sensitive to the presence of disfluencies in speech, but the characteristics to which they attend may vary across different ages. Twenty-two month olds are well on their way to grammatical competence, and may even be combining words productively. Therefore, these older infants may detect disfluency based on properties of the language at which they are already competent, such as the repetition of function words, as well as the overall prosodic properties of the utterances. These two types of cues together may allow older infants to exclude from grammatical analysis both utterances and prosodic breaks which are ill-formed due to disfluency. On the other hand, younger infants are entirely reliant on the prosodic characteristics of the speech signal, and their ability to determine whether an utterance or prosodic break is a legitimate source of information about the grammar is limited to this aspect of the speech stream.

If so, these younger infants might be limited to infant-directed speech as a source of *reliable* input, while older infants may be able to make use of prosodic and structural information in both infant- and adult-directed speech -- infant-directed speech because the prosodic information is highly reliable, and adult-directed speech because the infants can detect disfluency, and reject it as input. This suggests a trade-off. On the one hand, infant-directed speech contains less prosodic information about grammatical structure, because it is shorter and less complex. On the other hand, adult-directed speech contains a greater wealth of information (and constitutes the majority of the ambient linguistic input), but these data are less reliable as input. This might suggest that

infants begin by paying attention to the simpler, infant-directed speech, and only later (though well before they attain productive competence) access the adult-directed ambient input.

An additional puzzle lurking in the background of this discussion is: Why is infant-directed speech fluent and error-free? On first glance, the answer is simple. Mothers wish to be understood by their infants and therefore speak more precisely and fluently. However, this idea doesn't really hold up. Normal speakers don't "choose" to make speech errors and disfluencies in every day conversation. Speech errors are just that - errors. Most of the time, we are even unaware that we are producing them. Therefore, how can mothers choose to *not* make speech errors? One possibility is that because infant-directed speech is both slower and shorter than typical adult-directed speech, there are simpler fewer chances for errors. However, there is another reason for disfluency in adult conversational speech - the interaction of the second speaker. Conversational speech is a process of give and take - a speaker monitors in an on-going fashion the behavior of the listener to determine whether they are being understood. Spoken and gestural activity on the part of the listener can actively alter the intended output of the speaker - thereby creating disfluency. Many of the disfluencies generated in the transcript from MOT4 were of this type. MOT4 was clearly responding to head nods and gestures of comprehension from the listener, and interrupting herself to respond to overlapping speech by her conversational partners. In infant-directed speech, the partner is less active. Mothers may respond to the behavior of their infants, but since infants by definition do not understand their mother's speech, they are unlikely to respond in a way that will alter the mother's intended speech output.

Our answer to Chomsky's [1] and Newport et al.'s [8] contradictory assertions is that they are both right. The speech input environment of the child (and the infant) is complex and varied, both within a child's, and across different children's, experience. This input contains some instances of simple, error-free highly fluent utterances targeted directly to the infant in a quiet home environment. And it contains many noisy, highly disfluent, complex utterances for which the infant is simply along for the ride. We have only just begun to ask to what extent each of these types of input contributes to the language acquisition process. The current study suggests that infants may at least bring some tools to the table to help them differentiate "good", fluent input, from "bad", disfluent input.

7. Acknowledgements

The authors are extremely grateful for the long-term participation of the three mothers in the transcription study. We wish to thank the many participants in the preference experiment, as well as our adult raters. Thanks also to three undergraduate RAs, Rina Foygel, Megan Blossom and Greg Fay, for their work in transcribing and preparing speech samples, and to Katherine White for comments on this paper.

8. References

[1] Chomsky, N. 1962. Explanatory Models in Linguistics. In *Logic, Methodology and Philosophy of Science*, Ernest Nagel, Patrick Suppes, and Alfred Tarski (Eds.), pp. 528-550, Stanford, CA: Stanford University Press. (cited from Thomas, 2004)

[2] Fernald, A. 1985. Four-month-olds prefer to listen to motherese. *Infant Behavior and Development, 8,* 181-195.

[3] Kemler Nelson, D.G., Jusczyk, P.W., Mandel, D.R., Myers, J., Turk, A., & Gerken, L.A. 1995. The headturn preference procedure for testing auditory perception. *Infant Behavior and Development, 18,* 111-116.

[4] MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

[5] Mandel, D.R., Jusczyk, P.W. & Kemler Nelson, D.G. 1994. Does sentential prosody help infants organize and remember speech information? *Cognition, 53,* 155-180.

[6] Mandel, D.R., Kemler Nelson, D.G., & Jusczyk, P.W. 1996. Infants remember the order of words in a spoken sentence. *Cognitive Development, 11,* 181-196.

[7] Nazzi, T., Kemler Nelson, D.G., Jusczyk, P.W., & Jusczyk A.M. 2000. Six month olds' detection of clauses embedded in continuous speech: Effects of prosodic well-formedness. *Infancy, 1,* 123-147.

[8] Newport, E.L., Gleitman, H., & Gleitman, L.R. 1977. Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C.E. Snow & C.A. Ferguson (Eds.), *Talking to children: language input and acquisition*, pp. 109-149. New York: C.U.P.

[9] Soderstrom, M., Seidl, A., Kemler Nelson, D.G., & Jusczyk, P.W. 2003. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49,* 249-267.

[10] Soderstrom, M., Kemler Nelson, D.G., & Jusczyk, P.W. 2005. Six-month-olds recognize clauses embedded in different passages. *Infant Behavior and Development, 28,* 87-94.

[11] Soderstrom, M., Blossom, M., Morgan, J.L., & Foygel, I. in preparation. *The speech environment of 6-9 month old infants*

[12] Weijer, J. van de (2002). How much does an infant hear in a day? In J. Costa and M. João Freitas (Eds.). *Proceedings of the GALA2001 Conference on Language Acquisition*, pp. 279-282. Lisboa: Associação Portuguesa de Linguistíca.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 163–164.

A Cross-Linguistic Look at VP-Ellipsis and Verbal Speech Errors

Ellen Thompson

Florida International University, U.S.A.

Abstract

This paper argues that consideration of spontaneous speech errors provides insight into cross-linguistic analyses of syntactic phenomena. In particular, I claim that differences in the distribution of non-parallel VP-Ellipsis constructions in English and German, as well as variation in the spontaneously-occurring verbal speech errors, is explained by a parametric analysis of variation in the inflectional systems of the two languages.

1. A Hybrid Theory of Inflection

Lasnik (1995) argues for a parametric analysis of verbal inflection according to which a language may exhibit (i) a unitary system, where main and auxiliary verbs come either pre-inflected in the lexicon, or are derivationally constructed out of syntactically separate stem and affix, or (ii) a hybrid system, where main and auxiliary verbs are derived differently.

2. VP Ellipsis

2.1. VP Ellipsis in English

Evidence for this parametric analysis comes from VP-Ellipsis constructions. English permits parallel VP-Ellipsis, as in (1a), as well as non-parallel Ellipsis, as in (1b). It is not clear how ellipsis is resolved in (1b), since resolution requires identical forms. Assuming that English main verbs are constructed out of syntactically separate stem and affix, (1b) is possible because there is a point in the derivation of the clause where the stem and inflection are separate units, and therefore *sleep* is available as the ellipsis antecedent, as shown in (1c).

(1) a. Mary won a prize, and John did too.
b. John slept, and Mary will too.
c. John [past] sleep, and Mary Mary will ~~sleep~~ too

Warner (1986) notes that non-parallel VP ellipsis is not licensed with auxiliary verbs, as shown by (2a). This is due to the fact that there is no point in the derivation of an auxiliary verb where the verb is separate from inflection; there is no form *have* which can resolve the ellipsis site.

(2) a. *John has left, but Mary shouldn't
Impossible reading: 'John has left, but Mary shouldn't have left'
b. *John [has] left, but Mary shouldn't ~~has left~~

2.2. VP Ellipsis in German

Following the analysis of Winkler (1997), German allows VP-Ellipsis, as shown in (3). In addition, I show that German permits both non-parallel VP-Ellipsis with main verbs, as well as non-parallel VP-Ellipsis with auxiliary verbs. Following the analysis of Lasnik, this indicates that German inflection is unitary; both main and auxiliary verbs are formed by adding inflection onto the stem.

(3) …weil Leon die Aufgabe lösen kann, und auch PETER
…because Leon the task solve can and also Peter
'…because Leon can solve the task and so can Peter'

3. Verbal Speech Errors

3.1. Spontaneous Speech Errors in English

I assume, following Fromkin 1971, 1973, 1980, 1988, Cutler 1982, that occurring spontaneous speech errors are constrained to those errors which are statable in terms of the linguistic system. As Fromkin (1988:121) notes, "…spontaneously produced speech errors reveal deviations in the units and rules" of language.

The analysis outlined above of English non-parallel VP Ellipsis constructions predicts the different behavior of main and auxiliary verbs in spontaneous speech errors. For example, an irregular main verb may appear in speech errors as a regular form, as in (4a)-(b) (the intended utterance appears to the left and the spoken utterance to the right of the arrow; all English errors are from the UCLA Speech Error Corpus; see http://www.mpi.nl/world/corpus/sedb/). However, we do not observe similar errors with irregular auxiliary verbs.

(4) a. the last I knew about that ->
the last I knowed about that
b. … and the objects that would be locally bound ->
… and the objects that would be locally binded

In addition, as is seen in the examples in (5a-b), main verbs may switch position, in which case inflectional material may be stranded and each verb surface with the inflection of the other verb. It is also possible for the inflectional material itself to switch position between two verbs, as in (5c). Errors with auxiliary verbs appearing in these error patterns are not attested in the data.

(5) a. We've learned to love mountains ->
We've loved to learn mountains
b. It goes to show -> It shows to go
c. I saw him digging up those bulbs ->
I see him dugging up ...

Observe that in movement errors the inflectional item itself may move from the verb to another item of the utterance, as in (6a-b). However, these movement errors seem to be restricted to main verbs.

(6) a. He kind a tends ta ... -> He kinds a tend ta ...
b. If she wants to come here ... ->
If she want to comes here

It is possible to find errors with the main verb deleted, and the inflection of the verb stranding, as in (7a). In contrast, when an auxiliary verb is targeted for deletion, the whole unit

is affected, as shown in (7b). In sum, the UCLA Speech Error Corpus contains thirteen examples of main verb separated from inflection, and zero of auxiliary verb separated.

(7) a. As I keep suggesting -> As I keeping
b. He doesn't seem happy now -> He not seem happy now

We can account for these errors by claiming that they result when the derivation is accessed before the verb and inflection have become a unit. We therefore predict that auxiliary verbs do not appear in these error patterns, since there is no point in the derivation of an auxiliary verb at which the verb and inflection are separate units.

3.2. Spontaneous Speech Errors in German

In German, we observe that, like in English, we find examples of main verbs switching position, with inflectional material of the verbs stranded, as in (8) (example from Bierwisch 1982:32). We also find examples of inflectional material of main verbs switching position, as in (9) (examples hereafter from Wiedenmann 1992).

(8) Ich kann nur über die Teile sprechen, die ich kenne- >

I can only about those parts speak that I know

Ich kann nur über die Teile kenn-en, die ich sprech-e

I can only about those parts know-INF that I speak-1.sg

'I can only speak about those parts that I know.'

(9) … daβ dein Zimmer komm-st, und du räum-t ->

… that your room come-2.sg.pre and you clean-3.sg.pre

… daβ dein Zimmer kommt, und du räumst

'… that you come and clean your room'

However, unlike the pattern in English, we also observe reversal errors where the inflection of a main and an auxiliary verb switch position, as in (10).

(10) die ich endlich mal weg-räum-te woll-en ->

the I finally away-clear-past will-infin

die ich endlich mal weg-räum-en woll-te

the I finally away-clear-infin will-past

In addition, we find movement errors where the auxiliary verb inflection separates from the stem and appears attached to another auxiliary verb, as in (11). (This example plausibly involves movement of inflection from *werden* to *muβ*, followed by reversal of the auxiliary stem and *schen*.)

(11) man schen werden muβ -> man wird schen mussen

one see will must-3.sg.pre -> one will-3sg.pre see must

'One must see.'

4. Conclusion

The different behavior of German and English auxiliary verbs is predicted if we assume that German main and auxiliary verbs are not distinct – both are constructed out of syntactically separate stem and affix. Therefore, German auxiliary verbs, as well as main verbs, allow separation and manipulation of inflection and verbal stem in errors, as well as in non-parallel VP Ellipsis constructions.

