Proceedings of

DiSS 2013

The 6th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
21–23 August 2013

TMH-QPSR

Volume 54(1)

Edited by
Robert Eklund

< This page intentionally left blank >

Proceedings of

DiSS 2013

The 6th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
21–23 August 2013

TMH-QPSR
Volume 54(1)

Edited by
Robert Eklund

Conference website: http://www.diss2013.org
Proceedings also available at: http://roberteklund.info/conferences/diss2013

Cover design by Robert Eklund
Front cover photo by Jens Edlund and Joakim Gustafson
Back cover photos by Robert Eklund

Proceedings of DiSS 2013, The 6th Workshop of Disfluency in Spontaneous Speech
held at the Royal Institute of Technology (KTH), Stockholm, Sweden, 21–23 August 2013
TMH-QPSR volume 54(1)
Editor: Robert Eklund
Department of Speech, Music and Hearing
Royal Institute of Technology (KTH)
Lindstedtsvägen 24
SE-100 44 Stockholm, Sweden

ISBN 978-91-981276-0-7
eISBN 978-91-981276-1-4
ISSN 1104-5787
ISRN KTH/CSC/TMH--13/01-SE
TRITA TMH 2013:1

Printed by Universitetsservice US-AB, Stockholm, Sweden, 2013

Preface

Following the successes of the previously organized Disfluency in Spontaneous Speech workshops held in Berkeley (1999), Edinburgh (2001), Göteborg (2003), Aix-en-Provence (2005) and Tokyo (2010), the organizers are proud to present DiSS 2013, held at the Royal Institute of Technology (KTH), Stockholm, Sweden, in August 2013.

As was the case with the previous workshops, a wide variety of papers addressing disfluency from an equally varied array of disciplines is included.

The organizers would like to extend their thanks to everyone who helped organize this event, including the Scientific Committee members and, of course, all the contributors.

Stockholm, August 2013

Jens Edlund
Robert Eklund
Joakim Gustafson
Sofia Strömbergssson

Committees

Program and organization

Jens Edlund
KTH Royal Institute of Technology, Sweden

Joakim Gustafson
KTH Royal Institute of Technology, Sweden

Sofia Strömbergsson
KTH Royal Institute of Technology, Sweden

Robert Eklund
Linköping University, Sweden

Scientific committee

Martine Adda-Decker
LIMSI CNRS, France

Jens Allwood
Göteborg University, Sweden

Elisabeth Ahlsén
Göteborg University, Sweden

Dale Barr
University of Glasgow, Scotland

Herbert H. Clark
Stanford University, USA

Martin Corley
Edinburgh University, Scotland

Yasuharu Den
Chiba University, Japan

Jens Edlund
KTH Royal Institute of Technology, Sweden

Robert Eklund
Linköping University, Sweden

Jean E. Fox Tree
University of California, Santa Cruz, USA

Joakim Gustafson
KTH Royal Institute of Technology, Sweden

Robert Hartsuiker
University of Ghent, Belgium

Peter Heeman
Oregon Health and Science University, USA

Rebecca Hincks
KTH Royal Institute of Technology, Sweden

David House
KTH Royal Institute of Technology, Sweden

Robin Lickley
Queen Margaret University, Scotland

Sieb Nooteboom
Utrecht University, The Netherlands

Elizabeth Shriberg
Microsoft, USA

Sofia Strömbergsson
KTH Royal Institute of Technology, Sweden

Marc Swerts
Tilburg University, The Netherlands

Shu-Chuan Tseng
Academia Sinica, Taiwan

Åsa Wengelin
Göteborg University, Sweden

Table of contents

Plenary talks

Presented papers

Plenary Talk

Conceptions of disfluencies

Herbert H. Clark

Stanford University, USA

For most of us, a disfluency is any feature of an utterance that deviates from the ideal delivery of that utterance. It is a scientific ragbag of a category that includes pauses, prolonged words, self-repairs, repeats, *uh* and *um*, restarts, slips of the tongue, stutters, and various other phenomena. What holds the category together is that we take its members to be evidence of the "intrinsic troubles" people have in speaking. Still, there have been two approaches to the study of these troubles. One has focused on failures in communication. The idea is that people in conversation monitor for such failures and, when they find them, repair them. The second tradition has focused, instead, on success and failure together. The idea here is that not only do people repair things that have gone wrong, but they display and acknowledge things that have gone right. I will argue that these two views lead to distinct accounts of what disfluencies are and how people deal with them.

Plenary Talk

Disfluency in speech: the listener's perspective

Martin Corley

University of Edinburgh, Scotland

Disfluencies in spontaneous speech have the potential to affect listeners in at least two ways: They may impact upon the moment-to-moment process of determining the speaker's intended meaning, and they may influence the listener's lasting impression of what was said. In this talk, I outline what we know about each of these types of effect, focusing on three sources of evidence. Evidence from a series of eyetracking and ERP studies shows that listeners update their predictions of what is likely to be uttered following hesitation disfluencies; and that they pay more attention to words which are uttered immediately post-disfluency. Participants in the ERP studies are more likely to later recognise having heard words which occur immediately post-disfluency, suggesting a link between short-term processing differences (in prediction and attention) and their longer-term consequences (in memory). Evidence from change detection studies confirms that words encountered post-disfluency are better encoded, and allows us to examine the range of signals that might be considered as "disfluent". Evidence from feeling-of-knowing studies shows that listeners have reduced confidence in the veracity of statements that are disfluent, showing that disfluency affects the listener's metalinguistic as well as linguistic representations.

Disfluency and discursive markers: when prosody and syntax plan discourse

Julie Beliao & Anne Lacheret

MoDyCo, Department of Linguistics, Paris-Ouest University, France

Abstract

Hesitations, interruptions within phrases or within words are common in spontaneous speech. Those phenomena are widely known to be observable from a prosodic point of view through disfluencies. From a syntactic point of view, many studies already established that discursive markers such as *hm*, *oh*, *I mean*, etc. are representative of spontaneous speech. In this study, we demonstrate through a joint corpus-based analysis that these prosodical and syntactical features are correlated, without however being equivalent. More precisely, the lack of either disfluencies or discursive markers is consistently shown to be representative of a planned discourse.

Index Terms: disfluency, discursive marker, genres

1. Introduction

Corpora of spontaneous speech are characterized by a massive presence of disfluencies phenomena, whose function is yet to be better described (hesitation, self-repair, formulation, memory search, malaise, style effects, trademarks facilitating the right to speak, etc.). Despite the apparent irregularity of these phenomena ant their diversity, constants emerge through observation of corpora.

The disfluencies represent a very heterogeneous class. In this study, we distinguish between those that correspond to numerous acoustic markers (extensions, crushing registry, etc.), and those that are expressed only morphosyntactically (mainly rehearsals and unfinished segments without associated acoustic markers). The former are the prosodic *disfluencies* which we denote "hes" and the latter are denoted *discursive markers*, abbreviated "DM".

From a prosodic point of view, Corley shows in [1] that repetitions, also called disfluencies or hesitations, are not always accidental and are used by the speaker as a communicative tool. In fact, they are even advocated as a particular way to plan discourse. Apart from varying with the planning of discourse, disfluencies were also shown to depend on the content of speech [2]. For example, longer or complicated statements are much more likely to contain disfluencies [3,4]. Similarly, a speaker unfamiliar with the subject he is talking about will tend to produce disfluencies [5,6].

The importance of disfluencies for the syntactic analysis of spontaneous speech was highlighted by Dell [7], who proposed a syntactical model for natural language that explicitly accounts for disfluency as a fundamental phenomenon in spontaneous speech.

In this study, we build on those considerations and first focus on the identification of a possible correlation between prosodic disfluencies and syntactic discursive markers. To this purpose, we proceed to a large-scale statistical analysis of the Rhapsodie corpus of spontaneous speech, which is independently annotated both in syntax and prosody. Then, we show that the rate of disfluencies and DM is indeed representative of both the type (planned or spontaneous) and context (private or public) of discourse.

This paper is organized as follows. In section 2, we describe the context of this study and its material, namely the Rhapsodie Corpus. In section 3, we present the statistical analysis we performed along with its results. Finally, some conclusions are drawn in section 4.

2. Corpus

The aim of this section is to present the corpus and the various transcription studies that have been conducted during 3 years within the Rhapsodie project [8]. This project aims at providing and testing on large-scale of constructions a new prosodical and syntactical transcription and annotation system. A total of 57 samples were gathered – such as existing samples from corpora [9,10,11] but also new ones – with a wide topological coverage.

The speech database covers various discourse genres and speaking styles and comprises about 3 hours of continuous speech, monologues and dialogues, private vs. public, face-to-face vs. broadcasting, more or less interactive, descriptive vs. argumentative vs. procedural samples.

In this study we present an analysis of disfluencies and discursive markers which are often considered as similar phenomena. Each of those two phenomena has been annotated independently from the other using either only formal syntactical criteria or only acoustical/prosodical criteria.

2.1. Syntactic analysis

2.1.1. Syntactic annotation

Combining the syntactic model proposed by the Aix School [12] and the pragmatic model developed within the Lablita experience [13], two levels of syntactic cohesion have been annotated within Rhapsodie: microsyntax (i.e., syntactic cohesion guaranteed by government) and macrosyntax (i.e., syntactic cohesion guaranteed by illocutionary dependency). Microsyntax describes the kind of syntactic relations which are usually encoded through dependency trees or phrase structure trees.

Macrosyntax, which is of interest to us here, can be understood as an intermediate level between syntax and discourse. This level describes the whole set of relations holding between all the sequences that make up one and only one illocutionary act. The annotation of macrosyntax is essential to account for a number of cohesion phenomena typical of spoken discourse and in particular of French spoken discourse, because of the high frequency of paratactic phenomena that characterizes this language.

A complete annotation and a functional tagging of pile structures are also available [12,13,14]. More generally, a complete categorical and functional tagging for every word was achieved in the Rhapsodie corpus, including discourse markers, which are integrated into the syntactic representation at the macrosyntactic level.

2.1.2. Discursive markers annotation protocol

Discursive markers (DM), also called "associated illocutionary units" [14], are considered as macro-syntactic units. They often come as series of impaired verbal constructions, such as *huh*, *well*, *uh*, *so*, *hem*, etc. These units, which we denote with quotation marks "" are equipped with an illocutionary operator but they do not convey information content that is added to the content of knowledge shared by the interlocutors. They are a special case of illucutionary units described below.

An Illocutionary Unit (IU) is any portion of discourse encoding a unique illocutionary act: assertions, questions, and commands [15,16]. An IU expresses a speech act that can be made explicit by introducing an implicit performative act such as "I say", "I ask", "I order". A test for detecting the Illocutionary Units that make up a discourse consists of the introduction of such performative segments (see below). A segmentation in IUs is particularly important for the study of the connection of prosody and syntax, which is the goal of Rhapsodie, because these units are prosodically marked. For example, consider the following statements:

(1) c'est fils de la Sarce "je crois"
it's son of Sarce "I think"
[Rhap-M011, Corpus Avanzi [17]]

(2) ils sont deux Argentins "hein"
they are two Argentinians "eh"
[Rhap-D2003, Broadcast Corpus]

(3) je lui ai dit "ben" "tu vois" je vends des livres
I told him "well" "you know" I sale books
[Rhap-D2001, Corpus Mertens [18]]

Segments "je crois" (*"I think"*), "tu vois" (*"you know"*), "hein" (*"eh"*) and "ben" (*"well"*) are equipped with an illocutionary operator that permits to recognize them as assertions (*I think, you know*) or exclamations (e*h*). They share internal characteristics with the nuclei, such as being segments organized around a finite verb or reduced to an interjection. However, these segments do not convey information content that is added to the content of knowledge shared by the interlocutors: they can be deleted, for example, without any state of knowledge being changed. They do not have a descriptive function, but rather a function of modal change (as in the first example) or interactional regulation (as in the following two examples). From this point of view, we can say that they lack illocutionary strength in the true sense of the term. Indeed, they are not proper illocutionary acts addressed at an interlocutor, who can not deny or question the content of these segments.

2.2. Prosodic analysis

2.2.1. Prosodic annotation

For prosody, Rhapsodie annotators built on the theoretical hypothesis formulated by the Dutch-IPO school [19] stating that, out of the total information characterizing the acoustic domain, only some perceptual cues selected by the listener are relevant for linguistic communication [20,21]. On this basis, they decided to manually annotate only three perceptual phenomena characterizing real productions: prominences, the cornerstone of the sentence-prosodic segmentation [22,23], pauses and disfluencies [24].

Starting from this annotation, a prosodic structure was automatically generated, organized around rhythmical and melodic components. For each constituent of the structure, prototypical-stylized melodic contours were computed. First, perceptual syllabic salience in speech contexts was annotated using a gradual labeling distinguishing between strong, weak or zero prominences. Second, all prosodic segments were annotated, including disfluencies and different kinds of pauses (silent pauses vs "uh" or syllabic hesitation in the proximity of pauses). Third, prototypical-stylized melodic contours were generated for units of different sizes and domains. The availability in the Rhapsodie Treebank of the contour of a large number of prosodic and syntactic units allow the user to build various lexicons of intonation shape in an extremely flexible way according to his/her research goals [25].

In more general terms, it should be highlighted that these annotation choices have allowed us to identify the primitives of prosodic structure independently from any reference to syntax or pragmatics, and to provide all the elements needed for a complete prosodic analysis of linguistic units.

2.2.2. Disfluencies protocol annotation

An element that breaks the flow phrase in the speech chain, like stumbling voice, is called a disfluency or hesitation (hes) in this study. It can take different types and often exhibits an excessive syllabic elongation. Moreover, it can often consist in a repetition of morphemes or in interruptions of words or sentences. Every syllable perceived as disfluent is marked with a specific tag, annotated H. The annotation of disfluencies is carried out manually under PRAAT [26]. Each encoding step performed sequentially on batches of data. The number of listening for annotating one batch is limited to 3.

The most classic disfluencies are:

- Interruptions: *it's not far* | ***you*** | ***I'm going***
- Repeated segments: it is not far **you you go**
- Hesitations: "Uh"
- Excessive syllabic lengthening (not corresponding to a boundary structure): ***we:::::::ll***

These phenomena are not mutually exclusive but can be combined. If disfluence is simple (one disfluent syllable), the disfluent syllable is annotated "H" in the dedicated tier. If it is combined, i.e. concerns several successive intervals, then all the corresponding syllables are tagged as disfluent.

3. Methodology and statistical analysis

In this study, we focus on the correlation of the number of DM with the number of disfluencies over all samples from the Rhapsodie corpus. To this purpose, we propose a statistical analysis focusing on two separate aspects.

A first analysis focuses on the average number of disfluencies and DM per minute (hes/min and DM/min). Studying these scatter-plot showing one vs the other across all samples, we performed a correlation study, to be developed in section 3.1.

Then, disfluencies and DM may simply be synchronized. If that would be the case, it would mean that the corresponding annotations were strongly influenced by each other. In section 3.2, we show through a synchronization analysis that it was actually not the case.

Finally, the average number of disfluencies and DM of the samples are used to identify those corresponding to planned or spontaneous speech.

3.1. Correlation of disfluencies and DM

A first remarkable fact is that there are approximately half DM compared to disfluencies in the whole Rhapsodie corpus, as can be seen in Table 1.

Table 1: *Basic numbers from the corpus*

Units	count
Syllables	45192
Words	33 182
Disfluencies	3460
DM	1818

However, having half as many DM than disfluencies does not mean they are uncorrelated. Indeed, when plotting hes/min vs DM/min for all samples clearly show a strong correlation between them.

To confirm this fact, we performed a linear regression whose results are displayed in figure 1. With an infinitesimal *p*-value of $p=1.8\times10^{-15}$, the null-hypothesis corresponding to decorrelation can safely be rejected. A correlation of 0.8 was observed between hes and DM.

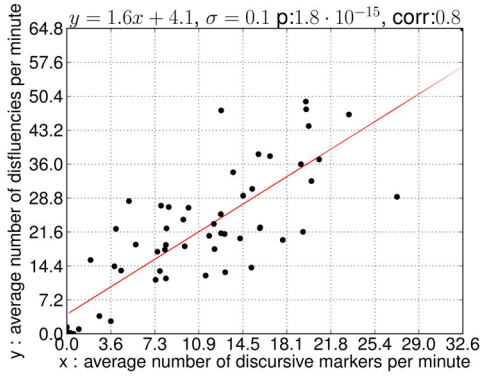


Figure 1: Average number *of disfluencies and discursive markers per minute for each sample (black dots)*

3.2. Independenceof DM and disfluency annotations

This strong correlation between DM/min and hes/min means that these phenomenon are related. Still, this may actually be due to a deterministic interaction such as a bias in the annotation.

In order to check for this hypothesis, we performed a synchronization analysis and display in Figure 2 the ratio of both hes and DM that have the same temporal support.

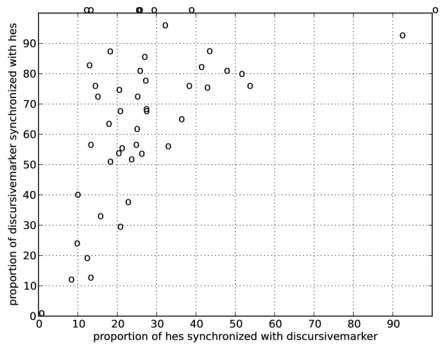


Figure 2 : *Display of synchronization between disfluencies and discursive markers.*

The low proportion (between 10% and 50%) of disfluencies which are synchronized with DM confirms that these units have been annotated independently from syntactic considerations, unlike DM.

On the contrary, there are a bit more (between 40% and 80%) of DM which are synchronized with disfluencies, proving that annotation of DM probably involved disfluencies in some way. However, this tendency looks rather weak considering the scatter-plot displayed in Figure 2, and may also simply be due to the higher number of disfluencies than DM, causing an increase in the probability of their synchronization.

3.3. When average number variation means planning

In this section we show how the joint presence of DM and disfluencies within a sample of the corpus is related to the corresponding type of discourse. To this purpose, we display for each sample its position along the regression line found in Figure 1. That way, a small value indicates a lack of both disfluencies and DM, whereas a high value indicates their joint important within the sample. The resulting barplot can be found in Figure 3.

Considering this figure, we can see that samples corresponding to planned speech contain very little disfluencies and DMs whereas the repartition of semi-spontaneous and spontaneous samples is more random. The same results pertain to public or private speech, the former including less DM and disfluencies than the latter. Hence, public speech may appear more planned, which seems natural. It should be emphasized that the two first samples are the corpus' shortest (less than 60 words) and don't contain any DM and disfluencies.

4. Conclusions

It is widely acknowledged in prosodic studies that there is a strong relation between disfluencies and speech planning. Similarly, many syntactical studies have established that discursive markers are typical of spontaneous speech.

However, there was no study we were aware of that performed a joint intonosyntactical analysis on a large-scale corpus to study how prosody and syntax agree on the same data. In this paper, we demonstrated that the density of disfluencies in a sample is indeed strongly correlated to the density of discursive markers, even if the two notions are showed not to be equivalent. It is hence our belief that a joint analysis of prosody and syntax may lead to a better understanding of spontaneous speech.

Since hesitations are the most frequent type of speech disfluence in many languages, it is possible that the majority of the synchronized cases falls within the class of hesitations. It would be and interesting perspective to examine more finely the degree of synchronization for sub-classes of disfluencies and discursive markers.

5. Acknowledgements

We warmly thank Antoine Liutkus for his English review and for his help in figures display as well as the Rhapsodie Consortium.

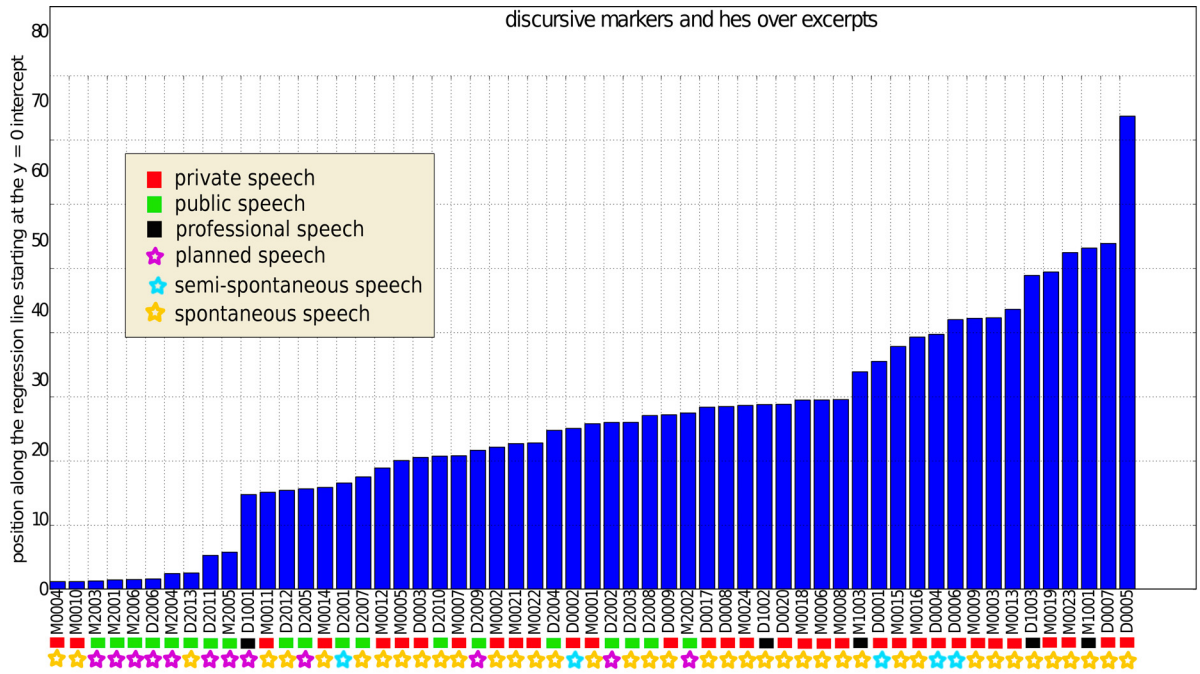


Figure 3: *Rate of disfluencies and discursive markers (Y-axis) for the 57 samples of the Rhapsodie corpus (X-axis), along with a description of the content of each sample in terms of public/private/professional/planned/spontaneous speech.*

6. References

[1] M. Corley, O.W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um", *Language and Linguistics Compass* 4, pp. 589–602, 2008.

[2] S. Schachter, N. Christenfeld, B. Ravina, F. Bilous, "Speech disfluency and the structure of knowledge", *Journal of Personality and Social Psychology* 60, pp. 362–267, 1991.

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

[4] E. Shriberg, "Disfluencies in Switchboard", *Proceedings International Conference on Spoken Language Processing*, Addendum, pp.11–14, Philadelphia, 1996.

[5] H. Bortfeld, S.D. Leon, J.E. Bloom, M.F. Schober, S.E. Brennan, "Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender Language and Speech", *Language and Speech* 43(3), pp. 229–147, 2000.

[6] S. Merlo, L. Mansur, "Descriptive Discourse: Topic Familiarity and Disfluencies". *Journal of Communication Disorders* 37, pp. 489–503, 2004.

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

[8] A. Lacheret, S. Kahane, P. Pietrandrea, "Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French", *Studies in Corpus Linguistics*, Amsterdam, Benjamins, 2013.

[9] S. Branca-Rosoff, S. Fleury, F. Lefeuvre, M. Pires. *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*, 2009.

[10] B. Laks, J. Durand, C. Lyche, Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. In *Phonologie, variation et accents du français Hermès*, pp. 19–6, 2009.

[11] M. Avanzi, A.-C. Simon, J.-P. Goldman, A. Auchlin, "Un corpus de français parlé annoté pour l'étude des proéminences, c-prom. Actes des 23èmes journées d'étude sur la parole", Mons, Belgique, 2010.

[12] C. Blanche-Benveniste, "Un modèle d'analyse syntaxique 'en grilles' pour les productions orales", *Anuario de Psicologia* Liliane Tolchinsky (coord.), vol. 47, Barcelona, pp. 11–28, 1990.

[13] E. Cresti, *Corpus di italiano parlato*, Florence: Accademia della Crusca, 2000.

[14] E. Bonvino, F. Masini, P. Pietrandrea, "List Constructions: a semantic network", *Troisième Conférence Internationale de l'AFLiCo*, Nanterre, 2009.

[15] S. Kahane, P. Pietrandrea, "Les parenthétiques comme Unités illocutoires associées. Une perspective macrosyntaxique", *Linx* 61, pp. 49–70, 2012.

[16] C. Benveniste, "Problèmes de linguistique générale", volume 1 of Coll. TEL, Gallimard, 1966.

[17] J.R. Searle, "The classification of illocutionary acts", *Language in Society* 5, pp. 1–24, 1996.

[18] M. Avanzi, *L'interface prosodie/syntaxe en français. Dislocations, incises et asyndètes*, Bruxelles, Peter Lang, 2012.

[19] P. Mertens, *L'intonation du français : de la description linguistique à la reconnaissance automatique*, Thèse de Doctorat, Université de Louvain, 1987.

[20] J. Hart, R. Collier, A. Cohen, *A perceptual study of intonation, an experimental phonetic approach to speech melody*, Cambridge University Press, 2006.

[21] A. Lacheret, F. Beaugendre. *La prosodie du français*, Paris, CNRS, pp. 62, 1999.

[22] C.W. Wightman, "ToBI or not ToBI?", *Proceedings of Speech Prosody*, Aix-en-Provence, France, pp. 25–29, 2002.

[23] J. Buhmann, J. Caspers, V.J. van Heuven, H. Hoekstra, J.-P. Mertens, M. Swerts, "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus", *Proceedings of LREC 2002*, Las Palmas, pp. 779–785, 2002.

[24] F. Tamburini, C. Caini, "An automatic System for Detecting Prosodic Prominence in American English Continuous Speech", *International Journal of Speech Technology* 8, pp. 33–44, 2005.

[25] A. Lacheret, N. Obin, M. Avanzi, "Design and evaluation of shared prosodic annotation for spontanesous French speech: from expert knowledge to non-expert annotation", *48th Annual Meeting of the Association for Computational Linguistics*, 4th Linguistic Annotation Workshop, Uppsala, Sweden, 2010.

[26] V. Aubergé, G. Bailly, "Generation of intonation: a global approach", In *Proceedings of the European Conference on Speech Communication and Technology*, Madrid, pp. 2065–2068, 1995.

Pauses following fillers in L1 and L2 German Map Task dialogues

Malte Belz & Myriam Klapi

Humboldt-Universität zu Berlin, Germany

Abstract

Fillers and pauses in spoken language indicate hesitations. Filler type (*uh* vs. *um*) is believed to signal a minor or major following speech delay in L1. We examined whether advanced speakers of L2 German use pauses following filler type (*äh* vs. *ähm*) in the same way as native speakers do. Two Map Task corpora of L1 and L2 were contrasted with respect to speaker role, filler type and the exact time interval of fillers and pauses. Speaker role influenced the disfluency patterns in L1 and L2 in the same way. Filler type had no impact on the length of the following pause, but the time interval patterns differed significantly. Longer filler intervals are followed by longer pauses in L2 and by shorter pauses in L1. These results suggest that filler type in German is not used to indicate the length of the following delay. Advanced learners seem to have adopted this pattern of use, but cannot overcome their hesitations as fast as native speakers, probably due to their less automatised speech production.

Index Terms: Fillers, Pauses, Spontaneous speech, L1, L2, Map Task, German, Disfluencies, Contrastive Analysis

1. Introduction

In this paper we examine whether German native (L1) and non-native speakers (L2) use the fillers *äh* and *ähm* and their following pauses in the same way.

Fluency in spontaneous speech is not constantly achieved. Disfluencies in spontaneous speech are frequent and have been analysed both for L1 ([1–6], inter alia) and L2 speakers [7, 8]. Among the most commonly studied disfluency categories are fillers and pauses. Fillers, like the English *uh* and *um*, or the German *äh* and *ähm* – sometimes also called filled pauses or hesitations [5, 9] – are well-described, although many studies propose different vantage points [10]. Together with pauses (also known as silent pauses or unfilled pauses [5]) fillers are central to the research of hesitation phenomena in L1 and L2 speech production [11–13]. Taken together, they constitute about 78% of the overall occurrence of disfluencies in spontaneous speech [5]. When examined separately, the use of fillers and pauses is often related to hesitation and repair phenomena. Consequently, a combined use raises the question of whether a more serious problem in speech production has been encountered. As learners have to deal with non-native speech processing, they may experience a working memory capacity overload [14]. Therefore, deviations in delay behaviour can be expected, and interesting implications can be drawn from this phenomenon when applying contrastive analysis.

Prior research has made evident that native speakers often use fillers and pauses in order to find time for processing decisions [15]. According to [7] and [14], limitations in L2 proficiency cause patterns of error and repair which are different from those of native speakers. Hence, one would expect that, because learners of a foreign language have to process a higher cognitive load, they will differ from native speakers in their filler and pause patterns. The question which arises at this point is whether this is true regarding advanced L2 speakers. Do they use fillers with following pauses as L1 speakers do? In order to evaluate hypotheses of this kind, it is essential to conduct contrastive research, thus enabling a comparison to native speakers. Differences in the use of the patterns described above may relate to less automatised speech production processes and monitoring in L2 [16–18].

Clark & Fox Tree [9] examined the English fillers *uh* and *um* in combination with following pauses. Their results suggest that filler types affect the length of their respective following pauses. *Uh*s preceding pauses signal a minor delay, whereas *um*s preceding pauses signal a major delay. Example 1 illustrates their use:

(1) ich sags dir 0.6 s ähm 1.7 s also du musst äh 0.4 s nach äh re/ rechts hoch

I'm telling you 0.6 s um 1.7 s well you have to uh 0.4 s go uh upwards to the right

(BeMaTaC_L1_2013-01/2012-01-19-A, 04:32)

The L1 speaker in the example above inserts *ähm* as well as a 1.7 s pause before giving directions. Before specifying the exact direction, the speaker inserts *äh* together with a shorter pause (0.4 s). From the example above, it may be assumed that *uh* and *um* in Map Task dialogues behave differently with respect to their following pause. In genres like interviews, however, the differences observed for post-filler pauses may not be perceived quite as clearly [19].

No quantitative study of combined fillers preceding pauses has yet been made in a German L1/L2 Map Task setting. The present approach attempts to bridge this gap by demonstrating a contrastive corpus analysis of two recently constructed corpora. Our hypotheses are the following:

1) It is often implicitly assumed that fillers in different languages show similar patterns in use, when their form seems to be identical (Engl. *uh/um* vs. Germ. *äh/ähm*). As a first step, it is crucial to examine whether the length of the following pause in German is influenced by the filler type (*äh* vs. *ähm*), as it is in English [9]. Our contrastive analysis suggests that the two filler types deviate in relation to the length of their following pauses. This prediction is relevant for both L1 and L2 speaker categories and should be observable in both groups.

2) According to the given experimental Map Task design, the speaker role (i.e. instructor vs. instructee) is expected to affect the length of pauses. Instructors take up the highest amount of speaking time (see section 2.1). We expect this effect to be observable for both L1 and L2.

3) If type of filler turns out to be the only influence on the following pause length, then the filler length is not anticipated to affect the length of the pause. Though no evidence for a similar use of filler categorisation preceding pauses has been found for German yet, we expect fillers to behave in the English way, as stated by [9]. As the proficiency of learners in our data exceeds intermediate levels, we expect them to adopt the native-like pattern.

2. Method

2.1. Corpora

The Berlin Map Task Corpus (BeMaTaC) [20, 21] and the Hamburg Map Task corpus (HAMATAC) [22] both use a Map Task design [23], where one speaker (= instructor) instructs another speaker (= instructee) to reproduce a route on a map with landmarks. This design is suitable for multilevel linguistic research, as it enables spontaneous dialogues elicited in a controlled context [21]. BeMaTaC has been inspired by HAMATAC and follows the same experimental design, enabling comparable and contrastive studies of native and non-native German.

BeMaTaC (version 2013-01) consists of 12 dialogues, 16 native German speakers and 11192 tokens in the relAnnis format. In order to conduct the present research we accessed BeMaTaC via ANNIS [24], an open-source browser-based search and visualisation tool for deeply annotated corpora.

HAMATAC (version 0.2 [2011-09-30]) consists of 24 dialogues (21433 words) by 24 advanced learners of German. For lack of a standardised L2 proficiency test, we rely on the meta-data, consisting of learners with an advanced proficiency level (20 out of 24). Participants' native languages covered a wide range (Romance, Slavic, Persian and Non-Indo-European languages). We extended the corpus with further annotation layers. The corpus was converted with the SaltNPepper converter [25] to the relAnnis format.

2.2. Data

All fillers preceding pauses were extracted from these two corpora (Table 1), every instance linked to its metadata role (instructor vs. instructee) and subject ID. The exact filler interval time as given in the transcriptions were extracted and calculated in L1 and L2, as well as for pauses in L1. Pauses annotated in HAMATAC were extracted from the vocal transcription tier as given in deciseconds. Zero-length pauses were not considered as relevant instances and therefore not taken into account.

Table 1: *Frequencies of fillers preceding pauses*

Fillers preceding pauses	BeMaTaC		HAMATAC	
	äh	ähm	äh	ähm
Actual numbers	34	42	108	142
In %	44.7	55.0	43.2	56.8
In % of overall words or tokens	0.67		1.16	
In % of overall fillers	27.74		28.09	

The L1 data were extracted via ANNIS using the token boundaries we obtained from a PRAAT transcription [26]. The L2 data could not be extracted quite as easily. Therefore, we exported the PRAAT voice transcription tier and calculated the exact filler interval times.

Describing a path through a Map Task is rather challenging since instructor and instructee cannot see each other. A high working memory load can be expected, especially if subjects have to conceptualise their message in a foreign language. As we expected higher hesitation levels, we did neither apply a cut-off of length for fillers nor for pauses.

2.3. Model

Since there are individual differences in disfluency length distribution, we applied a linear mixed-effects model to the data, which allows us to treat subject IDs as random effects while looking for significant patterns between fixed effects. We started with a full model, including as many fixed effects and their interactions as technically possible. Then we reduced the complexity of the model stepwise by comparing their AIC and performing log-likelihood tests. The remaining fixed effects which seem to predict pause length are language type (L1 vs. L2), role (instructor vs. instructee) and interaction of language with filler length (see Table 2).

3. Results

The findings are summarised in Table 1. L2 speakers use the described phenomenon nearly twice as much as L1 speakers (1.16% vs. 0.67%). We observe that in both groups approximately every fourth instance is followed by a pause (27.74% vs. 28.09%). We see that pause length differs with respect to the preceding filler type (Figure 1). Pauses following *äh* exhibit a large variance in L1 ($\bar{t} = 0.57$ *s*, $\hat{\sigma} = 0.41$ *s, median = 0.43 s*) and in L2 ($\bar{t} = 0.95$ *s*, $\hat{\sigma} = 1.33$ *s, median = 0.6 s*), whereas pauses following *ähm* have a more narrow variance, both in L1 ($\bar{t} = 0.63$ *s*, $\hat{\sigma} = 0.43$ *s, median = 0.59 s*) and L2 ($\bar{t} = 0.91$ *s*, $\hat{\sigma} = 0.73$ *s, median = 0.7 s*).

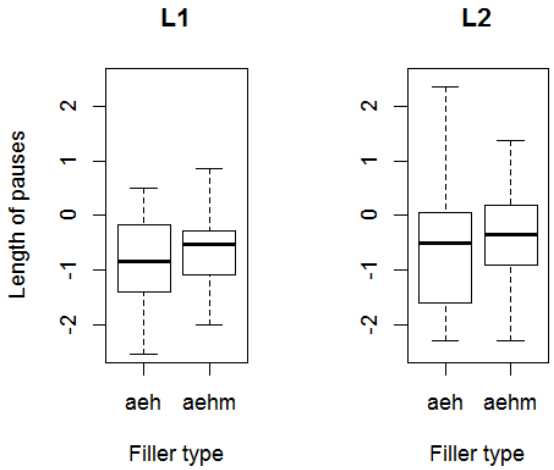


Figure 1: *Filler type and variance of pause length in L1 and L2 (logarithmic scale).*

Differences can be seen regarding the overall length variance of pauses depending on filler types. Nevertheless, no significant results were found regarding the interaction between filler type and pause length, either for L1 or for L2, as will be shown below.

As far as we can tell, filler type (*ähm* vs. *äh*) has no significant impact on the length of the following pause ($Df = 282$, $p < 0.96$), nor does interaction between language and filler type ($Df = 282$, $p < 0.28$). These effects were therefore excluded from the model, among others.

As expected, the instructor role exhibited a significant effect ($Df = 288$, $p < 0.043$) compared to the instructee role for both language types. The main effect of filler length does not appear to be interpretable in a way that makes sense to us due to its strong interaction with language type (L1 vs. L2). However, the interaction between L2 and filler length was significant ($Df = 288$, $p < 0.0016$). This indicates that filler

length had a different effect on the duration of the unfilled pause for L2 compared to L1. More specifically, L2 speakers tend to produce longer pauses following respectively longer filler intervals. Thus, the longer a filler stretches in time, the longer the following pause seems expected to be. This effect does not depend on filler type, as shown in Figure 2.

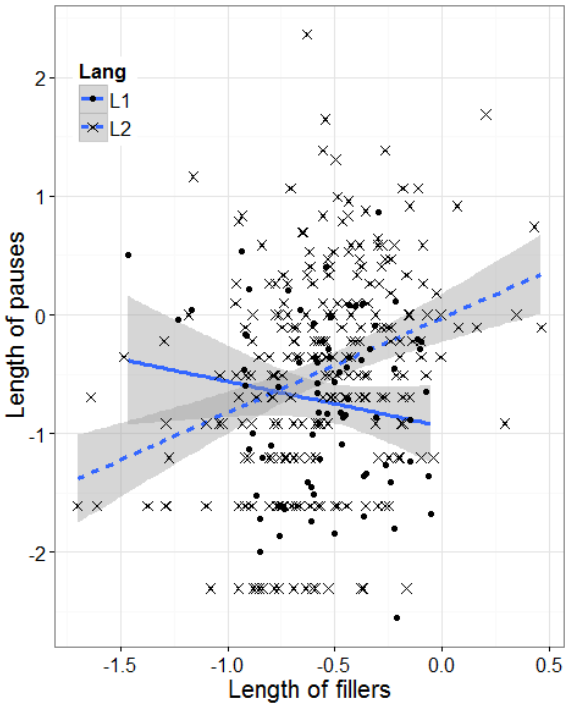


Figure 1: *Interaction of pauses preceding fillers in L1 and L2 (logarithmic scale).*

Random slopes were ruled out after a χ^2-test calculation, which proved not to be significant. To avoid spurious correlations of fixed effects, the logarithms of filled pauses were centered. Model coefficients are illustrated in Table 2.

Table 2: *Results of the linear model.*

	Value	Std. Error	DF	t	p
(Intercept)	–1.15	0.23	288	–5.09	0.000
Lang	0.82	0.22	33	3.77	0.000
Role	0.32	0.16	288	2.04	0.042
log(FPlength)	–0.24	0.26	288	–0.90	0.370
Lang:log(FPlength)	0.98	0.31	288	3.20	0.001

4. Discussion

First, advanced learners of German insert pauses after fillers with a similar frequency in Map Task dialogues when compared to native speakers (27.74% in L1, 28.09% in L2). These similarities between native speakers and learners are expected when stipulating an advanced L2 level. They herewith verify the average competence level of L2 speakers in HAMATAC.

Second, we expected pauses in L1 to differ systematically in duration when used after *äh* and *ähm* respectively. Nonetheless, the results failed to reach a significant level, as has been calculated above. This is confirmed for both groups. This implies that there is no difference between *äh* and *ähm* regarding their following delays, being inconsistent with the findings of [9] for English.

Figure 1 might suggest differences in pause length for both filler types. However, these narrow discrepancies proved not to be significant. This was found by applying the best linear mixed-effects model as described above. It excludes the interaction of filler type with language as a fixed effect. Therefore, no significant difference of how advanced learners use pauses following *äh* and *ähm* compared to native speakers was detected, suggesting that advanced learners of German do not behave differently from German native speakers regarding the combination of filler types with following pauses.

A significant effect is found for the dependence of length of pauses on the speaker role, hence verifying hypothesis 2. Filled pauses are significantly longer when speaking as instructor than when speaking as instructee. This finding is consistent with the results of Bortfeld et al. [1], who also showed an influence of the factor role on the number of disfluencies produced. This holds for both L1 and L2 and might reflect higher cognitive demands that instructors have to deal with.

As to the third hypothesis, our anticipation regarding the length of *äh* and *ähm* predicted no effect on the length of the following pause. This prediction was surprisingly falsified. The model exhibits that pause length is influenced significantly by the presence of the interaction between L2 speakers and filler length. This finding suggests that pauses may indeed be dependent on the time it takes a speaker to articulate a filler. The duration of *äh* and *ähm* may therefore be interpreted as a signal of an upcoming planning pause. Hence, the implication which becomes evident is that when L2 speakers of German take longer to utter a filler, they somehow signal that they need a longer planning phase and tend to insert a longer silent pause. Since the finding for native speakers of German is opposite regarding the dependent pause (the longer the filler, the shorter the pause), it is suggested that L2 speakers show a deviating pausing behaviour with respect to fillers. This finding might also suggest that speech production in Map Task descriptions is hard to process for learners, despite their high level of competence in German.

5. Summary and Conclusion

The implications of this pilot study are two-fold. The current results indicate that the role of participants (i.e. instructor vs. instructee) within the Map Task significantly influences their disfluency patterns, confirming the findings of Bortfeld et al. [1]. This holds both for L1 and L2, providing us with a more comparable and thus more reliable environment for contrastive research.

We did not find a correlation between filler type and length of the following pause between L1 and L2, what one would have expected for learner speech production, namely a non-nativelike use of fillers with pauses. Since there is no such evidence, our finding suggests either that learners have adopted the use of these patterns at an earlier stage, or that there is no difference in the distinctive filler types. We argue for the latter, thus implying that filler type seems not to affect German learners concerning the process of planning in speech production. Our results imply an observable difference in the use of delays when compared to English. Even though no direct comparison has been made in this study, it is possible that the use of delays combined with fillers follows a language-specific pattern.

Our findings suggest that L1 and L2 speakers have different pausing behaviours depending on the time spent for uttering a filler, regardless of filler type. These results show that German learners deviate significantly from German native speakers in using this specific disfluency pattern, which might be related to less automatised speech processing and monitoring in non-native speech production, as described by Levelt [17] and Declerck & Kormos [18]. With the objective of identifying differences in learner speech disfluencies and L2 acquisition, a more fine-grained stratification of proficiency control may result in the emergence of a new measure for automatisation in L2 speech production.

6. Acknowledgements

We are grateful to Anke Lüdeling, Felix Golcher and Amir Zeldes for their continuous support. We thank Robert Eklund for his encouragement in contributing to this workshop.

7. References

[1] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender", *Language and Speech*, vol. 44, no. 2, pp. 123–147, 2001.

[2] J. E. Arnold, M. Fagnano, and M. K. Tanenhaus, "Disfluencies Signal Theee, Um, New Information", *Journal of Psycholinguistic Research*, vol. 32, no. 1, pp. 25–36, 2003.

[3] E. E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Unpublished Dissertation, University of California, Berkeley, 1994.

[4] E. E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies", *JIPA*, vol. 31, no. 1, pp. 153–169, 2001.

[5] R. Eklund, "Disfluency in Swedish human–human and human–machine travel booking dialogues", Dissertation, Department of Computer and Information Science, Linköpings Universitet, Linköping, Sweden, 2004.

[6] F. Ferreira and K. G. D. Bailey, "Disfluencies and human language comprehension", *Trends in Cognitive Sciences*, vol. 8, no. 5, pp. 231–237, 2004.

[7] J. Kormos, "Monitoring and Self-Repair in L2", *Language Learning*, vol. 49, no. 2, pp. 303–342, 1999.

[8] C. L. Rieger, "Disfluencies and hesitation strategies in oral L2 tests". In: *Proceedings of DiSS'03*, Disfluency in Spontaneous Speech Workshop, pp. 41–44, 2003.

[9] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking", *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[10] M. Corley and O. W. Stewart, "Hesitation Disfluencies in Spontaneous Speech: The Meaning of um", *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[11] S. S. Reich, "Significance of pauses for speech perception", *Journal Psycholinguist. Research*, vol. 9, no. 4, pp. 379–389, 1980.

[12] R. Griffiths, "Pausological Research in an L2 Context: A Rationale, and Review of Selected Studies", *Applied Linguistics*, vol. 12, no. 4, pp. 345–364, 1991.

[13] P. Trofimovich and W. Baker, "Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech", *Stud. Sec. Lang. Acq.*, vol. 28, no. 1, pp. 1–30, 2006.

[14] L. Temple, "Second language learner speech production," *Studia Linguistica*, vol. 54, no. 2, pp. 288–297, 2000.