This work thus supports the claim that spontaneous speech errors pattern differently depending on the structural properties of the language, and it provides evidence for Lasnik's (1995) division of languages into inflectionally unitary and inflectionally hybrid systems. Although research in speech errors has investigated language variation in the phonological domain (Berg 1987, Wells-Jensen 1999), variation in syntactic structures remains little-explored.

5. Acknowledgements

I would like to thank Bo Lisa Yang for her assistance with the German data.

6. References

[1] Berg, Thomas. 1987. A Cross-Linguistic comparison of slips of the tongue. Bloomington: Indiana University Linguistics Club.

[2] Fromkin, Victoria. 1988. Grammatical Aspects of Speech Errors. In Linguistics: The Cambridge Survey, Vol 2, Frederick Newmeyer, ed., Cambridge: Cambridge University Press.

[3] Lasnik, Howard. 1995. Verbal morphology: Syntactic Structures meets the Minimalist Program. In Evolution and Revolution in linguistic theory: Essays in honor of Carlos Otero, ed. Hector Campos and Paula Kempchinsky, 251-275. Washington, D. C.: Georgetown University Press.

[4] Pfau, Roldand. 2000. Features and Categories in Language Production. Ph.D. dissertation, Dept. of German Language and Literature, University of Frankfurt/Main.

[5] Wells-Jensen, Sheri. 1999. Cognitive Correlates of Linguistic Complexity: A Cross-Linguistic Comparison of Errors in Speech, Ph.D. Dissertation, SUNY-Buffalo.

[6] Wiedenmann, Nora. 1992. A Corpus of German Speech Errors. Instituts fur Phonetik und Sprachliche Kommunikation der Universitat Munchen (FIPKM) 30, S1-S77.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 165-168.

Acoustic-phonetic decoding of different types of spontaneous speech in Spanish

Doroteo T. Toledano[a], Antonio Moreno Sandoval[b], José Colás Pasamontes[a], Javier Garrido Salas[a]

[a]Human-Computer Technologies Laboratory (HCTLab), Escuela Politécnica Superior.
[b]Laboratorio de Lingüística Informática (LLI-UAM)

Universidad Autónoma de Madrid

Abstract

This paper presents preliminary acoustic-phonetic decoding results for Spanish on the spontaneous speech corpus C-ORAL-ROM. These results are compared with results on the read speech corpus ALBAYZIN. We also compare the decoding results obtained with the different types of spontaneous speech in C-ORAL-ROM. As the most important conclusions, the experiments show that the type of spontaneous speech has a deep impact on spontaneous speech recognition results. Best speech recognition results are those obtained on speech captured from the media.

1. Introduction

Currently, spontaneous speech processing is one of the most active research lines in speech technology, and in particular in speech recognition. In the last years, the National Institute of Standards and Technology (NIST) [1] has launched a new series of competitive evaluations under the name Rich Transcription in which spontaneous speech processing (and particularly disfluency detection) is a key topic. Unfortunately that program does not include Spanish as one of the languages of interest. There are a few Spanish research groups that are conducting research in the field of spontaneous speech processing [2]. However, this field is still largely unexplored in Spanish.

In this paper we present initial results of our research using the C-ORAL-ROM corpus [3]. This corpus is a spontaneous speech corpus covering several languages. One of the main features of the corpus is that spontaneous speech is classified under several categories. Section 2 describes this corpus as well as some adaptations that have been necessary to automatically process it. The experiments described are acoustic-phonetic decodings performed using Hidden Markov Model (HMM) as acoustic models. These models have been trained using a read speech corpus in Spanish, ALBAYZIN [4]. Section 3 describes the training of the HMMs used. Section 4 presents the acoustic-phonetic decoding results, both for read speech and spontaneous speech and compares them. This section also compares results on different types of spontaneous speech, as defined in the C-ORAL-ROM corpus. Finally, section 5 summarizes the most important conclusions as well as future research lines.

2. Description of the C-ORAL-ROM corpus

C-ORAL-ROM is a multilingual corpus that comprises four romance languages: Italian, French, Portuguese and Spanish. In our work we have used the Spanish sub-corpus, which contains around 300.000 spoken words. From a sociolinguistic point of view, speakers are characterized by their age, gender, place of birth, educational level and profession. From a textual point of view the corpus is divided into the parts shown on Table 1 [5].

Table 1: Distribution of words in C-ORAL-ROM.

<table>
<tr><td colspan="4">Informal
150.000 words</td><td>Formal
150.000 words</td></tr>
<tr><td colspan="2">Familiar
113.000</td><td colspan="2">Public
37.000</td><td>Formal in natural context
65.000</td></tr>
<tr><td rowspan="2">Monologs
33.000</td><td rowspan="2">Dialogs/
Convers.
80.000</td><td rowspan="2">Monologs
6.000</td><td rowspan="2">Dialogs/
Convers.
31.000</td><td>Formal on the media
60.000</td></tr>
<tr><td>Telephone conversations
25.000</td></tr>
</table>

Table 1 shows that the main division is balanced between formal speech and informal speech. For informal speech a division is considered between speech in a familiar/private context and speech in a public context. The first group is further classified into monologs, dialogs and conversations with three or more speakers. The second group is similarly classified into monologs, dialogs and conversations. Regarding formal speech, a division has been made between speech in natural context and speech on the media. The former includes political speeches, political debates, preaching, teaching, professional expositions, conferences, speech in business contexts and speech in legal contexts. Speech on the media (also referenced in this article as *broadcast news*, which is the usual name for this kind of speech in the context of automatic speech recognition) includes news, sports, interviews, meteorology, science, reports and talk shows. Telephone conversations, although initially considered under the formal speech category in C-ORAL-ROM, have very particular features and is more similar to informal speech than to formal speech. For these reasons we have considered telephone conversations under the category of informal speech on a subdivision of its own.

These divisions and subdivisions of C-ORAL-ROM will allow us to compare the acoustic-phonetic decoding results using different types of spontaneous speech.

C-ORAL-ROM contains 183 recordings totaling over 40 hours of speech. There are basically three type of recordings depending on their duration: 7-10 minutes, 15 minutes and 30 minutes. These recordings were too long for their automatic processing. For that reason, we extracted each spoken utterance (between pauses) on a separate file using the existing C-ORAL-ROM manual segmentation. This manual segmentation has been essential to perform the experiments described in this paper.

Table 2: Divisions of C-ORAL-ROM.

Informal	*Familiar/Private*	*Monolog*
		Dialog
	Public	*Conversation*

Formal		
Formal in natural context	*Media (Broadcast News)*	*Telephone*
Political speech	*News*	
Political debate	*Sports*	
Preaching	*Interviews*	
Teaching	*Meteorology*	
Professional exposition	*Scientific*	
Conferences	*Reports*	
Business	*Talk shows*	

2.1. Phonological transcription

In order to compare acoustic-phonetic decoding results, a reference phonological transcription is required in advance. C-ORAL-ROM did not include that phonological transcription, including only an orthographic one. For that reason, the phonological transcription was generated from the orthographic one, making use of a simple phonological transcriptor based on rules. This transcriptor uses a minimum set of phonemes for Spanish (23 phonemes). Obviously, such a simple transcriptor does not allow to obtain a correct transcription in all cases. However, we consider that the precision achieved is good enough to obtain significant acoustic-phonetic decoding results.

3. Training of the HMMs for acoustic-phonetic decoding

The Hidden Markov Models used to perform the acoustic-phonetic decoding were trained on the ALBAYZIN corpus using the Hidden Markov Model ToolKit (HTK) software [6]. The front-end used for feature extraction was the advanced distributed speech recognition front-end defined by the ETSI standard ETSI ES 202 050 [7]. This front-end includes mechanisms for robustness against channel (convolutive) distortion and additive noise. Basically the mechanism used for noise robustness is a double Wiener filter that estimates and substracts the noise spectrum. The one used against convolutive distortion is cepstral mean normalization (CMN).

The set of phonemes used in all experiments is the minimum set of 23 phonemes in Spanish. We also consider models for initial, final and intermediate silences. We trained both context-dependent and context-independent models. We started training seed context-independent models using 600 of the 1200 utterances of ALBAYZIN that were phonetically labeled and segmented by hand (the other 600 were reserved for adjustment and evaluation purposes). Next we used those seed models to train context-independent models with 3500 utterances from the training set of ALBAYZIN. We trained models with up to 150 Gaussians per state. However, we observed that results improved very slightly using over 65 Gaussians, so we decided to use that number of Gaussians per state. From the context-independent models, we trained context-dependent models and then performed state tying making use of an algorithm based on a decision tree. The models resulting from the state tying contained a total number of states of 2079. Given that the context-independent models contained a total of 26x3x65=5070 Gaussians, we chose to use context-dependent models with a complexity similar to that used in the context-independent ones. This way we can compare context-independent and context-dependent models. Following this reasoning we chose to use context-dependent models using 2 Gaussians per state, which implied a total of 2079x2=4158 Gaussians.

4. Acoustic-phonetic decoding results

The test we performed consisted of the evaluation of the accuracy of the acoustic phonetic decoding achieved with the models. In other words, we tried to determine the phonemic recognition accuracy using just the acoustic models, without any other kind of lexical or grammatical restriction. The only restrictions imposed were that each utterance should start and end with a silence. For the case of the context-dependent models, we also imposed that the contexts should be respected.

In order to evaluate the results, we aligned the phonemic string obtained from the decoder and the reference phonemic string obtained from the phonological transcriber (section 2.1). Using this alignment the percentage of phones correctly recognized (%C) and the phonemic decoding accuracy (%A) were computed. The phonemic decoding accuracy is the percentage of phones correctly detected minus the percentage of inserted phones.

4.1. Acoustic-phonetic decoding of read speech

It is important, before proceeding to further analysis of the results, to have an idea of the precision reached by the acoustic models under optimal conditions. These optimal conditions mean in our case read speech recorded under the same acoustical environment and conditions as the training speech. To assess that optimal performance we have made an acoustic-phonetic test on a subset of 300 utterances of the ALBAYZIN corpus. These utterances were phonetically segmented and labelled by hand and were not used in the training phase.

Using this test set, the acoustic-phonetic decoding with context-independent models reached %C = 81.07% correct phonemes and A = 76.56% phonemic accuracy. These results were evaluated using as reference phoneme strings the phoneme labels produced by the automatic transcriber based on the orthographic transcription. In order to check the validity of this phonemic string as reference string we also evaluated the same results comparing against the manually annotated reference phonemic labels. These results (%C = 81.36% and %A = 76.24%) are very similar to those using the automatically generated phonemic transcription. This justifies our evaluation of the acoustic-phonetic transcription of the C-ORAL-ROM corpus using an automatically generated reference phomemic labelling (there is not manually verified phonemic annotation for the C-ORAL-ROM corpus yet).

The former results were always using context-independent HMMs. If we use context-dependent models we obtain %C = 83.88% and %A = 74.55% when comparing against the automatic phonemic transcription. If we compare against the manual phonemic transcription results are very similar, %C = 83.79%, %A = 73.01%. Results obtained with context-dependent HMMs are also very similar to those obtained with context-independent HMMs.