[15] H. H. Clark and T. Wasow, "Repeating Words in Spontaneous Speech", *Cognitive Psychology*, vol. 37, pp. 201–242, 1998.

[16] W. J. M. Levelt, "Monitoring and self-repair in speech", *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

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

[18] M. Declerck and J. Kormos, "The effect of dual task demands and proficiency on second language speech production", *Bilingualism*, vol. 15, no. 4, pp. 782–796, 2012.

[19] D. C. O'Connell and S. Kowal, "Uh and Um Revisited: Are They Interjections for Signaling Delay?", *J. Psycholinguist. Res.*, vol. 34, no. 6, pp. 555–576, 2005.

[20] L. Giesel, M. Klapi, D. Krüger, I. Nunberger, O. Rasskazova, and S. Sauer, "A deeply annotated multimodal map-task corpus of spoken learner and native German", *Proc. DGfS*, Potsdam, 2013.

[21] S. Sauer and A. Lüdeling, "BeMaTaC: A Flexible Multilayer Spoken Dialogue Corpus for Contrastive SLA Analyses". In: *ICAME* 34, 2013.

[22] T. Schmidt, H. Hedeland, T. Lehmberg, and K. Wörner, HAMATAC – The Hamburg MapTask Corpus. http://www.exmaralda.org/files/HAMATAC.pdf

[23] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC Map Task Corpus", *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[24] A. Zeldes, J. Ritz, A. Lüdeling, and C. Chiarcos, "ANNIS: A Search Tool for Multi-Layer Annotated Corpora", in *Proceedings of Corpus Linguistics*, 2009.

[25] F. Zipser and L. Romary, "A model oriented approach to the mapping of annotation formats using standards". In: Proceedings of the Workshop on Language Resource and Language Technology Standards, *LREC 2010*, Malta, 2010.

[26] P. Boersma, "Praat, a system for doing phonetics by computer", *Glot International*, vol. 5, no. 9, pp. 341–345, 2001.

8. URLs

BeMaTaC
http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematac

HAMATAC
http://vs.corpora.uni-hamburg.de/corpora/z2-hamatac/public/index.html

HESITA(tions) in Portuguese: a database

Sara Candeias[1], Dirce Celorico[1], Jorge Proença[1], Arlindo Veiga[1,2], Fernando Perdigão[1,2]

[1] Instituto de Telecomunicações, Coimbra, Portugal
[2] Electrical and Computer Engineering Departement, University of Coimbra, Portugal

Abstract

With this paper we present a European Portuguese database of hesitations in speech. Under the name of HESITA, this database contains annotations of hesitation events, such as filled pauses, vocalic extensions, truncated words, repetitions and substitutions. The hesitations were found over 30 daily news programs collected from podcasts of a Portuguese television channel. The database also includes speaking style classification as well as acoustical information and other speech events. Statistic analysis of the hesitation events in terms of their occurrence is presented. Insights into the process of human speech communication can be extracted from this database, which encloses relevant information about how Portuguese speakers hesitate. The HESITA database is freely available online to the research community.

Index Terms: hesitations, disfluency, prepared speech, spontaneous speech, annotation, hesitation corpus

1. Introduction

It is commonly agreed that hesitations (synonym here for disfluencies) characterize spontaneous speech and play a fundamental role in its structure, reflecting aspects of the language production and the management of inter-communication [1], [2] and [3]. Across several corpora, studies as [2], [5], [8] have shown that hesitation-like events occur frequently at high rates per word during the speech production. In the last decade, a growing number of works on language processing have focused on hesitation events underlining the importance of gathering knowledge on these type of events for successful speech technology development (see [2], [6], [11], [17] and [31], as examples). Regular features of those events have been accepted as an important parameter to take into account both in automatic speech recognition (for more robust language and acoustic models [10], [11], [12]) and in speech synthesis (to improve the naturalness of the speech [13]).

Although some theories and models have arisen in an attempt to explain the phenomenon and its benefits for communication purposes, hesitation phenomenon remains as a linguistic challenge. Hence, they appear to be regulated by language specific constraints and they perform a linguistic universal role in the speech structure, systematically and meaningfully [14–17].

Since hesitation events are crucial to facilitate natural language processing tasks, several studies have attempted to verify which properties may provide clues to their recognition. Phonetic and prosodic properties and contextual distributions are shown to give significant cues in [11], [15], [16], [23] and [18], respectively. Studies on different languages, such as English [19], [20], Swedish [5], Mandarin [21] and French [8], have attempted to distinguish linguistic properties between filled pauses and extension events, mainly in order to pursuit the linguistic reasons of why extensions cannot be eliminated at a pre-processing module. Others, e.g. [22], point out lexical and syntactic principles, which may link up repetitions with word cut-offs. To detect repetitions, acoustic features including duration [23] and some syntactic cues [24] have been frequently used.

For European Portuguese there are also various linguistic studies on hesitations that have attempted to provide significant knowledge on the topic and claiming the regular trend of it. Regarding filled pauses, works such as [25–27] can be mentioned as first works on the subject. In [7] and in [10], fundamental frequency and duration of filled pauses are presented as characteristics that contribute for on-line planning efforts either in spontaneous speech or in oral reading. Other works on the topic for European Portuguese can be found in [9] and in [6], [11]. Although the classification of filled pauses is not the main topic of these last two works, it shows that such hesitation events are responsible for the distinction between unplanned versus planned speech.

With this paper we intend to present a European Portuguese database of hesitations in speech. Under the name of HESITA, this database contains annotations of hesitation events, such as filled pauses, vocalic extensions, truncated words, repetitions and substitutions. Additionally, other acoustical characteristics such as environment condition, speaking styles and speaker were annotated as well. We believe that these multiple annotation layers provide a wide range of opportunities for studying the structure of the human speech communication process, under the domain of either speech technology development or linguistic descriptive works.

In section 2 we concisely describe the components of the HESITA database. Section 3 provides statistics about the distribution of the hesitation events, illustrating their phonetic forms and relation with speaking styles. Section 4 presents a brief discussion, mainly focusing on the HESITA application

2. The HESITA database

The HESITA database comprises manually annotated hesitation events in 30 daily news programs collected from podcasts of a European Portuguese television channel (about 27 hours of speech). The audio was downsampled from 44.1 kHz to 16 kHz sampling rate and the video information was discarded. It contains studio and out of studio recordings as well as some telephone sessions. Prepared (read) speaking style is dominant, since most of the speech encompasses utterances of anchors and professional speakers (14 hours). However we can frequently find spontaneous

speech segments in commentators, reporters, interviewers and interviewees (10 hours). Lombard speech appears as well, but with low representativeness (18 minutes, with only 12 events of hesitation).

Under the term of hesitation, the following categories were identified and annotated, closely following the notation presented in [2]:

- filled pauses (f)
- vocalic expressions (+)
- repetitions (r)
- substitutions (s)
- filler words (p)
- deletions (d) and
- insertions (i)

Only the speech segments were annotated in terms of hesitation events. Filled pause vocalizations were transcribed using the SAMPA phonetic alphabet for European Portuguese [28]. HESITA database (hereafter HESITA DB) also encompasses information regarding audio characteristics (background environments, such as studio, street, speech overlapping, noise and music) and acoustic events (non-speech events, such as music, jingles, laughter, coughing or clapping). Respiratory and other events, such as noise from cars or wind, were also taken in consideration in the annotation procedure. Speaking style and speaker information are included in the annotation labels as well.

All the annotations were performed by using the Transcriber software tool [29]. See an example in Fig.1, in which (SP_STU_E1_JM) exemplifies an annotation of speech with noise-free environment (SP), in a spontaneous speaking style (STU) with low level of spontaneity (E1) and from a male journalist (JM). (SP_STU_E3_M) exemplifies an annotation of speech with noise-free environment (STU), in a spontaneous speaking style with high level of spontaneity (E3) and from a male speaker (M). (SP_OVR_E3_M) represents the annotation of an audio segment speech with noise-free environment (SP) but with overlapping (OVR), in a spontaneous speaking style with high level of spontaneity (E3) and from a male speaker (M). We can also verify the annotation of some hesitation events, including repetitions (r), extensions within a word (w+) and filled pauses (f). Phonetic symbols attest extended vowel sounds or vocalic fillers. The presence of a respiratory event is annotated as (res).

In the database, each news program is associated with a WAV audio file and a TRS text file (containing the manual transcriptions in the Transcriber format).

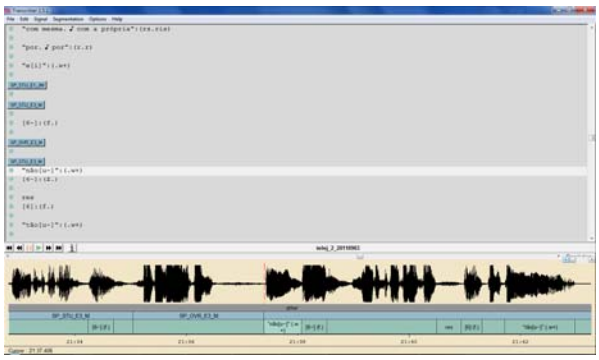


Figure 1: *Examples of audio segments annotation, using the Transcriber software tool.*

3. Hesitation patterns

Considering all the segments that were annotated accordingly to the presence of hesitations, we can see in Table 1 how the hesitation patterns are distributed. A total of 4608 events were observed in which filled pauses (f.) and vocalic extensions within a word (.w+) are the most common, achieving 36.5% and 23.4% of the hesitation events, respectively. The most common hesitation events are somewhat similar to what has been observed for English or Swedish, as reported in [2] and [5], respectively. A similar distribution was previously stated for European Portuguese in [4] and [10].

In Tables 1 and 2, figures are given for the relative occurrence of the most frequent hesitation events observed in the HESITA DB. The left column in Table 1 and Table 2 gives the information about each hesitation pattern. Pattern models display the way that the hesitation occurs, indicating the order of the words before and after the so-called "repair-point". This point marks the place from which the hesitation is corrected and the fluency is restored. For instance, the pattern (r.r) indicates that a word r was repeated as repair or reinforcement; in the pattern (s-.s), the word s was cut and then substituted; in (r2.r) the same word r was repeated twice and finally restored; in (rs-.rs) the word r was repeated and word s was cut and, then substituted with correction.

More complex hesitation patterns are present in the HESITA DB, although not very frequently. For instance, the hesitation with the transcription "*que vo-.que.que.que voltam.que.que possam*" has the following pattern which shows embedded hesitations: ((rs-.(r2.r)s).(r.r)s).

3.1. Hesitations across speaking styles

According to overall figures that have been given for other languages, in general hesitation events occur mainly in spontaneous speech [2], [5], [8], [9]. The same trend is observed in the present database, in which the occurrences of such events in spontaneous speech count 4406 against 188 in read (prepared) speech and 12 in Lombard speech (2 additional events were in noisy segments and thus not further classified).

The total of 188 hesitations observed in 14 hours for read (prepared) speaking style results in a rate of 0.22 hesitations per minute. The 4406 hesitation events in 10 hours of spontaneous speech result in a rate of 7.34 hesitations per minute, which reveals a tendency in fluency also verified for other languages (see [30], for instance).

Considering the gender in our database, this variable appears to not affect fluency rates in spontaneous speech: female and male speakers produce similar rates of hesitations, with 7.72 and 7.26 hesitations per minute, respectively.

It has also been noted by [14] for other languages, that the density of hesitations in speech varies with the speaking style, and in prepared speech corpora vocalic fillers are relatively infrequent, not exceeding 0.7% of the speech data. We corroborate this finding; in our case, in read (prepared) speech, fillers accounts for less than 0.2% of the speech duration.

Table 1.*Top 10 most frequent hesitation patterns.*

Patterns	# Events	% Events
(f.)	1681	36.5
(.w+)	1078	23.4
(r.r)	376	8.16
(p.)	213	4.62
(r+.r)	165	3.58
(s-.s)	95	2.06
(s.s)	84	1.82
(rr.rr)	72	1.56
(r2.r)	45	0.98
(rs-.rs)	43	0.93
others	756	16.4

Table 2 shows the distribution of the 5 most common hesitation patterns in the read (prepared) speech, with the high frequency of vocalic expressions (.w+) (39.36%) just followed by filled pauses (f.) (32.45%). Although the difference between the occurrences of vocalic extensions and filled pauses is not so expressive, it is possible that the choice for the extensions reflects the fact that vocalic fillers tend to be more stigmatized in a prepared speech context. Repetitions in read or prepared speech become residual. The relative frequency rates for substitutions are higher in the prepared speech than in spontaneous speech (9.57% vs. 3.61%), proving that they are more adequate for communicative strategy mainly in what the fluency of speaking is concerned.

Table 2. Top 5 most frequent *hesitation patterns for read (prepared) speech.*

Patterns	# Events	% Events
(.w+)	74	39,36
(f.)	61	32,45
(s-.s)	10	5,32
(s.s)	8	4,26
(rs-.rs)	4	2,13

3.2. Phonetic form of filled pauses

During the annotation procedure, we found the two most common phonetic forms for filled pauses: the near-open central vowel [ɐ] ([6] in SAMPA) and the mid-central vowel [əə] ([@] in SAMPA), representing 48% and 20% of the all filled pauses, respectively. Table 3 shows the ranking of the ten most used in HESITA DB.

Table 3. *Phone distribution of filled pauses (top10 most frequent).*

Phone	N. events	% (Perct.)
[ɐ]	808	48,07
[ə]	344	20,46
[ɐ̃]	155	9,22
[ɐm]	73	4,34
[ũ]	46	2,74
[ɐũ]	31	1,84
[ɐə]	28	1,67
[ɐɐ]	21	1,25
[əɐ]	13	0,77
[ɐ̃m]	9	0,54

The distributon shown in Table 3 also supports the view that the vocalizations preferred by Portuguese speakers are around central vowels, corresponding to the reduced vowels in an unstressed position (/a/ vs. /i/, /e/, /ɛ/, respectively). There is also a slight inclination for the high back rounded nasal vowel [ũ] as well (around 3%). Also a nasal preference is evident in the DB: see [ɐ̃], [ɐm] and [ɐ̃m] or [ũ] in Table 3. Our point here is not to associate a meaning to the filler sounds. However, there is strong empirical evidence that speakers use all of them for playing a structuring role in the speech. The choice for a vocalic sound rather than other appears to be, at least in some contexts, motivated by the behavior of neighbor phonetic segments, neutralizing in some way the phonetic difference of the vocalic fillers.

3.3. Segmentation of hesitations

The annotation of the hesitation events closely follows [2]. It encompasses the initial and final temporal marks and the corresponding label contains the pattern and the orthographic transcription. The repair-point was also marked temporally, showing the instant were the hesitation is corrected and when the fluency on speech is recovered. It has been verified that the period of time that corresponds to the beginning of the hesitation to its repair-point is much larger (0.61 seconds in average) than the period of time between the repair point and the end of the hesitation correction (0.34 seconds in average). This matches what was found in previous studies, such as in [31]. These trends concerning the distribution and duration of hesitation events may be analyzed as manifestations of planning effort as well.

4. Discussion and conclusion

Browsing the literature in the area (e.g. [2], [5], [8]), it has been strongly evidenced that speakers use hesitations as part of their speech structure and in order to achieve a better synchronization with the interlocutors. Various scientific domains more directly interested in gathering knowledge for better identifying salient information in human speech communication, such as the linguistic or clinical/therapeutic areas covering speech fluency, can beneficiate of the analysis of the hesitation distribution along the speech, matching the complementary distribution of such events with the speaking styles, speakers or with the acoustical environment as example. Also, some relevant observation about the temporal characteristics (duration of segments) can be pointed out from the HESITA DB. Different vocalic choices for filled pauses can also be associated, at least in some contexts, with different functions and meanings of expressiveness, such as doubt, denial or agreement.

Studies on hesitations have recently gained in importance to increase the usability of speech systems, by overpassing the challenges proposed by the presence of such phenomena in continuous speech. We also believe that automatic language processing could benefit from a richer representation of the audio signal that incorporates speaking styles information, in order to breakdown errors in the automatic speech recognition or to improve automatic conversational speech summarization, for instance. Detection of hesitation events also provides the segmentation of multimedia data into consistent parts, as

claimed in [11]. It leads also to important applications such as the identification of speech segments to train acoustic models for speech recognition in a more cost-effective way.

Our goal in this paper was to present a database for European Portuguese, which contains a large and rich variety of speech data events and is mainly focused on the hesitations. We thus expect that this database can be a relevant base of work for further studies regarding a variety of speech phenomena. The HESITA database is available through the Meta-Net [32] as well as in the project page [33].

5. Acknowledgements

This work is funded by FCT and QREN projects (PTDC/CLE- LIN/11 2411/2009; TICE.Healty13842) and partially supported by FCT (Instituto de Telecomunicações multiannual funding PEst-OE/EEI/LA0008/2011).

Sara Candeias is supported by the FCT grant SFRH/BPD/36584/2007.

6. References

[1] W.J.M. Levelt, *Speaking. From Intention to articulation*, Cambridge, Massachusetts: The MIT Press, 1989.

[2] E. Shriberg, "Preliminaries to a theory of speech disfluencies", Ph.D. dissertation, University of California, 1994.

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

[4] H. Moniz, et al., "On Filled Pauses and Prolongations in European Portuguese", in *Interspeech '07*, ISCA, Antwerp, Belgium, pp. 2645–2648, 2007.

[5] R. Eklund, "Disfluency in Swedish human–human and human–machine travel booking dialogues", PhD dissertation, Institute of Technology, Linköping University, 2004.

[6] A. Veiga, et al., "Prosodic and Phonetic Features for Speaking Styles Classification and Detection", in *Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science*, Toledano, D.T., Ortega, A., Teixeira, A., Gonzalez-Rodriguez, J., Hernandez-Gomez, L., San-Segundo, R., Ramos, D. (eds.), 2012. vol. 328, pp. 89–98, Springer.

[7] A. I. Mata, "Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologia, Resultados e Implicações Didáticas", Ph.D. dissertation, Faculdade de Letras, Universidade de Lisboa, 1999.

[8] M. Candea, "Contribution à l'Étude des Pauses Silencieuses et des Phenomenes Dits «d'Hesitation» en Français Oral Spontané – Étude sur un Corpus de Récit en Classe de Français", Ph.D. dissertation, Universite Paris III – Sorbonne Nouvelle, 2000.

[9] H. Moniz, "Contributo para a Caracterização dos Mecanismos de (Dis)Fluência no Português Europeu", M.S. thesis, Faculdade de Letras, Universidade de Lisboa, 2006.

[10] A. Veiga, et al., "Characterization of hesitations using acoustic models", in *Proc. of the 17th International Congress of Phonetic Sciences, ICPhS XVII*, Hong Kong, pp. 2054–2057, 2011.

[11] A. Veiga, et al., "Towards Automatic Classification of Speech Styles", in *Lecture Notes in Artificial Intelligence (LNAI)*, H. Caseli et al. (Eds.), Springer-Verlag Berlin Heidelberg, 7243, pp. 421–426, 2012.

[12] Y. Liu, et al., "Enriched speech recognition with automatic detection of sentence boundaries and disfluencies", in *IEEE Transaction on Audio, Speech, and Language Processing* 14, pp. 1526–1540, 2006.

[13] J. Adell, et al., "On the Generation of Synthetic Disfluent Speech: Local Prosodic Modifications used by the Insertion of Editing Terms", in *Interspeech '08*, Brisbane, Australia, pp. 2278–2281, 2008.

[14] I. Vasilescu, et al., "Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech". in *Proc. Interspeech '06*, Pittsburgh, PA, USA, pp. 1850–1853, 2006.

[15] I. Vasilescu, et al., "Perceptual Salience Of Language-Specific Acoustic Differences In Autonomous Fillers Across Eight Languages," in *Interspeech'05*, Lisboa, pp. 1773–1776, 2005.

[16] M. Candea, et al., "Inter- and Intra-Language Acoustic Analysis Of Autonomous Fillers." in Proc. DISS'05, Aix-en-Provence, France, pp. 47–51, 2005.

[17] R. Eklund and E. Shriberg, "Crosslinguistic disfluency modeling: a comparative analysis of Swedish and American English human–human and human–machine dialogs", in *International Conf. on Spoken Language Processing*, Sydney, Australia, 6, pp. 2631–2634, 1998.

[18] H.H. Clark and J.E. Fox Tree, "Using uh and um in spontaneous speaking", *Cognition* 84, pp. 73–111, 2002.

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

[20] A. Bell, et al., "Effects Of Disfluencies, Predictability, And Utterance Position On Word Form Variation In English Conversation", in *Journal of the Acoustical Society of America* 113(2), pp. 1001–1024, 2003.

[21] T.-L. Lee, et al., "Prolongation in spontaneous Mandarin", in *Interspeech'04*, Jeju Island, Korea, pp. 2181–2184, 2004.

[22] S. Henry and B. Pallaud, "Word fragment and repeats in spontaneous spoken French", in *Disfluency in spontaneous speech workshop, DiSS'03*, R. Eklund (ed.), Göteborg University, 3–8 Sept. 2003, pp. 77–80, 2003.

[23] E. Shriberg, "Acoustic properties of disfluent repetitions", in *Proc. ICPhS*, Stockholm, Sweden 4, pp. 384–387, 1995.

[24] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech", *Cognitive Psychology* 37, pp. 201–242, 1998.

[25] M. J. R. Freitas, "Estratégias de Organização Temporal do Discurso", M.S. thesis, Faculdade de Letras, Universidade de Lisboa, 1990.

[26] M. R. Delgado-Martins and M. J. Freitas, "Temporal structures of speech: reading news on TV", in *ETRW '91*, Barcelona, 19-1–19-5, 1991.

[27] M. C. Viana, "Para a Síntese da Entoação do Português", Graduate research thesis, Universidade de Lisboa, 1987.

[28] J.C. Wells, "SAMPA computer readable phonetic alphabet", *Handbook of Standards and Resources for Spoken Language Systems*, Gibbon, D., Moore, R. and Winski, R. (eds.), Berlin and New York: Mouton de Gruyter, Part IV, section B, 1997. (http://www.phon.ucl.ac.uk/home/sampa/)

[29] C. Barras, et al., "Transcriber: a free tool for segmenting, labeling and transcribing speech", in *Proc. 1st International Conf. on Language Resources and Evaluation (LREC)*, 1998, pp. 1373–1376. (http://trans.sourceforge.net/)

[30] H. Bortfeld Leon, et al., "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender", Language and Speech, 2001, 44, pp. 123–147.

[31] H. Moniz, et al., "Analysis of disfluencies in a corpus of university lectures", in *Proc. of ExLing, Athens*, Greece, 2012

[32] http://metanet4u.l2f.inesc-id.pt/repository/search/

[33] http://lsi.co.it.pt/spl/hesitation/downloads.html

Choosing a threshold for silent pauses to measure second language fluency

Nivja H. De Jong & Hans Rutger Bosker

Utrecht Institute of Linguistics OTS, The Netherlands

Abstract

Second language (L2) research often involves analyses of acoustic measures of fluency. The studies investigating fluency, however, have been difficult to compare because the measures of fluency that were used differed widely. One of the differences between studies concerns the lower cut-off point for silent pauses, which has been set anywhere between 100 ms and 1000 ms. The goal of this paper is to find an optimal cut-off point. We calculate acoustic measures of fluency using different pause thresholds and then relate these measures to a measure of L2 proficiency and to ratings on fluency.

Index Terms: silent pauses, number of pauses, duration of pauses, silent pause threshold, second language speech.

1. Introduction

In research on both native (L1) and L2 speech, silent pauses are an important feature to describe, characterize, and compare speech from different speakers, performing in different speaker tasks. However, there is a longstanding debate on what should count as a pause. In connected spontaneous speech, part of the speech signal involves silence every time an occlusive is produced. These silences in speech are not considered as pauses that reflect hesitation behavior and it has been assumed that the silences before occlusives can quite easily be removed from a silent pause count by setting a certain threshold. Goldman-Eisler (1968) proposes a threshold of 250 ms to distinguish between 'articulatory' (<250 ms) and 'hesitation' (>250 ms) pauses [1], and this threshold has been followed both in research on L1 and L2 speech.

More recently, however, using this boundary has been called into question [2, 3, 4]. Most of pauses within the 130 ms – 250 ms range cannot be attributed to articulation [2]. Pauses as short as 60 ms that are *not* part of occlusives have been reported [3].

In L2 research, applying a threshold to measure number and duration of pauses has also been used. Often, Goldman-Eisler is cited and the boundary of 250 ms is used [5, 6]. But in some studies, a lower cut-off point is used [7: 100 ms], or a higher cut-off point is used [8: 400 ms], even as high as 1000 ms [9].

The current paper is an attempt to find the optimal cut-off point for the purpose of L2 research. We use two different strategies to find an optimal cut-off point. In L2 research, acoustic measures of fluency (such as number of silent pauses or speech rate) are used to compare speakers or performances of the same speakers in different tasks. These measures are thought to reflect automaticity in the L2 speech production processes. To find out which measures of fluency are indeed related to automaticity of L2 speech production, and to overall L2 proficiency, studies have related acoustic measures of fluency to subjective ratings on L2 proficiency [10], to subjective ratings on fluency [5, 11], but also to separate measures of L2 proficiency [6, 12].

In this paper, we calculate acoustic measures of fluency using different pause thresholds as lower cut-off points. We then evaluate (1) the relation between these measures of fluency and a measure of vocabulary knowledge as an approximate of overall L2 proficiency, and (2) the relation between the acoustic measures of fluency and subjective ratings. If we find that choosing a specific threshold leads to higher correlations either with L2 proficiency and/or with subjective fluency, this would argue for using this specific threshold in future L2 research on fluency.

Kirsner, Dunn and Hird [4] show that each individual may have his own criterion when distinguishing between short and long pauses, even fluctuating according to variables such as topic, task, time of day, and age. In the current paper, we will therefore also test whether a threshold that fluctuates per individual or speech sample improves the correlations between acoustic measures of fluency on the one hand, with L2 proficiency and perceived fluency, on the other.

2. Method

In what follows, we describe the data that were used in the present study. In short, the full corpus as described in [12] was used to evaluate the effect of different silent pause thresholds on the relation between acoustic measures of fluency with L2 proficiency (vocabulary knowledge); a subset of this corpus was used to evaluate the effect of different silent pause thresholds on the relation between acoustic measures of fluency with ratings.

2.1. Speech data

The corpus consisted of all L2-data from [12]. Fifty-one L2 speakers (24 Turkish L1 and 27 English L1) of Dutch performed eight speaking tasks in their L2. The total duration of speech in this L2-corpus was 9 hours and 43 minutes. For this data, orthographic transcriptions were made in CLAN [13]. Furthermore, silent pauses were detected by careful listening and by using the waveform (as shown in CLAN), and measured in milliseconds. The silent pauses were also classified with respect to their location, specifically whether the silent pauses occurred either within or between Analysis of Speech (AS) units [14]. AS-units can be described as utterances consisting of an independent clause or of a subclausal unit, together with the associated subordinate clause(s). In this study, we will report on measures of fluency based on pauses within AS units only.[1] In total, 10668 silent pauses within AS-units were identified.

Figure 1 shows the distribution of pause durations, after (natural) log transformation. Both [3] and [4] report most pauses to be falling in the "short pause" distribution (roughly under 200 ms), whereas in our distribution most pauses are longer.

[1] Calculating the acoustic measures using all silent pauses, rather than only those within AS-units, led to lower correlations across all analyses.

Our participants speak in their L2, which has probably caused a different distribution as found before in read and spontaneous L1 speech (as reported before in [3,4]). Secondly, pauses in our data were detected manually (the noise in our speech files was too variable to allow for automatic pause detection). Manual detection will lead to fewer very short pauses as compared to automatic detection.

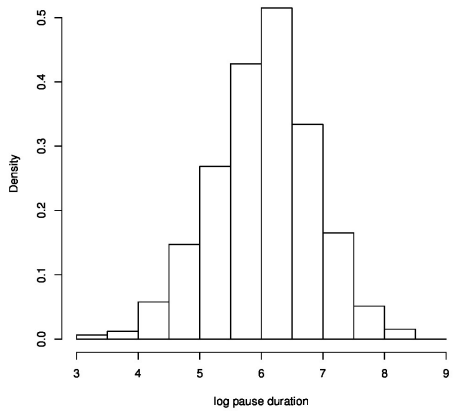


Figure 1: *Histogram of all pause durations in the (full) corpus.*

2.2. Perception data

A subset of the speech corpus was created for a listening experiment, as described in [11]. From 30 L2 speakers (15 L1 Turkish and 15 L1 English), speaking performances from three tasks were used. Twenty-second excerpts were taken from roughly the middle of these 90 speaking performances.

Twenty raters judged the speech samples on fluency on a 9-point equal appearing interval scale. From these ratings, so-called estimates were calculated. These estimates may be considered as mean ratings from the twenty judges, taking into account individual differences of raters, a general effect of order of presentation (raters became more strict towards the end of the experiment), and individual differences between raters with respect to this general order effect. For a complete description of how the perception data were obtained, we refer to [11]. The mean, sd, and range of these rating measures were 5.33, 1.51, and 1.34–8.50, respectively.

2.3. L2 proficiency data

In addition to performing the speaking tasks, the participants also carried out a productive vocabulary task with 116 items. We will use the scores on this task as a separate measure of L2 proficiency. Vocabulary knowledge has been shown to be a good predictor of overall proficiency [15]. Moreover, the same vocabulary test as was used in the current paper, has been shown to be a strong predictor of overall speaking proficiency [16]. The mean, sd, and range of these vocabulary scores were 56, 23, and 8 – 103, respectively.

2.4. Calculating fluency measures

We thus obtained two speech datasets: the full corpus of 51 speakers performing 8 speaking tasks (almost 10 hours of speech), and the subset of 90 roughly 20-second excerpts from thirty of these speakers (54 minutes of speech). To test what the impact may be of setting different thresholds of silent pauses on conclusions drawn in L2 fluency research, we calculated three acoustic measures of fluency that are strongly influenced by choosing different silent pause thresholds.

Choosing different thresholds will strongly influence some acoustic measures of fluency, but other measures of fluency will only change slightly, and for yet other measures of fluency, changing the threshold will not lead to any changes. For instance, speech rate, which is calculated by dividing number of syllables or number of phonemes by total time (including silent pauses), will not change depending on the chosen silent pause threshold because neither the number of syllables nor the total time will change. However, all measures of fluency that are calculated relative to phonation time will slightly change depending on the silent pause threshold, because the phonation time will change accordingly.

In the current paper, we focus on measures of fluency for which changing the silent pause threshold may lead to larger differences: i.e., we focus on the number and duration of silent pauses. For each participant (N = 51) or for each speech sample (N = 90), we calculated three measures of fluency: number of silent pauses per second total time, number of silent pauses per second phonation time, and mean (log) duration of silent pauses. These calculations were made using thresholds for the lower boundary of silent pause durations: at 20, 50, 100, and then at every 50 ms up to 1000 ms for the full corpus. The calculations for the smaller corpus of 90 20-second speech segments were made using the same thresholds, but in these samples the highest threshold was set to 400 ms, because higher thresholds would lead to missing data points (rendering comparisons across thresholds impossible).

Table 1 shows correlations between the fluency measures at a quite low (50 ms) and quite high (400 ms) threshold. It is not surprising that the correlation between the two frequency measures number of pauses per total time and number of pauses per phonation time are highly related (at both thresholds $r = 0.95$). At a boundary of 400 ms, we also see a strong correlation between number of pauses total time and mean duration of pauses ($r = 0.52$).

The correlations across the two thresholds were also carried out (not shown in Table 1). For the measure mean log duration, the measure calculated at 50 ms and 400 ms is highly related ($r = 0.95$), whereas for the two frequency measures the correlation between the measures calculated at the two different thresholds is less strong ($r = 0.65$ and $r = 0.60$ for the frequency measures calculated per total time and per phonation time, respectively). We can therefore expect that for the analyses in which we relate the measures of fluency to vocabulary knowledge and to ratings on fluency, we will not find changes for mean pause duration, as this measure hardly changes with different thresholds. For the frequency measures, on the other hand, we may expect to find differences.

Table 1: *Correlations between fluency measures in the ful corpus for measures calculated at thresholds 50 m and 400 ms.*

	50 ms			400 ms		
	1	2	3	1	2	3
Pauses / sec total time (1)	1	.95	–.06	1	.95	.52
Pauses / sec phon time (2)		1	–.33		1	.30
Log duration of pauses (3)			1			1

As stated in the Introduction, we also calculated measures of fluency with individualized thresholds. Kirsner and colleagues report individualized thresholds with a mean, standard deviation and range of 255, 83 and 98–490 ms, respectively [4]. They established these individualized thresholds by modeling the individual distributions with bi-Gaussian fits per individual. We will calculate individual thresholds in a

different way, for two reasons. The first reason is that our data do not follow clear bi-Gaussian distributions The second reason is that in our data we have information on articulation rates available: the faster the articulation rate, the shorter the articulation pauses must be. To calculate individual thresholds, we will therefore use a threshold for each individual (or speech sample) that is relative to the individual's articulation rate. For the full corpus, the mean threshold was set to 250 ms, and, relative to individual's articulation rate, an individualized threshold was calculated, with a range of individualized thresholds between 139–324 ms. For the small corpus, the individualized thresholds were also around 250 ms, ranging from 138–384 ms, now relative to each of the 90 speech samples' articulation rates.

3. Analyses

3.1. Relating measures of L2 fluency to L2 proficiency (thresholds 20 ms – 1000 ms)

We related the acoustic fluency measures, as calculated from the full corpus described above, to the measure of L2 proficiency (vocabulary knowledge). Figure 2 shows the Pearson correlations (on the y-axis) between the measure of L2 proficiency and the fluency measures (as shown by different lines), for the silent pause thresholds 20 ms – 1000 ms (on the x-axis).

For mean log pause duration, none of the Pearson correlations were found to be significant. For the two frequency measures of pauses, there is a rise in correlations from thresholds 20 ms (per total time: $r = –0.42$; per speaking time: $r = –0.39$) to 300 ms ($r = –0.48$ and $r = –0.53$, respectively), after which the correlations drop.

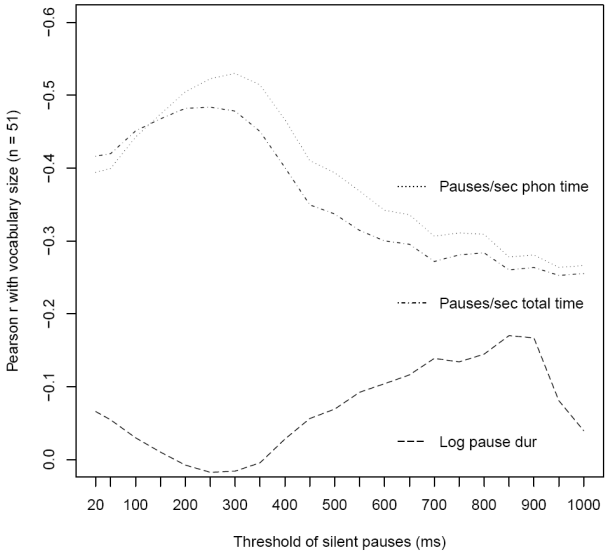


Figure 2: *Pearson correlations between vocabulary size and measures of fluency, calculated for different silent pause cut-off points.*

3.2. Relating measures of L2 fluency to perceived fluency (thresholds 20 ms – 400 ms)

For each speech sample of roughly 20 seconds ($N = 90$) in the small corpus, we calculated the acoustic measures of fluency as described above. Figure 3 shows the Pearson correlations (on the y-axis) between the measure of perceived fluency and the three acoustic fluency measures (as shown by different lines), for the silent pause thresholds 20 ms – 400 ms (on the x-axis).

As can be seen from this figure, changing the threshold from 20 ms to 400 ms does not lead to differences in correlations between the measure log pause duration and perceived fluency. For both frequency measures of pauses, however, we find that the higher the lower cut-off point for silent pauses, the higher the correlation between the resulting frequency measure (either number of silent pauses per total time or number of silent pauses per phonation time) on the one hand, and the ratings of fluency, on the other.

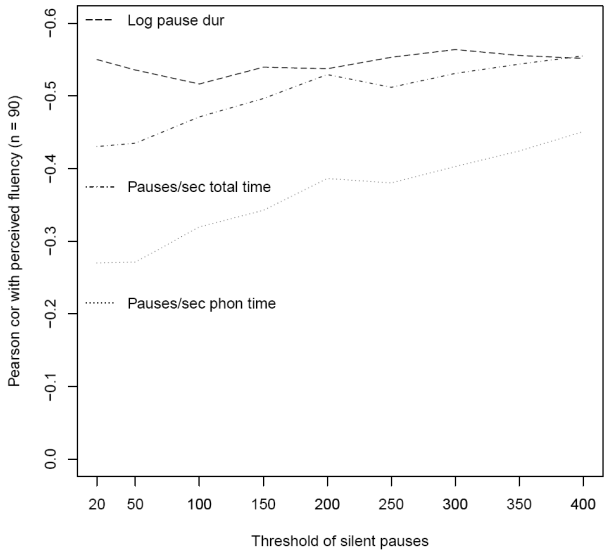


Figure 3: *Pearson correlations between ratings of fluency and measures of fluency, calculated for different silent pause cut-off points.*

3.3. Applying individual thresholds

As described above, we also calculated the measures of fluency when using individualized thresholds, relative to the articulation rates. For the correlations between acoustic measures of fluency and vocabulary knowledge, using these individualized thresholds, did not lead to different results from using the non-individual thresholds of 250 ms or 300 ms.

For the correlations between acoustic measures of fluency and the ratings, we did find differences, however: when number of pauses (per phonation time and per total time) was calculated using the individualized thresholds, the Pearson correlations were $r = –0.49$ and $r = –0.60$ respectively. These are higher than the correlations found for a non-individual threshold of 250 ms ($r = –0.38$ and $r = –0.51$ respectively). The correlation between perceived fluency and log duration of pauses when calculated with the individualized threshold, on the other hand, was lower ($r = –0.47$) than when it was calculated with a non-individual threshold of 250 ms ($r = –0.55$).

4. Discussion

From our results, a number of observations can be made. The first is that the correlations between log duration of pauses and vocabulary knowledge are always low and never significant (around $r = –0.1$; see Figure 2). The correlations between log duration of pauses and perceived fluency, on the other hand, are always much higher ($r = –0.55$; see Figure 3), irrespective of the threshold. This general discrepancy (low correlation between duration of pauses for measures of proficiency and high correlation for ratings of fluency) has been reported before [6,11]. What we can conclude from the present study, however, is that these findings are not dependent on a specific threshold.

Another finding from the current study is that the relation between vocabulary size and number of silent pauses *is* dependent on the chosen threshold. This relation is highest when a threshold of around 250–300 ms is used. We may conclude from this finding, that 250–300 ms is the optimal threshold for measuring the number of pauses (per total or per speaking time) with respect to studies that aim to investigate L2 proficiency. In other words, adding the number of pauses below 250 ms to counts obtained when a traditional cut-off point of 250 ms is used, leads to a measure of fluency that is less strongly related to L2 proficiency. Similarly, setting the threshold higher than 300 ms leads to lower correlations. We conclude from this finding that although many silent pauses are shorter than 250 ms (in our data between 22% and 27%), these pauses seem irrelevant when calculating measures of fluency that are related to L2 proficiency.

Such an optimal threshold for the number of silent pauses could not be found when relating the measures to perceived fluency; in this case the correlations get stronger as the threshold is higher. We could conclude from this finding that raters only take the number of long pauses into account (at least >400 ms) when judging on fluency. However, we propose another explanation for this finding. It follows naturally that a count of silent pauses with a high threshold is related to mean duration of silent pauses. If only the number of long pauses is counted, this count will be strongly related to the mean duration of pauses. Indeed, the correlation between mean log duration of silent pauses and the number of silent pauses when using a threshold of 400 ms is quite high (see Table 1: $r = 0.52$) and gets steadily higher if the threshold for counting pauses per second is raised (to $r = 0.77$ for a threshold 750 ms). The rise in correlations between number of pauses and ratings on fluency as the threshold is set higher, can therefore be explained by the fact that counting only long pauses is confounded with measuring mean duration of silent pauses (a measure that was strongly related to ratings on fluency).

In this study, we have also compared two frequency measures of pauses: number of pauses per second total time and number of pauses per second phonation time. For L2 proficiency, the correlations were almost the same for these measures (slightly higher when it was calculated per second phonation time). For the ratings, however, the correlations were higher when the acoustic measure was calculated per total time. We can explain this finding, again, by taking into account the intercorrelations between the acoustic fluency measures: the number of pauses per total time, especially as the threshold gets higher, is in fact a confounded measure of the number of pauses and their duration.

5. Conclusions

This study showed that a lower cut-off point for silent pauses of 250–300 ms leads to the highest correlation between the number of silent pauses and a measure of L2 proficiency (vocabulary knowledge). Such an optimal threshold could not be found for the mean (log) duration of silent pauses in relation to L2 proficiency: mean duration of silent pauses is not significantly related to L2 proficiency, no matter which threshold was chosen.

When relating the acoustic fluency measures to ratings on fluency, no clear optimum could be found. For mean duration of pauses, correlations between ratings and this measure were always high, irrespective of the threshold. For the number of silent pauses, the correlations became higher, as the threshold was set higher. This finding, however, can be explained by the fact that counting only long pauses (by setting the threshold high) is confounded with measuring the duration of pauses.

We therefore conclude that for the purpose of L2 research, the traditional cut-off point of 250 ms is a good choice and using a higher threshold than 300 ms has two disadvantages: (1) with respect to number of pauses, it leads to measures of fluency that are less representative of L2 proficiency, and (2) the acoustic fluency measures number of pauses and duration of pauses become confounded as higher thresholds are used.

6. Acknowledgements

The authors wish to thank research assistants Cem Keskin and Erica Bouma for measuring silent pauses; Cornelia Lahmann and Rasmus Steinkrauss for valuable discussion. This research was funded by Pearson Language Testing, grant "Oral fluency: production and perception" awarded to N.H. De Jong.

7. References

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

[2] A.E. Hieke, S. Kowal and D.C. O'Connell. "The trouble with 'articulatory' pauses", *Language and Speech* 26, pp. 203–214, 1983.

[3] E. Campione and J. Véronis. "A large-scale multilingual study of silent pause duration." *Proc. ESCA-workshop*, pp. 199–202, 2002.

[4] K. Kirsner, J. Dunn and K. Hird. "Fluency: Time for a paradigm shift", *Proc. DiSS '03*, pp. 13–16, 2003.

[5] C. Cucchiarini, H. Strik and L. Boves. "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech", *JASA* 111, pp. 2862–2873, 2010.

[6] N.H. De Jong, M.P. Steinel, A.F. Florijn, R. Schoonen and J.H. Hulstijn, "Linguistic skills and speaking fluency in a second language", *Applied Psycholinguistics*, online 2012.

[7] A. Riazantseva, "Second language proficiency and pausing: A study of Russian speakers of English", *SSLA* 23, pp. 497–526, 2001.

[8] T.M. Derwing, M.J. Munro, R.I. Thomson and M.J. Rossiter, "The relationship between L1 fluency and L2 fluency development", *SSLA* 31, pp. 533–557, 2009.

[9] N. Iwashita, "Features of oral proficiency in task performance by EFL and JFL learners". in M.T. Prior [Ed], Selected proceedings of the 2008 Second Language Research Forum, pp. 32–47, Cascadilla Proceedings Project, 2010.

[10] N. Iwashita, A. Brown, T. McNamara and S. O'Hagan, "Assessed Levels of Second Language Speaking Proficiency: How Distinct?", *Applied Linguistics* 29(1), pp. 24–49, 2008.

[11] H.R. Bosker, A.F. Pinget, H. Quené, T. Sanders and N.H. De Jong, "What makes speech sound fluent? The contributions of pauses, speed and repairs", *Language Testing* 30, pp. 159–175, 2013.

[12] N.H. De Jong, R. Groenhout, R. Schoonen and J.H. Hulstijn, "Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior." *Applied Psycholinguistics*, online 2013.

[13] B. MacWhinney, "The CHILDES project: Tools for analyzing talk". Mahwah, NJ: Erlbaum, 2000.

[14] P. Foster, A. Tonkyn and G. Wigglesworth, "Measuring spoken language: A unit for all reasons". *Applied Linguistics* 21, pp. 354–375, 2001.

[15] A. Zareva, P. Schwanenflugel and Y. Nikolova, "Relationship between lexical competence and language proficiency", *SSLA* 27, pp. 567–595, 2005.

[16] N.H. De Jong, M.P Steinel, A. Florijn, R. Schoonen and J.H. Hulstijn, "Facets of speaking proficiency", *SSLA* 34, pp. 5–34, 2012.