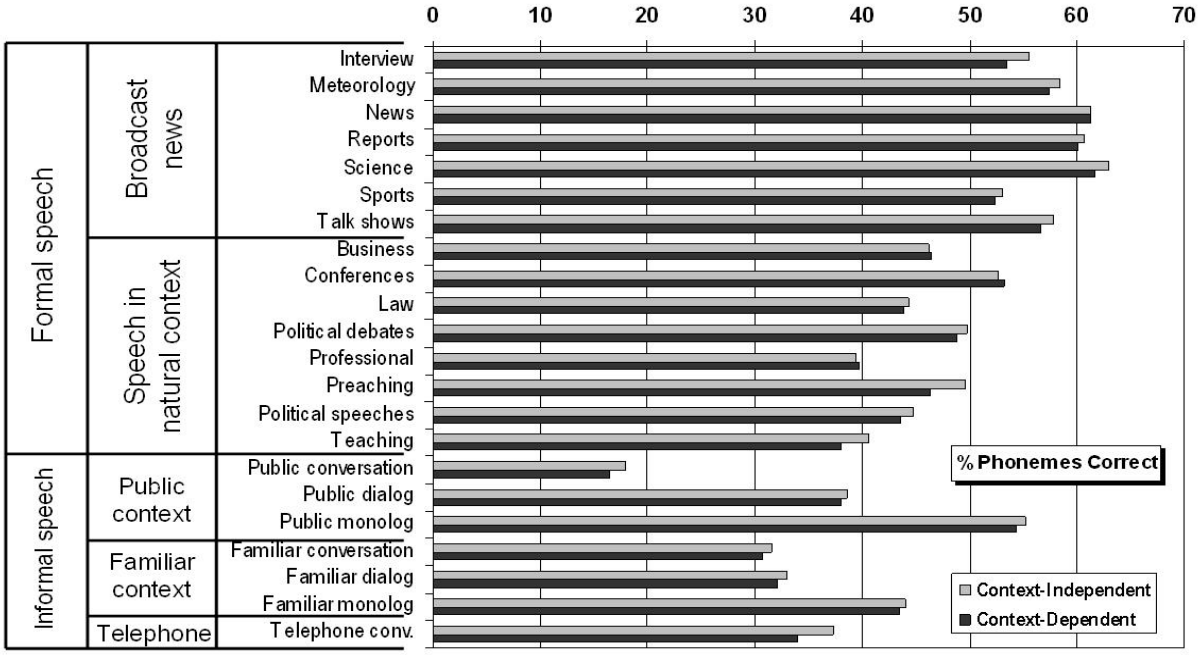


Figure 1: Phonemic decoding results (percentage of phones correct, %C) by subtype of spontaneous speech (see Section 2) of C-ORAL-ROM.

4.2. Acoustic-phonetic decoding of spontaneous speech

Once acoustic-phonetic decoding has been evaluated on read speech, we evaluate in this subsection the acoustic-phonetic decoding of spontaneous speech. If we perform the same test on the whole C-ORAL-ROM corpus using as reference phonemic labeling the automatically generated transcription, we get more modest results, as expected. In the case of using context-independent HMMs we obtain %C = 44.00% and %A = 25.71%. For the case of context-dependent HMMs results are again very similar, %C = 43.06% and %A = 25.07%.

Such a reduction in acoustic-phonetic decoding performance may be mainly due to the inherent difficulty that spontaneous speech presents for automatic processing. However, it would be misleading to consider that this is the only factor causing such a drastic decrease in phonetic decoding accuracy. Other factors that have an important impact on that reduction are the following:

- The channel mismatch between the speech used for training the HMMs and the speech on which the decoding was performed. ALBAYZIN is a head-mounted microphone, clean speech corpus, while C-ORAL-ROM is a corpus that includes speech recorded with different microphones on different acoustic environments (more or less noisy), speech taken from the media and even speech taken from telephone conversations. This channel mismatch is mitigated partially by the mechanisms of robustness against channel distortion and additive noise provided by the feature extraction front-end used [7]. However, its influence on the decoding results may still be important.

- The presence of noise with different characteristics and levels in C-ORAL-ROM. This effect is also mitigated, but not avoided, by the use of a front-end with mechanisms of robustness against noise [7].

- The mismatch between the characteristics of the speech used for training (ALBAYZIN) and testing (C-ORAL-ROM), both in type of speech and noise levels. It could be possible to perform a retraining or adaptation of the models using speech from C-ORAL-ROM. In this way, the acoustic models would be more adapted to the speed and level of the speech and the noise in C-ORAL-ROM, and presumably the phonetic decoding accuracy would increase.

All these factors limit the utility of the comparison between the phonetic decoding accuracy on read (Section 4.1) and spontaneous (Section 4.2) speech. However, even more interesting than this comparison is the comparison between the acoustic phonetic decoding accuracy on the different types of spontaneous speech in C-ORAL-ROM.

4.3. Comparison of phonetic decoding results on different types of spontaneous speech

Figure 1 shows the percentage of phones correctly recognized for each of the different types of spontaneous speech considered in C-ORAL-ROM, and briefly described in Section 2.

The comparison between context-dependent and context-independent acoustic models shows that results for both are very similar, although the context-independent ones are slightly better. This result could be due to the fact that our context-independent models are slightly more complex, since they include a larger overall number of Gaussians than the context-dependent models. Very likely an increase in the complexity of the context dependent models would produce an important improvement on the results shown here.

It is very interesting to see that there is a wide range of variation between the different types of spontaneous speech considered: from less than 20% phonemes correct for informal conversations in public context to over 60% phonemes correct for science programmes on the media.

In general, it can be observed that for conversations and dialogues results are among the worse obtained (around 30% phonemes correct for the whole group). Another subset related to them (in that it also contains dialogues and conversations) is the subset of telephone conversations for which results are similar. In all these cases it seems obvious that the interaction

(with frequent overlappings) among the speakers is the cause of the reduced phonetic decoding performance. In the case of the telephone conversations there also exists a clear mismatch between the characteristics of the speech used to train the acoustic models and that used to perform the phonetic decodings.

Regarding the informal monologs, it can be observed that in familiar context results are slightly better than for dialogs and conversations (slightly over 40%), while in public context results are clearly superior (close to 55% phonemes correct).

The subsets mentioned in the former paragraphs correspond to informal speech. It can be observed that, with the only exception of the monologs in public context (epubmn), results are always worse than those obtained with formal speech, both in natural contexts and on the media (broadcast news). Comparing these two big groups it can be realized that speech from formal situations in natural context tend to produce results worse (around 40% or 50% phonemes correct) than those observed on formal speech on the media, for which phonetic decoding results tend to be between 50% and 60% phonemes correct.

Comparing the different subsets within the formal speech on the media, interesting differences may be observed. Worse results are obtained with sports programmes, probably due to a less careful use of the language and exaggerated articulations as well as more overlappings between different speakers. Slightly better are the results obtained on interviews, where overlappings might also be very frequent. Following, and with intermediate results, are the results on meteorology programmes and talk shows. Finally, best results are attained on news programmes, reports and scientific programmes. It might be argued that this kind of programmes have a reduced number of overlappings as well as a more careful use of language, presumably with less difluencies.

As a final experiment, we have compared the results of the automatic phonetic decoding with the problems found by human experts when transcribing the recordings in C-ORAL-ROM. These difficulties were analyzed in [8]. In doing this comparison we can observe very significative coincidences. In particular, human transcribers found serious difficulties with the typical features of interaction in a spontaneous communication: overlappings, number of words per turn, and speaking rate. In this case the following intuition applied

Scale 1: Degree of formality

informal media formal

+difficult _____ - difficult

Scale 2: Number of speakers

conversation dialog monolog

+ difficult _____ - difficult

On the first scale, the more formal the speech type, the easier to transcribe it because more rethoric and discursive conventions are followed. Speech is more predictable and pronunciation is more careful.

On the second scale, the more speakers talking on a recording, the more difficult the transcription because of the need of distinguish among the different turns and speakers and of the need to take care of overlappings. With monologs this difficulty is reduced to the minimum.

These findings coincide basically with those obtained in phonetic recognition experiments: the easier recordings to transcribe are those on the media, produced by professional speakers that combine a good diction with experience of fluid elaboration within the linguistic rules. The more we move towards informal speech with more speakers, the more complex are both manual and automatic transcription.

5. Conclusions and future work

In this paper we have presented phonetic decoding results on spontaneous speech and have compared them to the results obtained on read speech with the same characteristics as the speech used to train the acoustic models. This comparison shows an overall relative reduction of the percentage of phones correctly recognized of about 50% when we move from read speech to spontaneous speech. Although the influence of the characteristics of the type of speech (read vs. spontaneous) on the results of phonetic decoding is clear, it is also true that these experiments are also influenced by other factors like channel mismatch and noise level mismatch between training and testing. This makes the comparison between phonetic decoding accuracy for read and spontaneous speech of limited utility.

Much more interesting is the comparison of phonetic decoding results on different types of spontaneous speech. Among the different types of spontaneous speech analyzed, best results are those with speech taken from the media. For this kind of spontaneous speech results show a relative worsening of only 25% (approximately) from the results obtained on read speech with the same characteristics used to train the acoustic models. This means that this kind of spontaneous speech is the easiest to process automatically among those analyzed. Following in order of complexity are the formal speech in natural contexts, the informal monologs, and finally the informal dialogs and conversations, for which overlappings and interruptions make the complexity of the automatic processing of this kind of speech much higher than for the former types. These results largely coincide with the experience of human transcribers.

As future work, we would like to extend this study to take into more detailed consideration the influence of aspects like the frequency of overlappings, interruptions and disfluencies in acoustic-phonetic decoding results.

6. References

[1] http://nist.gov/speech/tests/rt/

[2] Luis Javier Rodríguez Fuentes. Estudio y modelización acústica del habla espontánea en diálogos hombre-máquina y entre personas. Tesis Doctoral. Facultad de Ciencia y Tecnología Universidad del País Vasco.

[3] Cresti & Moneglia (eds.), C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, Amsterdam, John Benjamins, 2004.

[4] Climent Nadeu. ALBAYZIN, Universitat Politecnica de Catalunya, ETSET New Jersey, 1993.

[5] Moreno Sandoval, A. La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. Actas de la II Jornadas en Tecnologías del Habla, diciembre 2002, Granada.

[6] Young, S. et al., The HTK Book (for HTK Version 3.1), Microsoft Corporation, July 2000.

[7] Aurora Front-End manual. ETSI ES 202 050 V1.1.3 (2003-11).

[8] González Ledesma, A.; De la Madrid, G.; Alcántara Plá, M.; De la Torre, R.; Moreno-Sandoval, A. Orality and Difficulties in the Transcription of Spoken Corpora. Proceedings of the Workshop on Compiling and Processing Spoken Language Corpora, LREC, 2004, Lisbon.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 169-172.

The effects of filled pauses on native and non-native listeners' speech processing

Michiko Watanabe[1], Yasuharu Den[2], Keikichi Hirose[3] & Nobuaki Minematsu[1]

[1]Graduate School of Frontier Sciences, University of Tokyo, Japan
[2]Faculty of Letters, Chiba University, Japan
[3]Graduate School of Information Science and Technology, University of Tokyo, Japan

Abstract

Everyday speech is abundant with disfluencies. However, little is known about their roles in speech communication. We examined the effects of filled pauses at phrase boundaries on native and non-native listeners in Japanese. Study of spontaneous speech corpus showed that filled pauses tended to precede relatively long and complex constituents. We tested the hypothesis that filled pauses biased listeners' expectation about the upcoming phrase toward a longer and complex one. In the experiment participants were presented with two shapes at one time, one simple and the other compound. Their task was to identify the one that they heard as soon as possible. The speech stimuli involved two factors: complexity and fluency. As the complexity factor, a half of the speech stimuli described compound shapes with long and complex phrases and the other half described simple shapes with short and simple phrases. As the fluency factor phrases describing a shape had a preceding filled pause, a preceding silent pause of the same length, or no preceding pause. The results of the experiments with both native and non-native listeners showed that response times to the complex phrases were significantly shorter after filled or silent pauses than when there was no pause. In contrast, there was no significant difference between the three conditions for the simple phrases, supporting the hypothesis.

1. Introduction

Spontaneous speech, unlike written sentences or speech read aloud from written text, is full of disfluencies such as filled pauses (fillers), repetitions, false starts and prolongations. It has been reported that about six per 100 words are disfluent in conversational speech in American English [7]. It has been found that every 13 words are disfluent among female speakers and every 10 words are disfluent among male speakers in Japanese presentations [8]. In spite of their abundance in everyday speech not many empirical studies have been conducted into their effects on speech communication either in native or non-native languages.

Three general views are possible about the effects of disfluencies on listeners.

1) Disfluencies disturb listeners.
2) Disfluencies neither harm nor help listeners.
3) Disfluencies are helpful to listeners.

In a native language, listeners hardly seem to be disturbed by disfluencies. Fox Tree [3] found in her experiments using identical word monitoring task that existence of repeated words in a sentence did not affect reaction times to target words immediately after repetitions. This suggested that repetition of words had no effect on speech processing of listeners. Using the same methodology Fox Tree [4] tested the effects of two types of fillers, "um" and "uh", on native listeners' comprehension in English and Dutch. The author used the term "fillers" to refer only to the voiced parts of filled pauses. In both languages the time that listeners needed to monitor target words were shorter when "uh" was present immediately before the target words than when it was digitally excised. However, no difference was found between the two conditions with "um". The results indicated that "uh" was helpful to listeners while "um" neither helped nor hindered comprehension. In any case, her experiment showed no negative effect of fillers on listeners.