Lengthenings and filled pauses in Hungarian adults' and children's speech

Andrea Deme[1,2] *& Alexandra Markó*[1]

[1] Department of Phonetics, Eötvös Loránd University, Hungary
[2] Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary

Abstract

In the present paper vowel lengthenings and non-lexicalized filled pauses were studied in the spontaneous speech of children and adults (focusing more on the much less studied phenomenon: vowel lengthening). The results revealed different usage and appearance of lengthenings in the two age groups, therefore, differences in speech skills and strategies can be concluded. LEs and FPs differ mostly in their position in the speech session between the age groups, which has implications regarding different planning strategies of adults and children. We also draw conclusions regarding the methodological considerations in the issue of identifying vowel lengthening supporting a previously formulated conception.

Index Terms: lengthening, (non-lexicalized) filled pause, spontaneous speech, speech planning, discourse management.

1. Introduction

Although lengthenings and pauses (silent and filled) are the two most common disfluencies of spontaneous speech, lengthenings have received little attention so far in the literature, while silent and filled pauses are more often placed in the focus of research. Lengthenings (hereafter LE) are generally considered to be similar to non-lexicalized filled pauses (hereafter FP) as both use the least amount of energy and similar duration to afford the speaker time for thought and speech planning [1]. However, there is also a growing body of evidence suggesting that despite their obvious similarities, they also differ significantly in their acoustic properties as well as their positions and functions in discourse.

Previous findings have already implied that children may differ from adults in terms of the acoustic features, and the place of occurrence of LE [2] and the frequency of occurrence of FP [3], which might be the result of differences in speech planning skills and strategies. Based on these observations, the aim of the present study was to compare the patterns of LE and FP within and between two different age groups.

In our investigation different frequency of LEs and FPs between age groups was hypothesized, with children leaning toward lengthening, and adults being more prone to hesitations. Differences in phonetic and positional characteristics of both phenomena between adults and children also were assumed. The differences were thought to originate from the presumed differences of speech production strategies of the speakers of the two age groups.

2. Subjects, material and method

Spontaneous speech samples of 8 children (4 boys, 4 girls, aged 7 to 8 years) and 8 women (aged between 20 and 32 years) were used for analysis. The total duration of the speech material was approximately 1 hour.

For identifying prolonged vowels a perception test was used, introduced in [2]. As [2] points out, two approaches are often blended in the literature: to regard LE as a perceptual phenomenon, or separate lengthened vowels on the grounds of physical duration alone. Therefore, in our present study we adhere to the notion, which regards LE as perceptual phenomena; hence the designation of LEs was carried out with a listening test including 10 linguists. They were asked to listen to the speech samples through headphones, played by a computer, in a quiet room, and meanwhile to follow the transcription of the played texts presented on the screen (the transcription was orthographic, but without punctuation). The subjects' task was to mark those vowels in the written texts, which were considered lengthened based on their subjective auditory judgment. The vowels marked by at least 6 linguists were counted as lengthened. FPs in the texts were identified by means of auditory perception of the authors, together with the visual confirmation of a spectrogram.

The two phenomena (LEs and FPs) were analysed both in children's and adults' speech, and compared according to the following parameters:

- frequency of occurrence,
- duration,
- fundamental frequency (f_0 in semitones),
- position in the clause,
- the position in the "speech session"(separate part of the utterance bordered by silent pauses), and
- the presence of adjacent disfluencies.

Though the following parameters were only applicable to the inspection of LEs, they were included in the research, since they seemed to reveal important aspects of this (less studied) phenomenon. These additional features were:

- the duration of lengthened and non-lengthened vowels,
- the extent of lengthening,
- the rate of lengthened vowels,
- the LE's position within the word, and
- the ratio of word types (content vs. function) involved.

Praat 5.3 was used for acoustic analysis and SPSS 15.0 for statistical analysis (*t*-test, Pearson-correlation).

3. Results

3.1. Comparison of LE and FP between the two age groups

In the adults' corpus 70 LEs and 127 FPs were annotated. In terms of **frequency** of occurrences this means that a FP was produced on average every 17 seconds, and roughly one in every 115 vowels was marked as lengthened. In the children's material we found 67 LEs and 56 FPs, which means an FP every 32 seconds, and a LE occurred once in every 56 vowels. The rate of FPs seems to be similar in the two groups, but the rate of LEs in child speech is almost the double of the count calculated for adults.

When compared to the count of all vowels uttered, adults seem to prefer FPs to LEs, which in adults' speech seem to be relatively rare (0.44%) with a moderate standard deviation (0.33%). In child speech, however, LE seemed to appear more often and diverged (0.78±0.65%). The frequency as well as the deviation of FP was similar in the two groups (adults: 0.78±0.55%; children: 0.66±0.55%) (Figure 1).

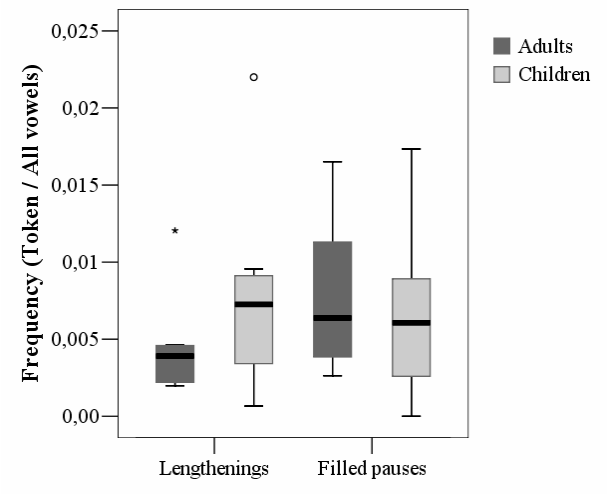


Figure 1: *Frequency of LE and FP in both age groups.*

There was no correlation observed between the ratio of FPs and LEs in the speakers' material. As we have seen, the inter-speaker variability is rather high, however the age groups do not differ in this point of view. Figure 2 shows that there are speakers among both children and adults who preferred LEs (e.g., A3, Ch1, Ch4), but the opposite preference can also be noticed (e.g., A2, A5, Ch8), while balanced ratios are shown as well (e.g., A1, A8, Ch6).

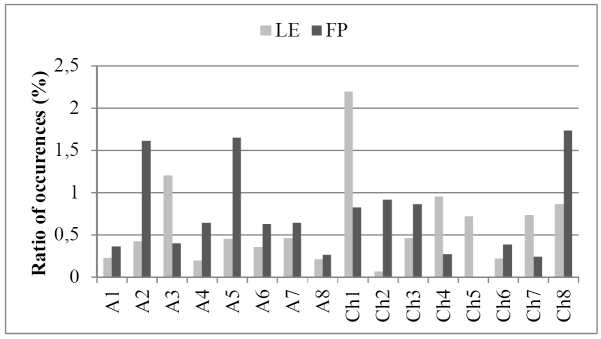


Figure 2: *Ratio of occurrences of LEs and FPs in the speakers' material in % of the total amount of speech sounds (A= adult, Ch= child).*

The pattern of the **duration**al differences between the two phenomena differed according to age groups (with greater variability for FPs in both): while in child speech LE might be slightly longer than FP on average (LE: 344±87 ms; FP: 319±189 ms), in adults it is the opposite (LE 286±105 ms; FP: 374±143 ms) (Figure 3).

The **f_0 values** of FP show a very systematic correspondence: they appear to be very similar within the two groups, but, as it could be presumed, children realized both of the phenomena on a higher fundamental frequency (LE for adults: 10±3 st, for children: 18±2 st; FP for adults: 9±3 st, for children: 19±4 st).

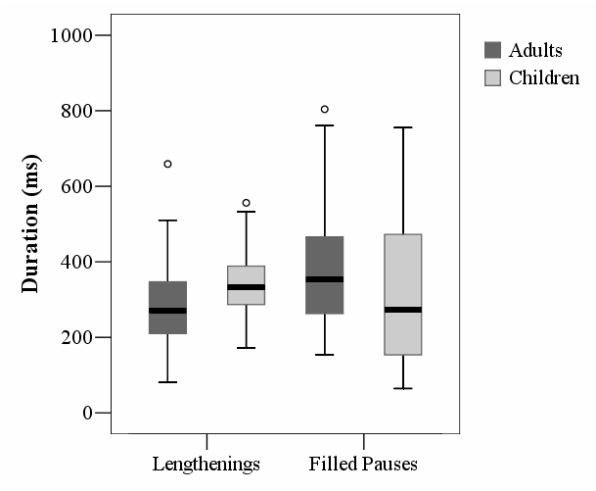


Figure 3: *Duration of LE and FP.*

With respect to the **position in the clause** LE and FP vary marginally in adults, and both phenomena seem to occur mainly on the boundaries (at the conjunction or at the beginning or ending) of the clauses (Figure 4). Contrarily, children tended to use LE and FP more separately: as the distribution of LE is leaning towards the boundary position, FPs seem to be equally distributed.

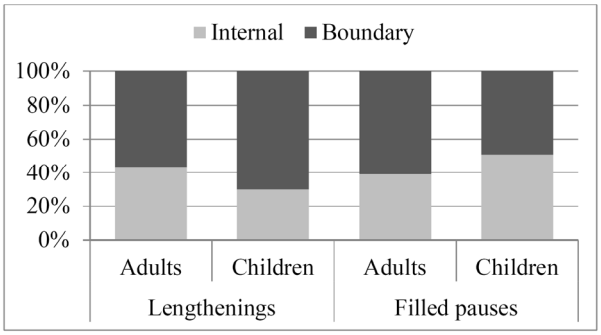


Figure 4: *Position of LE and FP in the clause.*

Position inside the speech sessions can be interpreted as follows. For an FP the position is determined with the help of the adjacent silent pauses (*initial* position is preceded by a pause, *final* position is followed, *isolated* is preceded and followed by a pause, while *medial* means the lack of adjacent pauses). Therefore the results suggest that in child speech half of the occurrences stand between pauses, while adults preferred the initial or final positions in a same ratio (Figure 5). For LE the position refers to the word's place in which the LE occurred (e.g. LE in the initial position is in the first word of the session, and isolated position means a one-word session). For LEs we could not find remarkable differences in the position patterns of the two age groups, both used LE mainly in the final position. This can be a marker of discourse management.

The context of both phenomena was analyzed to see whether any (other) **disfluency phenomena** (like false start, repetition, etc.) occur next to the given LE or FP. According to the results, LEs in adult speech are more likely to occur in the company of other disfluency phenomena than in child speech (yet almost 80% appear without adjacent disfluencies), whereas in child speech they seem to be a more independent means of discourse. In contrast, the pattern is the opposite for FPs, as FPs in children have the highest ratio of neighbouring disfluencies out of all cases (Figure 6).

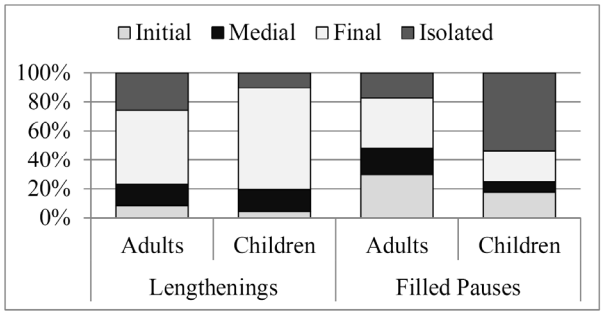


Figure 5: *Position of LE and FP in the speech sessions.*

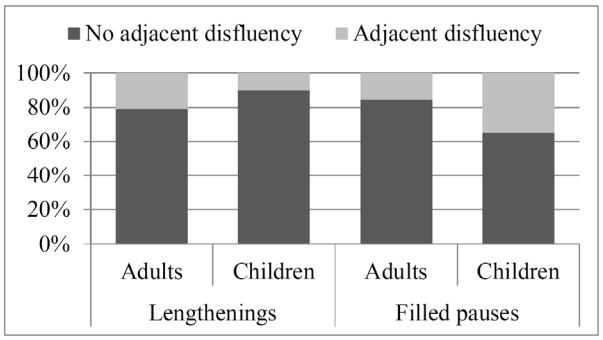


Figure 6. *Distribution of LEs and FPs in terms of adjacent disfluencies.*

The figure does not show those cases when a FP and a LE occurred together. In the children's material 10.4% of LEs were accompanied by a FP, this ratio is 24.3% in adults' speech. At the same time 12.5% of FPs were preceded or followed by a LE in children's speech, and similarly 12.2% in the adults' material.

3.2. Additional differences between adults and children in terms of the characteristics of LE

The **duration** of all vowels was measured in the corpus. The values of duration (as one might expect on the basis of the data, see Figure 7) showed positive correlation with the number of markings in both age groups (Pearson's $r = 0.514$, $p < 0.001$ for children; Pearson's $r = 0.645$, $p < 0.001$ for adults). In most cases, the physical length of the prolonged vowels was significantly longer than the values of the corresponding non-lengthened realizations in both age groups (*t*-test, $p < 0.05$), but at the same time, a relevant degree of overlap can be observed between the lengthened and non-lengthened clusters (Figure 7). This means that some of the non-marked vowels' duration exceeded some LEs' duration. (It should be noted that Hungarian sound system distinguishes phonologically short and long vowels [4], although the physical duration of short and long vowels naturally overlap [5].)

The **extent of lengthening** was determined as follows: the average duration of vowels, which in the perception test were marked as lengthened were compared to the average of those, which were not marked by at least 6 informants (Figure 8). The results showed differences in terms of vowel quantity, but the age groups differed only with respect to the long vowels. In case of (phonologically) short vowels a greater extent of lengthening was observed than for long ones; and the degree of prolongation was very similar in the two age groups (children: 374%, adults: 365%). For long vowels an average prolongation of 262% was found in the children's material, while 307% in adults' speech.

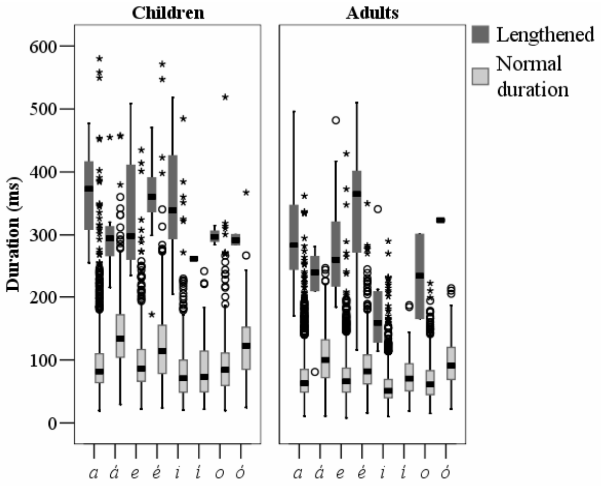


Figure 7: *Average duration of vowels (a* = [ɑ], *á* = [aː], *e* = [ɛ], *é* = [eː], *i* = [i], *í* = [iː], *o* = [o], *ó* = [oː]*).*

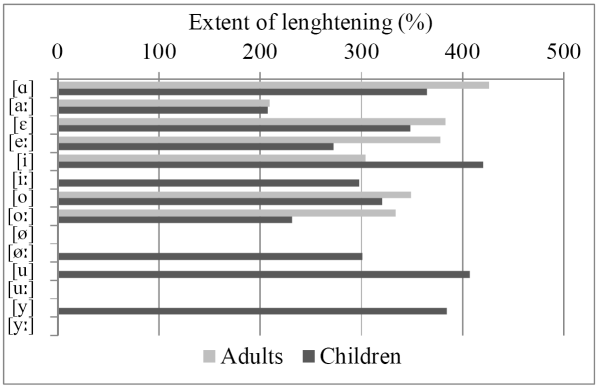


Figure 8: *The extent of lengthening.*

Ratio of lengthened vowels was defined as well. With respect to the vowel quality the differences between the two age groups were spectacular: while in children's speech all vowels of the Hungarian sound system were lengthened in a certain amount, in adults' speech this phenomenon was not documented in case of [ø øː u uː y yː] (Figure 9). It should be noted, that the ratio of the various vowels was similar in the speech samples (and in accordance with the frequency of speech sounds in Hungarian spontaneous speech reported by the literature [5]). Although [ø øː u uː y yː] are relatively rare in Hungarian speech, children lengthened a remarkable proportion of their occurrences (while adults did not at all).

In analysis of the LE's **position within the word** four categories were determined: *initial* means LE in the first syllable, *final* means LE in the last syllable, *medial* means a LE word internally, and *isolated* stands for one-syllable words, in which the vowel was lengthened (Figure 10). In adult speech final and isolated tokens occurred in the same ratio, while in child speech the frequency of isolated LEs exceeded 70%.

Finally the **type of the word** in which the LE occurred (content or function word) was compared between the age groups. The similarity of the proportion of isolated LEs (in Figure 10) and function words (in which LEs appear) (in Figure 11) is not surprising, as most of both of them can be traced back to the lengthened (one-syllable) articles and connectives, while in content words LEs tend to be positioned in the last syllable.

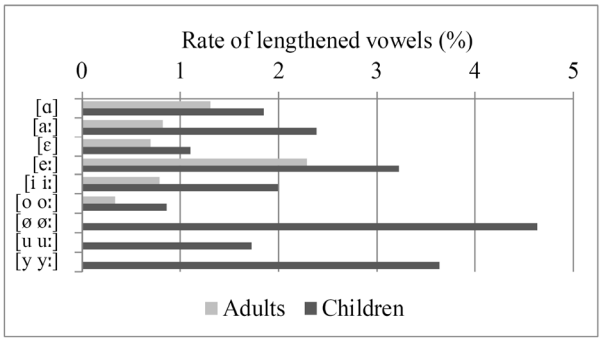


Figure 9: *Rate of lengthened vowels in terms of vowel quality (lengthened/all).*

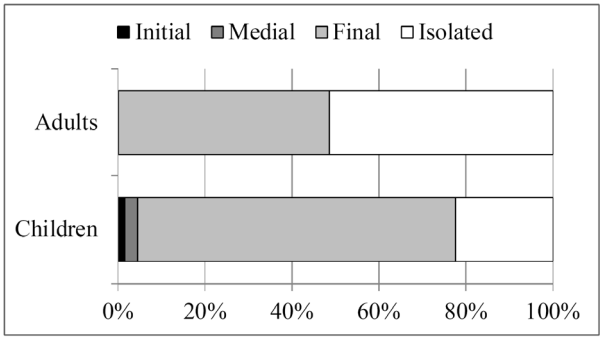


Figure 10: *Position of LEs within the word.*

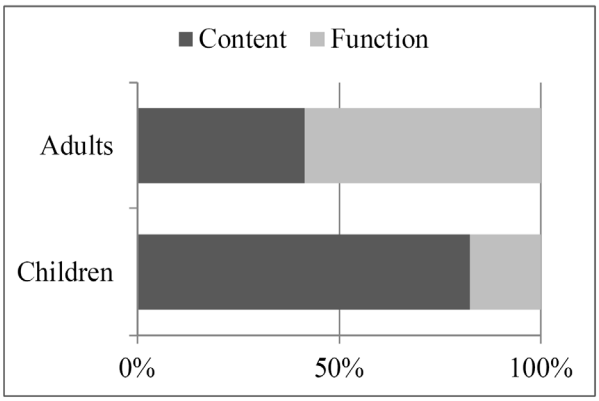


Figure 11: *Distribution of LEs in terms of word type.*

4. Discussion and conclusions

The above analysis revealed that occurrences of LE show different patterns in child and adult speech: despite the great inter-speaker variability children on average used LEs twice as often as adults do and the phenomenon in child speech affected all of the vowel qualities, but mostly the content words (while in adult speech the rare vowels were excluded and mostly the function words were involved, as described for Swedish adult speech as well [1]).

The greater variability of the duration of FPs in child speech might have some implications regarding latter acquisition (thus not yet stabilized strategy) using FP in the discourse.

Both phenomena occurred mainly on the boundaries of the clauses in adult speech, while in the children's material LE was more often in the boundary position, but FPs were equally distributed. On one hand, balance in distribution might denote a more incidental use of a speech unit. On the other hand, as the clause-internal position means a point of grammatical incompletion, pattern differences found in "filling up" this position with a non-lexical item, again, imply differences between the speech planning processes or discourse managing strategies of the speakers. This is especially true, if we accept the generally assumed idea about the function of FP: pausing, or time-gaining (for a basic summary of typical interpretation found in the literature see [6]).

According to the argument of [6], a clitic attached onto the previous element (and any additional lengthening of it) signals that the clitic was planned simultaneously with the element it is stuck to. Adapting their reasonable argument, we suggest that great differences found in position of LE and FP in the speech session indicate planning differences.

LE and FP diverge between the two age groups based on their position in speech sessions: session ending LEs found in both groups, and ending FPs found in adults imply a similar planning strategy while FPs in child speech, which form isolated utterances imply another. Furthermore, it can also be assumed that session final LE and FP can be a marker of discourse management.

Conclusions can also be drawn concerning the methodology of designating LEs applied in the paper. The duration of vowels showed positive correlation with the number of marks collected in the listening test and the lengthened vowels showed statistically longer durations from the normal realization. However, it was also revealed that in some cases lengthened vowels were shorter in duration. These findings support the suggestion introduced in [2], which states the necessity of differentiating the two approaches often blended in the literature: to regard LE as a perceptual phenomenon, or separate lengthened vowels on the grounds of physical duration alone.

Our analysis sheds light on the different usage of LE and FP within and between the two age groups, suggesting that LE and FP maybe a part of a globally different speech production strategy changing with age and can even have functions over and beyond providing time for speech planning (e.g. in discourse management).

5. References

[1] R. Eklund, "Prolongations: A dark horse in the disfluency stable", *Proc. on Disfluency in Spontaneous Speech (DiSS'01)*, Edinburgh, Scotland, UK, August 29–31, pp. 5–8, 2001.

[2] A. Deme, "Magánhangzónyújtások gyermekek spontán beszédében" [Lengthenings in the spontaneous speech of children], in Váradi, T. [ed], *VI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*, Budapest, Hungary, pp. 24–39. Online: http://www.nytud.hu/alknyelvdok12/proceedings12/deme2012.pdf, 2013.

[3] K. Menyhárt, "A spontán beszéd megakadásai az életkor függvényében", in Hunyadi, L. (ed.) *Kísérleti fonetika – laboratóriumi fonológia a gyakorlatban* [Experimental phonetics – laboratory phonology in practice], Debrecen: Debreceni Egyetem Kossuth Egyetemi Kiadója, pp. 125–138, 2003.

[4] P. Siptár, Törkenczy, M., *The Phonology of Hungarian*. Oxford: Oxford University Press, 2000.

[5] M. Gósy, *Fonetika, a beszéd tudománya* [Phonetics, the science of speech]. Budapest: Osiris. 2004.

[6] H.H. Clark and J.E. Fox Tree, "Using uh and um in spontaneous speaking", *Cognition* 84(1), pp. 73–111, 2002.

Anti-zero pronominalization: when Japanese speakers overtly express omissible topic phrases

Yasuharu Den[1] *& Natsuko Nakagawa*[2]

[1] Faculty of Letters, Chiba University, Japan
[2] Graduate School of Human and Environmental Studies, Kyoto University, Japan

Abstract

In this paper, we focus on cases where Japanese speakers overtly express a topic phrase that could have been omitted. We call this phenomenon *anti-zero-pronominalization* and hypothesize that this helps speakers gain time for planning a following utterance; anti-zero-pronominalization is another option to deal with cognitive load at the beginning of an utterance in addition to fillers and other speech disfluencies. Based on a quantitative analysis of a corpus of spontaneous Japanese dialogs, we investigate the difference between overt topic NPs and zero-pronouns. We show that i) the utterance is more complex when the topic is expressed as an overt NP than when it is expressed as a zero-pronoun; ii) turn-initial items such as fillers are produced less frequently when overt NPs appear than when zero-pronouns appear; and iii) the utterance becomes more complex when the last mora of the topic is more prolonged.

Index Terms: zero-pronouns, topic phrases, cognitive load, Japanese dialogs

1. Introduction

In Japanese, speakers can omit arguments of predicates when they are recoverable from the context. Unlike English, in which ellipsis is possible only in limited syntactic configurations such as coordinated structures, Japanese has rather weak constraints on the use of ellipsis, or *zero-pronouns*. Not only subjects but also objects and other arguments can be omitted regardless of syntactic configurations. Semantic constraints are also weak; in addition to first and second person pronouns, expressions referring to third persons and non-animate or abstract objects are subject to zero-pronominalization.

There have been a number of studies on Japanese zero-pronominalization in terms of pragmatic constraints of the use of zero-pronouns, models of zero-anaphora resolution, characteristics of zero-pronouns in real data, and so on [1–6]. Based on quantitative analyses of narratives and conversations, some researchers reported that in spoken Japanese about 70% of arguments in their data were zero-pronouns [2, 3, 6]. According to some scholars [7, 8], the choice between overt NPs and zero-pronouns is crosslinguistically determined by the cognitive status of the referent in question. In general, zero-pronouns are used when the referent is assumed to be activated in the hearer's mind, while overt NPs are used when the referent is not assumed to be activated.

There are, however, some cases in which speakers overtly express arguments of predicates even though the referent can be assumed to be activated. For instance, consider the following exchange:

(1) L: zyosee-wa i-nai-n-desu-ne
women-TOP be-NEG-N-POL-FP
There were no women, right?

R: zyosee-wa-ne i-masi-ta-kedomo (0.4)
women-TOP-FP be-POL-PAST-though

mata ryoo-ga tigat-te
as well dormitory-NOM be.different-and
There were women, but the dormitories were different.

(D01M0047:403.854-408.814)

In responding to L's question, R could have omitted "zyosee (*women*)" and simply said, "i-masi-ta-kedomo, . . . (*(Yes,) there were, ...*)," which is perfectly acceptable in this context because the expression "zyosee (*women*)" is mentioned in the immediate context. The speaker, however, repeats the topic phrase "zyosee-wa-ne." We call this phenomenon *anti-zero-pronominalization*.

One possible motivation for speakers to anti-zero-pronominalize omissible topic phrases is to gain time for planning a following utterance. By overtly expressing a topic phrase at the beginning of an utterance, the speaker can delay the production of the substantial content of the utterance. In this respect, we hypothesize that anti-zero-pronominalization has a similar function to speech disfluencies such as fillers [9, 10], word repetitions [11, 12], and prolongation [13].

In this paper, we elucidate factors behind the use of anti-zero-pronominalization in Japanese dialogs. In particular, we test hypothesis that overt topic phrases, as opposed to zero-pronouns, are used when the speaker's cognitive load in speech planning is relatively high. To test this hypothesis, we first extract from a corpus of spontaneous dialogs exchanges like (1), where a discourse entity overtly expressed in the first utterance is repeated in the second utterance by the other speaker and expressed either as an overt NP or as a zero-pronoun. We then examine whether or not the expression type (overt vs. zero) of the topic affects some production-related variables such as the duration of the substantial part of the utterance and the presence of prefaced items at the beginning of the utterance.

2. Methods

2.1. Corpus

We used the dialog subset of the core data of the *Corpus of Spontaneous Japanese* [14]. The data consists of 18 sessions of dialogs produced by 6 dyads, each of which participated in 3 different kinds of sessions: SPS-Int, APS-Int, and Task. In the SPS-Int dialog, one of the participants of each dyad, called 'interviewer,' interviewed the other participant ('interviewee') on the simulated public speech (SPS) the interviewee had given before the interview.

L: sono sanzyoo-no heya-ni hokani-wa donna kadenseehin-toka oi-te-ta-n-desu-ka
that three.tatami-GEN room-in other-TOP what.kind home.appliances-like have-PROG-PAST-N-POL-Q
What kind of home appliances did (you) have in that three-tatami room?

R: | a | (0.2) | sore-wa | kekkoo odoroku-hodo takusan hait-te |
| oh | | | that-TOP | very surprising-as alot had-and |
| **Preface** | | | **Topic** | **Body** |
Oh, (I) had surprisingly many of those things.

(D01M0019:225.268-231.874)

Figure 1: *The structure of the responding utterance.*

In the APS-Int dialog, the interviewer and the interviewee talked about the academic presentation speech (APS) the interviewee had given before the interview. In the Task dialog, each dyad performed a task-oriented dialog without roles of interviewer and interviewee.

The core data of the corpus is annotated with rich information, including phonetic segments, words, phrases, and clause units, and provided in the form of relational database [15]. Starting and ending times of units at each level of granularity are precisely identified, which enables us to conduct detailed analyses based on time information.

2.2. Annotation

We first extracted exchanges between two interlocutors, where a discourse entity overtly expressed in the first utterance was mentioned in the second utterance by the other speaker in the form of either an overt NP or a zero-pronoun. The detailed procedures are the following:

1. We identified clause units, which we regarded as utterances [16]. We combined two clause units that were temporally proximate if the first unit was response tokens, or other items that typically appear at the beginning of a new turn, e.g., "iya (*no*)" and "soo-desu-ne (*let me see*)."
2. We automatically extracted two consecutive utterances produced by two different speakers under the condition that the second utterance started no earlier than the beginning of the last phrase of the first utterance and no later than one second after the end of the first utterance.
3. We only retained utterance pairs where the first utterance initiated an exchange and the second utterance responded to it, such as question-answer pairs, and where a discourse entity overtly expressed in the first utterance was mentioned in the second utterance in the form of either an overt NP or a zero-pronoun. We disregarded the distinction between full NPs and pronouns, both regarding overt NPs. We excluded expressions referring to the interlocutors and overt NPs in the second utterance markded by non-topic particles such as "ga (NOM)" and "ni (DAT)."

These procedures left us 123 instances of utterance pairs: 46 overt and 77 zero expressions in the responding utterance (6 vs. 13 for APS-Int, 20 vs. 26 for SPS-Int, and 20 vs. 38 for Task). Due to the small number of cases in APS-Int, we merged the two interview subsets into a single session type.[1]

We next annotated the structure of the responding utterance:
i) the preface, in which utterance-initial fillers, response tokens, and other items mentioned above appear; ii) the topic, which is expressed as an overt NP and is repeated from the initiating utterance; and iii) the body, which forms the substantial content of the utterance (see Figure 1). The preface and the topic can be empty; an empty topic means that the discourse entity repeated in the responding utterance is expressed by a zero-pronoun.

2.3. Variables and predictions

The major factor of the analysis was the expression type of the topic in the responding utterance, overt NP vs. zero-pronoun. We examined its effect on two dependent variables. The first dependent variable was the duration of the body. If anti-zero-pronominalization helps the speaker gain time for planning the following utterance, the utterance body would be more complex when the topic is expressed as an overt NP than it is expressed as a zero-pronoun. The complexity of the body is simply approximated by its duration.

The second dependent variable was the presence or absence of the preface. If anti-zero-pronominalization has a similar function to fillers and other prefaced items, overt topics would show a complementary distribution to these items. That is, preface would be less often used when the topic is expressed as an overt NP than it is expressed as a zero-pronoun.

We also investigated whether there is a correlation between the duration of the body and the duration of the topic when the topic is overtly expressed. Since the speaker can gain more time for speech planning when the overt topic becomes longer, the duration of the topic may be positively correlated with the complexity of the body.

All duration variables were log-transformed and standardized before statistical analysis.

2.4. Statistical analysis

To test the above predictions, we employed linear mixed-effects models (Gaussian and logistic models depending on the types of the dependent variables) with random intercept for speakers. We included the session type, i.e., Interview and Task, as a fixed effect. We used lme4 and languageR packages of the R language for model fitting and calculation of MCMC *p*-values [17].

3. Results

3.1 Duration of the body vs. expression type

Figure 2 shows the distribution of the duration of the body depending on the expression type and the session type. The statistical results revealed significant effects of both the expression type and the session type (Table 1); the duration of the body was significantly longer for overt topics than for zero topics and significantly shorter for task-oriented dialogs than interviews.

[1] Keeping these two subsets as independent session types did not affect the conclusions obtained in this paper; the tendency of each session was the same.

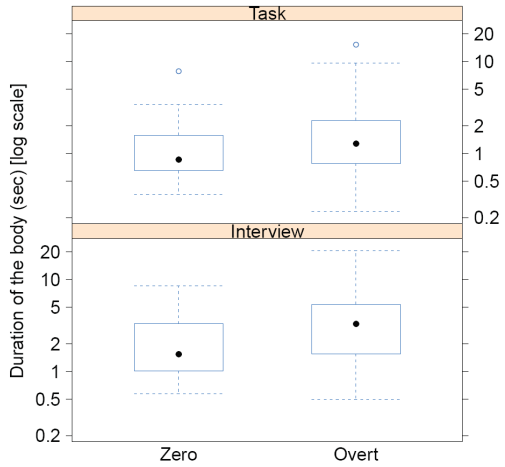


Figure 2: *Duration of the body relative to the expression type*

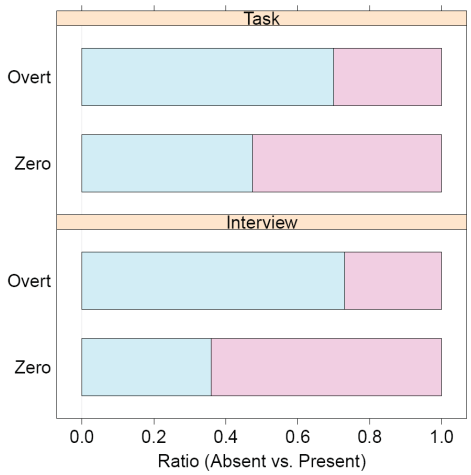


Figure 3: *Presence of the preface relative to the expression type*

Table 1: *Estimated parameters: Duration of the body vs. expression type. σs indicates the standard deviation at the speaker level and σ indicates the residual standard deviation.*

	Coef.	SE	t	MCMC p
(Intercept)	.191	.151	1.268	.237
ExpType:Overt	.426	.171	2.499	.016
SesType:Task	−.738	.181	−4.087	.000
				$\sigma_S = .171, \sigma = .907$

Table 2: *Estimated parameters: Presence of the preface vs. expression type. Note that this is a logistic ANOVA model.*

	Coef.	SE	z	p
(Intercept)	.475	.301	1.577	.115
ExpType:Overt	−1.294	.403	−3.207	.001
SesType:Task	−.268	.381	−.704	.481
				$\sigma_S = .000$

3.2. Presence of the preface vs. expression type

Figure 3 shows the ratio of the presence (right bars) vs. absence (left bars) of the preface depending on the expression type and the session type. The statistical results revealed only a significant effect of the expression type (Table 2); prefaces were significantly less often used for overt topics than for zero topics.

3.3. Duration of the body vs. duration of the topic

Figure 4 shows the scatter plot between the duration of the body and that of the topic depending on the session type, when the topic is expressed as an overt NP. The statistical results revealed a significant effect of the session type but no significant effect of the duration of the topic (Table 3); there was no reliable correlations between the duration of the body and that of the topic.

When we focus on the last mora of the topic, instead of the whole topic phrase, however, the situation dramatically changes. Figure 5 shows the scatter plot between the duration of the body and that of the *last mora* of the topic depending on the session type, when the topic is expressed as an overt NP. The statistical results revealed a significant effect of the duration of the last mora in addition to a significant effect of the session type (Table 4); the duration of the body became longer when the duration of the last mora of the topic became longer.

4. Discussion

So far we have shown the following.[2]

1. The duration of the body is longer when the topic is expressed as an overt NP than it is expressed as a zero-pronoun.
2. Preface is less often used when the topic is expressed as an overt NP than it is expressed as a zero-pronoun.
3. The duration of the body has no correlation with the duration of the whole topic phrase but has positive correlation with the duration of the *last mora* of the topic.

The first result suggests that anti-zero-pronominalization of the topic is a phenomenon closely related to the complexity of the utterance. When the speaker experiences a heavy cognitive load in producing an utterance, s/he tends to use an overt NP, instead of a zero-pronoun, to gain time for planning. In this respect, anti-zero-pronominalization has a similar function to fillers and other speech disfluencies.

The second result provides further support for this theory. Overt topics show a complementary distribution to fillers and other prefaced items, suggesting that the former can substitute for the latter, which has time-gaining function among others.

More interestingly, as shown in the third result, prolongation of the overt topic is enlarged when the substantial content of the utterance becomes complex. Watanabe and Den [13] found that the duration of the last mora of *wa*-marked topic phrases becomes longer as the complexity of the following clause increases, suggesting that the topic marker *wa* would be among time-gaining items. Our result is consistent with this, although about a quarter of the overt topics in our data did not end with *wa* or even lacked an overt topic marker. Thus, their result may be generalizable to topic phrases in general, not limited to *wa*-marked ones.

[2] In addition to these findings, we found that the duration of the body was consistently shorter in task-oriented dialogs than in interviews. This is natural considering that responses by interviewees are generally complex and tend to be long.

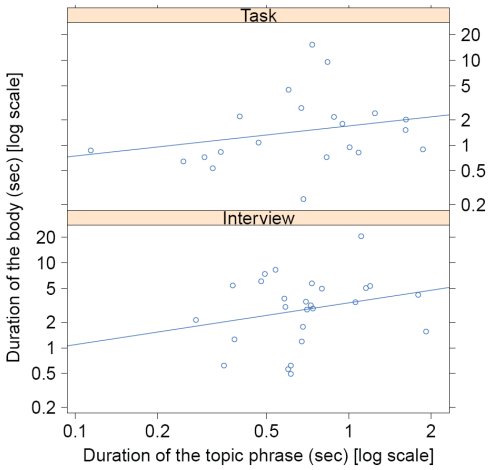


Figure 4: *Duration of the body as function of the duration of the topic. Regression lines are also shown.*

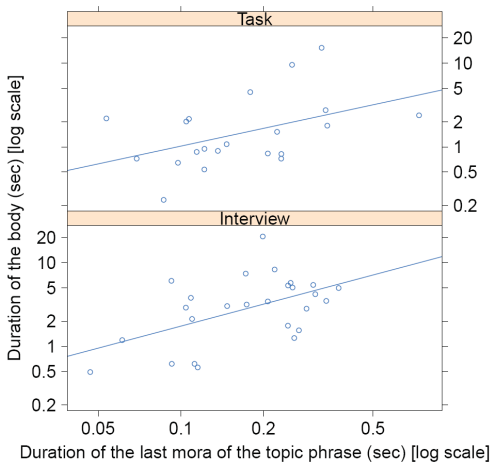


Figure 5: *Duration of the body as function of the duration of the last mora of the topic.*

Table 3. *Estimated parameters: Duration of the body vs. duration of the topic.*

	Coef.	SE	t	MCMC p
(Intercept)	.360	.242	1.486	.146
DurTopic	.223	.136	1.636	.109
SesType:Task	−.924	.341	−2.710	.015
				$\sigma_S = .372, \sigma = .879$

Table 4. *Estimated parameters: Duration of the body vs. duration of the last mora of the topic.*

	Coef.	SE	t	MCMC p
(Intercept)	.268	.166	1.610	.144
DurLastMora	.459	.127	3.626	.001
SesType:Task	−.616	.252	−2.442	.019
				$\sigma_S = .000, \sigma = .848$

As a line of studies have shown, Japanese has several options to deal with cognitive load at the beginning of an utterance, i.e., fillers [10], word repetitions [12], prolongation [13], and anti-zero-pronominalization (this paper). However, we do not know yet how speakers choose one, or more, option among these on a particular occasion. Investigation into factors behind this choice would be another step in speech disfluency research.

5. Acknowledgements

This work is supported by Grant-in-Aid for Scientific Research (B) "Constructing cognitive and communicative models for dialog based on utterance-unit annotations" led by Yasuharu Den.

6. References

[1] S. Kuno, *The structure of the Japanese language*. Cambridge, MA: MIT Press, 1973.

[2] P. M. Clancy, "Referential choice in English and Japanese narrative discourse," in *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*, W. Chafe, Ed. Norwood: Ablex, 1980, pp. 127–202.

[3] J. Hinds, "Topic continuity in Japanese," in *Topic continuity in discourse,* T. Givón, Ed. Amsterdam: John Benjamins, 1983, pp. 43–93.

[4] M. Kameyama, "Zero anaphora: The case of Japanese," Ph.D. dissertation, Stanford University, 1985.

[5] M. Walker, M. Iida, and S. Cote, "Japanese discourse and the process of centering," *Computational Linguistics*, vol. 20, no. 2, pp. 193–232, 1994.

[6] J. Fry, *Ellipsis and wa-marking in Japanese conversation*. New York: Routledge, 2003.

[7] T. Givón, Ed., *Topic continuity in Discourse*. Amsterdam: John Benjamins, 1983.

[8] J. K. Gundel, N. Hedberg, and R. Zacharski, "Cognitive status and the form of referring expressions in discourse," *Language*, vol. 69, no. 2, pp. 274–307, 1993.

[9] G. Beattie, "Planning units in spontaneous speech: Some evidence from hesitation in speech and speaker gaze direction in conversation," *Linguistics*, vol. 17, pp. 61–78, 1979.

[10] M. Watanabe, *Features and roles of filled pauses in speech communication: A corpus-based study of spontaneous speech*. Tokyo: Hituzi Syobo, 2009.

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

[12] Y. Den, "Are word repetitions really intended by the speaker?" in *Proc. of the ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*, Edinburgh, UK, 2001, pp. 25–28.

[13] M. Watanabe and Y. Den, "Utterance-initial elements in Japanese: A comparison among fillers, conjunctions, and topic phrases," in *Proc. of the DiSS-LPSS Joint Workshop 2010*, Tokyo, 2010, pp. 31–34.

[14] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003, pp. 7–12.

[15] H. Koiso, Y. Den, and K. Maekawa, "Nihongo-hanasikotoba-koopasu-RDB-no kootiku (Construction of relational database for the Corpus of Spontaneous Japanese)," in *Proc. of the 1st Workshop on Japanese Corpus Linguistics*, Tokyo, 2012, pp. 393–400.

[16] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara, "Identification of "sentences" in spontaneous Japanese: Detection and modification of clause boundaries," in *Proc. of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003, pp. 183–186.

[17] R. H. Baayen, Analyzing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press, 2008.

Self-repairs in German children's peer interaction – initial explorations

Laura E. de Ruiter

Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Abstract

Forty-nine self-repairs were extracted from a corpus of conversational speech of ten German children (mean age 5;1) with peers. The repairs were analysed using [1]'s classification and compared with his adult data. Children produced fewer appropriateness repairs than adults, but more covert repairs and more phonetic repairs. Like adults, children had a preference to interrupt themselves within-word only for error repairs. Unlike adults, children did not produce editing terms following interruptions.

1. Introduction

If there is one group of speakers that we may expect to be engaging in self-repair of their speech, it is young children. They are, after all, still in the process of learning language and more error-prone than adult speakers. Yet while self-repair in adult speech has attracted the interest of researchers since the 1970s – both from conversation-analytical and psycho-linguistic perspectives – children's use of self-repair has been studied to a much lesser extent.

Research in this area has mostly been concerned with disfluencies in general and has often been clinically motivated, using speech elicitation techniques such as imitation or modelled production [e.g., 2, 3] in order to find diagnostics for children at risk of developing speech impairments. However, in order to gain a better understanding how (typically developing) children monitor and repair their speech in real life, and to what extent their repairs pattern with those of adults, it would appear more useful to look at productions in more natural environments.

Recently, [4, 5] analysed children's speech when talking with their caretakers (note, however, that only active declarative sentences were analysed). [4, 5] categorized children's disfluencies as either stalls or revisions. "Stall" pertains to "all sentences disruptions that add no new phonological, lexical or grammatical material to a sentence" [4:819]. In contrast, "revisions" are changes in phonology, lexical choice or morphosyntax (ibid.) Analysing stall and revision patterns of children between 1;10 and 4;0, they found that the revision rate increased with age, while stall rate did not. Based on this contrastive pattern, [4, 5] argue that stalls and revisions are indications of two different phenomena. Stalls are assumed to be the result of incremental sentence production, where higher levels of processing (e.g., formulating) pass their results on to lower levels (e.g., articulation). When a speaker has already begun to speak, a "glitch" [4:820] at a higher level can force him or her to stall. Revisions, on the other hand, are corrections that are made because an error has been detected in overtly produced speech. It is assumed that certain deviations from the intended message can only be detected (and corrected) when speakers themselves hear them.

While the stall/revision dichotomy is an interesting framework for studying developments in sentence production, summarizing all types of overt repairs as revisions does lead to loss of some detail that may be informative about differences or similarities between adults' and children's speech production. For adult speakers, [1] distinguished among others between appropriateness repairs (A-repairs), such as the replacement of a more general term with a more specific one (e.g., *book – novel*) and error repairs (E-repairs), such as corrections of mispronunciations (e.g., *drim – drink*). He observed that speakers are more likely to interrupt their speech within a word when the word is erroneous than when it is merely less appropriate. [1] suggested that this has pragmatic reasons: "it is all right to interrupt a word which needs total replacement because it is erroneous, but it is not good practice to interrupt a correct word which only needs further specification" [1:63]. An interesting question for child language researchers is whether children do show the same tendency.

What is more, the stalls/revisions framework does not look at the distribution of so-called editing terms (or fillers) such as "uh" or "uhm"; utterances containing such terms are subsumed under revisions, along with repetitions (which may or may not contain editing terms themselves). But for adult speech, editing terms have been argued to be flags for upcoming delays [6], which also benefit the listener's speech comprehension [e.g., 7]. (Note, however, that fillers are not exclusively used for this purpose, see [8].) If fillers/editing terms play an important role in adult conversation, learning to use them in appropriate contexts is also something children need to do in order to become competent speakers of their language.