In contrast with the effects of repetitions and fillers, listeners' reaction times to target words were longer when the target words were preceded by false starts than when the false starts were cut out, indicating that false starts had negative effects even on native listeners.

Regarding disfluencies in non-native languages, some researchers have argued that they are the main obstacle for listeners' perception and comprehension of speech. Voss [9] asked German subjects to transcribe a stretch of spontaneous English and analysed the transcripts. He found that nearly one third of all perception errors were connected with disfluencies. Misunderstanding was due to either misinterpreting disfluencies as parts of words or to misinterpreting parts of words as disfluencies. Fukao et al. [5] reported that international students studying at Japanese universities had difficulties in coping with disfluencies in lectures. They were sometimes not able to distinguish filled pauses from words and had problems in processing ungrammatical sentences, repairs, omissions or speech errors in lectures.

On the other hand, it has been claimed that disfluencies are sometimes helpful to listeners. Blau [1] compared non-native listeners' comprehension of monologues between three conditions: (1) normal speed, (2) modified to include extra three second pauses inserted, on average, every 23 words, and (3) with similar pauses filled with hesitations such as "well", "I mean", and "uh". Comprehension of the filled pause version was significantly better than that of the normal version and slightly better than the silent pause version. The results indicated that filled pauses sometimes helped comprehension of non-native listeners.

Summarizing the discrepant results of previous research about the effects of disfluencies on non-native listeners, Buck [2] argued that disfluencies, as well as silent pauses, which slowed down the speech rate, helped comprehension of non-native listeners as long as disfluencies were recognised as disfluencies. If listeners failed to recognize disfluencies as such, they could have detrimental effects. However, as studies with native listeners showed, word repetitions and fillers "um", which slowed down the speech rate, measured by the amount of linguistic information conveyed per unit time, neither helped nor hindered comprehension [3], [4]. Buck's argument

needs more empirical support and detailed analysis of various types of disfluencies at different locations.

In the present research we have examined the effects of filled pauses at phrase boundaries on native and non-native listeners' ability to process speech. It has been reported that filled pauses amount to about 70 % of the total disfluencies in Japanese [8]. The Japanese language seems to have a wider variety of filled pauses than English and Dutch. "Ano", "e", "eto" and "ma" have been listed as the most frequent fillers both in dialogues and monologues [6], [10].

Corpus based studies of spontaneous Japanese showed that filled pauses tended to appear more frequently before relatively long and complex constituents. Watanabe et al. [12] showed that the probability of filled pauses occurring at clause boundaries increased in a roughly linear manner when expressed as a function of the number of words in the following clauses. Watanabe [11] carried out a study on Japanese phrases sandwiched between silent pauses longer than 200 ms, "Inter Pausal Units (IPUs)", which tended to be units shorter than clauses. These IPUs contained significantly larger numbers of morae, words and phrases when they were immediately preceded by fillers than when they were not.

Based on these findings we have inferred that listeners are making use of this tendency in occurrence of filled pauses when they process speech. In the experiments described below we have tested the hypothesis that filled pauses bias listeners' expectations about the upcoming phrase toward a relatively long and complex one. We have tested this hypothesis with native speakers of Japanese in Experiment 1 and with non-native speakers of Japanese in Experiment 2.

2. Experiment 1

2.1. Outline

A pair of shapes in the same colour was presented side by side on a computer screen, one a simple shape (circle, triangle or square) and the other a compound shape (two arrows attached to a paired shape. See Fig. 1). One second after a visual stimulus had appeared speech referring to one of the two shapes was played. Participants were instructed to press a button corresponding to a shape being referred to as soon as possible. The instruction given to the participants was as follows (translated from Japanese): "A woman is asking to bring a paper decoration in a certain colour and a shape. Which one is she asking for? Two pictures of paper appear on the computer screen. Please press either a left or right mouse button corresponding to the paper that she is asking for as soon as possible."

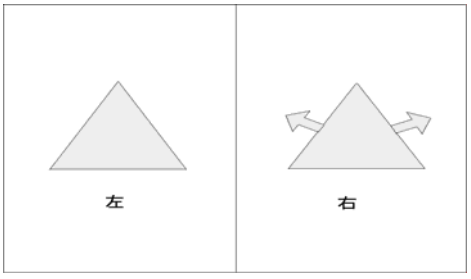


Figure 1: An example of visual stimuli. Visual stimuli always had a simple shape (round, square or triangular) on one side and a compound shape (with two arrows attached to the simple shape) on the other. The two shapes were always displayed in the same colour.

2.2. Speech stimuli

Each utterance contained a word describing a colour (we call it "a colour word") and a word describing a shape (we call it "a shape word") in this order as in "yellow and triangular". The speech stimuli involved two factors: 1) **complexity factor**: either a simple shape or a compound shape was referred to (we call the conditions, "simple condition" and "complex condition" respectively); 2) **fluency factor**: a shape word was immediately preceded by a filled pause, a silent pause of the same length as a filled pause, or no pause (we call the conditions, "filler condition", "pause condition" and "fluent condition", respectively). Examples of speech stimuli are given below with English translation. Fillers are in italic and phrases describing a shape are in bold.

An example of a simple phrase with a filler:
(1) Anone, tonari no heya kara kiirokute *eto* **sankaku no**
Look, next of room from yellow and *um* triangle of
kami mottekite kureru?
paper bring (auxiliary)
Translation: Look, could you bring a yellow and *um* **triangular** paper from the next room?

An example of a complex phrase with a filler:
(2) Anone, watashi no heya kara kiirokute *eto* **sankaku ni**
Look, I (genitive) room from yellow and *um* triangle to
yajirushi ga **tsuita** kami mottekite kureru?
arrows (nominative) attached paper bring (auxiliary)
Translation: Look, could you bring a yellow and *um* **triangular** paper **with arrows** from my room?

We assumed that filled pauses were more typical before a phrase describing a compound shape rather than before a phrase describing a simple shape because a phrase describing a compound shape was generally longer and more complex. We predicted that when a filler was uttered, listeners were more likely to expect a phrase describing a compound shape to follow. As a result, when a phrase describing a compound shape was actually uttered, listeners' response times to the phrase would be shorter than when there was no filler before the phrase. On the other hand, when a phrase describing a simple shape was uttered after a filler, listeners' response times to the phrase would not be shorter than when there was no filler because the filler was not in a typical location and therefore not a good cue to the type of phrase that followed.

We included the silent pause condition to examine whether silent and filled pauses of the same duration had different effects. As the other parts of the speech were kept constant, any difference should be attributable to whether the pause contained a voice or not.

Speech stimuli were created in the following way: one of the authors uttered sentences asking a supposed interlocutor to bring a sheet of paper of a certain colour and shape from a certain place. Although the test stimuli were presented to the speaker as a reading list, the speaker uttered sentences without looking at the list so that utterances sounded like natural, everyday speech. The speaker uttered 180 sentences. The utterances were recorded with an AKGC414B Studio microphone in an acoustically treated recording studio. The speech was sampled at 44 kHz and digitized at 16 bits directly onto a PC. All the utterances contained a filler "eto" immediately before a shape word. We called the original speech "a filler version". Original utterances were edited with speech analyzing software and two new versions were created: 1) a pause version: filled pauses were substituted by silence

with the same length as filled pauses; 2) a fluent version: filled pauses were edited out. Three sets of stimuli, each of which contained 180 sentences, were created so that only one of the three versions from the same utterances appeared in each stimuli set. The amplitude of speech stimuli was normalised.

2.3. Participants

Thirty university students who were native speakers of Tokyo Japanese took part in the experiment.

2.4. Procedure

The experiment was carried out in quiet rooms at Tokyo University and Chiba University in Japan. Participants were randomly assigned to one of the three stimuli sets. After eight practice trials the participants listened to 180 sentences. The order of stimuli was randomised for each participant. Speech stimuli were presented through stereo headphones. Sentences were played to the end no matter when the participants pressed a response button. Time out was set within 500 ms from the end of sentences. There were three second intervals between the trials. The experiment lasted about 40 minutes excluding the practice session and a short break in the middle.

Response times from the beginning of sound files were automatically measured. The onset of the first words describing a shape was marked manually referring to speech sound, sound waves and sound spectrograms. In the example sentences (1) and (2) the word onsets were marked at the beginning of /s/ in "**sankaku** (triangle)". Response times from the word onset were calculated by subtracting the word onset time from response times measured from the beginning of sound files.

The medians of correct response times from the word onset in each condition for each participant were calculated and the mean medians of six conditions were compared.

2.5. Results

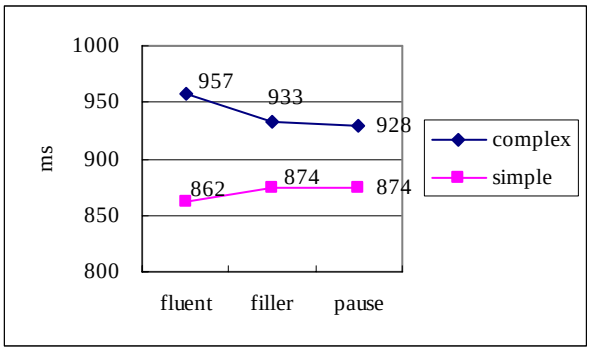


Figure 2: Japanese participants' mean response times from the onset of the first word describing a shape. 'Complex' means complex phrases and 'simple' means simple phrases.

Mean response times from the onset of shape words are shown in Figure 2. Two-way repeated measures analysis of variance (ANOVA) showed a main effect of complexity factor ($F(1, 29) = 76.051$, $p < .001$). A complexity-fluency interaction was significant ($F(2, 58) = 5.537$, $p < .006$). Post-hoc tests revealed that there was a significant difference between fluent-filler-pause conditions in the complex condition ($F(2, 28) = 6.533$, $p < .005$), but no significant difference in the simple condition ($F(2, 28) = 1.208$, $p = .314$). In the complex condition paired comparisons (alpha adjusted Bonferroni) showed significant differences between fluent-filler and fluent-pause conditions but no significant difference between filler-pause conditions ($t(29) = 3.329$, $p < .007$; $t(29) = 3.031$, $p < .015$; $t(29) = 0.492$, $p = 1.000$, respectively).

2.6. Discussion

Response times to complex phrases were shorter when a filled pause was present immediately before the phrase than when there was no preceding pause. On the other hand, there was no significant difference in response times to simple phrases between the filler and the fluent conditions. These results showed that existence of filled pauses accelerated listeners' responses to complex phrases but did not affect their responses to simple phrases, which was in accordance with our prediction and supported the hypothesis.

There was no significant difference in response times between the filler and the pause conditions in either the simple or the complex condition. This result indicated that the effects of filled pauses at phrase boundaries did not differ from the effects of silent pauses as long as the durations were the same.

3. Experiment 2

3.1. Outline and material

The outline and the material were the same as Experiment 1.

3.2. Participants

Thirty-eight native speakers of Chinese who had been staying in Japan for more than half a year and were fluent in everyday Japanese took part in the experiment. All the participants were either students or researchers at Chiba University or Tokyo University in Japan. Data from three participants were excluded from the analysis because they turned out to be bilingual speakers of Chinese and other languages. If the number of error trials, combined with trials which timed out, exceeded 18 for any participant, (i.e. exceeded 10% of presented trials), participants were excluded. That is, only participants scoring at least 162 out of 180 trials correct were considered for analysis. Five participants were excluded for this reason. This means that 30 participants were retained for analysis.

3.3. Procedure

The procedure was basically the same as that of Experiment 1. Most participants received eight practice trials, which was the same number presented to participants in Experiment 1. However, some participants did not press the response button until the utterance came to the end in the trial session. In each of these cases the participants were instructed not to wait until the speech ended, but to press the response button as soon as they knew the answer. They did four additional practice trials before starting the experiment.

3.4. Results

The ratio of the sum of errors and time outs of valid data was 2.6%. Mean response times of correct answers from the onset of shape words are shown in Figure 3.