Against this background, this study asks the following research questions:

1. What types of self-repairs do children produce in spontaneous peer-to-peer conversations?

2. Do children's self-repairs differ in structure or distribution from those of adult speakers? If so, in what way?

In order to answer these questions, I analysed self-repairs in German children's peer-interaction using [1]'s classification of repairs and compared them with the adult data reported by [1]. The data presented here is a subset of data that is currently being collected for an on-going project on children's acquisition of dialogue competence.

2. Method

2.1. Participants

The participants were ten German children (five boys, five girls) between 4;10 and 5;9 years (mean age: 5;1) who had been recruited through local nursery schools. All children were typically developing and had no reported speech/language impairments or other developmental deficits. Informed written consent was obtained from the caregivers.

2.2. Procedure

The children were recorded during free (i.e., unsupervised) activities at their nurseries, such as drawing or playing with building blocks and toy animals. The recordings were made in HDV 1080i format using a Canon XH G1s camera and a Sennheiser ME 80 microphone. The microphone was attached to a boom pole to capture the conversations from nearby and avoid recording too much ambient noise. Post-production was done using Final Cut pro 5 on a Power Mac G5 computer. The total number of analysed recordings was about 50 minutes long.

2.3. Transcription

The data were transcribed orthographically in [9], using modified GAT transcription rules [10].

2.4. Data selection and coding

Following [1] a stretch of speech was considered a self-repair if there was a hesitation or a pause (either with or without an editing term) that was followed by a repetition of parts of the original utterance (OU) or an altered version of the OU. Unfilled pauses that were not followed by a repetition or a replacement were not categorized as repairs. Repetitions that were used to hold the floor were not coded as repairs.[1] All 49 repairs were coded to be one of nine categories used in [1][2] (repairs are underlined):

(1) **AA-repairs**: appropriateness repairs that are intended to reduce potential ambiguity.

We	beginnen	in	het	midden	met...
You	*start*	*in*	*the*	*middle*	*with...*

in	het	midden	van	het	papier
in	*the*	*middle*	*of*	*the*	*paper*

(2) **AL-repairs**: appropriateness repairs used by speakers to change the level of reference, often to be more specific.

... met	een	blauw	vlakje,	een	blauw	rondje ...
... with	*a*	*blue*	*spot,*	*a*	*blue*	*disc ...*

(3) **AC-repairs:** appropriateness repairs used by speakers to make their utterance more coherent with the previous text.

Du	kannst	diese	Ladung	hier	da	reinkippen
You	*can*	*this*	*load*	*here*	*there*	*pour-in*

(4) **EF-repairs**: phonetic repairs.

Guck	mal	wie	weine	wenig...
Look		*how*	*lett*	*little...*

(5) **EL-repairs**: repairs of lexical errors (word substitutions).

Wann	wo	ist	die...
When	*where*	*is*	*the*

(6) **ES-repairs**: syntactic repairs.

Da	halt	werden	(...)	gehalten
There	*hold*	*are being*	*(...)*	*held*

(7) **C-repairs**: covert repairs, which consist of either just an interruption and an editing term or the repetition of one or more lexical items.

Weil	ich	weiß	ich	weiß...
Because	*I*	*know*	*I*	*know...*

(8) **D-repairs**: the speaker replaces the current message with a different one, often changing the linearization of events.

Der	fliegt	dann	der	rutscht	dann
He	*flies*	*then*	*he*	*slides*	*then*

und	dann	fliegt	er
and	*then*	*flies*	*he*

(9) **R-repairs**: repairs that were too confused to be assigned to one of the other categories (rest).

In addition, all repairs (with the exception of C-repairs, see below) were coded for:

- whether the repair was immediate or delayed (i.e. whether the repair occurred within or right after the to-be-repaired item – the *reparandum* – , or only several syllables later),
- the length of the delay (in syllables),
- whether the interruption occurred within a word or after a word,
- whether the retracing was immediate or anticipatory (i.e., whether parts of the OU preceding the repaired word were repeated),
- the length of the retracing span (in syllables),
- whether an editing term was used, and
- the position of the repair in the conversation as defined by [11] (i.e., in the same turn, in the transition space between turns or in the next turn).

As the *reparandum* of a C-repair is unclear, it is not possible to discern whether the repair was immediate, what the length of the delay is, whether retracing was immediate or what the length of the retracing span is. The same holds for D-repairs and R-repairs.

3. Results

3.1. Repair types

There were 49 self-repairs in the data. The absolute and relative frequencies of the different types are shown in Table 1.

[1] This refers to only one situation where a child was being interrupted by another child and then repeated one word several times in a loud voice until the other child had stopped talking.

[2] Examples (1) and (2) are taken from [1] as they did not occur in the child data. All other examples are taken from the corpus.

Table 1: *Absolute (N) and relative (%) frequencies of repair types.*

Repair type	N	%
AA	0	0
AL	0	0
AC	5	10.2
EF	6	12.2
EL	14	28.6
ES	1	2.0
C	18	36.7
D	3	6.1
R	2	4.1
Total	**49**	**100.0**

The most frequent repair type (36.7%) was C-repair. All E-repairs together account for 42.8%, with EL-repairs being the most frequent among them (28.6% of all repairs). Only 10.2% of all repairs were A-repairs, and children did not produce any AA- or AL-repairs at all. D-repairs and R-repairs were rather infrequent (6.1 and 4.1%, respectively).

Figure 1 shows the relative frequencies in comparison with the adult data reported by [1].

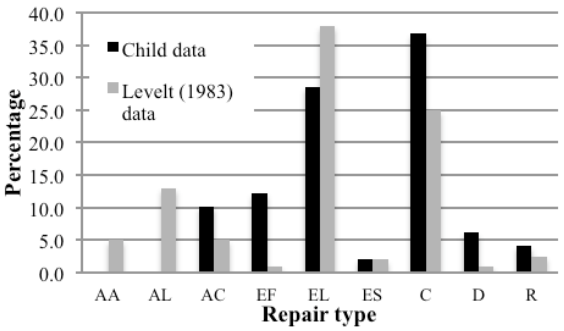


Figure 1: *Relative frequencies of repair types for child data and adult data reported by [1].*

The overall distributions are comparable in several respects. Syntactic repairs, for example, are rare in both the adult and the child data, and E-repairs are made more often than A-repairs in both groups. But three main differences can be observed: First, while the most frequent repairs in adults were lexical repairs (EL-repairs), children produced mainly covert repairs (C-repairs). Second, children corrected phonetic errors (EF-repairs) more often than adults. Third, ambiguity reducing (AA) and level (of reference) changing (AL) appropriateness repairs did not occur at all in the child data. I will come back to these differences in the Discussion.

3.2. Site of interruption

All but one of the 49 self-repairs occurred within the speaker's turn. In one case, it could be argued that the repair was done in the transition space, but given that the pause between OU/interruption and repair was exceptionally long (6.05 seconds), this item is somewhat difficult to classify. In any case, all self-repairs were also self-initiated.

In what follows, only the 26 A- and E-repairs will be considered, following the analysis in [1]. (Recall that sites of interruption cannot be determined for the other repair types.) Of those 26 repairs, 12 occurred within-word. This means that speakers aborted right in the middle of a word. In 14 cases, a repair occurred only after a word had been completed.

Were these disruptions produced immediately or with a delay? Note that these two variables are independent of each other: A within-word disruption may occur either within the *reparandum* itself or right after it (immediate repair), or only after the speaker has uttered other syllables following the *reparandum* (delayed repair). Likewise, a repair may be initiated after a word has been completed, be it the "trouble word" (immediate repair) or a following word (delayed repair). It turned out that only four out of 26 repairs were delayed, all other 22 repairs were immediate. The longest delay observed was three syllables.

How were these disruptions distributed across A- and E-repairs? The absolute numbers are given in Table 2.

Table 2: *Within-word and after-word interruptions (absolute frequencies) for appropriateness and error repairs, grouped by immediate vs. delayed repair.*

	Immediate		Delayed		Total
	Within-word	After-word	Within-word	After-word	
A-repairs	0	3	0	2	5
E-repairs	11	8	1	1	21
Total	11	11	1	3	**26**

It becomes clear that A-repairs never occurred within-word. In other words, children never interrupted themselves halfway into a word when they were making an appropriateness repair. In contrast, E-repairs occurred both within-word and after-word, with within-word interruptions slightly more often (for immediate repairs). The relative distribution is similar to that observed by [1]. I will return to these observations in the Discussion.

3.3. Editing terms

Only in two cases did children produce an editing term: one time "ähm" was used, in the other case "äh". Both editing terms occurred in C-repairs. The first case is provided in (10).

(10) ja weil das (.) ähm (--) das is nur der Boden.
yes because this (.) uhm (--) this is only the ground.

The near-absence of editing terms in the data is striking and will be returned to in the Discussion.

3.4. Restarting

When children restarted after an interruption, how far did they go back in the OU? (For reasons explained above, the analysis is again limited to A- and E-repairs.) The large majority of restarts were instant replacements (20). This means that the children did not retrace further than the *reparandum* itself. Note that instant replacements do not have to be immediate repairs. A speaker may initiate repair with a delay, but not retrace to an earlier word in the OU. In five cases children did retrace, but the retracing span was mostly one syllable only. Once a child made what [1] called a *fresh start*: she copied parts of the OU, but they were preceded by new material.

4. Discussion

When conversing with same-aged peers, five-year-old children monitor and repair their speech for (real or subjectively perceived) errors. Their behaviour is in many ways already similar to that of adults, but there are also differences that show that development in this area is not yet complete.

Like adults, children are more likely to repair real errors (E-repairs) than to make appropriateness repairs (A-repairs).

Children typically interrupted themselves within-word only when the word was erroneous, not when it was merely less appropriate. This is what [1] found for adult speakers. [1] explained this preference on pragmatic grounds, and if this was the case, the child data would suggest that children understand the difference between those two and their respective relevance for the discourse. But another explanation is possible as well: Perhaps detecting inappropriateness involves higher-level processes and simply takes more time than detecting real grammatical errors. In that case, the similarity of children and adults just reflects the way speech comprehension works in both groups.

Children did, however, produce fewer A-repairs overall than adults, and produced only one type of them. The higher incidence of AA- and AL-repairs in [1]'s data could have been task-induced (describing visual patterns to someone who cannot see them). Alternatively, children may be paying less attention to monitoring common ground (CG) and therefore notice fewer words that could be improved for the benefit of the interlocutor. For adults, it has been suggested that CG is not part of the initial utterance design and that monitoring one's utterances for CG violations is resource-dependent [12]. Given that children's language and memory capacities are still developing, these resources may not be available to them at all times. (This need not have a detrimental effect on their conversation, though, as ambiguity of reference accounts only for a fraction of misunderstandings, see [13].)

The finding that children produced relatively more phonetic repairs than adults (EF-repairs) is perhaps least surprising. Around the age of five, phonological development is not yet completed and articulation – even of function words – is less automatized than in adults [14].

A further difference between children and adults emerged in the proportion of C- and EL-repairs. Children produced more C-repairs, while adults produced more lexical replacements. Again, the reason for the dominance of EL-repairs in the adult data could be task-related: [1]'s subjects had to use many different colour terms and confused them frequently. Still, the high proportion of C-repairs in children is interesting, although the nature of C-repairs makes them difficult to analyse. [4] assumes that C-repairs (*stalls* in his terminology) are simply used to buying time during the course of sentence production. Another possibility is that – due to the still developing grammar – children's parser is less reliable in checking for grammaticality, and may produce more "false alarms" (on either inner speech or overt speech).

A striking feature of the children's repairs was the near-absence of editing terms. In contrast, [1] found that in adult speech, 58% of all repairs were preceded by an editing term. Relative dearth of filled pauses has also been observed in a picture story telling task with five-year-olds [15, 16], suggesting that this is a characteristic speech feature of that age. It seems that children have not yet learned to signal delays to their interlocutor – it would be interesting to study in more detail what effect this may have on their turn-management (with peers and with adults).

Finally, children's repairs were predominantly instant repairs, while adults have been found to retrace further in the OU and produce more fresh starts. This variance may be attributable to differences in short-term memory.

The findings presented here are based on a limited data set and invite further investigations, such as:

- When do children start approaching adult behaviour in the use of editing terms?
- Do children engage in more appropriateness repairs with increasing age?
- What is the influence of social context – does it matter to their self-repair behaviour whether children talk to same-aged peers or to adults?

By enlarging our database of conversational speech – from children of different ages and in different social contexts – we hope to find answers to some of these questions.

5. Acknowledgements

Thanks to Robert Eickhaus and Hüyesin Demir for helping collect the data. Thanks to all children, parents and nursery staff who supported this project with their participation.

6. References

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

[2] N. B. Ratner and C. C. Sih, "Effects of gradual increases in sentence length and complexity on children's dysfluency", *Journal of Speech and Hearing Disorders* 52(3), pp. 278–287, 1987.

[3] P. A. Gordon, H. L. Luper, and H. A. Peterson, "The effects of syntactic complexity on the occurrence of disfluencies in 5 year old nonstutterers", *Journal of Fluency Disorders* 11(2), pp. 151–164, 1986.

[4] M. Rispoli, "Changes in the nature of sentence production during the period of grammatical development", *Journal of Speech, Language and Hearing Research* 46(4), pp. 818–830, 2003.

[5] M. Rispoli, P. Hadley, and J. Holt, "Stalls and revisions: A developmental perspective on sentence production", *Journal of Speech, Language and Hearing Research* 51(4), pp. 953–966, 2008.

[6] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking", *Cognition* 84, pp. 73–111, 2002.

[7] J. E. Arnold, C. L. H. Kam, and M. K. Tanenhaus, "If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference resolution", *Journal of Experimental Psychology-Learning Memory and Cognition* 33(5), pp. 914–930, 2007.

[8] E. A. Schegloff, "Some Other "Uh(m)" s", *Discourse Processes* 47(2), pp. 130–174, 2010.

[9] T. L. A. Max Planck Institute for Psycholinguistics, "ELAN - Linguistic Annotator" 4.5.0 ed. Nijmegen, The Netherlands.

[10] M. Selting, P. Auer, B. Barden, J. Bergmann, E. Couper-Kuhlen, S. Günthner*, et al.*, "Gesprachsanalytisches Transkriptionssystem (GAT)", *Linguistische Berichte,* pp. 91–122, 1998.

[11] E. A. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation", *Language* 53(2), pp. 361–382, 1977.

[12] W. S. Horton and B. Keysar, "When do speakers take into account common ground?", *Cognition* 59(1), pp. 91–117, 1996.

[13] E. A. Schegloff, "Some sources of misunderstanding in talk-in-interaction", *Linguistics* 25(1), pp. 201–218, 1987.

[14] F. Wijnen, "Incidental word and sound errors in young speakers", J*ournal of Memory and Language* 31(6), pp. 734–755, 1992.

[15] L. E. De Ruiter, "Studies on intonation and information structure in child and adult German", PhD Thesis, Radboud University, Nijmegen, 2010.

[16] L. E. De Ruiter, "How German children signal information status in narrative discourse", accepted for publication.

Self-addressed questions in disfluencies

Jonathan Ginzburg[1], Raquel Fernández[2] & David Schlangen[3]

[1] Laboratoire de Linguistique Formelle (LLF) and CLILLAC-ARP and LabEx-EFL, Université Paris-Diderot, Sorbonne Paris Cité
[2] Institute for Logic, Language & Computation, University of Amsterdam
[3] Faculty of Linguistics and Literary Studies, Bielefeld University

Abstract

The paper considers self-addressed queries – queries speakers address to themselves in the aftermath of a filled pause. We study their distribution in the BNC and show that such queries show signs of sensitivity to the syntactic/semantic type of the sub-utterance they follow. We offer a formal model that explains the coherence of such queries.

1. Introduction

How to characterize the context associated with hesitations? For production, [3] claimed that fillers like 'uh' and 'um' should be treated as words with different distributions ('uh' more for short pauses,' um' for long pauses) and with discourse functions intended by the speaker. ([10] proposes a related account for Swedish glottalized filled pauses. Clark and Fox-Tree's hypothesis has more recently been strongly disputed for distributional differences [12] as well as with respect to speakers' intentions [4].

In this paper we consider a phenomenon that occurs in the aftermath of a filled pause, namely self-addressed queries, exemplified in ((1)):

(1) a. Carol 133 Well it's (pause) it's (pause) er (pause) what's his name? Bernard Matthews' turkey roast. (BNC, KBJ)

b. They're pretty ... um, how can I describe the Finns? They're quite an unusual crowd actually.

http://www.guardian.co.uk/sport/2010/sep/10/small-talk-steve-backley-interview

The question we investigate is whether such queries are essentially reflexive or show signs of sensitivity to the (the syntactic/semantic type of) the sub-utterance they follow.

Section 2 describes a corpus study we ran on the BNC to investigate this issue. The study demonstrates clearly a strong effect, with distinct distributions clustering around a small number of triggering contexts. After brief discussion of the results in section 3, section 4 provides a formal model in which we analyze the coherence of such moves, as part of an account of what we call *forwards-looking disfluences* – disfluencies where the moment of interruption is followed by a completion of the utterance which is delayed by a filled or unfilled pause (hesitation) or a repetition of a previously uttered part of the utterance (repetitions).

Conclusions and further work are provided in section 5.

2. Corpus study

We ran a corpus study on the BNC, using the search engine SCoRE ([14]) to search for all self-addressed queries. We searched using the pattern 'noun preceding 'er' or 'erm' preceding a wh word, adjacent to a verb.'. This yielded 692 hits, from this we manually selected all self-addressed queries, resulting in a corpus of 83 queries.

Representative examples are in (2) and the distribution is summarized in Table 1. Tables 2–6 provide a detailed summary of queries found, relative to triggering

(2) a. (**anticipating an N'**:) on top of the erm (pause) what do you call it?

b. (**anticipating a locative NP**:) No, we went out on Sat , er Sunday to erm (pause) where did we go?

c. (**anticipating an NP complement**:) He can't get any money se (pause) so so he can't get erm (pause) what do you call it?

d. (**anticipating a person–denoting NP**:) But you see somebody I think it was erm what's his name?

e. (**anticipating a person–denoting NP**: with erm, who was it who went bust?

f. (**anticipating a predicative phrase**: she's erm (pause) what is she, Indian or something?

Table 1: *Distribution of Self addressed questions in disfluencies in the British National Corpus*

categorial context	questions found
pre NP; prer _or verb _or NP and _	42
det _	20
locative prep _	12
be _	5
say _	4
Total self addressed questions	**83**

Table 2: *Distribution of Self addressed questions in pre NP context*

what's his/her name?	19
what do they/you call him/her/it?	13
who was it/the woman?	3
what's the other one?	3
what did you/I say?	2
what did it mention?	2
Total	**42**

Table 3*: Distribution of Self addressed questions in post Det context*

what do/did they/you call it/that/them?	14
what's it called?	2
what is it?	3
what am I looking for?	1
Total	**20**

Table 4: *Distribution of Self addressed questions in post Loc Prep context*

where is it?	3
where do they call that?	2
what's the name of the street/address?	2
what do they call X?	2
where do we go?	1
where did it say now?	1
what is it?	1
Total	**12**

Table 5: *Distribution of Self addressed questions in post-copular context.*

what is she/it?	3
what's the word I want?	1
what do you call it?	1
Total	**5**

Table 6: Distribution of Self addressed questions in post 'say' context

what did X say?	3
where did I get the number?	1
Total	**4**

3. Discussion

Table 1 indicates that self-addressed queries occur in a highly restricted set of contexts, above all where an NP is anticipated and after 'the'. Moreover, the distribution of such queries across these contexts varies manifestly: the anticipated NP contexts involve predominantly a search for a name or for how the person/thing is called with some 'who'-questions as well, whereas the post 'the' contexts only allow 'what' questions, predominantly of the form 'what does X call Y'; anticipated location NP contexts predominantly involve 'where' questions. The final two classes identified are somewhat smaller, so generalizations there are less robust – nonetheless, the anticipated predicative phrase and post 'say' context involve seem to involve quite distinct distributions from the other classes mentioned above.

4. Forward Looking Dysflencies in a dialogue model

4.1. Dialogue GameBoards

We start by providing background on the dialogue framework we use here, namely KoS (see e.g. [9, 8]). On the approach developed in KoS, there is actually no single context – instead of a single context, analysis is formulated at a level of information states, one per conversational participant. The dialogue gameboard represents information that arises from publicized interactions. Its structure is given in (3) – the *spkr,addr* fields allow one to track turn ownership, *Facts* represents conversationally shared assumptions, *Pending* and *Moves* represent respectively moves that are in the process of/have been grounded, *QUD* tracks the questions currently under discussion, though not simply questions qua semantic objects, but pairs of entities which we call *InfoStrucs*: a question and an antecedent sub-utterance.[1] This latter entity provides a partial specification of the focal (sub)utterance, and hence it is dubbed the *focus establishing constituent* (FEC) (cf. *parallel element* in higher order unification-based approaches to ellipsis resolution e.g. [6].)[2]

(3) DGBType$=_{def}$

$$\begin{bmatrix} \text{spkr: Ind} \\ \text{addr: Ind} \\ \text{utt-time : Time} \\ \text{c-utt : addressing(spkr,addr,utt-time)} \\ \text{Facts : Set(Proposition)} \\ \text{Pending : list(locutionary Proposition)} \\ \text{Moves : list(locutionary Proposition)} \\ \text{QUD : poset(Infostruc)} \end{bmatrix}$$

The basic units of change are mappings between dialogue gameboards that specify how one gameboard configuration can be modified into another on the basis of dialogue moves. We call a mapping between DGB types a *conversational rule*. The types specifying its domain and its range we dub, respectively, the *preconditions* and the *effects*, both of which are supertypes of DGBType.

Examples of such rules, needed to analyze querying and assertion interaction are given in (4). Rule (4-a) says that given a question q and ASK(A,B,q) being the LatestMove, one can update QUD with q as QUD–maximal. QSPEC is what characterizes the contextual background of reactive queries and assertions. (4-b) says that if q is QUD–maximal, then subsequent to this either conversational participant may make a move constrained to be q–specific (i.e. either About or Influencing q).[3]

(4) a. Ask QUD–incrementation

$$\begin{bmatrix} \text{pre} & : & \begin{bmatrix} \text{I} & : & \text{InfoStruc} \\ \text{LatestMove} = \text{Ask(spkr,addr,I.q)} & : & \text{IllocProp} \end{bmatrix} \\ \text{effects} & : & \left[\text{qud} = \langle \text{I.q,pre.qud} \rangle \ : \ \text{poset(InfoStruc)} \right] \end{bmatrix}$$

---

1 Extensive motivation for this can be found in [5, 8], based primarily on semantic and syntactic paralleism in non-sentential utterances such as short answers, sluicing, and various other fragments.

2 Thus, the FEC in the QUD associated with a wh-query will be the wh-phrase utterance, the FEC in the QUD emerging from a quantificational utterance will be the QNP utterance, whereas the FEC in a QUD accommodated in a clarification context will be the sub-utterance under clarification.

3 We notate the underspecification of the turn holder as 'TurnUnderspec', an abbreviation for the following specification which gets unified together with the rest of the rule:

$$\begin{bmatrix} \text{PrevAud} = \{\text{pre.spkr,pre.addr}\} & : & \text{Set(Ind)} \\ \text{spkr} & : & \text{Ind} \\ \text{c1} & : & \text{member(spkr, PrevAud)} \\ \text{addr} & : & \text{Ind} \\ \text{c2} & : & \text{member(addr, PrevAud)} \\ & & \wedge\ \text{addr} \neq \text{spkr} \end{bmatrix}$$

(4) b. QSpec

$$\begin{bmatrix} \text{pre: } \left[\text{qud} = \langle \text{i, I} \rangle \text{: poset(InfoStruc)}\right] \\ \text{effects : TurnUnderspec} \\ \wedge_{merge} \begin{bmatrix} \text{r : AbSemObj} \\ \text{R: IllocRel} \\ \text{LatestMove = R(spkr,addr,r) : IllocProp} \\ \text{c1 : Qspecific(r,i.q)} \end{bmatrix} \end{bmatrix}$$

4.2. Forwards-looking disfluencies and self-addressed queries

Our starting point is the account developed within the KoS framework for Clarification Requests (see e.g. [13, 7]): in the aftermath of an utterance *u* a variety of questions concerning *u* and definable from *u* and its grammatical type become available to the addressee of the utterance. These questions regulate the subject matter and ellipsis potential of CRs concerning *u* and generally have a short lifespan in context. We argue that disfluencies can and should be subsumed within a similar account, a point that goes back to [16]: in both cases (i) material is presented publicly, (ii) a problem with some of the material is detected and signalled (= there is a 'moment of interruption'); (iii) the problem is addressed and repaired leaving (iv) the incriminated material with a special status, but within the discourse context. Concretely for disfluencies – as the utterance unfolds incrementally questions can be pushed on to QUD about what has happened so far, as with Backwards Looking Disfluencies (BLDs) (e.g. *what did the speaker mean with subutterance u1?*) or what is still to come, as with Forwards Looking Disfluencies (FLDs) (e.g. *what word does the speaker mean to utter after sub-utterance u2?*).

We specify FLDs with the update rule in (5) – given a context where the LatestMove is a forward looking editing phrase by A, the next speaker – underspecified between the current one and the addressee – may address the issue of what A intended to say next by providing a co-propositional utterance:[4,5]

(5) Forward Looking Utterance Rule:

$$\begin{bmatrix} \text{preconds : } \begin{bmatrix} \text{spkr : Ind} \\ \text{addr : Ind} \\ \text{pending} = \langle \text{p0,rest} \rangle \text{: list(LocProp)} \\ \text{u0 : LocProp} \\ \text{c1: member(u0, p0.sit.constits)} \\ \text{LatestMove}^{content} = \text{FLDEdit(spkr,u0) : IllocProp} \end{bmatrix} \\ \text{effects : TurnUnderspec} \wedge_{merge} \\ \begin{bmatrix} \text{MaxQud} = \\ \begin{bmatrix} \text{q} = \lambda x \text{ MeanNextUtt(pre.spkr,pre.u0,x)} \\ \text{fec} = \{\} \end{bmatrix} \text{: InfoStruc} \\ \text{LatestMove : LocProp} \\ \text{c2: Copropositional(LatestMove}^{content}\text{,MaxQud)} \end{bmatrix} \end{bmatrix}$$

Rule (5) differs from its BLD analogue, in two ways. First, in that the preconditions involves the LatestMove having as its content what we describe as an FLDEdit move, which we elucidate somewhat shortly. Words like 'uh', 'thee' will be assumed to have such a force, hence the utterance of such a word is a prerequisite for an FLD. A second difference concerns parallelism: for BLDs it is intuitive that parallelism exists between reparandum and alteration (with certain caveats), given that one is replacing one sub-utterance with another that is essentially of the same type. However, for FLDs there is no such intuition—what is taking place is a search for the word after the reparandum, which has no reason to be parallel to the reparandum. Hence in our rule (5), the FEC is specified as the empty set. To make this explicit, we assume that 'uh' could be analyzed by means of the lexical entry in (6):

(6)
$$\begin{bmatrix} \text{phon : uh} \\ \text{cat = } \textit{interjection} \text{ : syncat} \\ \text{dgb-params : } \begin{bmatrix} \text{spkr : IND} \\ \text{addr : IND} \\ \text{MaxPending : LocProp} \\ \text{u0 : LocProp} \\ \text{c1: member(u0, MaxPending.sit.constits)} \\ \text{rest : address(spkr,addr,MaxPending)} \end{bmatrix} \\ \text{cont} = \left[\text{c1 : FLDEdit(spkr,addr,MaxPending)}\right] \text{: Prop} \end{bmatrix}$$

We demonstrate how to analyze (7):

(7) A: Show flights arriving in uh Boston. [18]

After A utters u0= 'in', she interjects 'uh', thereby expressing FLDEdit(A,B,'in'). This triggers the Forward Looking Utterance rule with MaxQud.q = λx MeanNextUtt(A,'in',x). 'Boston' can then be interpreted as answering this question, with resolution based on the rule used to interpret (elliptical) short answers.

Similar analyses can be provided for (8). Here instead of 'uh' we have lengthened versions of 'the' and 'a' respectively, which express FLDEdit moves:

(8) a. And also the- the dog was old. [2]
b. A vertical line to a- to a black disk [11]

Let us return to consider what the predicate 'FLDEdit' amounts to from a semantic point of view. Intuitively, (9) should be understood as 'A wants to say something to B after u0, but is having difficulty (so this will take a bit of time)':

(9) FLDEdit(A,B,u0)

This means we could unpack (9) in a number of ways, most obviously by making explicit the utterance-to-be-produced *u*1, representing this roughly as in (10):

(10) $\exists u1[\text{After}(u1,u0) \wedge \text{Want}(A,\text{Utter}(A,B,u1))]$

Moving on finally to (dysfluent) self addressed queries of the kind described in section 2, on our account such queries are licensed because these questions are co-propositional with the issue 'what did A mean to say after u0'.

---

4 This rule is inspired in part by Purver's rule for *fillers*, (91), p. 92, ([15]). Given that our rule leaves the turn ownership unspecified we unify FLDs with fillers.

5 CoPropositionality for two questions means that, modulo their domain, the questions involve similar answers. For instance 'Whether Bo left', 'Who left', and 'Which student left' (assuming Bo is a student) are all co-propositional. In the current context co-propositionality amounts to: either a CR which differs from MaxQud at most in terms of its domain, or a correction – a proposition that instantiates MaxQud.

Self addressed queries also highlight another feature of KoS's dialogue semantics: the fact that a speaker can straightforwardly answer their own question, indeed in these cases the speaker is the "addressee" of the query. Such cases get handled easily in KoS because turn taking is abstracted away from querying: the conversational rule QSpec, introduced earlier as (4-b), allows either conversationalist to take the turn given the QUD-maximality of q. This contrasts with a view of querying derived from Speech Act Theory (e.g. [17]) still widely as- sumed (see e.g. [1]), where there is very tight link to intentional categories of 2-person dialogue ('. . . Speaker wants Hearer to provide an answer . . . Speaker does not know the answer . . . ').

5. Conclusions

In this paper we offer the first detailed corpus study of self-addressed queries that occur in the aftermath of filled pauses. We show that such queries show marked signs of sensitivity to the (the syntactic/semantic type of) the sub-utterance $u0$ they follow. We then offer a formal model from which the possibility for such queries follows directly. An obvious next step is to study differences between the distribution we found in the BNC and that occurring in other languages, as surface syntax in particular of NPs seems to be a significant factor, as does predicational structure.

6. Acknowledgements

This work has been partially funded by the Labex EFL (ANR/CGI).

7. References

[1] N. Asher and A. Lascarides, *Logics of Conversation*. Cambridge: Cambridge University Press, 2003.

[2] J. Besser and J. Alexandersson, "A comprehensive disfluency model for multi-party interaction", in Proceedings of SigDial 8, pp. 182–189, 2007.

[3] H. Clark and J. FoxTree, "Using uh and um in spontaneous speech", *Cognition*, vol. 84, pp. 73–111, 2002.

[4] M. Corley and O. W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of 'um'", *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[5] R. Fernández, "Non-sentential utterances in dialogue: Classification, resolution and use", Ph.D. dissertation, King's College, London, 2006.

[6] C. Gardent and M. Kohlhase, "Computing parallelism in discourse," in *Proc. IJCAI*, pp. 1016–1021, 1997.

[7] J. Ginzburg, "Situation semantics: from indexicality to metacommunicative interaction", in *The Handbook of Semantics*, K. von Heusinger, C. Maierborn, and P. Portner, Eds. Walter de Gruyter, 2011.

[8] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*. Oxford: Oxford University Press, 2012.

[9] J. Ginzburg and R. Fernández, "Computational models of dialogue", in *Handbook of Computational Linguistics and Natural Language*, A. Clark, C. Fox, and S. Lappin, Eds. Oxford: Blackwell, 2010.

[10] M. Horne, "Attitude reports in spontaneous dialogue: Uncertainty, politeness and filled pauses", in *From Quantification to Conversation*, L. Borin and S. Larsson, Eds. Gothenburg: Gothenburg University, pp. 309–318, 2008.

[11] W. J. Levelt, "Monitoring and self-repair in speech", *Cognition*, vol. 14, no. 4, pp. 41–104, 1983.

[12] D. C. O'Connell and S. Kowal, "Uh and Um Revisited: Are They Interjections for Signaling Delay?" *Journal of Psycholinguistic Research*, vol. 34, no. 6, pp. 555–576, 2005.

[13] M. Purver, "Clarie: Handling clarification requests in a dialogue system", *Research on Language & Computation*, vol. 4, no. 2, pp. 259–288, 2006.

[14] M. Purver, "Score: A tool for searching the bnc", King's College, London, Tech. Rep. TR-01-07, 2001.

[15] M. Purver, "The theory and use of clarification in dialogue", Ph.D. dissertation, King's College, London, 2004.

[16] E. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation", Language, vol. 53, pp. 361–382, 1977.

[17] J. Searle, *Speech Acts*. Cambridge: Cambridge University Press, 1969.

[18] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies", Ph.D. dissertation, University of California at Berkeley, Berkeley, USA, 1994.

Acoustic and linguistics features related to speech planning appearing at weak clause boundaries in Japanese monologs

Hanae Koiso[1] *& Yasuharu Den*[2]

[1] National Institute for Japanese Language and Linguistics, Japan
[2] Faculty of Letters, Chiba University, Japan

Abstract

In this paper, we focus on weak clause boundaries in Japanese monologs in order to investigate the relationship of the length of constituents following weak boundaries to three acoustic and linguistic features: 1) occurrence rate of fillers, 2) occurrence rate of boundary pitch movements, and 3) degree of lengthening of clause-final morae. We found that all these features were significantly correlated with the length of following constituents. Most importantly, boundary pitch movements had an additional effect that can be distinct from the effect of clause-final lengthening. These results suggest that Japanese speakers have *earlier-occurring* items that help them deal with cognitive load in speech planning, in addition to fillers and other clause-initial disfluencies.

Index Terms: fillers, boundary pitch movements, clause-final lengthening, Japanese monologs

1. Introduction

In spoken Japanese, successive clauses are frequently linked up, resulting in a long stretch of them in a sentence or utterance [1]. Since Japanese is a predicate-final language, this clause chaining consists of more than one finite or non-finite predicate clause, each followed by a conjunctive particle, and terminates at a clause with a finite predicate. Thus, Japanese speakers can formulate utterances with an arbitrary and in principle unlimited number of clauses by manipulating the final elements of these clauses.

Clauses in Japanese can be classified into those with strong boundaries and those with weak boundaries, according to their degree of dependency on the subsequent clauses. Watanabe [2] has found that the rate of fillers at strong clause boundaries is higher than the rate at weak clause boundaries. In addition, Koiso [3] has found that the boundary pitch movements (BPMs) of accentual phrases occur more frequently at strong clause boundaries than at weak clause boundaries.[1] These facts suggest that, in the course of clause chaining, fillers and BPMs are used at similar locations, that is, where speech planning mainly takes place. However, this does not necessarily mean that BPMs have a relation to planning difficulty as fillers do, since BPMs generally involve the lengthening of segments, which is itself considered to be related to planning difficulty [4].

There is another fact that may indicate the possibility that clauses with weak boundaries are also locations for speech planning. Watanabe [2] has reported a tendency that the longer the following clause, the more frequently fillers occur at weak clause boundaries, but not at strong clause boundaries. Also, Watanabe and Den [5] have found that the duration of a filler *e* is positively correlated with the length of the following clause at weak clause boundaries only. These results suggest that weak clause boundaries may also serve as locations for speech planning when the following clause is long and complex.

In this paper, we focus on weak boundaries within utterances and investigate how the complexity of the following constituents is related to three acoustic and linguistic features: fillers, clause-final lengthening, and BPMs, any of which may serve to gain time for speech planning. For this purpose, we conduct a quantitative analysis of a large-scale corpus of spontaneous Japanese monologs, and show the relation between the length of the following constituents and these features.

2. Method

2.1. Data

We used the *Corpus of Spontaneous Japanese* [6] in the present analysis. We selected 177 monologs in its "Core" data set (CSJ-Core), which consists of "Academic Presentation Speech" (APS, 70 monologs) and "Simulated Public Speech" (SPS, 107 monologs). APS consists of live recording of academic presentations covering meetings of scholars in engineering and the humanities and social sciences, while SPS consists of public speeches of about 10–12 minutes given by laypeople on everyday topics like "my most delightful memory" and "the town I live in" in front of small, friendly audience. Including both APS and SPS, there were 78 female and 99 male speakers, ranging in age from their early 20s to their late 60s.

2.2. Annotation

CSJ-Core contains a variety of hand-corrected annotations, including clause units, *bunsetsu* phrases (see below), long- and short-unit words, phonetic segments, dependency structures, and prosodic information. Clause boundaries are automatically detected by the CBAP-csj program [7] and classified into one of the following three categories based on their degree of completeness as a syntactic and semantic unit and on that of dependency on the subsequent clause:

Absolute Boundary (AB) corresponds to the sentence boundary in the usual sense.

Strong Boundary (SB) is the boundary of a clause that is relatively independent of the subsequent clause.

Weak Boundary (WB) is the boundary of a clause that is relatively dependent on the subsequent clause.

[1] Note, however, that weak boundaries defined in Watanabe [2] and Koiso [3] are different from each other; weak boundaries in Koiso are all located within utterances, whereas those in Watanabe correspond to utterance-final position approved by some pragmatic criteria.

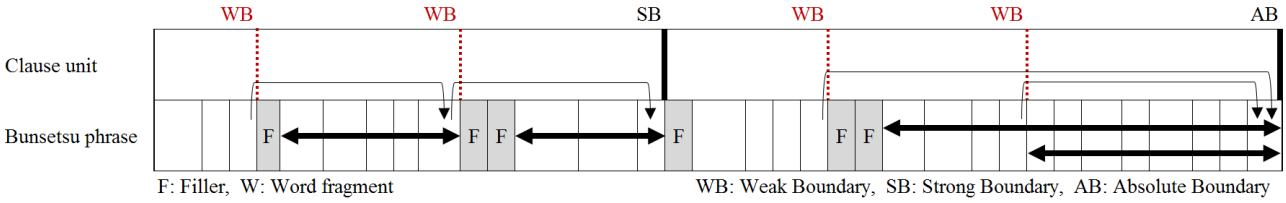


Figure 2: *The lengths of constituents following WBs. The arcs indicate dependency between the clauses ending with WBs and the clauses they modify. The length of the constituents was measured by the duration of the interval from the left edge of the phrase immediately following the WB in question to the right edge of the clause modified by the WB clause. Fillers (F) appearing at the beginning of this interval were excluded from this calculation.*

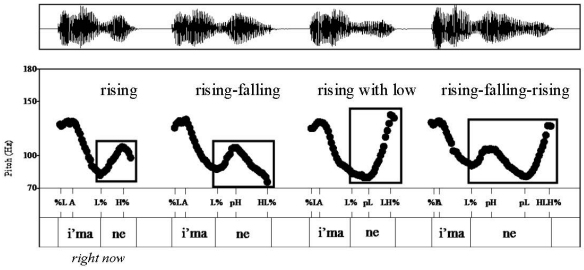


Figure 1: *Four main types of BPMs in Japanese: a simple rising tone (H%), a rising–falling tone (HL%), a rising tone with a sustained low (LH%), and a rising–falling–rising tone (HLH%). BPMs are indicated by squares. Note that BPM is always preceded by a low tone (L%) marking the right edge of an accentual phrase.*

On the basis of these categories, *clause units* are defined as clauses ending with either absolute or strong boundaries. The clause units identified were carefully checked by expert labelers. WBs located within these hand-corrected clause units are regarded as utterance-internal breaks – they are the target of the present study. Clause units are segmented in terms of *bunsetsu* phrases, which consist of one content word possibly followed by one or more function words. Dependency structures between *bunsetsu* phrases were also labeled, again by expert labelers.

In addition to these syntactic and morphological annotations, CSJ-Core was also annotated in term of prosody using the X-JToBI scheme [8]. Among the labels of X-JToBI, we focus on the final boundary tones of accentual phrases, which are a fundamental prosodic unit in Japanese. The right edge of an accentual phrase is always characterized by a low tone (L%), which can be followed by an additional tone such as a simple rising tone (H%), a rising–falling tone (HL%), a rising tone with a sustained low (LH%), or a rising–falling–rising tone (HLH%) (see Figure 1). These extra movements at the right edge of accentual phrases are called *boundary pitch movements* (BPMs).

2.3. Variables for stastistical analysis

The dependent variable of the analysis was the length of the constituent following a WB, which was taken to function as a rough estimate of the cognitive load imposed on speakers by speech planning at that boundary. The length of the constituent was measured by the duration between the start of the phrase immediately following the WB and the end of the clause modified by the WB clause (see Figure 2). Fillers appearing at the beginning of the interval were excluded from this calculation.

Three acoustic and linguistic features at WBs were considered as factors that may affect the dependent variable:

1. presence or absence of fillers that immediately follow the WB clause (Filler);
2. presence or absence of a BPM at the end of the WB clause (BPM); and
3. duration of the last mora of the WB clause (DurLastMora).

The duration variables, that is, the dependent variable and the third independent variable, were log-transformed and standardized before statistical analysis.

2.4. Target of the analysis

We excluded from the analysis clauses ending with WB that did not modify any other clause or that did not coincide with accentual phrase boundaries. Of 19,186 WBs in the data, 16,245 instances were retained for the analysis.

3. Results

The top row in Figure 3 shows the distribution of the duration of constituents following a WB depending on the presence or absence of following fillers and BPMs and the scatter plot between the duration of constituents and the duration of the last mora. The duration of constituents was longer in the presence both of following fillers and of BPMs, and also when the duration of the last mora of the WB clause lengthened.

Furthermore, when we restricted WB clauses to major subclasses, namely, causal clauses (1,390 instances), conditional clauses (1,613 instances), and *te*-marked clauses (5,520 instances), the same tendencies were replicated (see the second, third and fourth rows in Figure 3).

In order to statistically test the effects of the three variables, we applied linear mixed-effects models with random intercept for speakers. The (means of the) estimated parameters and the *p*-values were calculated using Markov Chain Monte Carlo (MCMC) sampling implemented in lme4 and languageR packages of the R language [9].

Table 1 shows the results. All three independent variables had significant effects on the duration of constituents following WBs. This result was consistent throughout the four data sets (that is, all WBs and causal, conditional, and *te*-marked subclasses.