Two way repeated measures analysis of variance (ANOVA) revealed that there were significant main effects of the complexity factor and the fluency factor ($F(1, 29) = 8.555$, $p < .007$; $F(2, 58) = 5.155$, $p < .009$, respectively). There was a significant complexity-fluency interaction ($F(2, 58) = 6.274$, $p < .003$). Post-hoc tests revealed that there was a significant difference between fluent-filler-pause conditions in the complex condition ($F(2, 28) = 7.867$, $p < .002$), but no significant difference in the simple condition ($F(2, 28) = .957$, $p = .396$). In the complex condition paired comparisons (alpha adjusted Bonferroni) showed significant differences between fluent-filler and fluent-pause conditions but no significant

difference between filler-pause conditions ($t(29) = 3.793$, $p < .002$; $t(29) = 3.219$, $p < .009$; $t(29) = 1.631$, $p = .341$, respectively).

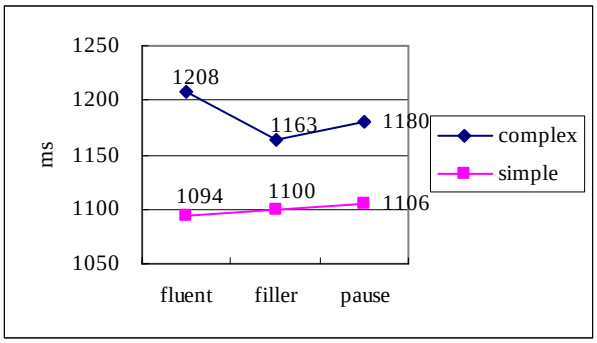


Figure 3: Chinese participants' mean response times from the onset of the first word describing a shape. 'Complex' means complex phrases and 'simple' means simple phrases.

3.5. Discussion

Although it took Chinese speakers 237ms longer on average to respond to correct figures than Japanese speakers, response times of Chinese speakers in the six conditions showed a similar pattern to those of Japanese speakers. Response times to complex phrases were shorter when a filled or silent pause preceded the phrase than when there was no preceding pause, while there was no significant difference in response times to simple phrases between any conditions. The results indicated that listeners tended to expect a longer and complex phrase to follow when they heard a filled pause, supporting the hypothesis.

The results agreed with Blau [1]'s results in that fillers, as well as silent pauses, helped non-native listeners' speech processing. Our results indicated that advanced language learners had acquired native like strategies in processing filled pauses.

4. Conclusion

The present research showed that filled pauses before long and complex phrases helped both native and non-native listeners' processing of speech by indicating complexity of the following phrase. Although our research was limited to the effects of one type of filler "eto", the results demonstrated that filled pauses at phrase boundaries were not harmful, at worst, and sometimes helpful to listeners. This is in accordance with the results of Fox Tree [4]'s study on fillers in English and Dutch. Our research provided information about the contexts in which filled pauses were helpful to listeners.

Our study with Chinese subjects suggested that advanced language learners were processing filled pauses in a way similar to native speakers. Namely, filled pauses at phrase boundaries seemed useful for non-native listeners as well as native listeners to predict complexity of the following phrase.

5. Acknowledgment

We thank Prof. Max Coltheart and Dr. Sallyanne Palethorp at Macquarie University for their kind advice in planning the experiment. This study was partly supported by JST/CREST the Expressive Speech Processing project.

6. References

[1] Blau, Eileen, Kay. 1991. More on comprehensible input: The effect of pauses and hesitation markers on listening comprehension, from ERIC database. Paper presented at the Annual Meeting of the Puerto Rico Teachers of English to Speakers of Other Languages (San Juan, PR, November 15, 1991).

[2] Buck, Gary, 2001, *Assessing Listening*, Cambridge: Cambridge University Press.

[3] Fox Tree, Jean Eleonore, 1995, The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language* 34, pp. 709-738.

[4] Fox Tree, Jean Eleonore, 2001. Listeners' uses of *um* and *uh* in speech comprehension. *Memory and Cognition*, 29 (2), pp. 320-326.

[5] Fukao, Yuriko, Sumiko Mizuta & Kazuo Ohtsubo. 1991. Development of teaching material for advanced learners of Japanese to improve their listening skills for lecture comprehension. Paper presented at the autumn meeting of the Society for Teaching Japanese as a Foreign Language (in Japanese).

[6] Murakami, Jinich. & Shigeki Sagayama. 1991. A discussion of acoustic and linguistic problems in spontaneous speech recognition, *Technical report of IEICE*, SP91-100, NLC91-57: pp.71-78 (in Japanese).

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

[8] The National Institute for Japanese Language. *The Corpus of Spontaneous Japanese* homepage. http://www2.kokken.go.jp/%7Ecsj/public/6_1.html

[9] Voss, B. 1979. Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, 22(2): pp. 129-44.

[10] Watanabe, Michiko. 2001. An analysis of usage of fillers in Japanese Lecture-style speech, *Proc. the Spontaneous Speech Science and Technology Workshop*, Tokyo. pp. 69-76 (in Japanese).

[11] Watanabe, Michiko. 2003. The constituent complexity and types of fillers in Japanese. *The Proc. of the 15th ICPhS*, pp. 2473-2476, Barcelona, Spain.

[12] Watanabe, Michiko, Yasuharu Den, Keikichi Hirose & Nobuaki Minematsu. 2004. Types of clause boundaries and the frequencies of filled pauses. *Proc. the 18th General Meeting of the Phonetic Society of Japan.* pp. 65-70. (in Japanese).

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 173-178.

Gesture marking of disfluencies in spontaneous speech

Yelena Yasinnik, Stefanie Shattuck-Hufnagel* & Nanette Veilleux***

* Massachusetts Institute of Technology, Cambridge, MA USA
** Simmons College, Boston, MA, USA

Abstract

Speakers effectively use both visual and acoustic cues to convey information in speech. While earlier research has concentrated on the association of visual cues (provided by gestures) with fluent prosodic structure, this study looks at the relationship between visual cues, prosodic markers and spoken disfluencies. Preliminary results suggested that speakers preferentially perform gestures in the eye region in spoken disfluencies, but a more careful frame-by-frame analysis capturing all gestures revealed that movements of the eye region (blinks, frowns, eyebrow raises and changes in direction of eyegaze) occur with high frequency in both fluent and non-fluent speech. The paper describes a method for frame-by-frame labelling of speech- accompanying gestures for a speech sample, whose output can then be combined with independently derived labels of the prosody. Initial analysis of 3 minute samples from two speakers reveals that one speaker produces eye movements in association with disfluencies and the other does not, and that this tendency does not result from alignment of brow gestures with pitch accents.

1. Introduction

This work focuses on the interaction among a) spoken disfluencies, b) visual cues provided by upper face movements, and c) prosodic prominence i.e. the perceptual salience provided by intonationally-cued phrase-level prominences called pitch accents. This relationship is of interest because it will tell us something about the underlying speech production planning process (i.e. about how speakers signal to listeners that a disfluency has occurred) and also because it may be of practical use in devising recognition systems that take advantage of such cues, as well as in developing on-screen personas that behave in a natural-looking ways. It is also possible that different kinds of disfluencies are associated with different kinds of gestural markers; such a finding would be particularly useful in automatic recognition and would have interesting implications for planning behavior. This study describes a labelling method for obtaining fine-grained information about the temporal relations among these three kinds of phenomena, and describes several preliminary results that illuminate the gesture-disfluency relationship.

2. Background

Initial studies of the relevant cues to prosodic prominence and phrasing in speech focused largely on acoustic cues such as f0, duration and amplitude. More recently, research on visual cues to prosodic events has contributed to our understanding of the complexity of human speech understanding and speech production. For example, visual cues from a speaker's gestures have been found to accompany the acoustic markers of prosodic events in speech, such as the phrase-level prominences called pitch accents. In particular, a number of investigators have reported an association between eyebrow raising and pitch accents (Keating et al. [5]). Yasinnik et al. [11] found that non-facial gestures produced by other non-speech articulators, such as the hand or head, are also associated with prosodic prominence; their study examined a certain type of gesture, which they define as a "hit" (i.e. a movement with an abrupt end point). Cavé et al. [2] reported an association between eyebrow movement and F0 rises, and showed that the two cues do not universally co-occur. Their results suggest "that eyebrow movements and fundamental frequency changes are not automatically linked (i.e., they are not the result of muscular synergy), but are more a consequence of linguistic and communicational choices".

One such communicational choice, possibly cued by both acoustic and visual means, is prosodic prominence. House et al. [4] found that the perception of prominence could be influenced by accompanying visual cues. In twelve acoustically identical stimuli, the presence/absence of head-nods and eyebrow movements of a computer-generated talking head influenced listeners' perceptions of which word was most salient. Although head-nods seemed to influence listeners more than eyebrow movements, eyebrow movements also significantly enhanced the perceived prominence of the closest pitch accented word. Similarly, Krahmer et al, [8] found that eyebrow movements did indeed contribute to determining which of two adjacent words was prominent, although in this case the presence/absence of an acoustic marker (pitch accent) was a more effective cue. Granström [3] found visual (again, facial) cues can indicate to human listeners whether an automatic system (as a travel agent talking head) has correctly understood the speaker. Taken together, these findings raise the possibility that spoken disfluencies may be marked gesturally, to help the listener identify the disruption.

3. Method

3.1. Database

Samples were excised from commercially available DVD recordings of 4 academic lecturers. Professional lectures were selected because they provide a ready source of large amounts of semi-spontaneous speech produced by speakers who are practiced communicators and recorded with high-quality audio. The main disadvantage of this corpus for our purposes is the fact that in some regions the speaker's body is either turned aside or not entirely visible because the producer chose to use a close-up frame or because other graphic material for the course is displayed instead. This disadvantage is easily overcome, however, because of the availability of many hours of recordings from each speaker. All 4 lecturers were male and appeared to be speakers of American English (M1am, M2am, M3am, M4am). The samples were approximately 40 minutes (M1), 9 minutes (M2), 11 minutes (M3), and 7.5 minutes (M4). Corresponding video and sound samples were transferred to a MacIntosh computer for gesture and speech labelling.

3.2. Preliminary Observation to Determine Relevant Gestures

The initial phase of the research involved watching the video files of all four speakers using iMovie and iDVD, locating speech errors and documenting any gestures – head, face, hands, and body - that occurred in the region containing the error, any editing remarks and the correction if one was made. No precise speech-gesture alignment was done at this point, but we noted the non-speech gestures that occurred in conjunction with each error.

The gestures we observed were sorted into categories according to their type, e.g. blink, shake, etc. Six main categories emerged:

- **Eyebrow raise**
- **Frown**
- **Blink**
- **Eyegaze transfer**
- **Shake (of the hand or head)**
- **Freeze –** an abrupt stop in the train of gesturing that was not preceded by relaxation (see below for discussion of gesture segment types, such as preparation, stroke, hold and relaxation).

Two aspects of this initial set of observations about disfluency-associated gestures are notable. First, many of the gestures involve the eye region: brow raises, frowns, blinks and eyegaze transfers. Two-thirds of the gestures that occurred in 59 disfluent regions (49/72) involved the eyes. Second, there was a striking absence of a type of gesture which occurs commonly in fluent speech, i.e. movements of the head or hands characterized by short sharp end points, that we have designated as 'hits' (Yasinnik et al. [11]). The movements of the head or hands that were observed in disfluent regions using this informal method were not the usual single-movement 'hits'; instead, we observed shakes (i.e. repeated short movements back and forth) or gestures that were temporarily frozen. It is possible that the gestures that were 'frozen' in disfluent regions would have been hits if completed, but in error regions they were not completed. Based on these preliminary observations, we formed two hypotheses about the relationship between disfluencies and speech-accompanying gestures in these four male speakers of American English: 1) disfluent regions in the speech are marked by gestures involving the eye region, and 2) they are not marked by hits. The remainder of this paper is focused on the first hypothesis.

These initial observations raised questions which require a more fine-grained method of analysis, one that permits investigation of the detailed alignments among disfluencies, speech-accompanying gestures and prosodic elements such as pitch accents. Two questions in particular arise about how to interpret the findings from the coarse-grained analysis. First, is the predominance of gestures that involve the eye region in disfluent regions unusual, or do eye movements occur freely and often throughout the speech samples? Second, is this predominance due to the fact that eye movements occur on pitch-accented syllables and thus error corrections (which are likely to include contrastively pitch accented syllables) are also likely to be associated with eye movements?

The labelling method we adopted for this more fine-grained analysis was developed in earlier studies of the alignment of gestures with aspects of prosodic structure such as prominences (pitch accents) and constituent boundaries (e.g. intonational phrase boundaries). In this method, the video file and audio file for the sampled lecture are separated, and labelled independently (by different labelers). The sound file is labelled for word alignments, prosody (e.g. intonational phrases and pitch accents) and disfluencies, and the video file for frame-by-frame gestural events, capturing the various subsections of a gesture, such as the preparation stage, the gesture itself (with optional hold), the relaxation stage and optional pause before the next gesture (McNeill, [9]). Aligning these two sets of independently transcribed labels provides a way of testing for co-occurrence of the gestures with the speech at the syllable-by-syllable level, and this analysis can be carried out with confidence that the results are not determined by any perceptual bias e.g. toward aligning gestural strokes with auditorily prominent syllables, etc.

3.3. Labelling

Because this aspect of the work is very time consuming, we selected shorter three-minute samples from two of the original speakers, M2am and M4am, for fine-grained labelling. These samples contained a substantial number of disfluencies, 16 for M2 and 16 for M4, and minimal video disruption from lecture-related graphics. For each sample, the sound file was transcribed and the words aligned with the wave form and spectrogram using xwaves; this representation formed the basis for prosody labelling of pitch accents and boundary related tones using the ToBI system, and the disfluent regions. The video file was labelled frame by frame for the onset of each subsection of each gesture. The labelling method for both files is described in some detail here, because it gives a flavor of the level of precision that is captured by these labels, and provides a clear picture of some of the challenges that can arise with labelling speech and gestural phenomena.

Gesture labelling. Gestures were labelled in the video file, without listening to the sound, using Anvil 4.5.2 [7]. This software allows the creation of multiple time- and video-aligned tiers for labelling gestures performed by different articulators: hands, head, eyes, and eyebrows. The tiers were displayed with tick marks for every second of the video and smaller tick marks equivalent to one video frame - $1/30^{th}$ part of a second. Every frame corresponded to one image in the video.

Within each tier, the labeller marked a beginning and an end of a region and designated it for a certain gesture with a label and an optional comment. The criteria for marking the onset and offset of a gesture were adopted from criteria for onsets and offsets of hand gestures described in Yasinnik et al. [11]:

- Onsets were marked at the frame where shape of the articulator began to change (e.g. narrowing of eyes during a blink), or the articulator's position began to change (e.g. head turn during a head shake), or the articulator's location began to change (e.g. hand moving from one place to another) which often was accompanied by blurring of the image in the video frame. The blurring provided a useful clue to the location of the onset video frame.
- Offsets were marked at the frame where the articulator stopped moving (this generally corresponded to a clearer image than in the surrounding frames), or at the frame just before the next change in articulator shape or position, which usually signaled the relaxation stage, but sometimes was simply an onset of the next gesture.

The labels in all tiers were saved together as an Anvil annotation file, which provided a text summary of time-stamps for onsets and offsets of gestures, as well as the labels for all marked gestures.

Reliability. Two labellers separately labelled the hand gestures in a 30-second segment of one of the video files. For each gesture, several phases were identified: preparation, stroke (sometimes followed by hold) and relaxation (for related discussion see McNeill [9] and Kendon [6]).

Out of 85 and 84 labels provided by Labeller 1 and Labeller 2 respectively, 71 markings (84%) agreed on a label and a time-stamp within one frame and 75 markings agreed within three frames. Inter-labeller disagreements arose largely from two sources: a) Labeller 2, who was more familiar with this speaker's gestures from earlier labelling experience, understood that during a pause in gesturing, this speaker often tapped his hands together while in neutral position, and labelled this region as a pause, while the Labeller 1 labelled each tap as a separate gesture event, and b) during one diagonal upward hand movement (accompanied by shaking), Labeller 2 broke the gesture into two shakes in different spatial locations, separated by a preparation stage. The limited nature of these disagreements suggests that more specific definitions of gestural onsets and offsets will increase this already high level of agreement.

Prosody labelling. The prosody of the spoken utterances was independently labelled from the sound files, by an experienced labeler using the ToBI system (http://www.ling.ohio-state.edu/~tobi/ame_tobi) from sound files displayed in Praat (see Boersma [1]). Aspects of the labels used here include pitch accent locations (i.e. the words and syllables marked with intonational phrase-level prominence) and intonational phrase boundaries. Experienced ToBI labellers transcribed the words (including the symbol PAU for each perceptually-noticeable silence) in the Praat TextTier file "words", using the waveform and spectogram to align the words with the sound. They also transcribed the pitch accents and boundary-related tones of the utterances in corresponding "tones" and "breaks" TextTier files, while listening to the sound files and viewing time-aligned f0 tracks, which had been created with the Praat pitch-tracking algorithm. The "words", "breaks", and "tones" TextTier files provide a time-stamp for the beginning and ending of each transcribed word, part of a word – if the word is cut off, and of each pause, as well as a time-stamp and label for each marked tonal element and intonation phrase boundary.

Disfluency labelling. In the miscellaneous TextTier, disfluent regions were assigned one of the four categories summarized below; for a related disfluency classification scheme see Shriberg [10].

- **Filled pause** – contains filler words *um* or *uh*, (e.g. *about that uh equation*);
- **Bobble** - contains repetition of a word or part of a word, (e.g. *many k- complications, forms the- the basis*); this category might be called a stutter, but this term has wide use in the literature for a type of speech pathology, and in addition, the two renditions of the target were sometimes separated by a filled pause, e.g. *mol-uh-molecule*, which is not the intuitive sense of what the term 'stutter' means;
- **Substitution** - contains substitution of a linguistic element often followed by a correction, i.e. an incorrect word, syllable, or sound, often drawn from nearby context (e.g. *har- artists had, ungovernmentally sh- sanctioned*)
- **Other disfluencies** such as a change of plan after a phrase has started, (e.g. *They, in fact, have all- …, they actually had to join the group*) or one whose type is ambiguous (e.g. *the b- tennis ball*, where the word fragment *b-* can reflect either a change of plans, i.e. a plan to say *the ball* replaced by a plan to say *the tennis ball*, or a linguistic substitution like *the bennis tall*, that was interrupted and corrected by the speaker before it was fully pronounced).

The 'miscellaneous' TextTier file provides a time-stamp for the ending of the most perceptually disfluent word, or part of a word, within each error. Disfluencies could be divided into the standard subsections of reparandum (region requiring repair), editing and repair, although defining the extent of the repair in cases of a change in plans is not straightforward.

3.4. Alignment

This set of labels provides a complete record of the gestures made by the speaker during the sample, and also reveals the precise alignment of gestures, prosodic prominences and disfluencies. Preliminary analysis of these data for the two 3-minute samples focused on two questions: (1) is the preponderance of eye-region gestures in disfluent regions a characteristic of disfluencies, or does it arise because eye-region gestures occur with high frequency throughout both fluent and disfluent speech? (2) do eye-region gestures of a particular kind, eyebrow hits, align differently with the prosodic elements of spoken utterances than other kinds of speech-accompanying gestures, and (3) do eyebrow hits occur preferentially in association with disfluencies?

4. Results

Results from the broad marking of gestural events in disfluent regions for the four speakers were described above. We noted the predominance of eye-region gestures (eyebrow raise, frown, blink, or eyegaze transfer), and absence of hand or head 'hit' gestures except for occasional shaking; any ongoing head or hand movements that might have been hits froze during the error interval. In addition, only 10 of the 59 errors (17%) were not marked by any gesture, and most of these non-gesture-marked errors were Filled Pauses (7 / 10). We note that these errors are by definition marked in the acoustic signal (in the form of *um* or *uh*), which is consistent with the hypothesis that eye-region gestures during disfluent regions are less likely in circumstances where the speech signal does not provide unambiguous cues to the occurrence of a disfluency, or does not provide them early enough in the disfluent event to prevent initial confusion on the part of the listener.

(1) How prevalent are eye gestures for these speakers? Analysis of the fine-grained gesture labels for three minutes of speech for each of two speakers (M2am and M4am) confirmed the general prevalence of eye movements in disfluent regions obtained from informal observation: 70+% of errors co-occurred with an eye gesture of some kind. 70% (7/10) for M2am, and 73% (22/27) for M4am. However, these results must be compared to the number of eye gestures accompanying equivalent non-disfluent speech. Since there is some difficulty in determining precisely what an appropriate comparison interval in fluent speech would be, we used a simple 3-word interval as an approximation since errors generally involve about three words. This provides a rough impression of how widespread eye gestures are throughout the samples. In the M2am sample, there were 313 eye gestures distributed over about 820 fluent words or an average of 1.14 eye gestures for every three word interval. Similarly, there were 188 eye gestures distributed over about 500 fluent words in the M4am sample or an average of 1.13 eye gestures for every three word interval. This indicates that eye gestures are frequent, commonly occurring in both fluent and disfluent speech.

This finding suggests that eye gestures as a class may be too broad a category to gainfully map to errors. Results from the coarser labels suggested that one type of eye movement, brow raising, is particularly common in disfluent regions, and Keating [5] has reported that brow movements can be associated with pitch accents. As a preliminary to testing the hypothesis that brow movements occur preferentially in conjunction with disfluencies vs. with pitch accents, we examined the distribution of gestural 'hits' by three different articulators: hands, head and eyebrows.

(2) Do gestural 'hits' of eyebrows, head and hands align with pitch-accented syllables? If some eye movements occur disproportionately often in disfluencies, it may be because they tend to occur on pitch-accented syllables, and errors (or at least corrections) are likely to contain contrastively accented syllables. If this is the case, then we expect to find that eyebrow hits occur more reliably on accented syllables than do hits produced by other articulators such as the head or hands. To test this hypothesis, we examined the 30 millisecond time segment of each hit frame in the sound waveform of the lecture. Many aspects of the alignments are of interest, including the hit frame's location in relation to the corresponding word, pitch accents, location of pitch accents in the syllabic structure of the word, and the phrasal structure. Here we focus on whether the 30-ms frame of each gestural hit overlapped with the duration of a pitch-accented syllable.

This alignment can take one of four forms: (1) *Early PAcc* if the hit frame occurs in the syllable preceding the speech prominence; (2) *Yes* if the frame occurs in the same syllable as the prominence; (3) *Foot* if the hit frame occurs in the syllable following the prominence and if that syllable is weak; and (4) No if the hit frame location has no relation to any syllable marked with a pitch accent prominence. An additional comment of *Maybe* was added to each label when the corresponding pitch accent label was marked as uncertain (i.e. *? In the ToBI system), indicating that the labeler was unsure whether there was a pitch accent. NB: although the foot is often defined as a strong syllable plus up to two weak following syllables within a word, e.g. *WOman*, we adopted a unit more like the Abercrombian foot, for which the weak syllable can be part of the following word (e.g. *send him* in *send him out*, *bake a* in *bake a pie*, or even *wrote a-* in *wrote about*).

The timing of hits and pitch accents were strongly correlated for this sampled speaker. The very high Pearson's Index, 0.98, established that gesture hits and pitch accents correlate positively with greater than 99% confidence ($p < 0.01$). Overall, 53% of hits occurred in pitch accented syllables, 24% occurred in the second syllable of the foot, and 13% occurred in other non-prominent parts of the speech signal. This data is consistent with other previously sampled speakers, but a closer examination of the alignment results revealed some interesting differences between gestural articulators.

While the correlation for hand hits with pitch accents was high (59% aligned with the accented syllable and an additional 19% with the second syllable of the foot, **Figure 2**) and head hits also showed a high correspondence (44% on the accented syllable and 42% on the following weak syllable, **Figure 1**), the eyebrow hits produced very different results: hits aligned equally with the accented syllable, the following weak syllable and with no accent-related syllable (**Figure 3**). This difference suggests that the relatively tight relationship between pitch accents and gestural hits observed for head and hand gestures is not observed for eyebrow gestures, at least for this speaker's sample. This supports the hypothesis that when brow gestures occur in disfluent regions, it is not just because they tend to be associated with pitch accents.