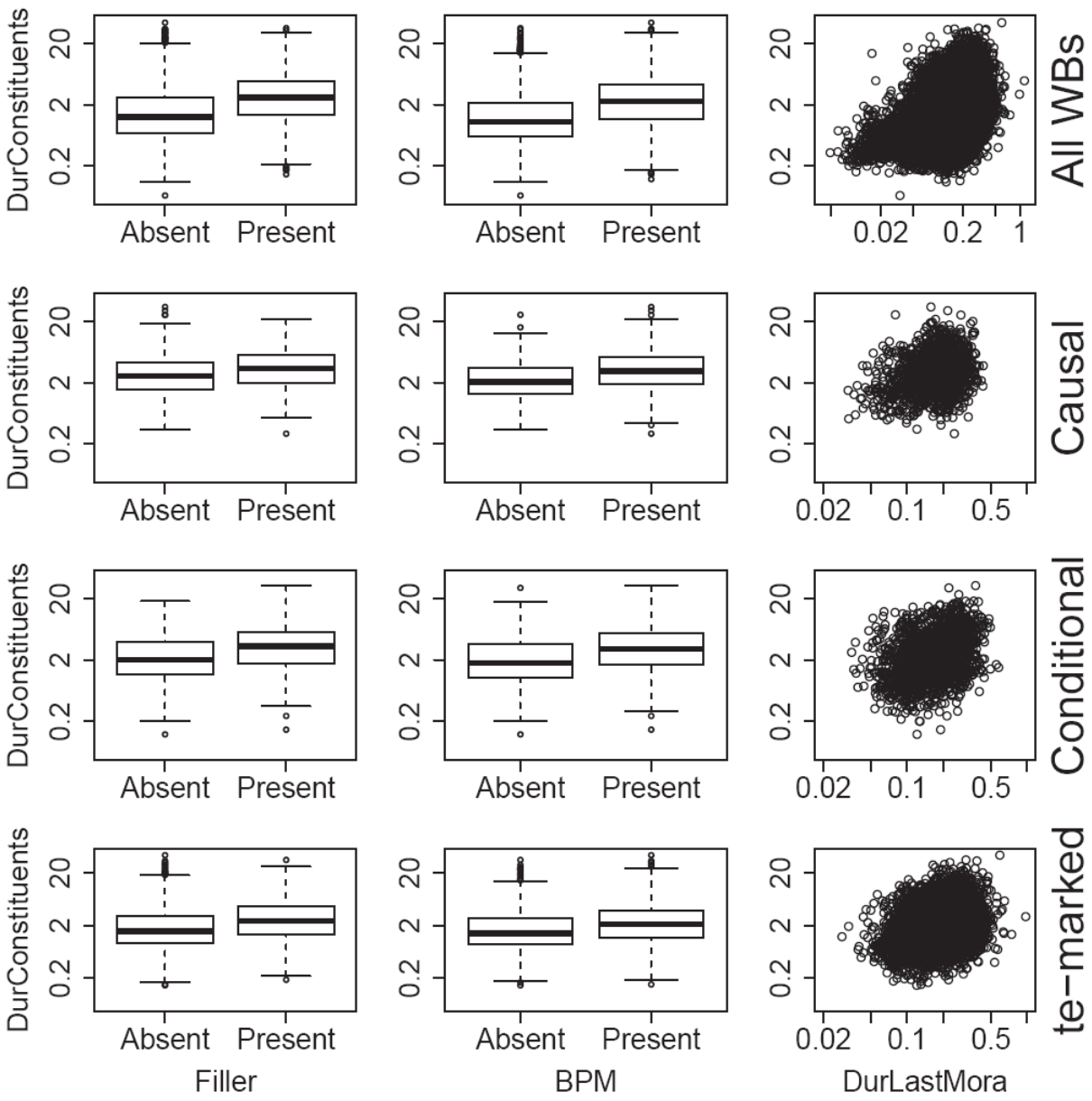


Figure 3: *The relations between the duration of constituents following WBs and the three independent variables (*`Filler`*, BPM, and DurLastMora). The top row shows the results for all WBs, and the other three rows show the results for causal, conditional, and te-marked subclasses, respectively.*

Table 1: *Estimated parameters of the linear mixed-effects models applied to WBs all together and to causal, conditional, and te-marked subclasses. σs indicates the standard deviation at the speaker level and σ indicates the residual standard deviation. The means and p values of the coefficients were obtained by using MCMC sampling.*

	Fixed-effect	MCMC Mean	MCMC p
All WBs	(Intercept)	−.200	.0001
	`Filler`	.333	.0001
	`BPM`	.293	.0001
	`DurLastMora`	.357	.0001
		σ_S = .112, σ = .861	

	Fixed-effect	MCMC Mean	MCMC p
Causal	(Intercept)	.370	.0001
	`Filler`	.199	.0001
	`BPM`	.192	.0002
	`DurLastMora`	.227	.0001
		σ_S = .119, σ = .725	

	Fixed-effect	MCMC Mean	MCMC p
Conditional	(Intercept)	.228	.0001
	`Filler`	.199	.0006
	`BPM`	.168	.0018
	`DurLastMora`	.308	.0001
		σ_S = .139, σ = .847	

	Fixed-effect	MCMC Mean	MCMC p
te-marked	(Intercept)	−.058	.0001
	`Filler`	.189	.0001
	`BPM`	.128	.0001
	`DurLastMora`	.336	.0001
		σ_S = .115, σ = .841	

4. Discussion

We found that all the three independent variables in this study had significant effects on the length of constituents following WBs. The duration of the constituents was longer when there were fillers immediately following than when there were not; duration also increased in the presence of BPMs at the boundaries; and finally, the duration of the constituents lengthened when the duration of the last mora of the WB clause did. These results suggest that fillers, BPMs, and clause-final lengthening are used when Japanese speakers are about to produce a long, complex clause and, hence, have an increased need to manage cognitive load in their speech planning.

The same results were replicated in all the analyses of the three major subclasses of the data, that is, in causal, conditional, and *te*-marked clauses. This means that these tendencies were widely and consistently observed across various types of WBs.

Fillers and word repetitions are reported to occur very frequently at utterance-initial positions, where they help speakers gain time for speech planning [10, 11]. In Japanese, these linguistic devices for speech planning have been found also to occur at clause-initial positions [12, 2]. Our result on fillers following WBs is consistent with these previous findings. However, the other two features found to be relevant to speech planning (that is, BPMs and clause-final lengthening) are not located at the beginning of a new clause but at the end of the preceding clause. In this sense, it can be stated that Japanese speakers have *earlier-occurring* items, as well as clause-initial ones, to deal with cognitive load in speech planning.

The result regarding BPM is the most surprising here. BPM generally involves the lengthening of the last mora of the clause, which bears the complex tones involved. Since this lengthening is considered in and of itself to be related to planning difficulty, it is tempting to infer that it may be the *duration*, and not the *presence,* of BPMs that is relevant to speech planning. However, this is not the case: *both* BPMs themselves and the lengthening of the morae bearing them have significant effects on the length of constituent after WBs. This suggests that the presence of BPMs *per se* has some relation to the complexity of the following constituents.

One possible interpretation of the correlation of BPMs to constituent complexity is that BPMs are in fact related to the continuity of utterances and that the continuity in turn is related to the complexity. As Koori [13] pointed out, rising tones in Japanese tend to occur at syntactically and semantically deep boundaries, and indicate continuing or on-going speech. When the constituent following the WB is long and complex, the boundary will be relatively deep. On these occasions, speakers may use BPMs to indicate the continuation of the utterance being construed. We will explore this possibility in future research.

5. Acknowledgements

This work is supported by Grant-in-Aid for Collaborative Research Project of NINJAL "Empirical study on the role of prosodic features in conversations" led by Hanae Koiso, and Grant-in-Aid for Scientific Research (B) "Constructing cognitive and communicative models for dialog based on utterance-unit annotations" led by Yasuharu Den.

6. References

[1] S. Iwasaki and T. Ono, "'Sentence' in spontaneous spoken Japanese discourse," in *Complex sentences in grammar and discourse*, J. Bybee and M. Noonan, (Eds), Amsterdam: Benjamins, 2001, pp. 175–202.

[2] M. Watanabe, *Features and roles of filled pauses in speech communication: A corpus-based study of spontaneous speech*. Tokyo: Hituzi Syobo, 2009.

[3] H. Koiso, "Nihongo-hanasi-kotoba-koopasu-o motii-ta hukugoo-kyookai-ontyoo-no hatugen-keizoku-hyoozi-kinoo-no kentoo (Continuation function of boundary pitch movements in the Corpus of Spontaneous Japanese)," in *Proceedings of the 2nd Workshop on Japanese corpus linguistics*, pp. 221–230, 2012.

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

[5] M. Watanabe and Y. Den, "Utterance-initial elements in Japanese: A comparison among fillers, conjunctions, and topic phrases," in *Proceedings of the DiSS-LPSS Joint Workshop 2010*, Tokyo, 2010, pp. 31–34.

[6] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA and IEEE Workshop on Spontaneous speech processing and recognition*, Tokyo, 2003, pp. 7–12.

[7] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara, "Identification of "sentences" in spontaneous Japanese: Detection and modification of clause boundaries," in *Proceedings of the ISCA and IEEE Workshop on Spontaneous speech processing and recognition*, Tokyo, 2003, pp. 183–186.

[8] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. J. Venditti, "X-JToBI: An extended J_ToBI for spontaneous speech," in *Proceedings of the 7th International Conference on Spoken Language Processing (INTERSPEECH 2002)*, Denver, CO, pp. 1545–1548, 2002.

[9] R. H. Baayen, *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press, 2008.

[10] H. H. Clark, "Speaking in time," *Speech Communication*, vol. 36, pp. 5–13, 2002.

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

[12] Y. Den, "Are word repetitions really intended by the speaker?" in *Proceedings of the ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*, 2001, pp. 25–28.

[13] S. Koori, "Onsee-no tokutyoo-kara mi-ta bun (Sentence characteristics in terms of phonetic aspects)," *Nihongogaku*, vol. 15, no. 9, pp. 60–70, 1996.

Prediction of F0 height of filled pauses in spontaneous Japanese: a preliminary report

Kikuo Maekawa

National Institute for Japanese Language and Linguistics, Japan

Abstract

F0 values of filled pauses (FP) in the Corpus of Spontaneous Japanese were analyzed to examine the mechanism by which the F0 heights of FP were determined. Statistical analyses of the F0 values of FP occurring in between two full-fledged accentual phrases (AP) revealed correspondence between the occurrence timing of FP and the F0 height. Based upon this finding, 5 models of F0 prediction were proposed. Comparison of the mean prediction errors revealed that the best prediction was obtained in a model that linearly interpolate the phrase-final L% tone of the immediately preceding AP and the phrase-initial %L tone of the immediately following AP. This finding suggests that the F0 of FP was specified at the level of phonetic realization rather than phonological prosodic representation.

1. Introduction

Frequent occurrence of filled pauses (FP hereafter) is one of the most salient characteristics of spontaneous speech. There's a wide consensus among the researchers that FP play positive roles in the processing of spontaneous speech. The supposed cognitive roles of FP include prognosis of the perplexity of upcoming word [1], or the complexity of the upcoming clause [2], marking of discourse structure [3], discourse management [4], indication of the degree of factuality of university lectures [5], etc. There are also speech analytic studies on the phonetic characteristics of FP ([6] among others), and, applications-oriented studies including synthesis of dialogue speech [7], recognition of spontaneous speech [8], etc.

Despite its cognitive importance, mechanisms of FP production are left mostly untouched in the study of speech production. In the study of speech prosody, for example, existing theories of prosodic structure do not pay any attention for the intonational or other prosodic characteristics of FP [9]. The lack of scientific knowledge in this field poses, accordingly, serious limitations on the design of prosodic annotation schema for spontaneous speech.

In the X-JToBI annotation scheme, which was proposed for the prosodic annotation of spontaneous speech [10], FP are treated as a special kind of accentual phrase (AP hereafter) whose pitch height is specified tonally either as FH ('filler-high') or FL ('filler-low'). This binary labeling, however, was not proposed on a firm theoretical basis. It is rather a simple extrapolation of established knowledge about the prosody of Japanese that L and H are required for the specification of linguistic contrast and pragmatic information. There is no a priori reason to believe that FP are specified with respect to binary, or whatever, tonal opposition.

In the rest of this paper, corpus-based analyses of FP will be conducted in terms of their location in utterance, timing with respect to adjacent AP, and, F0 height, to know if it is possible to predict the F0 height of FP from their occurrence environment.

2. The data

The 'Core' part of the Corpus of Spontaneous Japanese (CSJ hereafter), which is X-JToBI annotated, was used for analyses [11]. 44 hours of speeches containing about half a million words are included in the CSJ-Core. FP in the CSJ-Core are marked not only in the X-JToBI annotation, they are also marked in the speech transcriptions. Since the criteria of FP recognition are not identical in the prosodic annotation and speech transcription, the total number of FP do not coincide in the prosodic annotation and transcription. The main difference stems from the treatment of a FP (/de/ see Table 1) occurring in the beginning of utterance, which is treated as a FP in prosodic annotation, while it is treated as an ordinary conjunctive in speech transcription. In the present study, FP were recognized according to the criteria of the X-JToBI scheme. The total number of FP analyzed in this study was 35,164.

As for the textual property, 160 different textual shapes were recognized in the speech transcription of the FP in CSJ-Core. Since this classification is too detailed for the present analyses, FP were reclassified into 23 classes based upon the similarity of their segmental shapes. These classes were further reclassified into 8 classes. The results of two-way classifications are shown in Table 1 as Class1 and Class2 respectively. Note that FP whose occurrence frequencies were less than 10 were omitted from the classifications.

Table 1: *Textual classification of FP.*

Class1	Class2	Example
A	A	/a/ /aQ/
AH		/aH/
AN		/ano/ /anoH/ /aHno/ /aHnoH/ etc.
DE	D	/de/ /te/ /Nde/
E	E	/e/
EH		/ee/ /eH/
ET		/eHto/ /eHtoH/ /eHQto/ /eQto/ etc.
KN	K	/kono/
KO		/kou/
M	M	/ma/
MH		/maH/
MO		/moH/
N	N	/N/
NH		/N:/
NT		/N:to/ /Nto/ /N:toH/ /N:Qto/ etc.
UN		/uHN/ /uN/ etc.
SN	S	/sono/ /sonoH/ etc.
U	V	/u/
UH		/uH/
I		/i/
IH		/iH/
O		/o/
OH		/oH/

Not all FP in the CSJ-Core are suitable for the present study, because it is often impossible to measure the F0 value of FP. FP that meet the following 3 criteria were chosen: (a) The F0 value of the FP in question is reliably measurable, (b) The F0 values of the phrase-final L% tone of the AP that immediately precedes the FP in question is reliably measurable, and, (c) The F0 value of the phrase-initial %L tone of the AP that immediately follows the FP in question is reliably measurable. F0 values and their reliability information could be extracted from the XML files containing the X-JToBI annotation data.

In the X-JToBI data of CSJ-Core, the F0 value of a FP was represented by the F0 value measured near the center of the FP duration. This measurement criterion was adopted on the assumption that the F0 pattern of FP in Japanese is nearly flat unlike the ordinary AP in Japanese (see sections 4.1 and 6).

As the result, 4,892 FP were chosen for analysis. Note that the cases where more than two FP occurred consecutively were excluded from the data. The F0 values of the FP and AP were log-transformed and z-transformed for each speaker.

3. Analysis

3.1. Location of occurrence in clause

Occurrence location of FP in a clause is examined in the first place. The frequencies of Class1 FP were computed for each word positions in a clause, beginning from the first up to, whenever possible, 15th position. The results revealed crucial difference between the Class1 DE and all other FP. More than 80 % of DE occurred as the first 'word' of clauses, while the distributions of other FP have their peaks in the clause-medial positions.

3.2. Timing of occurrence

Timing of FP occurrence was analyzed with respect to the adjacent AP. The relative timing of the beginning of a FP is called relPosit and defined as (T3-T1)/(T2-T1), where T1 is the ending time of the AP that immediately precedes the FP in question, T2 is the beginning time of the AP that immediately follows the FP in question, and, T3 is the beginning time of the FP in question (Cf. Figure 4 below). This index distributes in the interval $0.0 \leq \text{relPosit} < 1.0$. When relPosit=0.0, there is pause between the preceding AP and the FP. When relPosit takes a positive value, there is a pause, and the larger the relPosit, the closer the beginnings of FP is to the following AP. Note that relposit can't be equal to 1.0 because a FP has its own duration.

Figure 1 is a boxplot showing the distributions of relPosit computed for each of the Class1 FP having the frequency higher than 10. Here again, DE is unique in that its distribution is concentered near the higher end of ordinate (relPosit). Similar distribution is found in the case of N. On the other hand, the distributions of AH and AN are concentrated near the origin of ordinate. And, the distributions of FP like A, M, MO, O, and U are widely dispersed across the whole ordinate.

3.3. F0 height

3.3.1. Relation between the timing and F0

There is a loose correlation between the mean F0 heights and the mean relPosit values as shown in Figure 2. The abscissa and ordinate of the figure are the relPosit value as divided into 10 classes and the z-transformed value of log-transformed F0 value (F0Logn, hereafter). F0Logn value stays nearly constant within the lower range of abscissa, but increases considerably toward the higher end. Figure 3 shows the result of the same analysis as applied individually to some Class2 FP that have high frequency of occurrence (N>49). All classes other than A and D show the same tendency as in Figure 2, and in the case of class A, F0Logn increases in the area where relPosit > 0.9.

3.3.2. Prediction models of F0 height

Prediction of the F0Logn values was conducted based upon the findings in the previous subsection. Results of 5 different prediction models were compared (See Figure 4). In Model 1, the F0 value of the AP-final L% tone in the preceding AP is copied as the F0 value of the FP. In Model 2, on the contrary, the F0 value of the AP-initial %L tone of the following AP is copied as the F0 of FP. The relPosit values play no role in these models. Model 3 is a hybrid of Models 1 and 2.

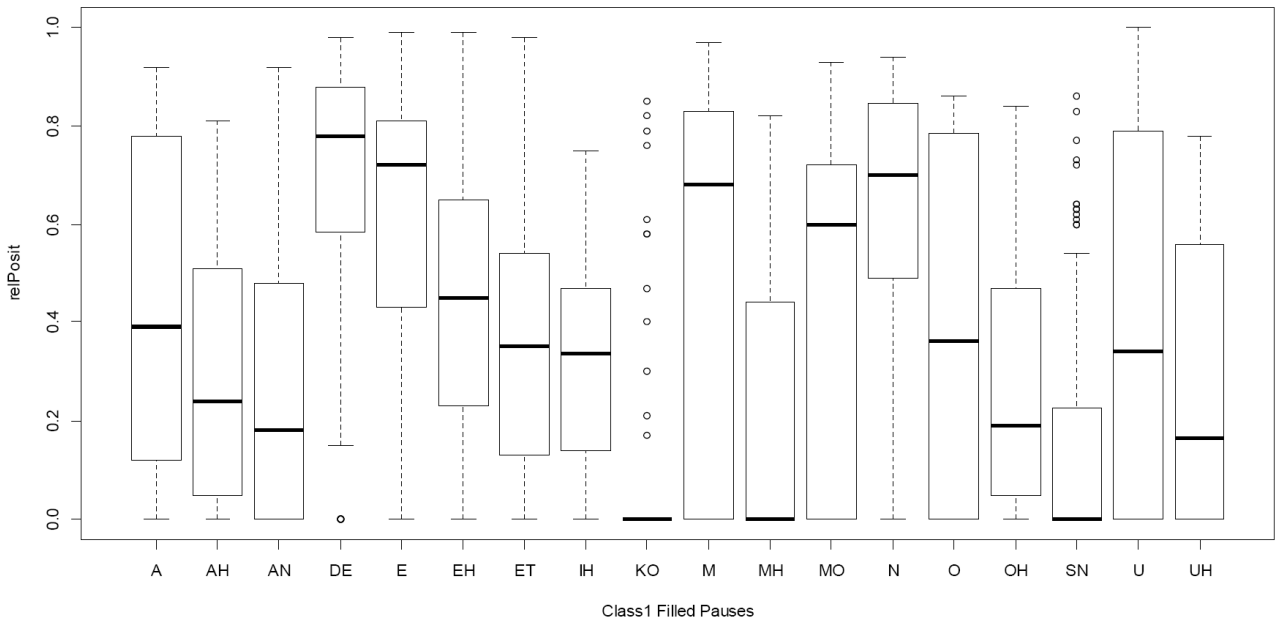


Figure 1: *Distributions of the timing of occurrence (relPosit values) of Class1 FP (N>8).*

In this model, the value preceding L% is copied if the relPost < 0.7; the following %L is copied otherwise. In Model 4, the F0 value of the FP is determined by the interpolation between the values of preceding L% and following %L. Lastly, Model 5 is a hybrid of Models 1 and 4; Model 1 is applied for FP whose relPosit values are smaller than 0.7, and the F0 interpolation of L% and %L is applied for all other FP. Note that in this model, L% is regarded to be copied to the location where relPosit = .7, and the F0 interpolation is conducted between the copied L% and the following %L as shown in Figure 4.

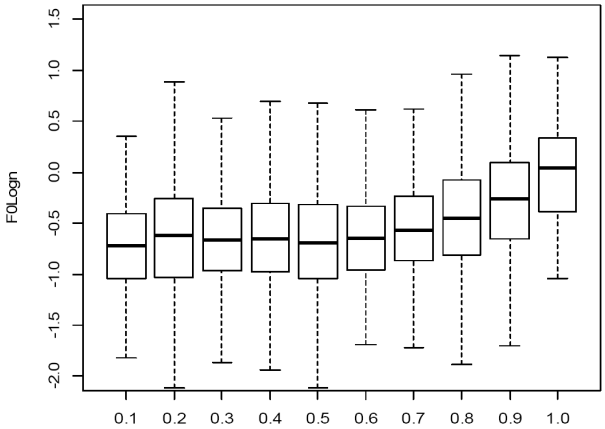


Figure 2: *Relation between the timing and F0 value.*

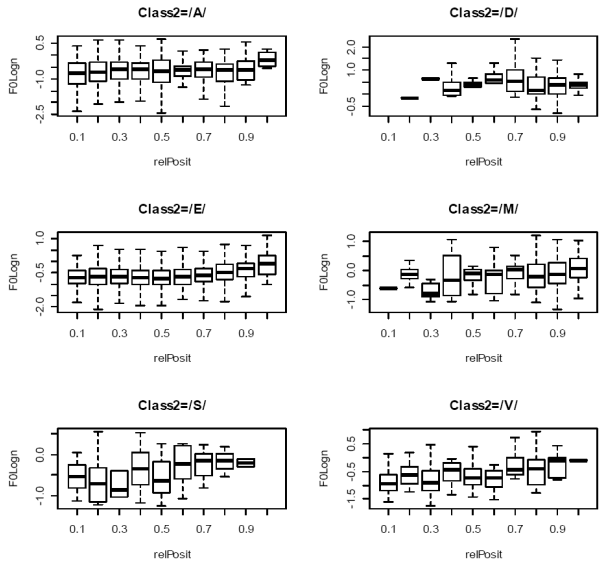


Figure 3: *Timing-F0 relation in individual FP.*

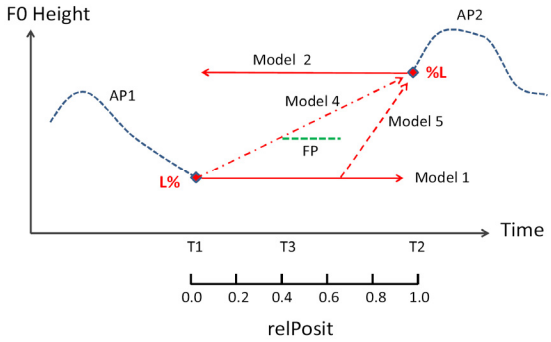


Figure 4: *Schematic representations of 5 prediction models.*

3.3.3. Evaluation of the models

The results of F0 prediction were summarized in Table 2. The RMS errors of prediction were shown in both linear (F0Hz) and log-normalized (F0Logn) values. For each FP, the model showing the least RMS error value is indicated by bold digits, and, the models of two lowest RMS errors are indicated by shaded cells. The FP were classified in terms of Class2. The best prediction was obtained either by Model 4 or 5 in all FP classes. The performances of Models 4 and 5 are very close in all FP except for DE, where Model 4 is superior to Model 5.

4. Discussion

4.1. Evaluation of the precision of prediction

The superiority of Models 4 and 5 over other models are the matter of relative comparison. As shown in the last row of Table 2, the overall RMS prediction errors of the 2 best models (4 or 5) were 0.64 and 0.63 in F0Logn (or, equivalently, 22.2 and 23.0 in Hz). 'Absolute' evaluation of these values is needed. There are reasons to believe that these performances are not bad ones.

First, Figure 5 shows the correlation between the F0 in Hz of FP and the F0 estimated by Model 4. The correlation coefficient is 0.809 and highly significant (t = 96.2074, df = 4890, p-value < 2.2e-16).

Second, Figure 6 compares the density distributions of RMS estimation errors by Model 4, F0 ranges of FP (i.e. the difference between the maximum and minimum F0 in a FP), and the F0 ranges of ordinary AP. F0 are log-normalized. It can be seen from this figure that distribution of the F0 ranges of FP is much narrower than that of ordinary AP. It is also clear that the range of estimation error is even smaller than that of FP. These facts suggest that, from the statistical point of view, it is expected that most of the estimated F0 values are in the fair vicinity of observed F0 values. As a matter of fact, it turns out that 53% of estimated F0 values (in Hz) locate within the ranges of the target FP, and 74% of them locate ± 10 Hz of the ranges.

Table 2: *Comparison of the RMS errors in the F0 estimation.*

Filled Pause	N	Model 1		Model 2		Model 3		Model 4		Model 5	
		F0Hz	F0Logn	F0Hz	F0Logn	F0Hz	F0Logn	F0Hz	F0Logn	F0Hz	F0Logn
AN	411	26.9	0.64	37.0	0.87	28.6	0.70	25.7	0.65	**26.1**	**0.62**
DE	118	76.6	1.83	28.7	0.54	38.9	0.77	**28.5**	**0.52**	46.2	1.02
E	769	26.5	0.76	27.3	0.71	27.0	0.69	20.6	0.58	**18.4**	**0.50**
EH	1790	24.4	0.72	31.4	0.88	25.1	0.72	22.1	0.69	**22.1**	**0.64**
ET	221	30.0	0.79	35.5	0.86	30.5	0.77	**24.4**	**0.65**	28.0	0.73
M	194	31.6	0.91	22.7	0.57	23.0	0.60	**18.0**	**0.49**	21.3	0.59
ALL	3975	27.6	0.77	30.7	0.81	26.2	0.70	22.2	0.64	23.0	0.63

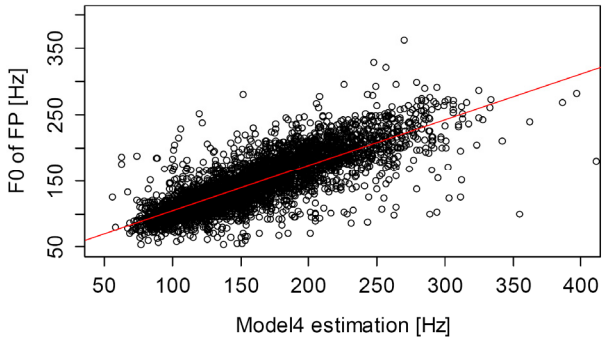


Figure 5: *Correlation between the observed and estimated F0 in Hz. Estimation is by Model 4. Regression line is overlaid.*

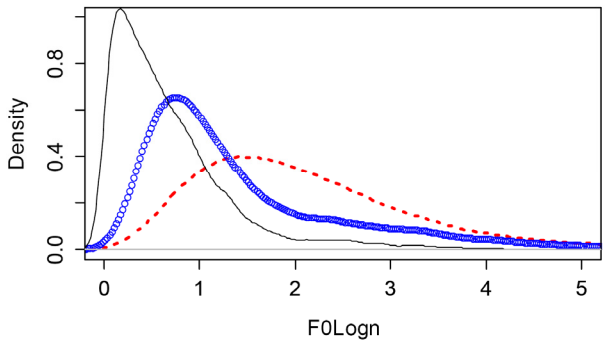


Figure 6: *Comparison of the density distributions of estimation error by Model 4 (real line), F0 range of FP (circle), and F0 range of ordinary AP (dotted line). F0 values are log-normalized.*

4.2. Phonological assessment of models

Five prediction models used in the current study can be classified into two types from a point of view of phonology. Models 1, 2, and 3 are 'phonological' models in that they are based upon the copying (one of the most typical phonological manipulations) of boundary L tones. In these models, a phonological tone is supposed to be associated to FP.

Model 4, on the other hand, is a 'phonetic' model in that it does not include phonological manipulation of tones, and, the FP are not supposed to have any tonal association. The F0 height of the FP is computed on the basis of purely interpolation between the preceding and following boundary L tones. Lastly, Model 5 is a hybrid of 'phonological' and 'phonetic' models.

The results of model comparison revealed that purely 'phonological' models are much inferior to 'phonetic' models in terms of the performance of prediction. From this, it is perhaps safe to conclude that the use of 'FL' and 'FH' labels in the X-JToBI annotation was a non-essential convention. Although there is possibility that these labels reflects native speakers' tendency in the perception of the relative height of FP, the judgments seem to be predictable from the FP's occurrence timing (i.e. relPosit).

To examine the validity of this hypothesis, the relation between the mean relPosit value and the rate of 'FH' labeling was examined for Class1 FP that have frequency higher than 8 and occurring in the environment where %L of the following AP is higher than the L% of the preceding AP.

The Class1 FP that showed the lowest rates of 'FH' label included IH (0.0%, *N*=9), O (0.0%, *N*=20), OH (0.0%, N=33), and UH (0.0%, N=8). And the relPosit values of these FP concentrate in the lower end of ordinate in Figure 1. On the contrary, the FP that showed the highest rates of 'FH' included DE (53.5%, *N*=86), M (28.4%, *N*=130), and MO (21.6%, *N*=28). The relPosit values of these FP tend to concentrate in the higher end in Figure 1.

The overall correlation between the mean relPosit value and the mean rate of FH label is 0.523, and statistically significant (t = 2.605, df = 18, p-value = 0.01791). Note that the correlation is not very high because there are many Class1 FP whose relPosit values scatter widely along the ordinate of Figure 1. Exactly the same tendency is observed when all FP are analyzed (i.e. including the cases where %L is not higher than L%), but the correlation coefficient becomes 0.466 (the correlation, however, is still significant at 0.05).

5. Concluding remarks

The present study revealed the 'phonetic' nature of the F0 height of FP in Japanese. This finding coincides largely with the conclusion of [12] that analyzed the intonation of clause-internal FP in English, but the present finding covers the FP in clause-initial positions as well. The remaining problems include analysis of the cases where more than two FP occur consecutively, finer comparison of Models 4 and 5, inspection of cases where large prediction errors were observed, reexamination of the assumption that F0 is nearly flat within a FP, and so forth. Pilot examination of F0 prediction models by means of synthetic speech is currently underway.

6. Acknowledgements

This study is supported by the Kakenhi grant no.23520483 to the present author. It is also supported by a NINAJL project "Basic Research on Corpus Annotation".

7. References

[1] M.H. Siu and M. Ostendorf "Modeling disfluencies in conversational speech", *Proc. ICSLP '96*, pp. 386–389, 1996.

[2] M. Watanabe, K. Hirose, Y. Den and N. Minematsu. "Filled pauses as cues to the complexity of following phrases", *Proc. INTERSPEECH 2005*, pp. 37–40, 2005.

[3] M. Swerts, "Filled pauses as markers of discourse structure." *Journal of Pragmatics* 30, pp. 485–496, 1998.

[4] T. Sadanobu and Y. Takubo, "Danwa ni okeru shinteki monitaa kikou", Gengo Kenkyu, 108, pp. 74–93, 1995

[5] S. Schachter, N. Christenfeld, B. Ravina and F. Bilous. "Speech disfluency and structure of knowledge", *Journal of Personality and Social Psychology* 60 (3), pp. 362–367, 1991.

[6] E. Shriberg, "Phonetic consequences of speech disfluency." *Proc. ICPhS 1999*, pp. 619–622,1999.

[7] J. Adell, A. Bonafonte, and D. Escudero. "Filled pauses in speech synthesis", *Speech Prosody 2010*, pp. 1–4, 2010.

[8] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira. "Toward the realization of spontaneous speech recognition." *Proc. ICSLP 2000*, pp. 518–521, 2000.

[9] J. Pierrehumbert and M. Beckman. *Japanese Tone Structure*. The MIT Press, 1988.

[10] K. Maekawa, H. Kikuchi, Y. Igarashi and J. Venditti. "X-JToBI: An extended J_ToBI for spontaneous speech", *Proc. ICSLP 2002*, pp. 1545–1548, 2002.

[11] K. Maekawa, "Corpus of Spontaneous Japaese: Its design and evaluation", *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.

[12] E.E. Shriberg and R. J. Lickley. "Intonation of clause-internal filled pauses", *Phonetica* 5, pp. 172–179, 1993.

Analysis of parenthetical clauses in spontaneous Japanese

Takehiko Maruyama

National Institute for Japanese Language and Linguistics, Japan

Abstract

In this paper, I will discuss the functional aspects of parenthetical clauses and sentences in spontaneous Japanese monologues. Parentheticals can be defined as syntactic elements that are instantly inserted in the middle of an ongoing utterance to add supplemental information and thus interrupts the fluent flow of speech production. Examples of parenthetical clauses/sentences that appeared in the *Corpus of Spontaneous Japanese* were examined and then classified into three types. These types differ in their contextual functions, but share a commonality in that they present multiplex information simultaneously in the process of producing spontaneous speech.

Index Terms: parenthetical clause/sentence, Corpus of Spontaneous Japanese, contextual functions

1. Introduction

Since utterance productions are carried out linearly, a speaker must keep producing syntactically well-formed structures instantly in real time, and such a constraint sometimes causes disfluent phenomena in spontaneous speech. A parenthetical structure is one example of such disfluency, where the insertion of some syntactic unit interrupts the fluent flow of the ongoing utterance.

In parenthetical structures, various syntactic categories can be inserted into the original utterance (or "main clause"). Burton-Roberts [1] shows some examples of parentheticals:

(1) It was dawn, about quarter to six, when they arrived.

(2) It will stop raining, I expect, before Sunday.

(3) The main point —why not have a seat?— is outlined in the middle paragraph.

Example (1) shows the case of appositive parentheticals, (2) shows the insertion of a comment clause, and (3) shows a completely isolated utterance that is inserted within the main clause. In any case, parentheticals have no syntactic relation to the structure of the original utterance [1].

In the analyses of written language, parentheticals can be regarded as a kind of rhetorical technique. In spontaneous speech, however, they often interrupt the normal flow of the syntactic structures and thus bring disorder on the ongoing speech as in (3). And there is no previous studies that investigate the actual state of parentheticals in spontaneously spoken Japanese from qualitative and/or quantitative viewpoint.

Thus, following are the research questions for this paper: (1) investigate the frequencies of Japanese parenthetical clauses/sentences and their forms, and (2) examine them from the viewpoint of their contextual functions. To achieve these goals, it is necessary to examine the corpus of spontaneous speech, annotate the extent of parentheticals, and analyze the function of each example.

Section 2 introduces the Corpus of Spontaneous Japanese (CSJ) and the annotation criteria for parentheticals. Section 3 illustrates the results of the annotation and the classification of parentheticals into three types according to their contextual functions. Section 4 discusses the nature of parentheticals by applying the "disruption schema" proposed by Clark [2].

2. Data

2.1. CSJ: Corpus of Spontaneous Japanese

Released in 2004, CSJ is a large-scale and richly annotated spontaneous speech corpus of common Japanese [3]. It consists of 662 hours of speech including 7.5 million words, collected from 3,302 speeches by 1,417 speakers. Most of the speeches consist of spontaneous monologues, which are classified into two types: "Academic Presentation Speech (APS)" and "Simulated Public Speaking (SPS)." APS comprises live recordings of academic presentations in various academic societies. SPS, on the other hand, includes general speeches or comments by laypeople on everyday topics such as "a joyful memory of my life," "the town I live in," and "commentary on recent news." Most monologues in APS and SPS are 10–15 min long.

A speaker of a monologue is required to continue speaking spontaneously for a long time; therefore, various disfluencies can be observed. Thus, CSJ offers appropriate data to examine the process of the dynamic construction of natural speech in real time. Table 1 shows the data size of the "Core," a part of CSJ with richer annotations, which is used in this study.

Table 1: *Data size of "CSJ-Core".*

Type of talk	Sex	# of talks	# of words	Hours
APS	Male	46	137,821	11.75
	Female	24	80,339	9.66
SPS	Male	53	116,643	10.08
	Female	54	108,929	9.1
Total		177	443,732	41.30

2.2. Annotations

We have annotated 684 of the parenthetical clauses and sentences from CSJ-Core. The criteria for annotating parenthetical clauses/sentences are as follows:

1. There must be a dependency relation between the elements before and after the parenthetical.
2. The parentheticals must make no syntactic contribution to the main clause.
3. The parentheticals must end with sentence-final forms or conjunctive particles *ga*, *keredomo*, *keredo*, *kedomo*, and *kedo*.

Conjunctive particles *ga* and *ke(re)do(mo)* function as head words of coordinate clauses, which show high dependency from the main clause.

Table 2: *Frequency of parenthetical clause/sentence (per 100K words)*

		Total	*ga*	*keredomo*	*keredo*	*kedomo*	*kedo*	sentence
APS	Male	126.3	65.3	19.6	0.0	17.4	1.5	21.8
	Female	64.6	22.5	26.9	0.0	2.9	4.4	7.3
SPS	Male	185.8	31.9	29.8	0.7	29.0	54.4	20.3
	Female	148.7	34.8	46.4	4.4	12.3	32.7	9.4

Some examples of parenthetical clauses/sentences are shown below. Boldface shows the extent of parentheticals, : elongated point, and (**) pauses longer than 0.2 sec.

(4) *kanai wa:* (0.42) **ano: nanto ii masyooka** (0.99) *moo*
my wife -TOP FP how to say already
nobite masita ne
groggy -PAST
"my wife was, **uh how can I say that**, already groggy"

(5) *hoteru no (0.3) heya no naka mo sassoku* ***ano***
hotel -NOM room -NOM inside -OBJ soon FP
yoru tuita ndesu kedomo *chekku simasita*
night arrived check -PAST
"we checked, **uh we arrived there in the night**, inside the room of the hotel"

(6) *osake to (0.27) menyuu wa* ***sukunai ndesu ga***
liquor and menu -TOP few
syokuji ga oitearimasu
meal -SUBJ served
"liquor and, **though the menu is limited**, meals are served"

In (4), the first noun phrase, *kanai wa* (my wife), is dependent on the predicate *nobite masita ne* (was groggy), and the parenthetical sentence *anoo nanto ii masyooka* (uh, how can I say that) including the sentence-final form *masyooka* is inserted between the two. Examples (5) and (6) show cases of clauses *ano yoru tuita ndesu kedomo* (uh we arrived there in the night) and *menyuu wa sukunai ndesu ga* (though the menu is limited), which interrupt the flow of the main clauses. In each case, the parentheticals make no syntactic contribution to their main clauses, but are inserted instantly and thus suspend the original utterance, which brings disfluency to the flow of the ongoing narrative.

3. Analysis

3.1. Distribution of parentheticals

Previous studies found that there is clear prosodic/linguistic difference between APS and SPS, and male and female [4, 5]. In this study the entire CSJ-Core was divided into four groups, according to speech types by APS and SPS, and sex by Male and Female. Table 2 shows the frequency of parentheticals per 100K words in four groups.

Table 2 indicates significant differences among the four groups: parentheticals occur more frequently in SPS than APS, and more in the male group than the female group. Since SPS is a collection of casual speech, such a spontaneous speaking style may influence the frequent appearance of parentheticals.

Turning our attention to each conjunctive particle, *ga* appears more frequently than *ke(re)do(mo)* in the APS-male group, while it appears less frequently in the APS-female group. This result indicates that there is a different preference for using *ga*-clauses as parentheticals, depending on the speaker's sex. On the other hand, the four allomorphs of *ke(re)do(mo)* appear more frequently than *ga* in the SPS group. In particular, *keredomo*, *kedomo*, and *kedo* show noticeably different distributions between males and females and between APS and SPS. This result also indicates that there is some preference for choosing the *ke(re)do(mo)* allomorphs as parentheticals on the basis of sex and speaking style, especially in formal contexts.

Two questions arise here. One concerns the entire frequency of each conjunctive particle in the corpus, and the second relates to how frequently each clause is used as a parenthetical. Figure 1 shows the total frequency of each clause in the whole CSJ-Core per 100K words. Figure 2, on the other hand, shows the ratio that how much each clause is used as a parenthetical.

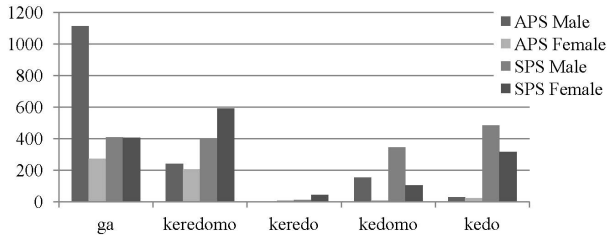


Figure 1: *Frequency of each clause (per 100K words).*

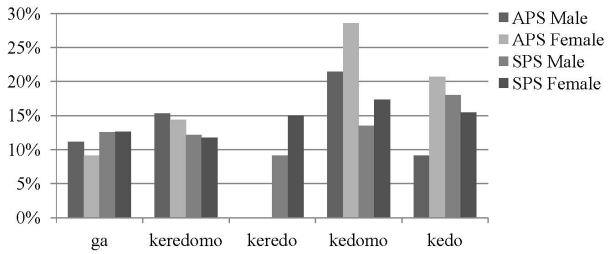


Figure 2: *Ratio of each clause that is used as a parenthetical.*

Figure 1 shows that the total frequencies of each clause are remarkably different from each other among the four groups. Thus, the different distribution observed in Table 2 can be regarded as it reflects the whole distribution of each clause. Figure 2 shows the different distributions of ratio that stands how much each clause is used as a parenthetical. The distributions in *ga* and *keredomo* are rather flat among the four groups, while the cases of *kedomo* and *kedo* vary widely. This indicates that the selection of *kedomo* and *kedo* as parenthetical clauses is affected by the speaker's sex and speaking style.

3.2. Classification of parentheticals

At this point, let us consider the motivation when parenthetical clauses/sentences are used in spontaneous monologues. A parenthetical interrupts the flow of an ongoing utterance, some syntactic unit is inserted, and then the original utterance restarts. With this procedure, the speaker shows multiplex information simultaneously; he/she may express his/her attitude or background knowledge on the topic, or add some explanatory information to the ongoing utterance. By examining each example and its contextual function, I classified parentheticals into three types, A, B, and C.

3.2.1. Type A: Explaining background knowledge

Parentheticals of type A are used to explain background knowledge or a presupposition related to understanding the current utterance. Example (5) corresponds to type A, where the parenthetical just supplements the information on time of arrival, which supports the understanding of the content of the utterance. I also classify example (4) as type A, which shows the speaker's attitude (hesitation) toward the utterance. Another example of type A is shown in (7), which is used to clarify the information on a projected slide.

(7) *ironna pataan wo* (0.3) ***koko ni kaitearu suuji wa***
various pattern -OBJ here written number -TOP
hindo desu ga: *takusan atsumete mimasita*
frequency a lot gather -PAST
"we gathered, **the number of written here shows the frequency**, a lot of various patterns"

3.2.2. Type B: Supplement to what the speaker has just said

Parentheticals of type B are used to supplement what the speaker has just said.

(8) *daigaku: no ninen no toki ni* ***e:to***
university -NOM 2nd year -NOM when FP
19: (0.7) ***96 nen no 7 gatu desu*** (0.53) *e:* (0.7) *nakama*
1996 -NOM July FP friend
to issyo ni kyanpu ni (0.31) *itta*
together camp went
"when I was in my second year at the university, **uh it was July of 1996**, I went to camp with my friends"

(9) *gakubu:* (0.21) **watasi kookakubu datta ndesu kedo**
department I engineering department
sotira no benkyoo wa (0.4) *hotondo siteorimasende*
there -NOM study -TOP hardly did not do
"department, **I used to belong to the engineering department**, I hardly studied there"

In (8) the speaker first said *daigaku ninen no toki ni* (when I was in my second year at the university), and then saw a need to supplement the precise year, thus inserting *eeto 1996 nen no 7 gatu desu* (uh it was July of 1996) with sentence-final form *desu*. Then, she restarted her original utterance. The speaker of (9) first uttered *gakubu* (department), and then suddenly noticed the lack of concrete information, so he instantly inserted *watasi koogakubu datta ndesu kedo* (I used to belong to the engineering department) as a supplementation.

3.2.3. Type C: Supplement to what the speaker is about to say

Parentheticals of type C provide prefatory information about what the speaker is about to say. In (6), the speaker tried to say *syokuji* (meal), and before that she inserted *menyuu wa sukunai ndesu ga* (though the menu is limited) as a proviso. Note that this *ga*-clause is inserted within a noun phrase, *osake to syokuji* (liquor and meal). Another example is shown in (10). The speaker provides a preface *ma kore wa toozen desu ga* (uh this is natural) before he states the conclusion *nagai* (long).

(10) *bimyoona mono yorimo* ***ma kore wa***
subtle than FP this -TOP
toozen desu ga (0.27) *e nagai keekoo ni aru*
natural FP long tendency be
"rather than something subtle, **uh this is natural though**, it tends to be long"

A total of 684 parenthetical clauses and sentences were examined and then classified into three types. Figure 3 shows the frequency of each type per 100K words in the four groups.

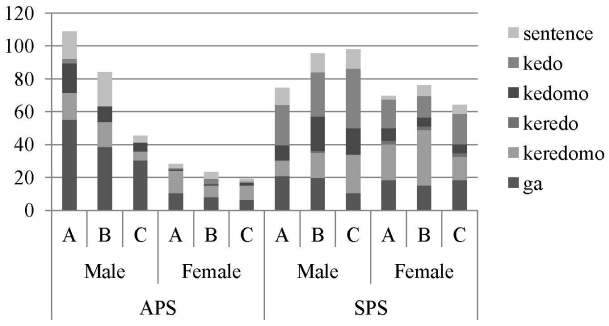


Figure 3: *Frequency of types A, B, C (per 100K words).*

Figure 3 shows a contrastive distribution between APS and SPS. In the APS group, type A's frequency is the highest and type C's is the lowest. On the other hand, in the SPS group, type C is the highest except for the females, and type A is the lowest. This result indicates that the speakers in APS tend to explain background knowledge, a presupposition, or the speaker's attitude by parentheticals, while in SPS, the speakers tend to add some supplementary comments instantly with parentheticals.