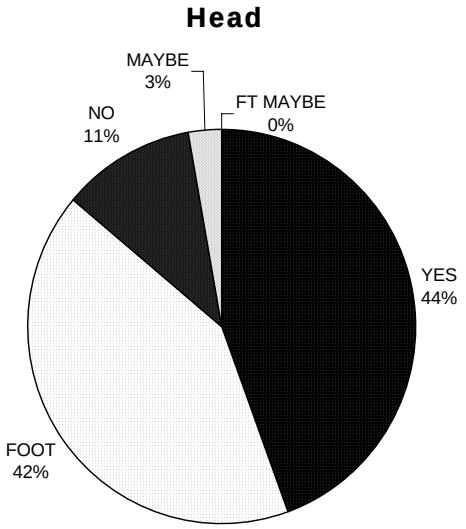


Figure 1: Percentage of head hit frames aligned with Pitch Accents (Yes), with the Foot (an unstressed syllable after the pitch accent), with a *? Pitch accent (Maybe) and with No pitch accent.

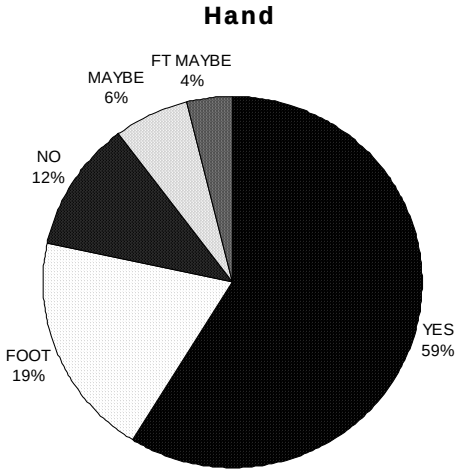


Figure 2: Percentage of hand hit frames aligned with Pitch Accents (Yes), with the Foot (an unstressed syllable after the pitch accent), with a *? Pitch accent (Maybe) and with No pitch accent.

We note also that the percentage of head hits aligned with the second syllable of the foot is large compared the percentage for hand hits. It is possible that this reflects the greater inertia of the relatively massive head, which may take longer to move into position.

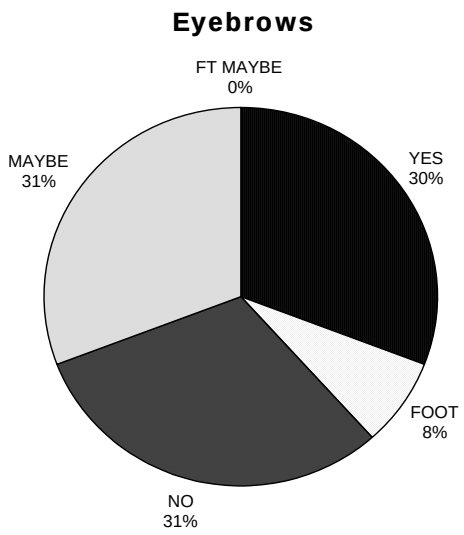


Figure 3: Percentage of eyebrow hit frames aligned with a Pitch Accent (Yes), with the Foot (an unstressed syllable after the pitch accent), with a *? pitch accent (Maybe) and with No pitch accent.

(3) Do eyebrow hits occur preferentially with disfluencies? Because eyebrow hits are distributed differently from hand and head hits in general in these two speech samples, we also analyzed their alignment with disfluent regions. To do this we determined the location of the 30 millisecond frame of each eyebrow hit in the speech waveform, and examined the nearby speech to see if a disfluency occurred. This analysis was confined to large, salient eyebrow hits; some eyebrow movements were rather small and difficult to distinguish from changes in light glinting off the spectacles of the speakers, and it was important to be sure we were looking at the distribution of eyebrow hits that would be salient to the watching listener. Results were strikingly different for the two speakers. Speaker M2am showed a tendency for eyebrow hits to occur in disfluent regions: of his 17 salient eyebrow hits, 11 occurred in conjunction with a disfluency and 6 did not. Speaker M4am, however, showed a very different pattern: of his 18 salient brow gestures, only 2 occurred near a disfluency and 16 did not. Since the number of disfluencies and the number of salient brow hits were similar for the two speakers (16 and 16 errors, 17 and 18 brow hits), these observations suggest that these two speakers are using eyebrow movements in different ways.

5. Discussion

The results described above confirm findings from earlier studies showing that speech-accompanying gestures are not randomly distributed in the speech signal, but in many cases are systematically aligned with speech events. In particular, the alignment of head and hand hits with pitch accents for speaker M2am is consistent with earlier work. However, this speaker does not align his brow hits with pitch accented syllables as reliably as his head and hand hits; instead, he tends to align them with disfluent regions. Thus, results for this speaker support the hypothesis that at least some types of eye-related gestures occur preferentially with disfluencies. However, the finding that speaker M4am does not align his brow hits with disfluent regions highlights the importance of surveying multiple speakers to determine which aspect of the gesture-disfluency relationship are general across speakers and which are idiosyncratic. It is possible that gesturing patterns are particularly variable across individuals, more so for example than some aspects of prosodic and syntactic usage.

A top priority for future work concerns the question of whether different types of disfluencies are marked by different kinds of gestures. The hypothesis that disfluencies that do not contain immediately obvious cues to their disfluent nature, such as substitutions, are preferentially marked by eye-region movements (because listeners will be focused more on the eye region of the talker) was not confirmed, since eye gestures occur freely throughout the speech samples we examined. However, it is still possible that errors that might initially mislead the listener into thinking they are part of fluent speech, such as word substitutions, are marked by certain types of eye gestures, while errors that are acoustically distinct (such as the repeated words/sounds of a bobble or the filler items of a filled pause) may be less consistently eye-gesture marked.

Another line of investigation will be to determine the extent to which all four speakers align their brow movements with pitch accents; initial observation suggests that Speaker M4am, whose brow hits were not associated with disfluencies, tended to align them with pitch accented syllables instead. If this initial impression is confirmed, it will reinforce the necessity of studying differences as well as similarities in the gestural marking of disfluencies by individual speakers. In addition, some of M2am and M4am brow movements were not discrete hits but raises that remained up for an extended interval. These gestures may be related to discourse structure.

Finally, we plan to pursue further the informal observation that there is a paucity of single hits in disfluent regions although these gestures are common in the fluent speech of these speakers.

6. Acknowledgements

Work supported by NIH/NIDCD grant RO1 DC00075, the Keith North Fund at the Speech Group at MIT/RLE, and MIT's Undergraduate Research Opportunities Program. Technical assistance and support from Helen Hanson and Michael Kipp are gratefully acknowledged, along with the time and effort invested in the project by Margaret Renick, Jessie Wang and Alicia Patterson.

7. References

[1] Boersma, P. 2001. PRAAT: A System for Doing Phonetics by Computer. *Glot International*, vol. 5, pp. 341-345.

[2] Cavé, Christian, Isabelle Guaïtella, Roxane Bertrand, Serge Santi, Françoise Harlay, Robert Espesser. About the Relationship Between Eyebrow Movements and F0 Variations. 1996. *Proc. ICSLP'96*, 3-6 October 1996, Philadelphia, PA, vol. 4, pp. 2175-2179.

[3] Granström B, House D. & M. G. Swerts. 2002. Multimodal Feedback Cues in Human-Machine Interactions. In B. Bel & I. Marlien (eds.), *Proc. Speech Prosody 2002 Conference*. Aix-en-Provence, France, pp. 347-350.

[4] House, David, Jonas Beskow & Bjorn Gransrom. 2001. Interaction of Visual Cues for Prominence. *Working Papers 49*, Dept. of Linguistics, Lund University, Sweden, pp. 62-65.

[5] Keating, P., M. Baroni, S. Mattys, R. Scarborough, A. Alwan, E. Auer, & L. Bernstein. 2003. Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *Proc. 15th International Congress of Phonetic Sciences*, 3-9 August 2003, Barcelona, Spain, pp. 2071-2074.

[6] Kendon, A. 1980. Gesticulation and Speech: Two Aspects of the Process of Utterance. In Mary Ritchie Key (Ed.), *The Relationship of Verbal and Nonverbal Communication*, The Hague: Mouton. pp. 207-227.

[7] Kipp, Michael. 2001. Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, pp. 1367-1370.

[8] Krahmer, Emiel, Zsofia Ruttkay, Marc Swerts & Wieger Wesselink. 2002. Pitch, Eyebrows and the Perception of Focus. *Proc. Speech Prosody 2002*, 11-13 April 2002, Aix en Province, France.

[9] McNeill, D. 1992. *Hand and Mind: What Gestures Reveal About Thought*. Univ. Chi.: Chicago.

[10] Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. *Proc. ICSLP'96*, 3-6 October 1996, Philadelphia, PA, USA, vol. 3.

[11] Yasinnik, Yelena, Margaret Renwick & Stefanie Shattuck-Hufnagel. 2004. The Timing of Speech-Accompanying Gestures with Respect to Prosody. *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, 11-13 June 2004, Cambridge, MA.

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 179–182.

A preliminary study of Mandarin filled pauses

Yuan Zhao & Dan Jurafsky

Stanford University, California, U.S.A.

Abstract

The paper reports preliminary results on Mandarin filled pauses (FPs), based on a large speech corpus of Mandarin telephone conversation. We find that Mandarin intensively uses both demonstratives (*zhege* 'this', *nage* 'that') and *uh/ mm* as FPs. Demonstratives are more frequent FPs and are more likely to be surrounded by other types of disfluency phenomena than *uh/mm*, as well as occurring more often in nominal environments. We also find durational differences: FP demonstratives are longer than non-FP demonstratives, and *mm* is longer than *uh*. The study also revealed dialectal influence on the use of FPs. Our results agree with earlier work which shows that a language may divide conversational labor among different FPs. Our work also extends this research in suggesting that different languages may assign conversational functions to FPs in different ways.

1. Introduction

Filled Pauses (FPs) are prevalent in Mandarin spontaneous speech and pose a major problem in Mandarin speech recognition. Although much work related to English and other languages has been done [1, 3, 5, 6], little empirical work on Mandarin FPs in spontaneous speech has been carried out. Previous work on Mandarin spontaneous speech, such as Tseng [7], mainly focuses on repairs and repetitions. The lack of research on FPs directly leads to the confusion of FP judgments and random use of characters in transcribing Mandarin speech. It is therefore necessary to identify the distinctive forms of FPs in Mandarin and to investigate their acoustic features and discourse functions.

The goal of the present work is to carry out a descriptive study of Mandarin FPs. More specifically, the study aims at identifying FPs and investigating their basic acoustic properties and distribution across syntactic units. In addition, it also examines the sociolinguistic variables that might influence speakers' use of FPs.

2. Method

The research was mainly based on the data drawn from LDC 98-HUB5 Mandarin corpus of telephone conversations, in which FPs such as uh and mm are hand-labeled by LDC. There are 37 conversations in the corpus, which have complete speaker information. Only the callers' information can be identified, which is shown in Table 1.

Table 1: Speaker information of 98-HUB5

Gender	Female	17
	Male	20
Dialect	Northern	21
	Southern	16
Age	Range (yr)	20-39
	Average (yr)	28.6
Education	Range (yr)	10-22
	Average (yr)	17.6

In addition, in order to see how FPs vary across corpora, another two LDC telephone conversation corpora, 2003-HUB5 and 96-CallHome, were also used for identifying distinctive forms of FPs in Mandarin. Demonstrative FPs were hand-labeled by the first author.

3. Results and discussion

3.1. How many FPs are there in Mandarin?

Four FPs are systematically used in three corpora, including *uh*, *mm*, and the demonstratives *zhege* (literally "this") and *nage* (literally "that"), as shown in Table 2.

Table 2: Occurrence of FPs in 98 HUB-5 per 1000 words

Demonstrative	*nage*	4.51
	zhege	2.17
	Total	**6.68**
Reduced Vowel	*uh*	2.55
	mm	1.46
	Total	**4.01**

In general, the total frequency of FPs (10.69 per 1000 words) seems to be smaller than in English. In the CallHome English corpus, for example, *um* occurs 7.15 times per 1000 words, and *uh* occurs 7.10 times per 1000 words.

Like Japanese and Spanish, besides reduced vowel FPs, Mandarin intensively employs demonstratives as FPs [4, 9]. In fact, demonstratives form the largest FP category in Mandarin. We also investigated other potential FPs, but found only these four (*nage/zhege/uh/mm*) seemed to act as true FPs. For example, we examined 200 instances of discourse particles transcribed as "oh"; we found that in each case particles with rounded vowels do not act as FPs, but as backchannels suggesting the addressee's attentiveness to the ongoing conversation.

3.2. Acoustic features of FPs

The basic acoustic features of the FPs, such as duration and pitch movement were measured from a sample of 100 FPs produced by male speakers, which are summarized in Table 3.

It was found that the duration of *mm* is over 1.5 times longer than that of *uh* and the difference is highly significant ($t=3.306$, $p<.002$, $MD=.2053$). In English, *uh* and *um* contrast mainly in the delays they initiate [1]. However, in the 100 samples of Mandarin *uh/mm* FP, less than 5 cases have pauses around them, which might indicate that Mandarin speakers tend to prolong the vowel or the nasal to fill an entire pause. This suggests that Mandarin FPs divide their labor according to the length of pause they need to fill: *mm* tends to fill a longer pause and *uh* tends to fill a shorter pause.

Table 3: Acoustic properties of FPs and comparison between FP and non-FP demonstrative

		uh	***mm***	**FP *nage***	**Word *nage***	**FP *zhege***	**Word *zhege***
Duration	Range (sec)	0.068-0.944	0.144-1.376	0.18-0.547	0.14-0.252	0.194-0.691	0.13-0.263
	Average(sec)	0.277	0.483	0.383	0.195	0.382	0.177
Pitch	Range (Hz)	97.6-224.1	90.2-198.6	98.8-187.0	101.2-269.0	101.0-319.8	79.7-261.2
	Average (Hz)	135.5	133.9	139.4	162.9	158.1	154.1

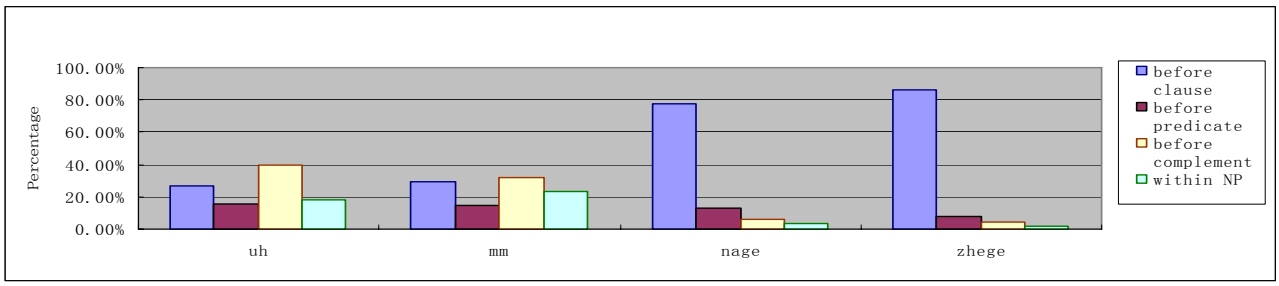


Figure 1: Distribution of filled pauses.

We also compared demonstrative FPs with their non-FP uses. It was found that the duration of both *zhege* and *nage* is significantly longer than their non-FP uses respectively (t=5.531, p<.001, MD=.2052) (t=6.278, p<.004, MD=.1882). In this case, FP demonstratives are around two times longer than their non-FP counterparts. No significant difference between the duration of *zhege* and *nage* was found. In addition, the statistics shows that the pitch of FP *nage* is significantly higher than that of non-FP *nage* (t=1.356, p<.024, MD=23.52).

3.3. Location of FPs

The second study examined the locations where FPs occur. FPs were found to occur mainly before syntactic constituents, such as before clauses, as shown in (a); before predicate VPs as shown in (b) and before complements as shown in (c). Another location where FPs frequently occur is within an NP constituent, such as between a modifier and a noun head, as shown in (d):

a. **nage** [$_S$ wo mingtian you shijian].
that I tomorrow have time
'mm, I have time tomorrow.'

b. wo **nage** [$_{VP}$ du-guo liang bian].
I that read-Asp two CL
"I, uh, read it twice'

c.[$_{VP}$ huainian **nage** [$_{NP}$ daxue shenghuo]].
miss that college life
"(I) miss, uh, the life in college.'

d. [$_{NP}$ gongzuo de **nage** [$_N$ mafan]]
works NOM that trouble
'The work's, uh, trouble'

The distribution of FPs occurring in different syntactic contexts was examined, as summarized in Table 4 and Figure 1.

Table 4: Distribution of FPs among syntactic units.

	Before Clause	Before Predicate	Before Complement		Within NP
			S	NP	
nage	26.4%	15.1%	6%	**34.0%**	**18.4%**
zhege	29.4%	14.7%	3.9%	**28.4%**	**23.5%**
uh	**77.7%**	12.6%	4.7%	1.5%	3.5%
mm	**86.3%**	7.6%	3.5%	0.9%	1.7%

Figure 1 suggests that most *uh* and *mm* are used clause initially. In addition, the occurrences of *uh* and *mm* decrease as the syntactic units become smaller and simpler. Grammatical weight theory predicts that speaker would encounter more planning problems before larger syntactic constituents [8]. Our result on FPs of *uh/mm* is in accordance with what grammatical weight theory predicts. In this case, the larger and more complex a constituent is, the more likely an *uh/mm* FP is to occur before it.

Contrasting with *uh/mm* FPs, demonstrative FPs are most likely to occur before complements, especially NP complements. In addition, unlike *uh* and *mm*, which are rarely used between a modifier and a noun head, demonstratives also occur frequently in this nominal environment. As shown in Table 4, 34% of *nage* and 28.4% of *zhege* occur before NP complement and 18.4% of *nage* and 23.5% of *zhege* occur before noun head. Demonstratives *nage* and *zhege* occur in nominal environments 52.4% and 51.9% of the time, respectively. Demonstrative FPs therefore contrast greatly with *uh/mm* in occurring in nominal-related disfluency environments. This suggests that demonstrative FPs may play a role in nominal search in Mandarin. We can formalize this suggestion as the "Nominal-Search Hypothesis": speakers are more likely to choose demonstratives as FPs when they encounter nominal-search problems in Mandarin

3.4 Disfluencies around FPs

Research such as [1, 2, 8] has shown that some FPs signal larger planning problems than others. One way to investigate which FPs represent bigger planning problems is to examine neighboring disfluencies occurring around the FPs. The distribution of disfluencies before and after the four types of FPs is summarized in Table 5.

Table 5: Disfluencies around filled pauses.

	before	after	repetition	Total
zhege	1.96%	23.5%	8.8%	**34.3%**
nage	0.99%	20.3%	8.0%	**29.3%**
uh	7%	7.8%	0	14.8%
mm	5.8%	4.9%	0	10.7%

The result in Table 5 suggests that demonstrative FPs are more likely to be accompanied by other types of disfluency phenomenon than *uh/mm*. For example, demonstrative FPs are often repeated several times and followed by other disfluencies. In comparison, *uh* and *mm* are not that likely to be accompanied by other type of disfluencies. The fact that demonstrative FPs are more likely than *uh* and *mm* to be followed by other disfluencies, and far more likely to be repeated, may suggest that demonstratives are more indicative of more serious planning problems. On the other hand, speakers may simply be using other strategies to indicate severe planning problems with *uh* and *mm*. For example, *uh* and *mm* could be prolonged to indicate severity, as suggested in [1]. This issue clearly calls for further investigation.

3.5 Social variables and filled pauses

Finally, the study examined speakers' demographic factors and their influence on the use of FPs. These factors include gender, age, education and place of growing up.

Figure 2 shows the distribution of FPs among individual southern and northern speakers. A t-test shows that place of growing up has a significant effect on the number of FPs being used. Speakers growing up in the south of China use significantly more FPs than those grow up in the north ($t=3.431$, $p<.002$). The main distinction lies in the total amount of *uh*. Southerner use significantly more *uh* as FPs than northerners ($t=2.888$, $p<0.002$). We found no significant effect in the use of nasal *mm* between northern and southern speakers.

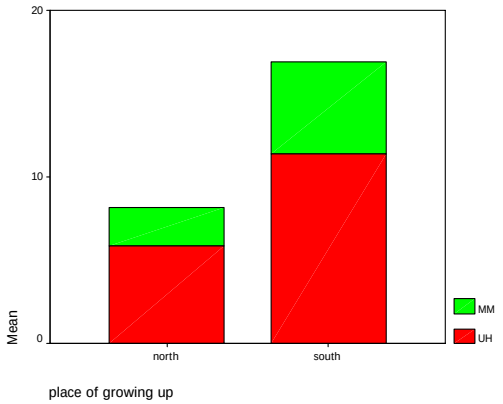


Figure 2: *uh/mm* used among southern speakers and northern speakers.

Sex, age and education do not have a significant effect on the use of FPs. Previous work on English FPs revealed that FP rates are related to demographic factors such as sex. Men produced significantly higher rates of FPs than women [6]. Our study does not find similar results in Mandarin; sex does not have a significant effect on the use of FPs in Mandarin ($t=-.315$, $p<.755$). Although sex alone does not have significant effect on the use of FPs, the interaction between sex and place of growing up turns out to be significant ($F=3.823$, $p<.019$). The mean number of FPs used by a southern female speaker is the highest and that of a northern male speaker is the lowest.

In addition, we found no significant effect of education level on the use of FP, although we did find a trend for speakers with more education to use fewer FPs than those with less education.

4. Conclusion

The paper presents preliminary results of a study on Mandarin FPs. Through comparing three corpora, four major FPs were identified. It revealed that apart from *uh* and *mm*, Mandarin speakers intensively use demonstratives as one major type of FPs. Demonstratives and *uh/mm* FPs differ in their distribution among syntactic contexts. Namely, demonstratives are more frequently used in nominal-searching environments, while *uh* and *mm* are more likely to be used clause-initially. The two types of FPs also contrast greatly in the amount of surrounding disfluencies. Demonstratives were found to be more likely accompanied by repetitions, false starts etc., while *uh/mm* were only rarely accompanied by other disfluencies. In addition, southern speakers tend to make use of more FPs in conversation than northern speakers, which may suggest dialectal influence on the use of FPs in Mandarin.

This paper provides only a preliminary picture of Mandarin FPs. Much more work needs to be done in the future.

5. Acknowledgements

This research was partially supported by Stanford University through a Deans' program summer research fellowship. Thanks to Huihsin Tseng and Rebecca Starr on help scripting.

6. References

[1] Clark, Herbert & Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speech. *Cognition, 84,* pp. 73-111.

[2] Clark, Herbert & Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology 37*, pp. 201-242.

[3] Eklund, Robert & Elizabeth Shriberg. 1998. Crosslinguistic disfluency modelling: A comparative analysis of Swedish and American English human-human and human-machine dialogues. *Proceedings of ICSLP'98*, 30 November-5 December 1998, Sydney, vol.6, pp. 2631-2634.

[4] Hayashi, M. and Yoon, K. to appear. "A cross-linguistic exploration of demonstratives in interaction: With particular reference to the context of word-formulation trouble." *Studies in Language.*

[5] Quimbo, Felix C M., Tatsuya Kawahara & Shuji Doshita. 1998. Prosodic analysis of fillers and self-repair in Japanese speech. *Proceedings of ICSLP'98*, 30 November-5 December 1998, Sydney, pp.3313-3316.

[6] Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. *Proceedings of International Conference on Spoken Language Processing*, Philadelphia, Addendum, pp. 11-14.

[7] Tseng, Shu-Chuan. 2003. Repairs and repetitions in spontaneous Mandarin. *Preceedings of DiSS'03,*5-8 September 2003, Goteborg, Sweden, pp.73-76.

[8] Wasaw, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change, 9,* pp.81-105.

[9] Wilkins, David P. 1993. Interjections as deictics. *Journal of Pragmatics, 18,* pp. 119-158.

AUTHOR INDEX

Web Page Design by Jean Véronis
http://www.up.univ-mrs.fr/delic/Diss05
Jean.Veronis@up.univ-mrs.fr

Proceedings compiled by Estelle Campione
Estelle.Campione@up.univ-aix.fr

Cover Design by Groupe Alteor
http://www.alteor.com

This Conference is sponsored by:

CONSEIL GENERAL

Région

Provence-Alpes-Côte d'Azur

UNIVERSITÉ DE PROVENCE
Aix-Marseille I

ATALA
Association pour le Traitement Automatique des Langues

International Speech Communication Association

Aix en Provence
L'OFFICE DE TOURISME

Aix en Provence
LA VILLE

GROUPE alteor
communication

Université de Provence

29, Av. Robert Schuman
13100 AIX-EN-PROVENCE
Tél. : 04 42 95 31 37