4. Discussion

4.1. Formula of parentheticals

The strategy of using parentheticals, which inserts isolated clauses or sentences in the middle of an ongoing utterance, sometimes causes a syntactically ambiguous structure and creates confusion and disorder for the listener. Nonetheless, we use parentheticals to add some comments while speaking, because such an insertion serves at that time to convey sufficient information to the listener in an efficient way. That is the nature of general disfluent phenomena and their repairs, such as filled pauses, repetitions, and self-repairs.

As many previous studies have already indicated, there are some characteristic structures around disfluencies, especially self-repairs, which can be formularized into three parts [6, 7, 8]. Clark [2] proposed a *disruption schema*, which also consists of three intervals, as shown in Figure 4. He also proposed the "suspension device (pause, word cut-off, elongation, nonreduction, filler)," "hiatus contents (no pause, pause, filler, editing expression, elongation, iconic gesture)," and "resumption type (continuation, repetition, substitution, deletion, addition)."

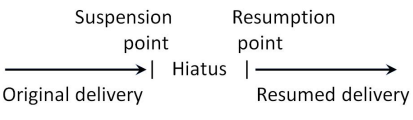


Figure 4: *"Disruption schema".*

Below, I will examine the formula of parenthetical clauses, applying the disruption schema to the examples.

4.1.1. Suspension

Suspending the ongoing utterances, various suspension devices appear in the examples cited above: elongations appear in (4) and (9); short pauses in (4), (6), (7), and (9); and fillers in (4), (5), (8), and (10). These can be regarded as devices to help listeners notice that the current utterance is suspended and that the deviation with a parenthetical has started (or, will start).

4.1.2. Hiatus

The whole extent of parenthetical clauses/sentences corresponds to "editing expressions" in the disruption schema. Example (11) shows an editing expression [2, p.274], which can be classified as a parenthetical of type B in this paper.

(11) We hear now a song from the new Columbia album featuring Very Jail... **Oops, I ought to be in jail for that slip...** of course, I mean Jerry Vale!

A parenthetical of type B makes an edit to what the speaker has just said, as in (11), which fits well with the case of self-repair. Type C is a case of a previous edit to what the speaker is about to say next. Type A edits the action of the utterance itself, because it adds a presupposition or the speaker's attitude with parentheticals.

4.1.3. Resumption

After the insertion of a parenthetical, the speaker must restart the original utterance. There are some linguistic strategies to mark the end of parentheticals, such as pauses in (4), (8), and (10); fillers in (8) and (10); and an elongation at the end of the parenthetical in (7).

Example (12) shows resumption with repetition, and (13) shows repetition and demonstrative *sotirano* (that), which refers to the preceding context. Both repetitions work as signals to show the point at which the original utterance is restarted.

(12) *ato* *syujin ni* ***kekkon siteru ndesu ga***
and husband -DAT I'm married
syujin ni *anda koto mo atte*
husband -DAT have knitted
"and for my husband, **I'm married**, I have knitted for my husband"

(13) *kaisya ga* ***e:to*** ***juugo nen*** ***tutometa ndesu kedo***
company -SBJ FP 15 years have worked
sotirano *kaisya ga* *iten simasite*
that company -SBJ moved
"the company, **I've worked for 15 years**, that company had moved"

In (14), the parenthetical sentence *kotira desune* (this is it) is inserted within a noun phrase, *kankee o* (relation -OBJ), and the isolated *o* is pronounced with a high pitch. In this case, the prominent pitch on the isolated particle works as a cue to restart the original utterance.

(14) *taimingu to* *simpukuhi* *kankee* ***kotira desune***
timing and amplitude -NOM relation this is it
o *moderu no* *koosoku jooken tosite* *kuwaeta*
-OBJ model -NOM as condition add
"the relation of timing and amplitude ratio, **this is it**, we added as a condition of the model"

Figure 5 shows the directions of types A, B, and C for adding comments to the main clause in the disruption schema.

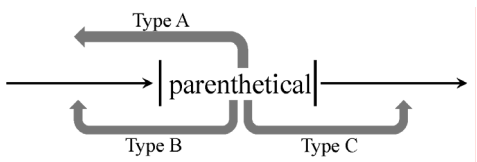


Figure 5: *Three types of parentheticals in the disruption schema.*

4.2. Further issue

I have examined parenthetical clauses/sentences and their contextual functions in this paper. As for issues that should be examined in the future, it is first necessary to investigate the prosodic characteristics around parentheticals, as is partly done in (14).

Second, the target of this study is limited to monologues. However, there may be different linguistic/interactional strategies of parentheticals in dialogues, such as "other-initiated" parentheticals. Extending the range of the target data is also an issue for further study.

5. Concluding remarks

This paper examined parenthetical clauses and sentences in spontaneous Japanese monologues, especially their quantitative and functional aspects. A total of 684 examples of parenthetical clauses/sentences were retrieved from the richly annotated corpus CSJ, and the distribution of each clause was shown from the viewpoint of the speakers' sex and the formality of their speaking style. I also investigated parentheticals and classified them into three types according to their contextual functions. Last, I examined the structure of the parenthetical with the "disruption schema," showing some common aspects they have with self-repairs.

The formal and functional variations of parentheticals reflect the speaker's mental process of linearizing some information in the discourse. The speaker constructs a multiplex information structure by parentheticals under the constraint that an utterance must be produced linearly. This can be regarded as an efficient strategy of spoken language to convey sufficient information to the listener in real time.

6. References

[1] N. Burton-Roberts, "Parentheticals", in E.K. Brown, [Ed], *Encyclopedia of Language and Linguistics*, pp. 179–182, Elsevier, 2005.
[2] H. Clark, *Using Language*, Cambridge University Press. 1996.
[3] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation", in *Proc. of ISCA and IEEEWorkshop on Spontaneous Speech Processing and Recognition*, Tokyo, pp. 7–12, 2003.
[4] K. Maekawa, "Final lowering and boundary pitch movements in spontaneous Japanese", in *Proc. of DiSS-LPSS Joint Workshop* 2010, Tokyo, pp. 47–50, 2010.
[5] K. Maekawa, "Discrimination of speech registers by prosody", in *Proc. of the 17th ICPhS*, Hong Kong, pp. 1302–1305, 2011.
[6] W.J.M. Levelt, "Monitoring and self-repair in speech", *Cognition*, 14:41–104, 1983.
[7] C. Nakatani and J. Hirschberg, "A Speech-first Model for Repair Identification and Correction*", Proc. of 31th Annual Meeting of ACL*, pp. 200–207, 1993.
[8] E.E. Shriberg, *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, U.C. Berkeley, 1994.

Automatic structural metadata identification based on multilayer prosodic information

Helena Moniz[1,2], Fernando Batista[1,3], Isabel Trancoso[1,4] & Ana Isabel Mata[2]

[1] Spoken Language Systems Lab – INESC-ID, Lisbon, Portugal
[2] FLUL/CLUL, Universidae de Lisboa, Portugal
[3] ISCTE – Instituto Universitário de Lisboa, Portugal
[4] IST, Lisboa, Portugal

Abstract

This paper discriminates different types of structural metadata in transcripts of university lectures: boundary events (comma, full stops and interrogatives), and disfluencies (repair). The disambiguation process is based on predefined multilayered linguistic information and on its hierarchical structure. Since boundary events may share similar linguistic properties, in terms of f_0 and energy slopes, presence/absence of silent pauses, and duration of different units of analysis, different classification methods based on a set of automatically derived prosodic features have been applied to differentiate between those events and disfluencies. This paper also performs a detailed analysis on the impact of each individual feature in discriminating each structural event. The results of our data-driven approach allow us to reach a structured set of basic features towards the disambiguation of metadata events. These results are a step forward towards the analysis of speech acts and their disambiguation from disfluencies.

Index Terms: disfluencies, automatic speech processing, structural metadata, speech prosody

1. Introduction

Enriching automatic speech transcripts with structural metadata [1, 2], namely punctuation marks and disfluencies, may highly contribute to the legibility of a string of words produced by a recognizer. This may be important for so many applications that the speech recognizer often appears integrated in a pipeline that also includes several other modules such as audio segmentation, capitalization, punctuation, and identification of disfluent regions. The task of enriching speech transcripts can be seen as a way to structure the string of words into several linguistic units, thus providing multilayered structured information which encompasses different modules of the grammar.

Different sources of information may be useful for this task, going much beyond the lexical cues derived from the speech transcripts, or the acoustic cues provided by the audio segmentation module (*e.g.,* speech/non-speech detection, background conditions classification, speaker diarization, etc.). In fact, one of the most important roles in the identification and evaluation of structured metadata is played by prosodic cues.

The goal of this paper is to study the impact of prosodic information in revealing structured metadata, addressing at the same time the task of recovering punctuation marks and the task of identifying disfluencies. The former is associated with the segmentation of the string of words into speech acts, and the later, besides other aspects, also allows the discrimination of potential ambiguous places for a punctuation mark. Punctuate spontaneous speech is in itself a quite complex task, further increased by the difficulty in segmenting disfluent sequences, and in differentiating between those structural metadata events. Annotators of the corpus used in this study report that those tasks are the hardest to accomplish, difficulty visible in the evaluation of manual transcripts, since the attribution of erroneous punctuation marks to delimit disfluent sequences corresponds to the majority of the errors. Furthermore, prosodic cues either for the attribution of a punctuation mark or for the signaling of a repair may be ambiguous [3, 4].

2. Related work

Recovering punctuation marks and disfluencies are two relevant MDA (Metadata Annotation) tasks. The impact of the methods and of the linguistic information on structural metadata tasks has been discussed in the literature. [5] report a general HMM (Hidden Markov Model) framework that allows the combination of lexical and prosodic clues for recovering *full stops*, *commas* and *question marks*. A similar approach was also used by [1, 6] for detecting sentence boundaries. A Maximum Entropy (ME) based method is described by [7] for inserting punctuation marks into spontaneous conversational speech, where the punctuation task is considered as a tagging task and words are tagged with the appropriate punctuation. It covers three punctuation marks: *commas*, *full stops*, and *question marks*; and the best results on the ASR output are achieved by combining lexical and prosodic features. A multi-pass linear fold algorithm for sentence boundary detection in spontaneous speech is proposed by [8], which uses prosodic features, focusing on the relation between sentence boundaries and break indices and duration, covering their local and global structural properties. Other recent studies have shown that the best performance for the punctuation task is achieved when prosodic, morphological and syntactic information are combined [1, 2, 9].

Much of the features and the methods used for sentence-like unit detection may be applied in disfluency detection tasks. What is specific of the latter is that disfluencies have an idiosyncratic structure: *reparandum*, interruption point, interregnum and repair of fluency [10, 11, 12]. The *reparandum* is the region to repair. The interruption point is the moment when the speaker stops his/her production to correct the linguistic material detected. Ultimately, it is the frontier between disfluent and fluent speech. The *interregnum* is an optional part and may include silent pauses, filled pauses (uh, um) or explicit editing expressions (I mean, no).

The repair is the corrected linguistic material. It is known that each of these regions has idiosyncratic acoustic properties that distinguish them from each other, inscribed in the edit signal theory [13], meaning that speakers signal an upcoming repair to their listeners. The edit signal is manifested by means of production of fragments, glottalizations, co-articulatory gestures and voice quality attributes, such as jitter (perturbations in the pitch period) in the *reparanda*.

Table 1: *Corpus properties and number of metadata events.*

Subset →	train+dev	test
Time (h)	28:00	3:24
number of words + filled pauses	216435	24516
number of disfluencies	8390	950
disfluencies followed by a repair	5608	720
number of full stops	8363	861
number of commas	22957	2612
number of question marks	3526	498

Sequentially, it is also edited by means of significantly different pause durations from fluent boundaries and by specific lexical items in the interregnum. Finally, it is edited via f_0 and energy contrastive or parallelistic patterns in the repair.

[14] present a statistical language model including the identification of POS tags, discourse markers, speech repairs, and intonational phrases, achieving better performances by analyzing those events simultaneously. Based on the edit signal theory, [11, 15] used CARTs to identify different prosodic features of the interruption point. [16, 1] used features based on previous studies and added language models to predict both prosodic and lexical features of sentence boundaries and disfluencies.

Our aim is to check if we are able to classify metadata structures based on the following set of features derived from the above-mentioned studies: pause at the boundary, pitch declination over the sentence, post-boundary pitch and energy resets, pre-boundary lengthening, word duration, silent pauses, filled pauses, and presence of fragments. By investigating how much one can classify and disambiguate in Portuguese, using just this set of very informative cues, we hope to contribute to the discussion of what are language and domain dependent effects in structural metadata evaluation.

3. Corpus

This study uses LECTRA[17], a corpus of university lectures transcribed for producing multimedia contents for e-learning applications. The corpus was divided into two main sets: *train+development* (89%), and *test* (11%). The sets include portions of each one of the courses and follow a temporal criterion, meaning the first classes of each course were included in the training set, whereas the final ones were integrated into both development and test sets. The data encompasses all the structural metadata events presented in Table 1.

One important aspect that characterizes Portuguese punctuation marks is the high frequency of *commas*, which in our corpus accounts for more than 50% of all events. In a previous study [3], where Portuguese and English Broadcast News are compared, the percentage of *commas* in the former is twice the frequency of the latter. The guidelines used for *commas* are the ones described in [18].

3.1. Integrating prosodic information

This work relies on information coming from the ASR output, manual transcripts, and the signal itself. After the speech recognition, all relevant manual annotations are transferred to the ASR transcripts, including all metadata events, by means of the NIST SCLite tool (http://www.nist.gov/speech). Durations of phones, words, and interword-pauses were extracted from the ASR output. Information regarding pitch (f_0) and energy (E) was not available in the ASR pipeline when this study started. For that reason, it has been directly extracted from the speech signal, using the Snack toolkit [19]. Energy and f_0 slopes were calculated based on linear regression.

Acoustic–phonetic parameters of segmental and supra-segmental units were, thus, automatically extracted to study structural metadata events. Organizing such information into hierarchies, meaning, into the smallest unit of analysis (phones or even sub-phone units) up to higher order constituents was crucial to the experiments conducted. At this point, the information extracted encompassed phones, syllables, words, sentence-like units, and speech-acts.

Our in-house speech recognizer [20], trained for the broadcast news domain, is totally unsuitable for the university lectures domain. The scarcity of text materials in Portuguese to train language models for this domain has motivated the decision of using the ASR in a force alignment mode, in order not to bias the study with the bad results obtained with an out-of-domain recognizer. For that reason, current experiments rely on force aligned transcripts that still contain about 0.9% of unaligned words (mainly due to low energy segments).

4. Predicting structural metadata events

Our experiments use a fixed set of purely automatic features, extracted either from the ASR output or from the speech signal itself. The features involve two words before the event and one word after the event, and characterize either a word or a sequence of two consecutive words. Features involving a single word include: pitch and energy slopes; ASR confidence score; word duration; number of syllables and number of phones. Features involving two consecutive words include: pitch and energy slopes shapes; pitch and energy differences; comparison of durations and silences before each word (*dur.comp*); and ratios for silences, word durations, pitch medians (*pmed.ratio*), and energy medians (*emed.ratio*). For example, *eslopes* : $RF_{cw,fw}$ is a shape feature that refers to the energy slope in the current (cw) and following words (fw), which is **R**ising in *cw* and is **F**alling in *fw*; $dur.ratio_{cw,fw}$ is a number between 0 and 1 that indicates the proportion of the duration of *cw* over the duration of *cw+fw*.

Our experiments were performed using the Weka toolkit [21] and distinct statistical methods have been applied, including: Naïve Bayes, Logistic Regression and Classification and Regression Trees (CART). The best results were consistently achieved using CARTs, closely followed by Logistic Regression. The remaining of this section shows the achieved results and performs an analysis on the most relevant features.

4.1. Results

Experiments aim at automatically detecting structural metadata events and at discriminating between those events, using mostly prosodic features (with the exception of two identical contiguous words). We have considered four different classes of structural elements, *full stop*s, *commas*, *question marks*, and disfluency repairs. Table 2 presents the best results achieved, using the standard metrics *precision, recall, F-measure* and *Slot Error Rate*. The best performance is achieved for *full stops,* confirming our expectation, since prosodic information is known to be crucial to classify those events in our language. The low results concerning *commas* are also justifiable, because our experiments rely on prosodic features, but *commas* depend mostly on lexical and syntactic features [9].

Table 2: *CART classification results for prosodic features.*

Class	**Precision**	**Recall**	**F-meas.**	**SER**
comma (,)	60.6	27.6	37.9	90.3
full stop (.)	64.1	67.6	65.8	70.2
question (?)	73.9	29.5	42.2	80.9
repair	60.8	13.1	21.6	95.4
weighted avg.	63.0	32.9	43.3	75.6

Table 3: *Confusion matrix between events.*

Classified as →	,	.	?	repair	*del.*
comma (,)	**718**	36	10	15	1823
full stop (.)	76	**579**	35	3	163
question (?)	27	225	**147**	4	95
repair	51	19	1	**93**	546
insertions	312	44	6	38	

The performance for *question marks* is mainly related to their lower frequency and to the multiple prosodic patterns found for these structures. Moreover, interrogatives in our language are not commonly produced with subject-auxiliary verb inversion, as in English, which renders the problem of identifying interrogatives even more challenging. The worse performance, specially affected by a low recall, is achieved for *repairs*. While prosodic features seem to be strong cues for detecting this class, the confusion matrix presented in Table 3 reveals that *repairs* are still confused with regular words. Table 3 also reveals that the most ambiguous class is, without doubt, interrogatives.

Our recent experiments as well as other reported work [10, 11, 12] suggest that *filled pauses* and *fragment*s serve as cues for detecting structural regions of a disfluent sequence. Supported by such facts, we have conducted an additional experiment using *filled pauses* and *fragments* as features. These features turned out to be amongst the most informative features, increasing the *repair* f-measure to 48.8%, and improving the overall f-measure to 47.8%. However, the impact of fragments is lower than the one reported by [11, 22] and this may be due to the fact that fragments in our corpus represent only 6.6% of all the disfluent types.

4.2. Most salient features

Equivalent experiments performed with Logistic Regression provide a good approximation to the impact of each feature. A first inspection on Table 4 suggests that two pairs of structural metadata events are prone to be classified as ambiguous: *full stops* and *question marks*; and *repairs* and *regular words*. However, a closer inspection reveals that a set of informative features stands out as determinant to disambiguate between such events, namely, pitch and energy shapes, duration ratios, and confidence levels of the units of analysis.

Features for the discrimination of a ***repair*** comprise: i) two identical contiguous words; ii) both energy and pitch increases in the following word and (mostly) a plateau contour on the preceding word; and iii) a higher confidence level for the following word than for the previous word. Reasoning about this, this set of features is showing that repetitions are being identified, that repair regions are characterized by prosodic contrast marking (increases in pitch and energy) between disfluency–fluency repair (as in our previous studies), and also that the repair identification has a high confidence level.

Table 4: *Top most relevant features, sorted by relevance.*

	Feature	**none**	**,**	**.**	**?**	**repair**
1	$pslopes : F\_{pw,cw}$			▪▪▪	▪▪▪▪▪	
2	$pslopes : --_{pw,cw}$			▪▪▪▪	▪▪▪▪▪	
3	$pslopes : R\_{pw,cw}$			▪▪▪▪	▪▪▪▪▪	
4	$conf_{cw}$			▪▪▪▪▪	▪▪▪▪	
5	$eslopes : RF_{cw,fw}$			▪▪▪▪▪	▪▪▪▪	
6	$eslopes : --_{pw,cw}$			▪▪▪▪▪	▪▪	
7	$eslopes : F\_{pw,cw}$			▪▪▪▪▪	▪▪▪	
8	$eslopes : R\_{cw,fw}$			▪▪▪▪▪	▪▪▪	
9	$eslopes : R\_{pw,cw}$			▪▪▪▪▪	▪▪	
10	$eslopes : RF_{pw,cw}$			▪▪▪	▪▪▪▪▪	
11	$eslopes : FF_{pw,cw}$			▪▪▪	▪▪▪▪▪	
12	$eslopes : RR_{cw,fw}$			▪▪▪▪	▪▪▪▪	
13	$eslopes : -F_{pw,cw}$			▪▪▪▪	▪▪▪▪	
14	$pslopes : RF_{cw,fw}$		▪	▪	▪▪▪▪▪	
15	$pslopes : F\_{cw,fw}$			▪▪▪▪▪	▪▪▪	
16	$pslopes : FF_{cw,fw}$			▪▪▪▪▪	▪▪▪	
17	$pslopes : R\_{cw,fw}$		▪	▪	▪▪▪▪▪	
18	$pslopes : RR_{cw,fw}$		▪	▪	▪▪▪▪▪	
19	$pslopes : FR_{cw,fw}$			▪▪▪▪▪	▪▪▪	
20	$bsil.ratio_{cw,fw}$	▪▪▪▪▪				
21	$bsil.comp :>_{cw,fw}$			▪▪▪	▪▪▪▪	
22	$emed.ratio_{cw,fw}$	▪	▪	▪▪▪	▪▪▪	
23	$bsil.ratio_{pw,cw}$	▪▪▪▪▪	▪▪			▪
24	$dur.ratio_{cw,fw}$		▪	▪▪▪▪▪		▪
25	$dur.ratio_{pw,cw}$	▪▪▪	▪▪	▪		▪
26	$emed.ratio_{pw,cw}$	▪▪	▪	▪▪▪		▪
27	$pslopes : -F_{pw,cw}$	▪▪▪▪	▪			▪▪▪
28	$pslopes : RF_{pw,cw}$	▪▪▪▪	▪▪			▪▪▪
29	$pslopes : FF_{pw,cw}$	▪▪▪▪	▪▪			▪▪▪
30	$pslopes : -F_{cw,fw}$	▪▪▪▪				▪▪▪▪▪
31	$eslopes : -F_{cw,fw}$	▪▪▪	▪▪			▪▪▪
32	$pslopes : --_{cw,fw}$	▪▪▪				▪▪▪▪▪
33	$equals_{pw,cw}$	▪	▪	▪▪▪		▪▪
34	$pslopes : -R_{cw,fw}$	▪▪▪				▪▪▪▪▪
35	$phones_{cw}$	▪	▪▪	▪▪	▪▪	▪
36	$bsil.comp :<_{cw,fw}$	▪▪▪	▪▪			▪▪
37	$bsil.comp :>_{pw,cw}$	▪	▪	▪▪	▪▪▪	▪
38	$eslopes : -R_{cw,fw}$	▪▪	▪▪			▪▪▪▪
39	$eslopes : --_{cw,fw}$	▪▪	▪▪			▪▪▪▪
40	$pmed.ratio_{pw,cw}$	▪▪	▪	▪▪	▪	▪
41	$eslopes : FR_{cw,fw}$		▪▪			▪▪▪▪▪
42	$pslopes : -R_{pw,cw}$	▪▪				▪▪▪▪▪
43	$eslopes : RR_{pw,cw}$	▪	▪			▪▪▪▪▪
44	$eslopes : -R_{pw,cw}$	▪	▪▪			▪▪▪▪▪
45	$eslopes : FR_{pw,cw}$	▪	▪▪			▪▪▪▪▪
46	$eslopes : F\_{cw,fw}$	▪	▪▪			▪▪▪▪▪
47	$pslopes : FR_{pw,cw}$	▪▪				▪▪▪▪▪
48	$pslopes : RR_{pw,cw}$	▪▪				▪▪▪▪▪
49	$equals_{cw,fw}$		▪		▪	▪▪▪▪▪
50	$eslopes : FF_{cw,fw}$	▪	▪▪			▪▪▪▪▪
51	$conf_{fw}$	▪	▪		▪	▪▪▪▪▪
52	$bsil.comp :=_{cw,fw}$	▪▪	▪	▪	▪	▪▪
53	$bsil.comp :=_{pw,cw}$	▪▪	▪▪	▪	▪	▪▪
54	$dur.comp :>_{cw,fw}$	▪	▪▪	▪	▪▪	▪
55	$dur.comp :<_{cw,fw}$	▪▪	▪	▪	▪	▪▪
56	$phones_{fw}$	▪	▪▪	▪▪	▪▪	▪
57	$pmed.ratio_{cw,fw}$	▪	▪	▪▪	▪▪	▪▪
58	$dur.comp :<_{pw,cw}$	▪▪	▪▪	▪▪	▪	▪
59	$dur.comp :>_{pw,cw}$	▪▪	▪	▪	▪▪	▪▪
60	$syls_{fw}$	▪▪	▪	▪	▪	▪▪

As for ***full stops***, the determinant prosodic features correspond to: i) a falling contour in the current word; ii) a plateau energy slope in the current word; iii) the duration ratio between the current and the following words; and iv) a higher confidence level for the current word.

Reasoning about this characterization, it is the one that most resemble the neutral statements in our language, with the canonical contour H+L* L%.

Question marks are characterized by two main patterns: i) a rising contour in the current word and a rising/rising energy slope between current and following words; and ii) a plateau pitch contour in the current word and a falling energy slope in the current word. The rising patterns associated with question marks are not surprising, since they commonly associated with interrogatives. The falling pitch contours have also been ascribed for different types of interrogatives, especially wh-questions in Portuguese.

Commas, as stated along the previous section, are the event characterized by fewer prosodic features. Being mostly identified by morphosyntactic features, they are not clearly disambiguated with prosodic features.

With regards to ***regular words***, the most salient features are related to the absence of silent pauses, explained by the fact that, contrarily to the other events, regular words within phrases are connected. The presence of a silent pause is a strong cue to the assignment of a structural metadata event.

The literature for Portuguese points out to an array of features relevant for the description of metadata events. With the exception of *commas*, the data-driven approach followed in this work allow us to reach a structured set of basic features towards the disambiguation of such events beyond the established evidences for language.

5. Conclusions

This paper reports experiments on a full discrimination of structural metadata events in a corpus of university lectures, a domain characterized by a high percentage of structural events, namely punctuation marks and disfluencies. Our previous work on automatic recovery of punctuation marks indicates that specific punctuation marks display different sets of linguistic features. This motivated the discrimination of the different SU types. Our experiments, purely based on prosodic features, achieved a considerable performance, further improved when the ground truth about filled pauses and fragments was also used. Moreover, based on a set of complex prosodic features, we were able to point out regular sets of features associated with the discrimination of events (*repairs*, *full stops*, and *question marks*).

Future experiments will extend this study to fully automatic speech recognition transcripts and evaluate how the discrimination between the punctuation marks and disfluencies is affected by the ASR errors. Future work will also tackle the inclusion of lexical and morphosyntactic features, which are expected to considerably improve the performance, specially for *commas* and *question marks*.

6. Acknowledgements

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under Ph.D grant SFRH/BD/44671/2008, projects PEst-OE/EEI/LA0021/2013 and PTDC/CLE-LIN/120017/2010, and by ISCTE–IUL.

7. References

[1] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies", *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, no. 5, pp. 1526–1540, 2006.

[2] M. Ostendorf and et al, "Speech segmentation and spoken document processing", *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 59–69, 2008.

[3] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts", *Transactions on Audio Speech and Language Processing*, no. 20, pp. 474–485, 2012.

[4] H. Moniz, F. Batista, I. Trancoso, and A. I. Mata, "Prosodic context-based analysis of disfluencies", in *Proc. Interspeech*, Portland, Oregon, 2012.

[5] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models", in *ASRU*, 2001.

[6] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics", *Speech Communication*, vol. 32, pp. 127–154, 2000.

[7] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," *in Proc. ICSLP*, 2002.

[8] D. Wang and S. S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues", in *Proc. ICASSP*, vol. 1, 2004.

[9] B. Favre, D. Hakkani-Tür, and E. Shriberg, "Syntactically-informed Models for Comma Prediction", in *ICASSP*, 2009.

[10] W. Levelt, Speaking. Cambridge: MIT Press, 1989.

[11] C. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech", *Journal of the Acoustical Society of America* (JASA), no. 95, pp. 1603–1616, 1994.

[12] E. Shriberg, "Preliminaries to a theory of speech disfluencies", Ph.D. dissertation, University of California, 1994.

[13] D. Hindle, "Deterministic parsing of syntactic non-fluencies", in *Proc. ACL*, 1983.

[14] P. Heeman and J. Allen, "Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue", *Computational Linguistics*, vol. 25, pp. 527–571, 1999.

[15] E. Shriberg, "Phonetic consequences of speech disfluency", in *Proc. ICPhS*, San Francisco, 1999.

[16] J.-H. Kim and P. C. Woodland, "Automatic capitalisation generation for speech input", *Computer Speech & Language*, vol. 18, no. 1, pp. 67–90, 2004.

[17] I. Trancoso, R. Martins, H. Moniz, A. I. Mata, and M. C. Viana, "The Lectra corpus – classroom lecture transcriptions in European Portuguese", in *Proc. LREC*, Morocco, 2008.

[18] I. Duarte, Língua Portuguesa, Instrumentos de Análise. Lisboa: Universidade Aberta, 2000.

[19] K. Sjölander, J. Beskow, J. Gustafson, E. Lewin, R. Carlson, and B. Granström, "Web-based educational tools for speech technology", in *Proc. ICSLP*, Australia, 1998.

[20] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in Portuguese", in *Proc. ICASSP*, 2008.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update", *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[22] J. Kim, S. E. Schwarm, and M. Ostendorf, "Detecting structural metadata with decision trees and transformation-based learning", in *HLT-NAACL*, 2004.

Which kind of hesitations can be found in Estonian spontaneous speech?

Rena Nemoto

Institute of Cybernetics at Tallinn University of Technology, Tallinn, Estonia

Abstract

This paper describes the acoustic characteristics of hesitations in Estonian spontaneous speech. We especially investigate duration, fundamental frequency, and first two formant analyses. Most frequent hesitations can be expressed by lengthened phonemes such as /*ää*/, /*ee*/, /*õõ*/, and /*mm*/. We compare lengthened phoneme hesitations with their related phonemes. The results from our preliminary hesitation study show (i) hesitations have longer duration and its range is spread; (ii) hesitations globally include lower pitch; (iii) hesitation formants are likely to be centralized or posterior and opened in comparison with related phonemes.

Index Terms: hesitation, Estonian, spontaneous speech

1. Introduction

Estonian disfluency has mostly been investigated at spoken language text level such as parsing [1] where the authors paid attention for repairs, repetitions and false starts, but not at acoustic level. In this paper, we investigate hesitations as a preliminary study. In some languages, hesitations are expressed by vowels (such as 'uh/um' or 'er' in English, '*euh*' in French, and '*eh*' in Spanish), and lengthened nasal consonants ('*mm*' in Mandarin) [2]. There is no particular hesitation word in Estonian like English and French. Instead, there are a lot of varieties with lengthened vowels or consonants, or mixing a vowel with a consonant, etc. Estonian has 9 vowels (cf. Figure 1) and 18 consonants. This paper tries to answer the following questions: (i) which vowels or consonants can be used if there is no particular hesitation word? (ii) Which differences can be found between lengthened phoneme hesitations and their related phonemes at acoustic level?

2. Corpus and methodology

For this study, we used the manually transcribed phonetic corpus of Estonian spontaneous speech of the university of Tartu [3]. We used 25 male and 26 female speakers in a corpus of monologues or dialogues, which contains about 15 hours for male and 13 hours for female speakers. Fundamental frequency (f_0) and two first formants (F1 and F2) were extracted every 5 milliseconds (ms) by using Praat software [4]. Measurements were averaged over phonemes. Phonemic duration was taken from the manually segmented corpus.

First we counted occurrences of words transcribed individually in lengthened vowels and consonants. Most frequent lengthened vowels and consonant are presented in Table 1. Figure 1 from [5] shows the Estonian vocalic system with vocalic hesitations expressed in added red line. Frequent vocalic hesitations did not contain close vowels such as [i, y, u]. The most frequent hesitation is *ee* for both male and female speakers with 55% of hesitation occurrences (male: 49% and female 62%). However, each speaker has his/her preference of hesitation use as shown in Table 2. Table 2 compares the hesitation occurrences between three speakers (two male and one female speakers), who contain longer speech duration than other speakers. MS1 is likely to more utter *õõ*, while two other speakers (MS2 and FS3) tend to more often employ *ee*.

Table 1. *Occurrences of hesitation.*

Hesitation	Male	Female	Total
aa	57	52	109
ää	78	102	180
ee	1,029	982	2,011
öö	48	65	113
õõ	572	185	757
mm	297	194	491
Total	2,081	1,580	3,661

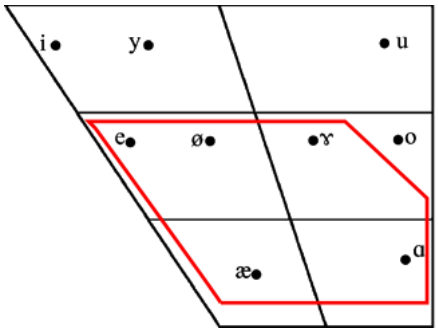


Figure 1: *Estonian vocalic system (vocalic hesitations in red).*

Table 2. *Hesitation occurrences of male speakers (MS1 & MS2) and a female speaker (FS3) with corpus duration in brackets.*

Hesitation	MS1 (45m.)	MS2 (51m.)	FS3 (50m.)
aa	2	6	2
ää	20	23	39
ee	169	311	623
öö	14	4	2
õõ	372	49	47
mm	22	22	3

3. Acoustic analysis

For this study, we compare duration, f_0, and two first formants (F1 and F2) between the hesitations and their related phonemes without considering quantity degree (short, long, overlong).

3.1. Duration

Figure 2 shows duration distribution of hesitations (left), and of related five vowels and one consonant (right). Mean hesitation duration reaches 306 milliseconds (ms) (median: 264 ms, standard deviation: 171 ms), whereas mean related phoneme duration is 78 ms (median: 67 ms, standard deviation: 49 ms). We notice that Estonian hesitations have also much longer durations than related phonemes as revealed

in the literature like other languages. The hesitation duration contour is more spread. The difference between hesitations and related phonemes turned out to be statistically significant ($p < 0.0001$) by Wilcoxon test using R software [6].

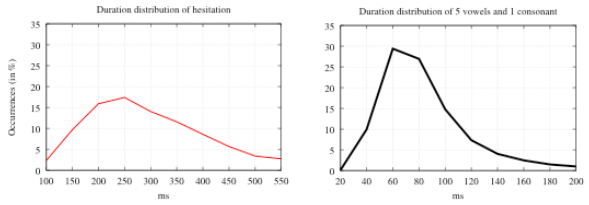


Figure 2: *Duration distribution of hesitations (left: 100–550 ms) and their related phonemes (right: 20–200 ms).*

3.2. Fundamental frequency (f_0)

As each speaker has different pitch, we chose three speakers, who had quite long speech (cf. Table 2), so as to compare f_0 between hesitation and their related vowels. Three frequent vocalic hesitations (*ää*, *ee*, *õõ*) and each related vowel were computed. Table 3 presents mean f_0 in Hz with over 80% of voicing ratios (in order to avoid extracting f_0 value errors) for each vocalic hesitation and its related vowel. Lower f_0 values for vocalic hesitations have been observed for all hesitations and speakers. Statistical analysis using Wilcoxon test showed significant differences of the f_0 values between vocalic hesitations and their related vowels ($p < 0.005$ for all pairs).

Table 3. *Mean f_0 (in Hz) of three speakers (MS1, MS2, FS3) in comparison with three vocalic hesitations and related vowels.*

Hesit./Vowel	MS1	MS2	FS3
ää/ ä (hesit./vow. occ.)	121/135 (20/266)	92/111 (23/327)	235/281 (39/306)
ee/e (hesit./vow. occ.)	126/129 (169/2955)	92/104 (311/1979)	244/264 (623/2034)
õõ/ õ (hesit./vow. occ.)	121/137 (372/405)	94/111 (49/232)	251/284 (47/304)

3.3. Formants

Last we measured two first formants (F1 and F2) through the same three speakers. Figure 3 illustrates F1 and F2 values for three vocalic hesitations (*ää*, *ee*, *õõ*) and their related vowels (*ä*, *e*, *õ*) of two male speakers (MS1 and MS2) and one female speaker (FS3).

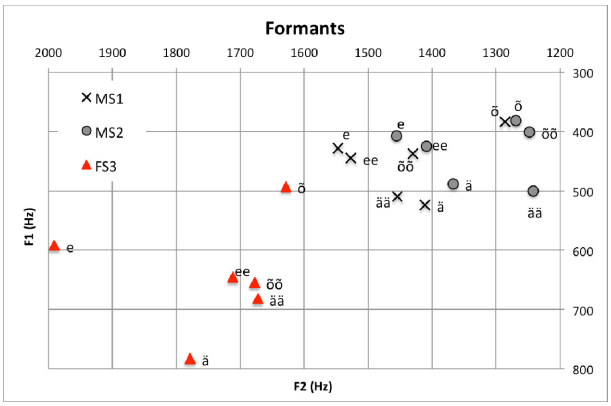


Figure 3: *F1/F2 mean values (in Hz) for three vocalic hesitations (ää, ee, õõ) and their related vowels of three speakers (MS1, MS2, and FS3).*

In comparison with vocalic hesitations and their related vowels, we can observe that MS1 has three vocalic hesitations centralizing to each other especially *õõ*, while MS2 globally contains more opened and posterior vocalic hesitations than the vowels. Three vocalic hesitations of FS3 are also centralized and very close to each other. We conducted statistical tests to verify if these formants are different between vocalic hesitations and related phonemes. As for FS3, both F1 and F2 of all pairs were significantly different ($p < 0.001$). There is no difference of F1 between the *ää*/*ä* pair for MS1 and MS2. However, the other pairs of F1 were significantly different ($p < 0.05$). As for F2, no significant difference is found in the pair *ee*/*e* for MS1 and the pair *õõ*/*õ* for MS2, whereas the others showed significant difference ($p < 0.01$).

4. Discussion

In this paper, we aimed at exploring the preliminary hesitation study of Estonian language. Our study focused on especially acoustic characteristics (duration, f_0, two first formants) of Estonian hesitations in a spontaneous speech corpus. As characteristics of hesitations in Estonian, we found that lengthened vowels such as /*aa*/, /*ää* /, /*ee*/, /*õõ* /, /*öö* /, and a lengthened consonant /mm/ were mostly used. The duration comparison between lengthened phoneme hesitations and their related phonemes showed that hesitations include longer duration and the duration range is spread. The result from f_0 analysis of three speakers revealed that vocalic hesitations have globally lower values. Two first formants of vocalic hesitations tended to be more centralized or posterior and opened than related phonemes. However the degree of centralization and aperture is dependent on speakers. In the future, we will study in-depth this point with more speakers.

5. Acknowledgements

This research was supported by the European Regional Development Fund (ERDF) through the Estonian Center of Excellence in Computer Science (EXCS) and the Estonian Ministry of Education and Research target-financed research theme No. 0140007s12.

6. References

[1] K. Müürisep and H. Nigol, "Shallow Parsing of Transcribed Speech of Estonian and Disfluency Detection", *Human Language Technology*, Springer Verlag, pp. 165–177, 2009.

[2] I. Vasilescu et al., "Vocalic hesitations vs vocalic systems: a cross-language comparison", *Proc. of ICPhS*, Saarbrücken, 2007.

[3] P. Lippus, "The acoustic features and perception of the Estonian quantity system, Ph.D. dissertation", Tartu University, 2011.

[4] P. Boersma and D. Weenink, "Praat: doing phonetics by computer", [Computer program], http://www.praat.org/.

[5] E.L. Asu and P. Teras, "Estonian", *Journal of the International Phonetic Association* 39, pp. 367–372, 2009.

[6] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.or, 2012.

Self-monitoring as reflected in identification of misspoken segments

Sieb Nooteboom & Hugo Quené

Utrecht Institute of Linguistics IuL OTS, The Netherlands

Abstract

Most segmental speech errors probably are articulatory blends of competing segments. Perceptual consequences were studied in listeners' reactions to misspoken segments. 291 speech fragments containing misspoken initial consonants plus 291 correct control fragments, all stemming from earlier SLIP experiments, were presented for identification to listeners. Results show that misidentifications (i.e. deviations from an earlier auditory transcription) are rare (3%), but reaction times to correctly identified fragments systematically reflect differences between correct controls, undetected, early detected and late detected speech errors, leading to the following speculative conclusions: (1) segmental errors begin their life in inner speech as full substitutions, and competition with correct target segments often is slightly delayed; (2) in early interruptions speech is initiated before competing target segments are activated, but then rapidly interrupted after error detection; (3) late detected errors reflect conflict-based monitoring of articulation or monitoring overt speech.

1. Introduction

Research reported in [2] and [4] on the articulation of speech errors elicited with metronome-controlled tongue twisters demonstrates that many segmental speech errors are not full, categorical, substitutions of one phoneme segment by another, but rather articulatory blends of two competing segments. In [4] the authors assume that in inner speech a segmental speech error arises when two competing segments are activated for the same slot, and that the activation of both competing segments is passed on to articulation, leading to conflicting articulatory gestures. From these findings we conclude that very likely misspoken segments carry acoustic and perceptual consequences of their discordant articulatory origin. In [9] it is argued that these properties of speech errors cannot be studied very well by means of auditory transcription, because, due to limitations of perception, articulatorily ambiguous speech sounds are either perceived as the misspoken segments or as the correct targets. However, very likely another measure of perceptual differences, such as reaction time in an identification task, would be sensitive enough to show subtle differences in perceptual clarity originating from competing articulatory gestures.

The basic idea of the research reported here is that potentially properties of self-monitoring for speech errors can be investigated indirectly through having listeners identify misspoken segments and their correct controls as elicited in SLIP experiments. Perceptual unclarity of these segments resulting from conflicting articulatory gestures would be reflected in (a) number of misidentifications and (b) average reaction times. The relation between perceptual clarity and self-monitoring can be established because of each misspoken segment it is known whether the speech error was or was not detected and repaired by the speaker, and, if detected, whether detection was early, as in *boo .. good beer* or late as in *bood geer ... uh good beer*. Because early detected errors in initial consonants very often are repaired when only part of the following vowel has been spoken, all speech sounds to be compared are to be presented in brief CV speech fragments. We have the following predictions:

(1) **Speech fragments excised from elicited segmental errors will on average have more misidentifications and longer reaction times than speech fragments from correct controls**.

This is because speech errors suffer from articulatory blending and correct controls do not.

(2) **Speech fragments excised from undetected errors will on average have more misidentifications and longer reaction times than speech fragments from detected errors.**

This prediction is based on the assumption that perceptually clear misspoken segments are easier to detect for the monitor as errors than perceptually unclear misspoken segments. This assumption follows from the comprehension-based monitor proposed in [3] plus the idea that error detection is a function of comparing the error form with the correct target form [7]. Recently a theory of production-based conflict monitoring is proposed, where the probability of error detection increases with the amount of conflict between competing segments [8]. From this theory one would make the inverse prediction.

(3) **Speech fragments excised from late detected errors will on average have more misidentifications and longer reaction times than speech fragments from early detected errors**.

The reason for this prediction is that in early detected speech errors (*boo ... good beer*) speech is initiated before the monitor has resolved the conflict between the two competing segments (cf. [7]: error detection follows speech initiation). This suggests that activation of the correct target segment often comes slightly later than activation of the error segment, and may come too late to have much effect on articulation before speech is initiated. In late detected errors both competing segments would be fully activated before activation is passed on to articulation and speech is initiated. Therefore speech will be often affected by conflicting articulatory gestures, and resulting speech sounds will suffer from perceptual unclarity,

From these three predictions it follows that we expect the following order of our four conditions in increasing number of misidentifications and increasing reaction times:

1) Speech fragments from correct controls

2) Speech fragments from early detected errors

3) Speech fragments from late detected errors

4) Speech fragments from undetected errors

2. Method

We selected 291 speech errors on initial consonants from the speech errors elicited with the SLIP technique [1] in two experiments described in [7]. The only selection criterion was that each of these speech errors had a correct control where no speech error was elicited, spoken by the same speaker. Thus we also had 291 correct controls. The 291 speech errors were either not detected and repaired by the speaker (158 cases), or early detected and repaired (as in *boo ... good beer*; 80 cases), or late detected and repaired (as in *bood geer ...uh... good beer*; 53 cases). From each speech error and each correct control a speech fragment was excised containing the initial consonant and 40 ms of the following vowel. All 582 speech fragments were presented over headphones to 21 listeners in a simple open response identification task, in which listeners were asked to press the key on a keyboard corresponding to the consonant sound they heard. Misidentifications defined as deviations (other than those stemming from misperception of voiced/voiceless) from an auditory transcription of the same material described in [7] were assessed for all 12222 responses and reaction times (RTs) were assessed for correctly identified speech fragments, and only for those stimuli that had less than 4 misidentifications. Misperceptions of voice were not counted as misidentifications because the voiced/voiceless feature in initial position is not very robust in Dutch, and misperceptions of voice in meaningless speech fragments are unpredictable.

3. Results

Figure 1 gives the relative frequencies of expected and observed misidentifications in the four conditions formed by the origin of the speech fragments.

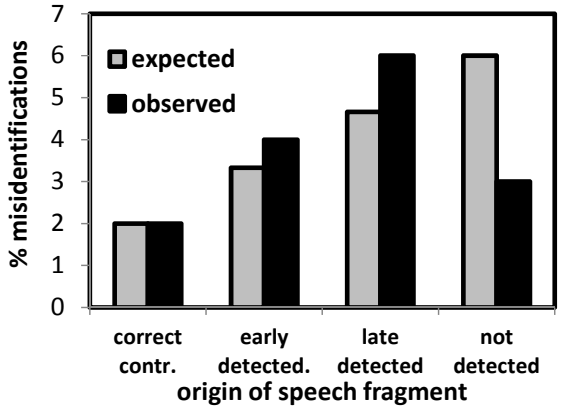


Figure 1: *Relative frequencies (in %) of expected and observed misidentifications of speech fragments for the four origins of speech fragments.*

The expected values were calculated by taking the range in the observed values, running from 2% to 6%, and dividing the other intervals equally, following the predicted order. Obviously, the order of observed values does not follow prediction, mainly because fragments from undetected errors have less, not more misidentifications than those from early and late detected errors. The responses were analyzed by means of mixed-effects logistic regression models with misidentification as a binomial dependent variable [10] and three planned orthogonal contrasts, viz. correct controls versus speech errors, undetected versus detected speech errors, and early versus late detected speech errors. The results of the best fitting model shows that, as predicted, fragments from speech errors have more misidentifications than fragments from correct controls ($p < .0001$), that fragments from undetected errors have less misidentifications than fragments from detected errors ($p < .0085$), and that late detected errors tend to have more misidentifications than early detected errors ($p < .0624$).

Figure 2 presents expected and observed average RTs for the four origins of speech fragments. Only RTs were included for correctly identified speech fragments, and only for those stimuli having not more than 3 misidentifications. Also 29 outliers were removed. There remained 11197 cases, 92% of all responses.

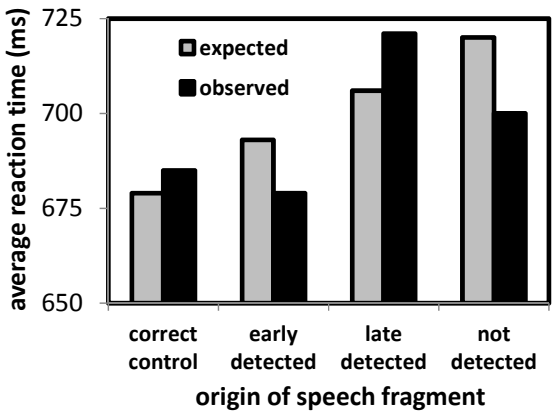


Figure 2: *Expected and observed RTs.*

Predicted values were calculated by starting from the range of observed values, running from 679 ms to 721 ms, and dividing the other intervals equally, following the predicted order. The observed order of average RTs does not follow the predicted order. After RTs were log-transformed to obtain a more normal distribution they were analyzed by means of mixed-effects regression models (LMM) with two crossed random effects, viz. pairs of stimuli (speech error paired with matching control) and listeners [10]. Again there were three planned orthogonal contrasts, viz. correct controls versus speech errors, undetected versus detected speech errors, and early versus late detected speech errors. The best fitting model did not include interactions. The model showed that there was no significant difference between fragments from correct controls and fragments from speech errors, no significant difference between detected and undetected errors, and a significant difference between early and late detected errors. Post hoc comparisons showed that fragments from early detected errors gave significantly *shorter* RTs than correct control fragments ($p < .0182$) and than fragments from undetected errors ($p < .0295$), and that fragments from late detected errors lead to significantly *longer* RTs than fragments from undetected errors.

4. Discussion

Systematic differences in percentages misidentifications and in RTs in a simple identification task with meaningless CV speech fragments can only reflect systematic differences in perceptual clarity of these speech fragments and in particular of the identified consonant segment. The basic assumption underlying the current research, based on the findings in [2] and [4] is that such differences in perceptual clarity of initial consonantal segments are caused by differences in the amount of articulatory ambiguity caused by conflicting articulatory gestures.

The percentages of misidentifications, interpreted in this way, suggest that, as one would expect from the results in [2] and [4], consonant segments resulting from speech errors on average suffer more from articulatory ambiguity than consonant segments in correct controls. They also suggest that consonants in detected speech errors suffer more from articulatory ambiguity than those in undetected speech errors. The effect is particularly strong for late detected speech errors, as if in these latter cases articulatory ambiguity positively correlated with probability that the segmental errors were detected and repaired in self-monitoring. If so, this is not predicted by the perception-based monitor comparing error form with the correct target [3], [7]. The result rather supports a theory of production-based self-monitoring in which the probability of error detection by the monitor increases with the amount of conflict between competing segments [8]. The consonant segments made in early detected speech errors apparently suffer on average less from articulatory ambiguity than segments made in late detected errors. This is in line with our suggestion that in early detected speech errors often speech is initiated before the competing correct target segment is fully activated. This would decrease the probability of articulatory conflict between the two segments and thus reduce the probability of perceptual unclarity caused by articulatory conflict.

It is noteworthy that the above interpretation of the relative numbers of misidentifications is based on only a tiny fraction of all responses. Only 3% of all responses deviated from the auditory transcription made by a single trained phonetician as described in [7] This means that such auditory transcriptions are fairly reliable as descriptions of what people actually perceive when hearing segmental speech errors. But this also means that we should be careful not to draw definitive conclusions from the handful of misidentifications. They are not necessarily representative of how the perceptual clarity of speech segments depends on their origin as correct controls, undetected, early detected and late detected speech errors. More weight should be given to RTs in the identification of those speech fragments that were not misidentified, and did not come from stimuli with more than 3 misidentifications. The reader may recall that the analysis of RTs was carried out on 92% of all responses. Here we assume that differences in RT reflect differences in perceptual unclarity that in turn stem from differences in the amount of articulatory conflict.

An important result is that identification of speech fragments from early detected consonantal errors does not take more but less time than identification of speech fragments from correct controls. This can only mean that on average these segments do not suffer at all from articulatory ambiguity. We interpret this as meaning that (a) speech errors start their life in internal speech as full substitutions of one segment by another, and that (b) often speech is initiated, i.e. activation is passed on to the articulators, immediately after activation of the error segment and before activation of the correct target segment. This is a reasonable hypothesis if we assume that the segment that is most strongly activated also arrives slightly earlier in the speech plan. Early detected speech errors are then precisely the cases where the error segment is more strongly activated than the correct target segment. When, slightly later, the correct target segment is also activated, the monitor can detect the error either on the basis of the amount of conflict between the two segments, or on the basis of perception-based comparison between the two segments. If the error is detected, the speech that is already initiated will be interrupted and a repair will be made. This nicely explains the frequent occurrence of early interruptions both in spontaneous speech errors [5] and in speech errors elicited with the SLIP technique [6], [7].

A second important result is that identification of speech fragments from late detected errors takes considerably more time than identification of speech fragments from early detected and undetected speech errors. We would not expect this on the basis of perception-based comparison between error form and correct target form, because a difference would be most easily detected when the perceptual distance is optimal. Our interpretation of the long reaction times for late detected errors is that these speech errors suffer relatively strongly from articulatory ambiguity and that the amount of articulatory conflict either directly or indirectly has increased the probability of error detection by the monitor. Directly if articulation is monitored for the amount of conflict between competing articulatory gestures, indirectly when the overt speech is monitored auditorily for unclear segments.

5. Conclusions

Speech errors start their life in inner speech as full categorical substitutions. These surface in overt speech in early interruptions where error detection follows speech initiation. Competition between segments surfaces as perceptual unclarity due to articulatory conflict. Probability of error detection increases with amount of conflict.

6. References

[1] B.J. Baars and M.T. Motley. “Spoonerisms: Experimental elicitation of human speech errors”. Journal Supplement Abstract service, Fall 1974. Catalog of selected documents in Psychology 3, pp. 28–47, 1974.

[2] L. Goldstein, M. Pouplier, L. Chen, E. Saltzman and D. Byrd. “Dynamic action units slip in speech production errors”. *Cognition* 103, pp. 386–412, 2007.

[3] W.J.M. Levelt, A. Roelofs and A. S. Meyer. “A theory of lexical access in speech production”. *Behavioral and Brain Sciences* 22, pp. 1–75. 1999.

[4] C.T. McMillan and M. Corley. “Cascading influences on the production of speech: Evidence from articulation". *Cognition* 117, pp. 243–260, 2010.

[5] S.G. Nooteboom. “Listening to one-self: Monitoring speech production”. In R. Hartsuiker, Y. Bastiaanse, A. Postma and F. Wijnen (Eds.), *Phonological Encoding and Monitoring in Normal and Pathological Speech*, pp. 167–186, 2005.

[6] S.G. Nooteboom. “Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech”. *Speech Communication* 47, pp. 43–58, 2005.

[7] S.G. Nooteboom and H. Quené. “Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors”. *Journal of Memory and Language* 58, pp. 837–861, 2008.

[8] N. Nozari, G. Dell and M. Schwartz. “Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production”. *Cognitive Psychology* 63, pp. 1–33, 2011.

[9] M. Pouplier and L. Goldstein. “Asymmetries in the perception of speech production”. *Journal of Phonetics* 33, pp. 47–75, 2005.

[10] H. Quené and H. Van den Bergh. “Examples of mixed-effects modeling with random effects and with binomial data”. *Journal of Memory and Language* 59(4), pp. 413–425, 2008.

Catogorizing syntactic chunks for marking disfluent speech in French language

Klim Peshkov, Laurent Prévot, Stéphane Rauzy & Berthille Pallaud

[1] Aix-en-Provence Université, Laboratoire Parole et Langage,
5 avenue Pasteur, Aix-en-Provence, France

Abstract

Disfluency is the first phenomenon one has to address when processing spontaneous speech. Efficient systems combining transcription-based and signal-based cues have been created for English. These systems generally use supervised machine learning models, trained over large annotated datasets combining signal and transcription. As for other languages, including French, the situation is complicated by the lack of resources. A few proposals based on filled pauses, truncated words and repetitions have been made for identifying disfluencies in French. In this paper, we propose a transcription-based approach to this task, with high-quality morpho-syntactic tags as input for identifying disfluent areas. Originally, we adopted a transcription-based approach for obtaining an independent way of characterizing disfluencies. This can be later compared and combined with prosodic cues. Our method consists in building syntactic chunks from our tagging and then classify these chunks into several categories, some of them being considered as disfluent. We apply our method to speaker style characterization, discourse genres zoning, as well as to dataset cleaning. Finally, an attempt is made to relate our disfluent chunks to a more standard description of disfluencies in order to open the way of a deeper integration of our work with the one of the *disfluency* community.

Index Terms: tagging, chunking, transcription-based approach, disfluencies, speaking style.

1. Introduction

Disfluencies are not our primary research objective, but we are facing them in various aspects of our research such as syntactic parsing of spontaneous speech, prosodic phrasing or discourse segmentation [1, 2].

There has been a substantial number of works on disfluency in English. First systematic study of the phenomenon was performed by Shriberg [3]. Recently, Besser provided a more fine-grained classification system for the disfluencies [4]. Virtually all disfluency detection systems make use of statistical modeling approaches, although some authors combine machine learning techniques with rule-based approaches [5, 6].

As for the knowledge sources, most of the studies work either exclusively on transcripts (or speech recognition outputs) [7, 8, 9, 10] or on multiple knowledge sources [6, 11, 5]. However, some authors have tested prosody-only approaches [12, 6]. In the comparison of disfluency detection systems presented in [6], the system based only on text-based cues outperforms prosody-only system, while the combination of prosody and text performs even better.

Works on disfluency in French are less numerous. A number of descriptive studies exist [13, 14]. Automatic detection of disfluencies applied to a highly-specialized domain of aeronautical communication was addressed by [15].

We do not offer here a new account on this well studied phenomenon. We are only trying to use all the information we have at the transcription level to identify disfluent areas.

Concerning the data, we work on a conversational corpus of long conversations (1 hour) in which two speakers tell each other personal stories [16]. Given the nature of the task, the fact that the pairs of interlocutors are good friends and the duration of the recording, this corpus is heavily loaded with disfluencies.

These disfluencies complicate most of the basic tasks we perform on this corpus, including, in particular, syntactic parsing, prosodic phrasing, and discourse segmentation. However, we have a high-quality morpho-syntactic annotation. The paper is an exploration of how close we can get to the disfluencies using only transcription-based sources of information, but integrating richer information than filled pauses, truncated words and repetitions as in [14]. The reason of not using signal-based cues is twofold. First, we would prefer to have an orthogonal characterization which we could later compare and combine with prosodic cues. Second, at this stage we do have a high quality transcription and tagging for the whole corpus while our prosodic analyses are partial and not as reliable. Therefore, for the time being, we prefer to rely exclusively on the transcription. We believe that it is a frequent scenario in the development of spoken language resources.

The paper is structured as follows. We will start by presenting how our chunks are created in section 2. Then, in section 3 we will propose a classification of these chunks that we consider to be interesting for approaching disfluent speech. Section 4 will apply our categorization to several sub-tasks related to disfluent speech. Finally before concluding, we will discuss the relations between our analyses and more standard accounts of disfluencies (section 5).

2. Building chunks

2.1. Morpho-syntactic tagging

The morpho-syntactic tags of the transcription have been obtained in three steps. In a first step, the enriched orthographic transcription has been filtered of annotations not containing syntactic content (filled pause, hesitation, truncation, laughter, ...). This filtered input was then proposed to a tagger [1]. This tagger is a stochastic tagger trained on written French texts. The morpho-syntactic information has been organized in an ad-hoc way in 50 tags. On this tagset, the performance of our tagger for written French is good (a F-measure score of 0.975 is obtained for the tagger output, version 2011). The tagger was slightly modified to account for the absence of punctuation marks in the input transcription. It was therefore allowed to the tagger to insert punctuation marks when appropriate (i.e. when this insertion increases the probability of the sequence of tags treated).

In a second step, an error analysis was performed on the output tags. Unknown words (i.e. the words not present in our initial lexicon) were listed and added to a lexicon specific to spontaneous speech. In our french corpus, the

more frequent phenomena are word reductions (e.g. *appart* for *appartment*), regional version or foreign words, and onomatopoeia. By comparing the lexical frequencies of our corpus versus their written French counterparts, we also established a list of potentially problematic words for the tagging task. Among those, oral discourse markers (e.g. *quoi*, (*what*), *enfin* (*in the end*), *bon* (*well*), *en fait* (*in fact*),…) were identified and their entries in the lexicon were modified. We associated the morpho-syntactic tag *Interjection* when these words are used with their discourse marker function.[1] This last modification concerns almost 10% of the tokens of the whole corpus.

In a final step, the new version of the tagger was applied to the filtered transcription inputs. From this new tagged output, a manual correction was performed on the 115, 000 tokens of the whole corpus. This manually corrected version was afterwards used as a gold standard to evaluate the performance of our tagger adapted for spontaneous speech transcription input. We obtained a F-measure of 0.948 which is a very good score, considering that no special treatment had been applied for dealing with disfluency phenomena.

2.2. Creating chunks

Analysis in chunks is an easy-to-implement and robust method for shallow syntactic analysis [17]. The main principle of chunking consists in including in one unit all the constituents situated to the left of each syntactic head. These units may be helpful for the detection of disfluencies in French, because some recurrent kinds of disfluencies will give rise to chunks with unusual morpho-syntactic patterns. We would like to stress that we do not plan to capture all kinds of disfluences using chunks, but only those that give rise to local perturbations of syntactic structure. For example, a very common disfuency in our corpus is a sequence of function words (1) or discourse markers (2) before a lexical word.

(1) les les les gens
the the the people

(2) ouais ben ouais je crois
yeah well yeah I thinks

Chunking our data is performed by a script using 26 rules. The rules are of two types. The first type specifies tags which are always added in the same chunk as the following token as illustrated in (3). Most of the function words belong to this category.

(3) determiner (*D*) + anything
preposition (*S*) + anything
conjunction (*C*) + anything
personal pronoun (*Pp*) + anything
auxiliary verb (*Va*)+ anything

The second type of rules specifies an ordered pair of POS tags which must be in the same chunk (4).

(4) adjective (*A*) + noun (*N*)
demonstrative pronoun (*Pd*) + verb (*V*)
adverb (*R*)+ adjective (*A*)
proper noun (*Np*) + proper noun (*Np*)
verb (*V*) + verb (*V*)

Any number of filled pauses and truncated words can appear inside a chunk if the rules specify that the last word before these elements must be in the same chunk as the first word after them. This means that the sequence *"Determiner (D) – filled pause (F P) – filled pause (F P) – Noun (N)"* will give rise to a single chunk: *DN* . Otherwise, they are attached to the beginning of the next chunk. The same approach is applied to the treatment of pauses inferior to 200 milliseconds. Chunks cannot span across pauses which length is above this threshold.

3. Categorizing chunks

3.1. Comparison spoken vs. written chunks

To classify our chunks we first compare them across corpora type by calculating their distribution both in spoken corpus and in a written corpus [18]. More precisely we computed a "spokenhood" ρ ratio for each chunk type as follows:

$$\rho_i = \frac{Spoken\ Frequency(i)}{Written\ Frequency(i)}$$

We split the list of chunks' types according to this ratio as illustrated in Table 1:

- $\rho \rightarrow \infty$: Appear in spoken data only, many disfluency related patterns should be there (DISFLUENT)
- $\rho > 17$:[2] Majority in spoken data (SPOKEN)
- $\rho < 17$: Majority in written data (CANONICAL)

Table 1: *Chunk comparison Spoken/Written*

Chunk	ρ	Example
I I I Pp V	∞	ouais ben ouais j' imagine yeah well yeah I imagine
I D N	109.75	bon ce truc well that thing
D A N	0.49	un vrai conflit a real conflict

3.2. Morpho-syntactic classification

The next step consisted in refining these categories based on the inner structure of the chunks. We started by simplifying the chunk tag set by applying a few rules (in the order presented below). Some of the rules are simplifications of some phrase structures, others are regular expression-style rules.

- C ⇝ I: assimilate conjunctions to interjection for the sake of simplicity since in this context they play a very similar role
- {V Vi}, {Va V} ⇝ V : simplification of verb structures, (*V* = Verb ; *Vi* = Infinitive verb ; *Va* = auxiliary verb);
- {Pp Px}, {Pp Po} ⇝ P: simplification of pronoun structures, (*Pp* = Personal Subject Pronoun ; *Px* = Reflexive pronoun ; *Po* = Personal Object Pronoun);
- {X X} , {XXX} ⇝ X+: simple regular expression patterning.

[1] The tag *Interjection* was found convenient because discourse markers, similarly to interjections, have no real syntactic function.

[2] The score of 17 has been decided after qualitative evaluation of the patterns.

Once these simplifications done, we manually tagged the 50 most frequent patterns of the SPOKEN and DISFLUENT categories and identified the following subcategories:

1. SPOKEN CANONICAL: It is a spoken form but there is no disfluency there (4.67% of the total number of chunks)
2. HESITATION: I+ Canonical, that is a sequence of discourse markers or conjunctions followed by a CANONI- CAL chunk (4.48%)
3. BC: Backchannels, defined as sequence of discourse markers, surrounded by silent pauses longer than 200 milliseconds (16.26%)
4. DM: Discourse markers [3]: I+ (6%)
5. INCOMPLETE: Chunks without a head (2.44%)
6. EXCESSIVE: Chunks with abnormally long patterns (0.93%)
7. RARE: Other chunks not belonging to any of the above categories. They are considered to be disfluent (0.03%)
8. FP/WF: This category is created automatically, including all chunks containing a filled pause or a word fragment (8.85%)

The morpho-syntactic analysis proposed that 12.24% of the chunks of our corpus were disfluent. For a comparison [3] reports 6,4% of disfluent tokens in the Switchboard corpus of spontaneous speech. It is quite comparable if we remember that our chunk are a little bigger than the tokens. Here we consider a chunk as disfluent if it belongs to one of the following categories: INCOMPLETE, EXCESSIVE, RARE or FP/WF. Non-inclusion of HESITATION, DM and BC into the definition of disfluency is motivated by the fact that from the perspective of our syntactic parser these categories do not break the syntactic structure.

3.3. Evaluation

The first evaluation we propose is a simple F-score measure against a manual annotation realized by one of the authors. The manual annotation was however at the token level while ours is at the chunk level. Therefore, for each proposed disfluent chunk we check whether the reference contains at least one disfluent token. The results are presented in Table 2.

Table 2: *Evaluation against manually annotated data.*

Version	Precision	Recall	F-score
Baseline (FP and WF)	72.5%	47.4%	57.3%
Morphosyntax alone	69.7%	56.2%	62.2%
Together	70.2%	57.7%	63.3%

Our baseline was obtained by taking *F P* and *W F* chunks (chunks including a filled pause or a word fragment) as disfluents.[4] The low recall when we used only morpho-syntactic cues, can be explained with the fact that, as we mentioned above, this method does not allow detecting disfluencies above the chunk level. Adding morpho-syntactic cues results in a significant improvement over the baseline. Moreover, the precision score is more important for our applications, such as removing disfluent zones from prosodic analyses. Indeed, we prefer not to have the entire set of disfluencies detected than to throw away too much clean data, mistakenly labeled as disfluent.

[3] DM includes discourse markers but also conjunctions because of one of our simplification rules.

[4] We also tried to include verbatim repetitions but it decreased the performance (drop of precision).

4. Applications

As said in the introduction, our research topic is not disfluencies *per se*, but rather their applications. In this section we present three applications relevant for our purposes.

4.1. Speaker style characterization

First of all, we are interested in speaker variability for a number of our studies. Although we have intuition about the fluency of all our speakers, we would like to be able to have several independent measures to describe their speaking style. (Dis)fluency is one of these dimensions. Such characterization will be used later for re-investigating results at different levels of analysis (in particular, phonetics, prosody). In Table 3 we present a comparison of our disfluency rate with an intuitive evaluation of the speakers performed by a linguist before our analysis[5] (*Flu-1* and *Flu-2* were characterized as fluent by the expert, *Dis-1* and *Dis-2* as disfluent). The distribution of canonical and disfluent chunks seems to coincide with the intuition of the expert. Therefore, our method may be useful to characterize speaker styles. However a multi-expert characterization and deeper statistical analyses are needed to confirm this.

Table 3: *Disfluency rate by speaker.*

Speaker	Flu-1	Flu-2	Dis-1	Dis-2	Avg
Canonical	60.7%	63.2%	49.8%	51.0%	55.9%
Disfluency	9.6%	6.8%	14.4%	11.4%	12.3%

4.2. Discourse activities segmentation

We also looked at the distribution of our categories across the discourse activities of the corpus (narrative sequences alternating with very open conversation). Our hypothesis was that narrative sequences are a little better planned and therefore should be less disfluent than the free conversation. However, we could not confirm this. Only a few individual speaker strategies seems to surface but not all in the same direction. Some speakers seem to be more disfluent while narrating while for some others it is the free conversation which is more disfluent. Indeed, a deeper qualitative assessment showed us that speakers have various strategies to deal while alternating these two discourse activities. The only tendency, clearly emerging from narrative versus non-narrative comparison, is the difference in BC category rate, which is obviously lower in the narratives: 0.75% in the narratives and 6.19% in the non-narrative zones.

5. From chunks to disfluencies

We would like to take advantage of our chunk categorization for performing a disfluency tagging more in line with standard accounts such as [3, 19, 4] . We present here structure of some detected disfluencies belonging to three big classes of disfluences from [4]: *Uncorrected*, *Deletable* disfluences and *Revision*.

Here in the Table 4, in the cases of *Uncorrected* and *Deletable* disfluencies, the system marked as disfluent the reparandum (RM), which is convenient for subsequent correction. The reparandum of the uncorrected disfluence is the chunk “sur”, labeled as INCOMPLETE.

[5] Unfortunately, only one expert has done this intuitive evaluation forbidding inter-coder agreement. However, the speakers presented in the table are the ones that, according to the expert, have a clear fluent / disfluent speech style.

The reparandum of the *Deletable* disfluence, "tout ça" belongs to DM category. However, sometimes the chunks are bigger than the disfluence and consequently some material before the reparandum and after the reparans (RS) can be marked as disfluent. The example of *Revision* is in fact a single chunk of the type EXCESSIVE.

Table 4: *Examples of detected disfluencies.*

Left context	RM	IM	RS	Right context
Uncorrected y a un truc there's something	sur about			(pause)
Deletable un peu stressée a little nervous	tout ça and all that			
Revision dans in	un a		une a	auberge hostel

6. Conclusion and future work

Although our results on disfluency marking are only preliminary, we believe that it is worth to develop our methodology further. First of all, there are some aspects we did not considered for this first study: repetitions and disfluent pauses. These categories are not very difficult to include in our approach. Overall, the contribution of the morpho-syntactic patterns seem thin, but our conclusion is that if one wants to go beyond pure (shallow) lexical approaches, our approach can be used to improve the results without including signal-based features.

A lot of the missed cases are not detected simply because some disfluencies occur on a longer stretches of text than chunk. One way to deal with this higher-level disfluencies is to detect lexical repetitions with a risk of introducing too much noise. But the repetition is not the only cue that can help in detecting them. Trying to use the syntactic information more efficiently, one can think of a slightly more complex improvement of the detection system. Each chunk has to be categorized according to the part of speech of its head into verbal chunks, noun chunks etc. After that we can try to define potentially disfluent sequences. Verbal chunks may be characterized with the information about verb valency. This would give us information about the well-formedness of the syntactic structure across the verbal chunk and its neighbours.

Finally, we would like to follow [10] and apply a Transformation-Based Learning approach [20] on our data and compare the results with the ones presented here.

7. References

[1] S. Rauzy and P. Blache, "Un point sur les outils du lpl pour l'analyse syntaxique du français", in *Workshop ATALA: Quels analyseurs syntaxiques pour le français ?*, Paris, France, 2009.

[2] K. Peshkov, L. Prévot, R. Bertrand, S. Rauzy, and P. Blache, "Quantitative experiments on prosodic and discourse units in the corpus of interactional data", in *Proceedings of SemDial 2012: The 16th Workshop on the Semantics and Pragmatics of Dialogue*, 2012.

[3] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California, 1994.

[4] J. Besser and J. Alexandersson, "A comprehensive disfluency model for multi-party interaction," in *Proceedings of the 8st SIGdial Workshop on Discourse and Dialogue*, 2008, pp. 182–189.

[5] S. Germesin, T. Becker, and P. Poller, "Hybrid multi-step disfluency detection", *Machine Learning for Multimodal Interaction*, pp. 185–195, 2008.

[6] Y. Liu, E. Shriberg, and A. Stolcke, "Automatic disfluency identification in conversational speech using multiple knowledge sources", in *Proceedings Eurospeech*, vol. 1. Geneva, Switzerland, pp. 957–960, 2003.

[7] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies", in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, pp. 405–408, 1996.

[8] P. A. Heeman and J. F. Allen, "Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue", *Computational Linguistics*, vol. 25, no. 4, pp. 527–571, 1999.

[9] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech", in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, pp. 1–9, 2001.

[10] M. Snover, B. Dorr, and R. Schwartz, "A lexically-driven algorithm for disfluency detection", in *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, pp. 157–160, 2004.

[11] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Comparing hmm, maximum entropy, and conditional random fields for disfluency detection", in *Proceedings of the INTERSPEECH '05*. Citeseer, 2005.

[12] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proc. Eurospeech*, vol. 5, 1997, p. 2383–2386.

[13] S. Henry and B. Pallaud, "Word fragments and repeats in spontaneous spoken French", in *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*, pp. 77–80, 2003.

[14] P. B. d. Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda, and P. Paroubek, "A quantitative study of disfluencies in French broadcast interviews," in *Disfluency in Spontaneous Speech*, pp. 27–32, 2005.

[15] J.-L. M. Bouraoui, "Analyse, modélisation, et détection automatique des disfluences dans le dialogue oral spontané contraint: le cas du contrôle aérien", Ph.D. dissertation, Université Paul Sabatier-Toulouse III, 2008.

[16] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID – corpus of interactional data – annotation et exploitation multimodale de parole conversationnelle," *Traitement Automatique des Langues*, vol. 49, no. 3, pp. 1–30, 2008.

[17] S. Abney, "Parsing by chunks", *Principle-based parsing*, vol. 44, pp. 257–278, 1991.

[18] G. Adda, J. Mariani, P. Paroubek, M. Rajman, and J. Lecomte, "L'action GRACE d'évaluation de l'assignation des parties du discours pour le français", *Langues*, vol. 2, no. 2, pp. 119–129, 1999.

[19] E. E. Shriberg, "Phonetic consequences of speech disfluency," DTIC Document, Tech. Rep., 1999.

[20] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging", *Computational Linguistics*, vol. 21, no. 4, pp. 543–565, 1995.

Acoustical characterization of vocalic fillers in European Portuguese

Jorge Proença[1], Dirce Celorico[1], Arlindo Veiga[1,2], Sara Candeias[1] & Fernando Perdigão[1,2]

[1] Instituto de Telecomunicações, Coinbra, Portugal
[2] Electrical and Computer Engineering Department, University of Coimbra, Portugal

Abstract

This study attempts to acoustically characterize the most common filled pause vocalizations (or vocalic fillers) in spontaneous speech in European Portuguese: the near-open central vowel [ɐ] and the mid-central vowel [ə]. For this purpose we analyzed the spectral information of the vocalic fillers by estimating their first two formant frequencies as well as their duration properties. The vocalic fillers are taken from a large corpus of European Portuguese broadcast news' speech. We also compared the vocalic fillers with lexical vowels possessing similar timbre. No formant variation trend was attained for the vocalic fillers and a great overlap of formant values is observed. These results provide a base of information for understanding the most common vocalic fillers in European Portuguese spontaneous speech.

Index Terms: filled pauses, vocalic fillers, formant estimation, spontaneous speech, hesitations.

1. Introduction

The interest in studying events that characterize the spontaneity of the speech has been increasing as the development of speech technologies grows. In this context, several studies on hesitations (so-called disfluencies) as well as vowel reductions have gained importance over the last years ([2,7,5,8], and [17,18] as examples, respectively). Among various hesitation phenomena, such as repetitions, truncated words or word extensions, filled pauses are, the ones more widely encountered in world's languages, mainly on spontaneous speech, [10], [19]. Relatively stable vocalic segments mostly fulfill these pauses. Occurring commonly without any lexical support, we refer to this type of fillers as vocalic fillers (VFs). Representing an insertion at any moment during spontaneous speech, VFs carry multiple functions, such as announcing upcoming discursive topics or planning and delaying speech, [2–4,9].

Although some works on VFs can already be found for Portuguese [6,10,11,13] and other languages [5,12,19], the most part of studies conducted on large spontaneous speech corpora comprises English or French languages. This paper presents a study which attempts to acoustically characterize the two most common vocalic fillers that occur in European Portuguese: the vocalic filler representing a similar timbre to the near-open central vowel [ɐ] and the one close to the timbre of the mid-central vowel [ə].

We chose the two first formant frequencies, F1 and F2, and filler duration as phonetic and prosodic parameters in search of reliable patterns of this type of fillers. Although all the studies mostly agree on characterizing these fillers as long and stable vocalic segments [10,19], their characterization for European Portuguese still needs further study. Additionally, we compare the vocalic fillers with the corresponding lexical vowels (LVs) possessing similar timbre, which are vowels produced in a context of complete words, in a similar way that was done for different languages, such as French, American English and European Spanish [16], [20,21]. Even though it has been noted that vocalic fillers manifest acoustical language-dependent characteristics, in fact they may not be necessarily acoustically equal to the lexical vowels with similar timbre, and, at least in some contexts, they appear to possess slightly different average positions in the triangle vowel area.

On the characterization of vocalic fillers, this study is an extension of our previous study [10] for European Portuguese where fundamental frequency and energy of VFs are presented. A better understanding of the structure of speech as well as insights on how to obtain filler acoustic models for use in automatic spontaneous speech recognition are ultimate goals of this work. We also believe that this work also promotes awareness of vocalic fillers in the phonetic studies of the language.

The rest of the paper is organized as follows. In the next section we briefly describe database selection. In section 3 we show how the estimation of formant frequencies of the sounds [ɐ] and [ə] belonging either to fillers or to lexicon was performed. In section 4 we present the main discussion of the achieved results. Finally, in section 5 the main conclusions are drawn and envisioned work is foreseen.

2. Database

For the present study we used two corpora: a corpus of hesitation events, the HESITA Database [22], and a European Portuguese corpus with no hesitations collected for control.

The corpus of hesitations was used to study filled pauses, which were manually annotated. This corpus comprises 30 daily news programs collected from a European Portuguese television channel podcast (about 27 hours of speech). It consists of 1152 vocalic fillers, exemplifying sounds similar to the near-open central vowel [ɐ] (808) and to the mid-central vowel [ə] (344). These two VFs were the most common in the database and the ones chosen for analysis. The next most common was the nasal [ɐ̃], with 155 occurrences.

The control corpus was used to estimate the acoustic characteristics of the vocalic sounds [ɐ] and [ə] occurring in a context of a complete word (the LVs): e.g. [ɐ] in <para> [pɐrɐ], ('for' in English) or [ə] in <devolver> [dəvolver] ('to give back'). It consists of recordings from 7 European Portuguese native adult speakers of sentences and command words, taken in a small office room. In each recording session, the same common set-up was used, which consists of a laptop computer and three microphones. For each session, a segmentation and phone-level transcription were automatically performed through forced alignment using in-house tools. The total number of extracted vocalic segments was 7426, in which we count 4411 [ɐ] and 3015 [ə].

Table 1 summarizes the overall [ɐ] and [ə] distribution by database and gender.

Table 1: *Relative frequency (#) of VFs and LVs by gender and timbre.*

Type	Gender	#[ɐ]	#[ə]
VFs	Male	605	301
	Female	203	43
LVs	Male	2674	1771
	Female	1737	1244

3. Formant frequencies determination

The first (F1) and second (F2) formant frequencies of the [ɐ] and [ə] vocalic fillers were automatically extracted using the Praat tool [14]. The base recommended ceilings for estimating five formants are 5500 Hz for female speakers and 5000Hz for male speakers but, through observation, these values can't always successfully estimate F1 and F2. A similar method to [15] was then applied, given the foreknowledge that different vowels and speakers need different formant ceilings for the automatic calculation. An iterative calculation of formants was performed in 10 ms steps using ceilings in the 4000–5500Hz range (for males) or 4800–6500 Hz (for females) in 50 Hz steps, followed by a selection of the optimal ceiling. As we do not keep speaker information for most of the news broadcast VFs, they were considered as if each belonged to a different speaker. Therefore, the ceiling that was selected as optimal for a given VF was the one that provided the smallest variance of the F1 and F2 pairs of values of that VF, calculated as the sum of the variances of 20log(F1) and 20log(F2). We empirically observed that wrong ceilings would usually fail on F1 or F2 for a couple of points, leading to sudden jumps and a larger variance than intrinsic F1 and F2 variations. One problem of this method is that a very high ceiling would provide smooth F2 values where the third formant frequency (F3) should be, and pass as optimal; but this was already taken into consideration and countered with the selected ranges of ceilings.

Other restrictions were applied before and after formant calculation. Utterances with high clipping of the audio signal were discarded. For each utterance, only the formant values where the energy level was above 10% of the maximum energy were considered, to specifically remove unvoiced transcription limits. Finally, utterances with highly variant formant values, probably indicating a failure in detecting F1 and F2 were not considered.

The same analysis was conducted for the lexical vowels [ɐ] and [ə] with an additional restriction of only considering segments of duration larger than 50 ms.

4. Results and discussion

After applying the restrictions mentioned in section 3, the number of vocalic segments kept was 520 [ɐ] and 244 [ə] VFs and 1517 [ɐ] and 385 [ə] LVs. A very large number of lexical vowels were cut from analysis; mostly the small-duration or low-energy segments, barely recognized during alignment and more drastically occurring for [ə], as a consequence of the nature of continuous speech that almost eliminates the already reduced vowel [ə] in certain cases. The durations of VFs and LVs are shown in Figure 1 and, as expected, VFs are generally longer.

Figures 2 and 3 place the extracted F1 and F2 mean values on a logarithmical scale for male and female speakers respectively, as usually done in similar cases [15]. The 'triangle' of [i], [ɛ], [a], [ɔ] and [u] vowels come from the median values of F1 and F2 taken from the read speech corpus. These values were calculated in a similar fashion to the method described, although with much more restrictive ceilings. They were included to show the centrality of [ɐ] and [ə]. Averagely, F1 is higher and F2 is lower for VFs than for LVs, but their distributions overlap. LVs show the biggest variances, which could be explained by the high dependence of phonetic context (and related coarticulation phenomenon) in both word and sentence production.

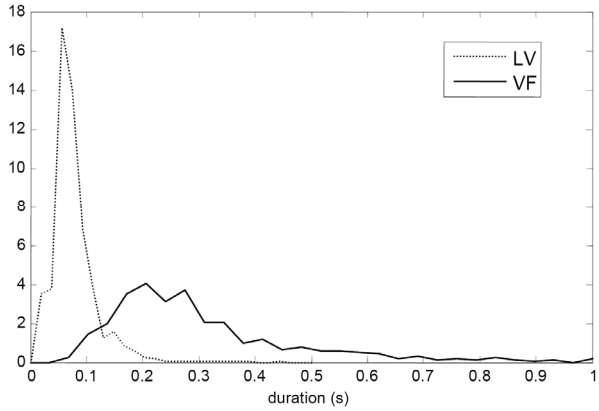


Figure 1: *Normalized histogram of the duration of VFs and LVs.*

The variation trend of F1 and F2 during each vocalic segment was also analyzed. Although their change can be non-linear, a linear fit was applied to each sequence of values and the variation rate was extracted from this fit. Fig. 4 shows these rates for F1 and F2, and it is observable that behaviors are highly variant, either positive or negative. Although the average is for a small negative change, no trend can be discerned. Furthermore, no correlation exists between F1 and F2 simultaneous variation. LVs also prove to be less stable than the noted VFs, as they achieve much higher variation rates.

The presented results also point out that [ɐ] and [ə] are often hard to distinguish. Each speaker could have its own personal preference on how to fill a pause vocalically. Choosing mainly sounds of the central vowels system, speakers appear to adapt the production with their own specific production, possibly even in a middle point of [ɐ] and [ə]. It would be interesting to perform an in-depth perceptual study to confirm that some vocalic fillers can be understood differently with and without context or for different listeners, which could coincide with most of the cases that overlap so far.

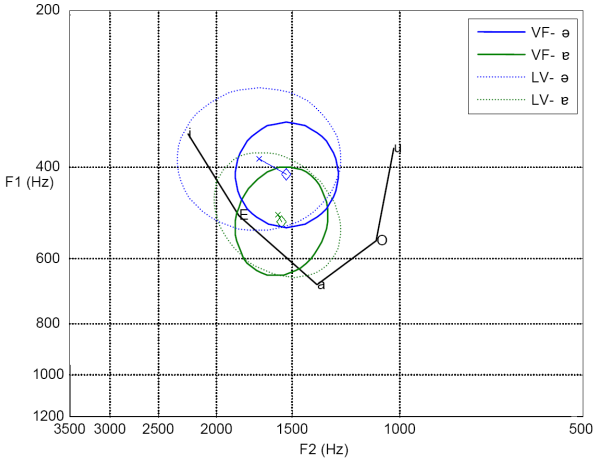


Figure 2: *Male speakers: F1 and F2 means and concentration ellipsoids for [ə] (blue) and [ɐ] (green) for VF and LV, including the male vowel 'triangle' from the LVs database.*

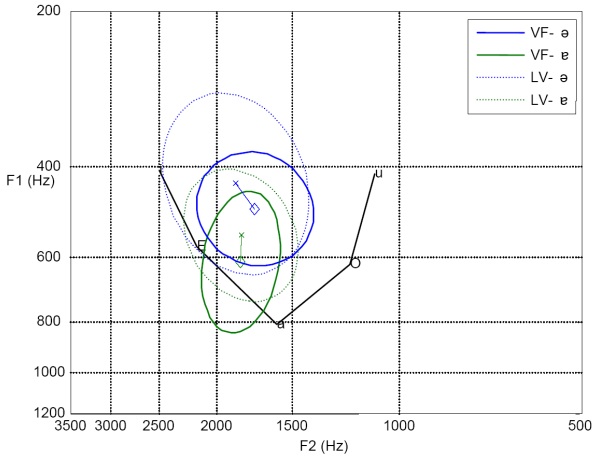


Figure 3: *Female speakers: F1 and F2 means and concentration ellipsoids for [ɐ] (blue) and [ə] (green) for VF and LV, including the female vowel 'triangle' from the LVs database.*

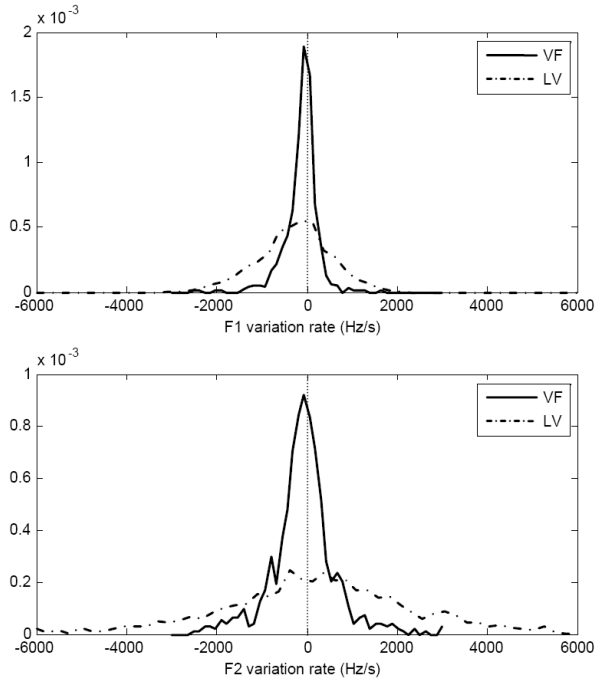


Figure 4: *Normalized histogram of the variation rates of F1 (top) and F2 (bottom) from a linear fit to each utterance, for VFs and LVs.*

5. Conclusions

The most common filled pause vocalizations (or vocalic fillers) in European Portuguese, [ɐ] and [ə], were characterized concurrently with their corresponding intra-word vowel productions (lexical vowels). Vocalic fillers were taken from a large corpus of European Portuguese broadcast news' speech (about 27 hours).

As expected, vocalic fillers are of longer duration and the lexical vowels are very short. Although the average of formant variation is for a small negative change, no specific trend was observed. Still, the variations of F1 and F2 indicate a higher stability of the fillers in comparison to the vowels. As to formant values, there is a small tendency for higher F1 and lower F2 in VFs. Fillers [ɐ] and [ə] are often hard to distinguish and each speaker could have its own specific production and may be strongly dependent on linguistic context. A perceptual study to confirm that some vocalic fillers can be understood differently with and without context or for different listeners would be interesting as future work.

We plan on extending this study to vocalic extensions, which are another class of filled pauses, as they may have similar behaviors in duration and formant frequencies as vocalic fillers. Based on the knowledge attained from this study, we also intend to develop an automatic detector of fillers and extensions from continuous speech.

6. Acknowledgements

This work is funded by FCT and QREN projects (PTDC/CLE-LIN/11 2411/2009; TICE.Healty13842) and partially supported by FCT (Instituto de Telecomunicações multiannual funding PEst-OE/EEI/LA0008/2011).

Sara Candeias is supported by the FCT grant SFRH/BPD/36584/2007.

7. References

[1] W. J. M. Levelt, *Speaking. From Intention to articulation*, Cambridge, Massachusetts, The MIT Press 1993, 1989.

[2] E. Shriberg, "Preliminaries to a theory of speech disfluencies", Ph.D. dissertation, University of California, 1994.

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

[4] H. Moniz, et al., "On Filled Pauses and Prolongations in European Portuguese", in *Proc. Interspeech '07*, ISCA, Antwerp, Belgium, 2007, pp. 2645–2648.

[5] R. Eklund, "Disfluency in Swedish human–human and human–machine travel booking dialogues", PhD dissertation, Institute of Technology, Linköping University, 2004.

[6] A. I. Mata, "Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologia, Resultados e Implicações Didáticas", Ph.D. dissertation, Faculdade de Letras, Universidade de Lisboa, 1999.

[7] M. Candea, "Contribution à l'Étude des Pauses Silencieuses et des Phenomenes Dits «d'Hesitation» en Français Oral Spontané – Étude sur un Corpus de Récit en Classe de Français", Ph.D. dissertation, Universite Paris III – Sorbonne Nouvelle, 2000.

[8] H. Moniz, "Contributo para a Caracterização dos Mecanismos de (Dis)Fluência no Português Europeu", M.S. thesis, Faculdade de Letras, Universidade de Lisboa, 2006.

[9] E. Shriberg, "Spontaneous speech: how people really talk, and why engineers should care", in *Proc. Eurospeech*, Lisboa, 2005.

[10] A. Veiga, et al., "Characterization of hesitations using acoustic models", in *Proc. of the 17th International Congress of Phonetic Sciences, ICPhS XVII*, Hong Kong, pp. 2054–2057, 2011.

[11] A. Veiga, et al., "Towards Automatic Classification of Speech Styles", in *Lecture Notes in Artificial Intelligence (LNAI),* H. Caseli et al. (Eds.), Springer-Verlag Berlin Heidelberg, 2012, 7243, pp. 421–426.

[12] I. Vasilescu, et al., "Perceptual Salience of Language-Specific Acoustic Differences in Autonomous Fillers Across Eight Languages", in *Interspeech '05*, Lisboa, 2005, pp. 1773–1776.

[13] M. J. R. Freitas, "Estratégias de Organização Temporal do Discurso", M.S. thesis, Faculdade de Letras, Universidade de Lisboa, 1990.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer", [Computer program, Version 5.3.42]. Available: http://www.praat.org/, retrieved on 2 March 2013.

[15] P. Escudero, et al., "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese", *J. Acoustical Society of America*, 126(3), pp. 1379–1393, 2009.

[16] I. Vasilescu, et al., “Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech”, in *Proc. Interspeech ’06*, Pittsburgh, PA, USA, pp. 1850–1853, 2006.

[17] S. Candeias, and F. Perdigão, “A realização do schwa no português europeu”, in II workshop on Portuguese description-JDP, 8th Symposium in Information and Human Language Technology (STIL), 2011.

[18] D. Braga, et al., “Back close non-syllabic vowel [u] behavior in European Portuguese: reduction or suppression,” in *Proc. ICSLP*, Seoul, 2001.

[19] R. Eklund, and E. Shriberg, E., “Crosslinguistic disfluency modeling: a comparative analysis of Swedish and American English human–human and human–machine dialogues”, in *Proc. Int. Conf. on Spoken Language Processing*, vol. 6, pp. 2631–2634, Sydney: Australian Speech Science and Technology Association, 1998.

[20] I. Vasilescu, and M. Adda-Decker, “On the Acoustic and Prosodic Characteristics of Vocalic Hesitations across Languages”. Available: http://perso.limsi.fr/Individu/madda/publications/PDF/IOS.pdf, retrieved on 2 March 2013.

[21] M. Candea, et al., “Inter- and intra-language acoustic analysis of autonomous fillers”, in *Proc. DiSS 2005*, Aix-en-Provence, France, pp. 47–51, 2005.

[22] Candeias, S., Celorico, D., Proença, J., Veiga, A. and Perdigão, F., “HESITA(tions) in Portuguese: a database”, *Proc. DiSS 2013*, ISCA endorsed Interspeech 2013 satellite workshop, August 21–23, 2013, KTH Royal Institute of Technology, Stockholm, Sweden.

The linguistic role of hesitation disfluencies: evidence from Hebrew and Japanese

Vered Silber-Varod[1] *& Takehiko Maruyama*[2]

[1] The Research Center for Innovation in Learning Technologies, The Open University, Israel
[2] National Institute for Japanese Language and Linguistics, Japan

Abstract

In this paper we examine a certain aspect of prosody-syntax interface, that of hesitation disfluencies (HD) that occur intra-phrases or intra-morphemes. Such cases were found in two spontaneous corpora of two syntactically distinct languages – Israeli Hebrew (IH) and Japanese. It was found that intra-phrasal hesitations in the two languages calls for different explanations, since in Japanese the noun (e.g., in NP) precedes the case marking particle while in IH the preposition (e.g., in PP) precedes the noun. In this paper we will present qualitative findings and suggest a unified view of the phenomenon of intra-phrasal HDs.

Index Terms: Hesitation disfluency, prosody-syntax interface, Israeli Hebrew, Japanese.

1. Introduction

In the present research we investigate the phenomenon of hesitation disfluencies (HD) in spontaneous Israeli Hebrew (IH) and Japanese. HDs are defined as prosodic manipulation of the speaker, produced by excessive elongation of a word final syllable (yet, see [1:59–70] for a detailed definition of the phonological realization of HDs in IH). For example:

Hesitation Disfluencies in spontaneous speech (1)

a. IH

[ani ani yexola lehavin et **ha:** et **ha:** tiskul]

I I can to.understand Acc. **the:** Acc. **the:** frustration.

'I can understand the frustration'

b. Japanese

[**juuitigatu:** no atama goro da to omou]

November: of beginning about COP QUOT think

'I think it was about the beginning of November**:**'

The cognitive function of HD has been dealt within several theories, with respect to the part that the mental lexicon plays in the speech process. In the study of disfluencies such as 'filled pauses' (FPs), a major approach views them as indicators of increased cognitive processing. Shriberg [2], for example, claims that disfluency rates depend on the length and complexity of the sentence [2:157]; Clark and Wasow [3] showed that in American English, complex syntactic structures predict repetitions of function words. Their findings led them to formulate the commit-and-restore model of repeated words [3:203] and to propose the Complexity Hypothesis. For example, they showed that speakers repeat the definite article *the* when it precedes a complex noun phrase; Roll et al. [4] concluded that, in Swedish, a disfluent *att* 'that' is evidence of cognitive processing of more complex syntactic structures.

The complexity theory was also used as explanation for FPs in Japanese. Watanabe et al. [5] examined spontaneous Japanese speech corpora and showed that constituents tend to be longer or more complex when they are immediately preceded by FPs than when they are not, they found that FPs cause listeners to expect that the speaker is going to refer to something that is likely to be expressed by a relatively long or complex constituent. Den [6] reported that clause-initial Japanese conjunctive *de* 'and/then' tends to be prolonged when a pause or fillers succeeds it, whereas clause complexity does not affect the duration of clause-initial '*de*'.

Such a perspective was also adopted for written text by Dresher [7], who showed that a kind of complexity approach can also be implemented on the Tiberian Hebrew accent system (*te'amim*). He demonstrated two cases of long subjects and relatively short verb phrases, where the main break (pause in Dresher's terminology) falls between the subject and the verb; and two other cases with short subjects, where the main break falls after the verb. Dresher [7] concluded that "The difference between the two types of cases has to do with the length and prosodic complexity (i.e. number of phrases) of the subject relative to the verb phrase" [7:25].

The syntactic complexity, though, is only one possibility to explain why speakers elongate certain linguistic increment. Givón [8] talks about possible mapping relations between linguistic and cognitive complexity as "Coding: More complex mentally-represented events are coded by more complex linguistic/syntactic structures; Processing-I: More complex mentally-represented events require more complex mental processing operations. … Processing-II: More complex syntactic structures require more complex mental processing operations." [8:283].

Taking the coding principle into account, HDs, assumed to be considered as "more coding material", might therefore be used for "less predictable" and "more important" information, and not always due to "more syntactically complex". Ariel [9] and Hudson [10] suggest a "larger chunk" includes *phonologically* larger, and not syntactically larger. Ariel [9] argues that the ease with which a piece of given information is processed reflects its degree of mental accessibility, and that representations of linguistic material and physically salient objects are assumed to be in the short-term working memory, as opposed to representations of encyclopedic knowledge, which are assumed to be in long-term memory. Two principal criteria of Accessibility Theory (AT) associated with specific degrees of accessibility may be relevant to explaining the predicted linguistic element after HDs: The first is *informativity*. Accessibility markers representing a low degree of accessibility incorporate more lexical information than those representing a high degree of accessibility (e.g. open lexical categories vs. closed categories). Second, the *attenuation* criterion (i.e., phonological size) states that all things being equal, the less accessible an entity referred to by an expression is, the larger the expression is phonologically. This criterion also refers to the difference between stressed and unstressed forms. Shorter and unstressed forms have a higher degree of accessibility (e.g., function words with CV syllabic structure, as /be/ 'in' in IH or /*ga*/ 'but' in Japanese) than longer and stressed forms (e.g., verbs, proper names, etc.).

Hudson [10] argues that Dependency Grammar theory allows us to count the number of active dependencies, defining

a dependency as active if either the head or the dependent are still awaited. An active dependency is satisfied as soon as the word concerned is encountered [10:275–279]. At that point, the burden on the working memory decreases and more space remains for continuous processing of information.

2. Research goal

Considering the different theories that were introduced above, the research goal of this paper is to find out the linguistic role of HDs, by taking an account of the *syntagmatic aspect* and the *function* HD plays. We will focus on specific syntactic structures – phrases, and their interface with HDs. Since phrases consist, both in Hebrew and Japanese, as well as in other languages, on a head and a nominal increment, a study of phrases' interface with HDs can shed light on the relations between *form*, the prolongated material, and *function*, the syntactic increment that is prolongated and the following increment(s). To sum up, in this paper we provide evidence from spontaneous spoken IH and Japanese corpora to tackle the question how theory can explain HDs on varied syntactical increment, since when we consider disfluent phenomena of prosody and syntax, "explicit procedure must be designed to deal with these problems", as mentioned before in Maruyama [11:782].

3. Data

3.1. Israeli Hebrew spontaneous corpus

The findings of IH presented in this paper are taken from 19 audio segments from 19 different recordings that were selected from CoSIH – Corpus of Spoken Israeli Hebrew [12]. The recordings, which were taken during 2001–2002, are of authentic Israeli Hebrew everyday conversations. Each dialogue consists of conversations between one core speaker and various interlocutors with whom the speaker interacted on that day. The research corpus consists of 31,760 word-tokens (over 6 hours of speech) of which 4,289 are word-types. 44% of the examined material is one-side telephone conversations; while 56% is face-to-face dialogues. The prosodic boundary tone inventory consists of 9,400 annotated boundary tones. The present research focuses on the 764 hesitation disfluencies that consist of 10.72% of all detected prosodic boundaries.

3.2. Japanese spontaneous corpus

The findings of Japanese presented in this paper are taken from the sample S01F0183 of CSJ (Corpus of Spontaneous Japanese) [13]. In this casual monologue a young woman makes her speech about the experience of travel to Hokkaido. In CSJ, elongated points are tagged by "<H>" (in this paper they will be tagged**:**). These are carefully heard and transcribed by transcribers who are well-trained. Final Boundary Tone (FBT) is also transcribed at each end of phrase. FBT tags include L%, H%, HL%, and so on.

Japanese is a pitch accent language, and each accentual phrase (AP) basically ends with low pitch (L%). But sometimes AP ends with high pitch, in that case H% is tagged. And sometimes (or more frequently by some speakers) low pitch comes after H% with elongation, which is tagged by HL%. HL% can be regarded as not the case of elongation with hesitation, but particular speaking style. So **:** tags with HL% are excluded from the analysis. Other excluded cases are of **:** at sentence-final particles, since they mark the end of S/S' and it is often elongated. These elongations do not reflect the hesitations.

The sample S01F0183 consists of 2,129 words, including total of 179 elongations. The examples discussed here regard the 46 cases of elongations, with pre- and post- context.

4. Word order

4.1. Israeli Hebrew word order

Among the "basic orders" found in languages of the world, Hebrew is said to prefer a SVO word order. Nevertheless, Israeli Hebrew word order is relatively free and all possible alternatives can appear in specific contexts, e.g. literature and poetry. In the verb system, Israeli Hebrew morphology is characterized by the non-concatenative Semitic type structure. One of the relevant issues concerning verbs is that verbs are also accompanied by affixes indicating tense, person, number, and gender. Rosén [14] suggested considering the preposition as forming one constituent together with the verb: "The preposition constitutes the government properties of the verb" [14:169–170]. Rosén presented an example of the prepositions /le/ 'to', /be/ 'in' and /al/ 'on', and noted that, with the occurrence of certain verbs, these prepositions have no substitution, and function as cases (such as the accusative case marker /et/ 'Acc.'). Nevertheless, Hebrew, as a "non-strict word-order" language, does not allow clitics and affixes at the phrase final position. Thus, the preposition stranding phenomenon does not occur in Hebrew. This characteristic of Hebrew means that we will not find prepositions in clause final position or in phrase final position (although this syntactic constraint is overruled in case of few coined idioms).

4.2. Japanese word order

Since Japanese is an agglutinative language, particles (case-marking particles, topic-marking particles, conjunctive particles, etc.) are put *after* noun or S'. It means that the function words of 'to' and 'that' are located *after* the content words, for example a noun or S' (2).

[*boku* ***wa*** *sueeden* ***ni*** *ikitai* ***to*** *omotta*] (2)
I topic Sweden DIR want to go **QUOT** think-PAST
'I thought that I want to go to Sweden.'

The structure of the sentence above is analyzed as (3).

[[*boku wa*] [[[*sueeden ni*] *ikitai*] ***to***] *omotta*] (3)

One can see /*wa*/ 'topic', /*ni*/ 'to', and /*to*/ 'that' are located after the noun/S'. No matter how the quoted S' becomes long and complicated, /*to*/ 'that' always comes after the whole structure of S'. Even so, as we will show the data later, particles are sometimes elongated.

5. Results

5.1. The elongated words

5.1.1. Israeli Hebrew results

The results of the present research on spontaneous Israeli Hebrew showed that HDs occur within syntactic units, even intra-phrases or intra-morphemes, for example: [le**:**-sader] 'to**:** arrange', and basically that these hesitated increments are function words. Moreover, according to the results, the elongated words vary in terms of complexity. For example, elongated definite article /ha/ 'the', that was found as the second most probable (p=0.938) elongated word in the corpus (149 such occurrences, as shown in Table 1), predicts a noun (i.e. simple structure) as its complement, while the subordinate conjunction /ʃe/ 'that' which predict a complex S' structure, was also found among the five most probable (p=0.875) elongated words (42 such occurrences, as shown

in Table 1). In brief, HDs were observed to split four types of dependency relationships. These are outlined below and ranked from the weakest to the strongest dependency:

The weakest type of dependency is of course when HD does not split a syntactic dependency, that is, when the word carrying the tone and the following word are not in any direct syntactic relationship (not in the same syntactic phrase or clause). HDs typically *follow* discourse markers, which are considered a "no dependency" type.

Coordinate structure dependencies occur between a conjunction and the adjacent syntactic structures (from increments to sentences). HDs are more likely to occur *after* the conjunction.

The subject-predicate dependency represents the cases where HDs occur between personal pronouns and verbs. This subject-predicate dependency is considered "stronger" than the two mentioned above, since in spontaneous IH they imply that HDs split intra-clausal structures.

Intra-phrasal sequences are the "strongest" type of dependencies that are broken by HDs, and indeed HDs were found to be the most likely to occur intra-phrasal.

The results of the distribution of elongated POS are presented in Table 1. The cases in lines 2, 4, 6, and 7 are considered intra-phrasal in IH.

5.1.2. Japanese results

In Japanese it was found that elongations in S01F0183 can be classified mainly into two types: A. elongations at the end of function words; B. elongations at the end of content words. Type A consists of conjunctives and particles, and type B mainly consists of nouns.

Type A-1) elongation of conjunctives:

The most frequent type was the case where a conjunctive was elongated at the beginning of an utterance [6]. Conjunctive */de/* 'then' appeared 38 times in the target data, and almost 40% of them (15 times) were elongtaed as '*de:*'. Examining the context, these elongations can be considered that the speaker utters '*de:*' at the beginning of an utterance as a discourse marker. Since this talk is a monologue, the speaker as to keep on producing her message. So this kind of elongation can be classified as a preventive device of silence; it signals that the speaker is under processing and has more to say. For example.

[*atatta toka itte:* ***de:*** *Hokkaido ryoko no tikketo*] (4)

won QUOT said and Hokkaido travel of ticket

'she said she won the quiz prize, and she had tickets to Hokkaido.'

Type A-2) elongation of particles:

Particles are sometimes elongated. There are some categories in Japanese particles, and major types are "conjunctive particle", "case-marking particle", and "topic-marking particle". In the case of "case-marking" and "topic-marking" particles, the elongations occur in the middle of the utterance, which seem to indicate that the further part of ongoing utterance is now under processing. "Conjunctive" particles are sometimes elongated too, but these particles appeared at the end of clauses, so it signals that the speaker will continue her next utterance, as the case of conjunctives.

Type B) elongation of nouns:

Nouns are sometimes elongated at the final syllable (rarely inside the noun). These cases are elongation within noun phrases (noun:+particle) or predicates (noun:+copula marker). This type can be considered as a signal of difficulty of fluent production of the ongoing utterance. The speaker confirms the validity of information.

In a subset of whole CSJ with 443,732 words that was retrieved for this research, we confirmed almost the same tendency of elongated POS as the cases observed in S01F0183.

The results of IH and Japanese distribution of elongated POS are presented in Table 1.

Table 1: *Distribution of the most probable elongated (HD) POS in IH and Japanese.*

	Elongated POS in CoSIH	Occurences	Elongated POS in CSJ	Occurrences
1	Conjunction	149	Noun	18
2	Definite article	149	Conjunction	16
3	Personal pronoun	111	Particle	9
4	Preposition	100	Adjective	2
5	Subordinate conjunction	42		1
6	Modifier	36		
7	Possessive particle	22		
8	the lexeme /hjj/ 'be'	17		
9	Existential lexeme (Auxiliary)	11		
10	Modal lexeme (Auxiliary)	10		

5.2. What follows HDs?

Results in IH show that there is wide variation in the *following* elements to the HDs, and this for itself is a simple argument that rejects complexity as an ultimate explanation for the HD phenomenon. Results showed no priority to syntactically complex structures in the expected following structures. Another argument are cases of elongated construct state nouns (5a–b) and infinitive prefixes (6), with the prediction of only simple following structures, a noun and a gerund, respectively.

Construct-state nouns in IH (5)

a. [ze jihje be m **sof e: ogust**].
'it will be at ehm **end of August.**'

b. [milxemet e**:** xami∫im ve ∫e∫].
war-of: fifty and six.
'the 1956 war.'

Infinitives in IH (6)

[holex **li: -krot**]?
'going **to: happen**?'

Such cases were also found in Japanese (7). In this case the speaker remembered the name of place *Nemuro*, and then she repeated the word within a noun phrase (*nemuro made*) with elongation.

Intra-phrasal HDs in Japanese (7)

[nemuro da **nemuro:** made kuruma de itte:]
'It was Nemuro, we went to **Nemuro:** by car.'

6. Discussion

The explanation that is taken to apply to the findings of IH corpus is that of syntactic planning coming before lexical planning [15, 16]. To this statement, another term – the placeholder [16] is added. In spontaneous speech, placeholders "mainly have a pronominal origin and serve as a preparatory substitute for a delayed constituent" [16:11] and placeholders are "among other lexical and grammatical resources that allow the speaker to refer to object and events for which the speaker fails to retrieve the exact name, or simply finds the exact name to be unnecessary or inappropriate" [16:11]. Both [15] and [16] assume the pronominal nature of disfluencies or placeholders. In this respect, HDs are prosodic morphemes which also have a pronominal nature. This is to say that IH speakers first utter the syntactic frame with the elongated word, the lexical element that "carries" the HD with its pronominal nature. The elongated word is expected to be followed by a syntactic increment or a target word. It is suggested in [1] that by the elongated production of function words the expected syntactic structure is already indicated, meaning the syntactic structure is thus complete. In other words, the following lexeme(s) to the HD does not contribute a fundamental increment to the *structure*, only to the *content*. This mechanism reduces the burden of the working memory, and thus enables processing of new information.

The question is whether the pronominal nature of HDs should be adapted to Japanese as well. In case of Japanese, not only function words but also content words (nouns) are sometimes elongated. In (8), three cases of elongation appeared; *de:*, elongation of the conjunctive (discourse marker) at the beginning of utterance, *sore: ga*, elongation at the final syllable of noun in noun phrase (noun+case marking particle), and *nido: me*, elongation within the compound noun.

[***de:***	*watasi*	*wa*	***sore:***	*ga*	*hokkaido*	*wa*	(8)
and:	I	TOP	**it:**	SUBJ	Hokkaido	TOP	
nido:	*me*	*datta n desu*		*keredomo*]			
second	times	Copula-PAST	though				

"and it was the second time for me to go to Hokkaido."

In this situation the speaker is under the constraint to produce a series of episodes linearly in real time (cf. linearization problem, [17]). Elongations at the end and within the nouns can be considered to be used; 1) to make a plan of further incremental utterance which is roughly designed at the beginning, or 2) to confirm the validity of information just mentioned by the noun. As shown in the cases of IH, these elongations also reduce the burden of the working memory.

As we have shown in Table 1, the distribution of elongated POS between IH and Japanese is different in some aspects, especially content words in Japanese. The difference may be caused by syntactic and/or pragmatic reason, which indicates that a typological study on disflunencies is possible.

7. Conclusions

To conclude, the comparative evaluation of the two languages which are different in their word order characteristics had shown that HD phenomenon is a matter of "processing time, cognitive complexity or mental effort". Syntactic complexity, if exists, can be explained as reflecting the "coding principle" [8] of the cognitive process.

According to the approach adopted here, what is common to all elongated words is the fact that they imply *continuity*, regardless of whether they are dependents of heads (i.e., nouns in NP) or heads of dependents (i.e., case marking particle in NP). What should be stressed here is that they share a [+dependency] syntactic feature. It can be said that what is actually elongated is not the word itself (or a syllable of the word), but this syntactic feature. The [+dependency] feature shows that "there is more to come", i.e. what are the communicative intentions of the speaker. It allows the speaker to think, if either the function word or the content word of a phrase are still awaited, by elongating structures. In our view, what is common to elongated grammatical elements is a [+dependency] feature.

We tried to connect between syntagmatic aspect and the linguistic function of HD, and to propose a pronominal approach to the phenomenon of HDs, which we explained in both languages. Nevertheless, it must be also related to cognitive process, as mentioned, or can be considered as a signal of difficulty of fluent production of the ongoing utterance.

8. References

[1] V. Silber-Varod, *The SpeeCHain Perspective: Prosodic–Syntactic Interface in Spontaneous Spoken Hebrew*, PhD dissertation, Tel Aviv University, 2011.

[2] E. Shriberg, "To 'errrr' is human: Ecology and acoustics of speech disfluencies". *JIPA* 31(1), pp. 154–169, 2001.

[3] H. H. Clark & T. Wasow, "Repeating words in spontaneous speech". *Cognitive Psychology* 37, pp. 201–242, 1998.

[4] M. Roll, J. Frid and M. Horne, "Measuring syntactic complexity in spontaneous spoken Swedish". *Language and Speech* 50 (2), pp. 227–245, 2007.

[5] M. Watanabe, K. Hirose, Y. Den and N. Minematsu, "Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners". *Speech Communication* 50(2), pp. 81–94, 2008.

[6] Y. Den, 2009. "Prolongation of clause-initial mono-word phrases in Japanese". In S.-C. Tseng (ed.), *Linguistic patterns in spontaneous speech*, Taiwan: Institute of Linguistics, Academia Sinica, pp. 167–192, 2009.

[7] B. E. Dresher, "The prosodic basis of the Tiberian Hebrew system of accents". *Language* 70(1), pp. 1–52, 1994.

[8] T. Givón, *The Genesis of Syntactic Complexity*. Amsterdam: John Benjamins Publishing, 2009.

[9] M. Ariel, "Accessibility theory: An overview". In T. Sanders, J. Schliperoord, and W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects*, Amsterdam: John Benjamins, pp. 29–87, 2001.

[10] R. A. Hudson, "Do we have heads in our minds?" In G.G. Corbett, N.M. Fraser and S. McGlashan (Eds.), *Heads in grammatical theory*, Cambridge: Cambridge University Press, pp. 266–291, 1993.

[11] T. Maruyama, "Speech segmentation by clausal and non-clausal boundaries in Japanese". *The Fifth International Conference on Cognitive Science*, Abstracts Volume 2: Workshop "Spoken discourse corpora as a window on cognitive mechanisms of speech production", Kaliningrad, Russia, pp. 782–783, 2012.

[12] The Corpus of Spoken Israeli Hebrew (CoSIH), available at: http://humanities.tau.ac.il/~cosih/english/

[13] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation". *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, p. 712, 2003.

[14] H.B. Rosén, Contemporary Hebrew. The Hague: Mouton, 1977.

[15] C. Blanche-Benveniste, "Linguistic analysis of spoken language: The case of French language". In Y. Kawaguchi, S. Zaima and T. Takagaki (eds.) *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: John Benjamins, 2006.

[16] V.I. Podlesskaya, 2010. "Parameters for typological variations of placeholders". In N. Amiridze, B. H. Davis, and M. Maclagan (Eds.), *Fillers, pauses and placeholders. Typological Studies in Language*, Amsterdam: John Benjamins. pp. 11–32, 2010.

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

Phrasal complexity and the occurrence of filled pauses in presentation speeches in Japanese

Michiko Watanabe

National Institute for Japanese Language and Linguistics, Japan

Abstract

Filled pauses are ubiquitous in everyday speech. I investigated whether linguistic complexity of upcoming phrases affects filler rate at phrase boundaries in presentation speeches in Japanese. Filler rate at phrase boundaries increased monotonically with complexity of the following phrases. However, when the following phrase was composed of more than 11 Bunsetsu-phrases, the filler rate did not show any constant increase. The results indicate that filler rate at phrase boundaries is closely related to cognitive load of local linguistic encoding and that the maximum planning span for linguistic encoding is about 10 Bunsetsu-phrases in Japanese monologues.

Index Terms: filled pause, Bunsetsu-phrase, linguistic complexity, planning load

1. Introduction

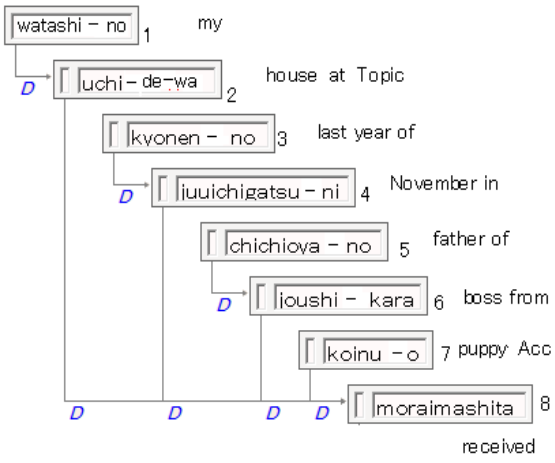
Levelt's [1] speech planning model assumes three stages of planning: 1) conceptualizing; 2) formulating (grammatical and phonological encoding); 3) articulating. Speech disfluencies such as filled pauses (fillers) and repetitions are claimed to occur when some troubles happen at any of these stages. A number of studies have investigated factors affecting the occurrence of disfluencies. Among the factors investigated are the degree of difficulty of lexical access to the word which speakers intend to utter, the range of choice in vocabulary that speakers can utilize, and the complexity of the linguistic constituents which immediately follow [2, 3, 4].

In this paper I focus on relevance of filled pauses to the complexity of the immediately following phrases. A previous study on Japanese filled pauses reports higher filler rate at syntactically stronger clause boundaries than at weaker ones [5]. Interestingly, the study also found clearer correspondence between filler rate and the complexity of immediately following clauses at weak clause boundaries than at stronger ones. No effect of the complexity of immediately following clause was observed at sentence boundaries, which are stronger than clause boundaries. Speakers are more likely to be engaged in global message planning at deep syntactic boundaries, such as sentence or coordinate clause boundaries, than at shallower ones. From the results of the previous study, I inferred that filler rate at syntactic boundaries reflects complexity of the immediately following linguistic units when a burden of global message planning is light. In other words, I expected that filler rate would show clearer correspondence to the complexity of the immediately following linguistic units at syntactic boundaries which are weaker than sentence or clause boundaries. To test this hypothesis, I investigated filler rate at phrase and weak clause boundaries as a function of the complexity of the immediately following phrases.

I focused on Bunsetsu-phrases as a unit of planning. A Bunsetsu-phrase (I call Bunsetsu, hereafter) is composed of one content word with or without function words. It is a component of clauses and sentences. Bunsetsu is a prosodic word and one or more Bunsetsu compose an accentual phrase. Interjectional particles can be added to the end of any Bunsetsu, and back-channeling is often given at or near the end of Bunsetsu in conversation. Bunsetsu has been regarded as one of the smallest planning units in Japanese [6].

I looked at dependency structure of Bunsetsu-phrases. Japanese is a head final language: A head Bunsetsu appears after all its dependent Bunsetsu-phrases. Figure 1 shows examples of Japanese sentences and their dependency structure. Sentence (b) has more complex structure than sentence (a). Each box represents a Bunsetsu. The number on the right of each box indicates id for the Bunsetsu in the sentence. Dependency relation is indicated by an edge from a dependent Bunsetsu to its head. The figure indicates that the last Bunsetsu, "moraimashita" has more dependent Bunsetsu in Sentence (b) than in sentence (a).

(a) I got a puppy from my friend.



(b) At my house we got a puppy from my father's boss last November.

Figure 1: *Examples of dependency structure of Bunsetsu-phrases in simple and more complex Japanese sentences.*

Speakers must have planned the head at least to certain extent when they start speaking a dependent phrase. When they want to convey rich information concerning the head, they need to plan and utter number of relevant dependents, and consequently the dependency structure will be complex, as in example sentence (b). When a given dependent phrase is far from its head and many other dependents are to be uttered before the head, the speaker needs to plan them while or immediately after he or she articulates the dependent. I hypothesized that the more dependent phrases to be planned between a given dependent phrase and its head, the higher the filler rate immediately after the dependent phrase, if filler rate corresponds to cognitive load of local linguistic encoding.

2. Method

2.1. Data

107 presentation speeches from the *Core* part of *The Corpus of Spontaneous Japanese* (CSJ) were used for analysis [7]. The presentations were given by paid volunteers to a small audience in informal settings. They are called simulated public speaking (SPS) in the corpus. 54 of them were given by female speakers and 53 by male speakers. All the speakers were those of Tokyo dialect. They talked about general topics such as "the happiest memory in my life" or "my town" for about 10 minutes. The speakers were instructed to prepare for a note for their speeches beforehand, but not to read out their manuscripts. All the speeches are transcribed and detailed linguistic information is given to the transcription.

2.2. Procedures

The corpus contains annotation of Bunsetsu boundaries and the head of each dependent Bunsetsu. There are 65,022 dependent Bunsetsu in total. I counted the number of Bunsetsu from each dependent Bunsetsu to its head. I regard the number as an index of complexity from the dependent to its head and call it "distance" for the dependent. In sentence (b) in Figure 1, for example, the first Bunsetsu, "watashi-no" depends on the second Bunsetsu. So the distance for the first Bunsetsu is one. As the second Bunsetsu "uchi-de-wa" depends on the eighth, the distance for the second is six. I grouped all the dependent Bunsetsu according to the distance to their heads. I computed filler rate immediately after each dependent Bunsetsu group, and examined whether there is any correspondence between the distance and the filler rate.

3. Results

Figure 2 shows filler rate immediately after each dependent Bunsetsu group as a function of the number of Bunsetsu to the head (distance). The rate for Bunsetsu groups the distance of which exceeds 20 is omitted because the number of samples is less than 50 and the rate is unreliable. The figure illustrates that there is a leap in filler rate between a Bunsetsu group with distance one and a group with distance two. However, the rate monotonically increases with distance after that until the distance reaches 11. When the distance is larger than 11, no consistent tendency is observed, though the rate is always higher than 25%.

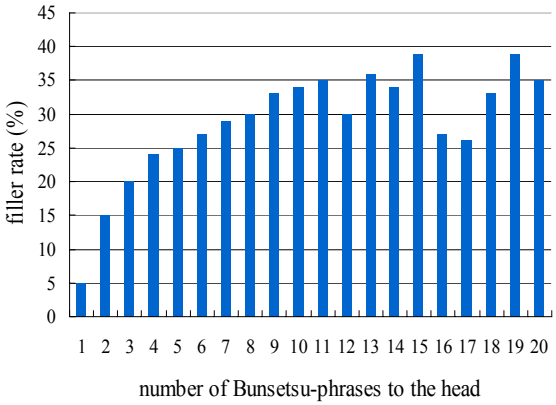


Figure 2: *Filler rate immediately after dependent Bunsetsu-phrases as a function of distance to the head.*

4. Discussion

The filler rate immediately after dependent Bunsetsu groups increased with complexity of the following phrases. This result is in accordance with the results of a study on repetition rates of articles and pronouns in English [4], and supports the hypothesis that filler rate reflects cognitive load of local linguistic encoding. However, no clear correspondence to complexity was observed when the distance to the head was larger than 11. This may be because speakers use additional fillers or other means, such as silent pauses, in the middle of dependency structure to gain time when they need to plan more than 11 Bunsetsu to convey a message. The results indicate that the maximum linguistic encoding span in Japanese presentation speech is approximately the length of ten Bunsetsu-phrases.

5. Acknowledgements

This research is supported by Grant-in-Aid for Scientific Research by JSPS (2012, Grant No. 24520494).

6. References

[1] W.J.M. Levelt, *Speaking*, The MIT Press: Cambridge, Massachusetts, 1989.

[2] J.E. Arnold, R. Altmann, M. Fagnano and M.K. Tanenhaus, "The old and thee, uh, new", *Psychological Science*, pp. 578–582, 2004.

[3] N. Christenfeld, "Options and Ums", *Journal of Language and Social Psychology* 113/2, 192–199, 1994.

[4] H.H. Clark and T. Wasow, "Repeating words in spontaneous speech", *Cognitive Psychology* 37, 201–242, 1998.

[5] M. Watanabe, *Features and Roles of Filled Pauses in Speech Communication – A corpus-based study of spontaneous speech* (Hituzi Linguistics in English No. 14), Hituzi Syobo Publishing, 2009.

[6] Y. Kitahara et al. (eds.), *Dictionary of Japanese Grammar* (in Japanese), 58, Yuseido, Tokyo, 1981.

[7] The National Institute for Japanese Language, Construction of The Corpus of Spontaneous Japanese, 2006.

Disfluencies and uncertainty perception – evidence from a human–machine scenario

Charlotte Wollermann[1], Eva Lasarcyk[2], Ulrich Schade[3] & Bernhard Schröder[1]

[1] German Linguistics, University of Duisburg-Essen, Germany
[2] Institute of Phonetics, Saarland University, Germany
[3] Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany

Abstract

This paper deals with the modelling and perception of disfluencies in articulatory speech synthesis. The stimuli are embedded into short dialogues in question-answering situations in a human–machine scenario. The system is supposed to express uncertainty in the answer. We test the influence of *delay*, *intonation*, and *filler* as prosodic indicators of uncertainty on perception in two studies. Study 1 deals with the effect of *delay* and *filler* on uncertainty perception. Results suggest an additive effect of the cues, i.e. the activation of both prosodic cues of uncertainty has a stronger impact on uncertainty perception than the deactivation of a single cue or of both cues. With respect to the effect of single cues, no significant difference can be observed. Study 2 investigates the impact of *delay* and *intonation* on perceived uncertainty. Again, a principle of additivity can be observed. Furthermore as modelled here, *intonation* has a stronger influence than *delay*. In both studies no correlation between the ranking of uncertainty and naturalness of the stimuli is found.

Index Terms: uncertainty, disfluencies, speech synthesis, speech perception.

1. Introduction

Given is a communicative situation with two conversational partners. A is asking a question to B and B is not certain with respect to her answer. This might be due to several reasons, e.g., B only partially knows the answer, B does not know how to formulate the message, B is grasping for information etc. In addition, uncertainty can be regarded as a complex phenomenon. In some works uncertainty is categorized as emotion [1, 2], in other works it is assumed to have a cognitive character [3]. In the context of question-answering situations, the following questions arise: How do speakers signal uncertainty prosodically with respect to answers? Which prosodic cues do hearers use for decoding uncertainty in answers?

2. Communication of uncertainty

In **human–human communication**, it was found that speakers mark uncertainty in question-answering situations by using *rising intonation*, *pauses*, *fillers*, and *lexical hedges* [4]. In [4] the authors used the *Feeling of Knowing* (FOK) paradigm for eliciting metamemory judgments. For taking the hearer's side into account, [5] defined the *Feeling of Another's Knowing* (FOAK). Results from their perception study provided evidence that the FOAK was influenced by the *intonation* and by the *form* of answers as well as by *pauses* and *fillers*. Furthermore, *smiles* and *funny faces* can serve as visual indicators of uncertainty [6].

In the context of **human–machine communication** however, it is less clear if these cues contribute to the perception of uncertainty in a comparable way. In [7] it is argued that the modelling of uncertainty can improve information systems by enriching expressive abilities. With respect to acoustic speech synthesis, *filled pauses* were modelled in [8] on the basis of a 'synthetic disfluent speech model'. For the implementation an unit selection synthesizer was used. The results of the perception study suggest no decrease of the system's naturalness. In the study of [9], utterances were selected from spontaneous conversational speech. Type and placement of *fillers* and *filled pauses* were predicted using a machine learning algorithm. For the synthesis, an unit selection voice was used. The evaluation shows no decrease of naturalness. Moreover, in [10] it was found that *filled pauses* occur more frequently in the human–machine corpora than in the human–human corpus. However, since it is less clear to what extent disfluencies interact with uncertainty perception, further research seems necessary. For investigating this question we use an articulatory speech synthesizer. A motivation is given in the following section.

3. Paralinguistic expression, prosody and articulatory speech synthesis

Natural speech is usually rich in both linguistic and paralinguistic information. Varying the paralinguistic level of a message to match one's needs can entail speech variations that can for instance involve voice quality, segmental duration, or fundamental frequency changes. Thus, modelling expressive speech can be challenging due to these prosodic variations. However, in order to achieve naturalness of the synthetic speech, variability needs to be considered [11].

To generate such highly variable speech, we use the articulatory synthesis system VocalTractLab [12]. The system produces utterances of high acoustic quality. It processes a timeline of articulatory gestures which are translated into trajectories of the articulators in the virtual three-dimensional vocal tract [13]. In an aerodynamic-acoustic simulation step, the speech signals are generated. Since the utterance is created 'from scratch', the system is very versatile and offers large degrees of freedom. The abovementioned paralinguistic demands on the manner of speaking can be integrated at the foundation of the utterance planning, and no post-hoc signal processing needs to be applied.

4. Previous work

In an intial investigation of modelling and perceiving of uncertainty [14], also the articulatory speech synthesizer of [12] was used. Four different degrees of intended uncertainty were generated by varying the cues *intonation* (rising vs.

falling), *delay* (present vs. absent), and the *filler* 'hmm' (present vs. absent). The scenario was a fictitious telephone dialogue between a weather expert system and a user. The answer of the system was marked by different degrees of uncertainty. Results show that the activation of all uncertainty cues has a stronger impact on the perceived uncertainty than *rising intonation* alone and *delay* combined with *rising intonation*.

In a follow-up study [15], all eight possible combinations of the three cues were used for conveying different degrees of uncertainty, and the stimuli were presented in a modified scenario, an interaction between a robot for image recognition and a user. The user showed pictures of fruits and vegetables to the robot and asked the robot, 'Was siehst Du?'/*What do you see?* The robot recognized the objects. Depending on a fictitious recognition confidence score, the system conveyed (un)certainty in its answer by using the cues mentioned above. Results provide evidence for additivity of all three uncertainty cues with respect to uncertainty perception. Compared to the effect of *rising intonation* and *filler*, the influence of *delay* was relatively weak. The following questions remain open: Does a much longer duration of the *delay* contribute more strongly to the perception of uncertainty? To what extent does the filler 'uh' effect the perception of uncertainty? We address these questions in the current paper. Therefore, we modify the material used in [15].

5. Material

Our stimuli consist of four different one-word phrases in German ('Melonen'/*melons*, 'Bananen'/*bananas*, 'Tomaten'/*tomatoes*, 'Kartoffeln'/*potatoes*) , each one is generated in eight different levels of uncertainty by varying *intonation* (rising vs. falling), *delay* (absent vs. present) and the *filler* 'uh' (absent vs. present). The variation of **intonation** takes place at the last syllable of each word: for *rising intonation* fundamental frequency increases to around 200 Hz, for *falling intonation* it decreases to around 70 Hz. The **delay** refers to the time between the user's question ('Was siehst Du?'/*What do you see?*) and the system's response ('Bananen', 'Tomaten', ...). In each case there is a default *delay* of 1000 ms. In the case of a long *delay* there are two subcases: i) when *filler* is absent the additional *delay* is 4000 ms, ii) when *filler* is present we have the default *delay* (1000 ms) + *filler* 'uh' (duration of 370 ms) + *delay* (3630 ms). For the **filler** we choose the particle 'uh' this time, since 'uh' is the *filler* which occurs most often in the Verbmobil corpus for German [16]. For distracting the subject from our interest we use four further items ('Bohnen'/*beans*, 'Paprika'/*sweet pepper*, 'Gurken'/*cucumber*, 'Knoblauch'/*garlic*).

6. Perception study I

The goal of this study is to test the impact of *filler* and/or *delay* on the perception of uncertainty and naturalness.

Table 1: *Cues of uncertainty. Left: Study I. Right: Study II.*

level	filler	delay	level	intonation	delay
U0	−	−	U0	−	−
U3	−	+	U3	−	+
U4	+	−	U8	+	−
U7	+	+	U11	+	+

6.1. Material and hypothesis

We use four different levels of intended uncertainty (cf. Table 1, left side). To explain the structure of the stimuli we use the example *bananas*:

U0: 'Was siehst Du?' [delay 1000 ms] 'Bananen.'

U3: 'Was siehst Du?' [delay 1000 ms] [delay 4000 ms] 'Bananen.'

U4: 'Was siehst Du?' [delay 1000 ms] ['uh' 370 ms] 'Bananen.'

U7: 'Was siehst Du?' [delay 1000 ms] ['uh' 370 ms] [delay 3630 ms] 'Bananen.'

The stimuli were divided into four sets. In each group we presented eight stimuli: The four items, and the four distractor items. Each stimulus occurred exactly once with respect to the overall data.

We assume that U0 yields to the lowest level of perceived uncertainty, whereas U7 leads to the highest level of perceived uncertainty. Based on our previous studies [14, 15], the following hierarchy is hypothesized: U0<U3<U4<U7.

6.2. Procedure

Subjects were 74 undergraduate students (62 f, 12 m) from the University of Duisburg-Essen, all of them native speakers of German. They were tested in four groups (g1: $N = 25$, g2: $N = 15$, g3: $N = 19$, g4: $N = 16$). For each group one set was presented. The dialogues were played back over loudspeakers. The procedure started with an example stimulus. For each dialogue, subjects had to judge the answer of the system on a questionnaire. There were two 5-point-Likert scales, and subjects were asked to judge how (un)certain the answer sounded and also how natural it sounded.

For statistical analysis, we use different tests. With the Kruskal-Wallis Rank Sum Test we test the overall difference between judgments with respect to uncertainty and naturalness, respectively. In a next step the Wilcoxon Signed Rank with Bonferroni correction is performed for single comparisons between the different levels. Finally, we calculate with the Spearman's Rho Test whether there is a correlation between the uncertainty ratings and the naturalness ratings.[1]

6.3. Results

Firstly, we present the results for the recipients judgments with respect to the perception of **uncertainty**. According to the Kruskal-Wallis Rank Sum Test the overall difference between judgments is $p < 0.0001$ (level of significance: 5%). Figure 1 shows the results in more detail. The Wilcoxon Signed Rank Test with Bonferroni correction (level of significance: $1/6 \times 5\%$) results in $p < 0.0001$ for all comparisons, but there is one exception. The judgments between U3 and U4 do not differ significantly from each other ($p > 0.008$).

In a next step we focus on the **naturalness** ratings. The Kruskal-Wallis Rank Sum Test does not show a difference between judgments when we look at the data overall ($p > 0.05$). For each of the four different levels of uncertainty the median is 3 (cf. Fig. 2). There is no significant difference between the judgments of the single levels as the Wilcoxon Signed Rank Test with Bonferroni correction shows (each time $p > 0.008$).

Furthermore, the Spearman's Rho Test computes a coefficient of 0.05, i.e. our data do not provide evidence for a correlation between the ratings of uncertainty and the ratings of naturalness.

[1] Results of the judgments for the perceived uncertainty exclusively were presented at the Workshop *Fluent Speech* in November 2012 in Utrecht (without publication).

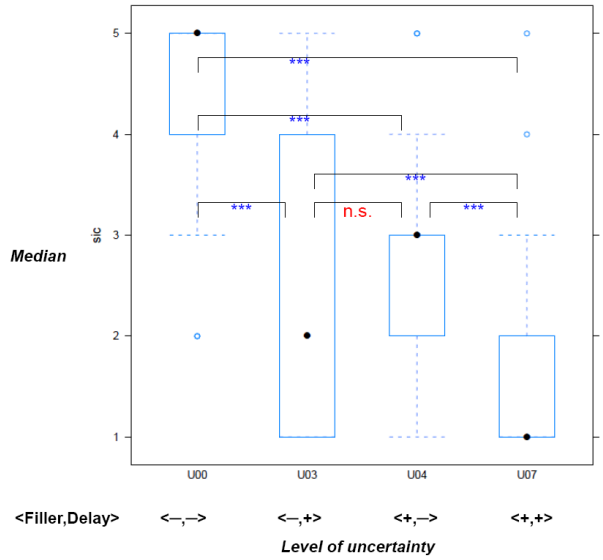


Figure 1: Study I: *Uncertainty judgments;* p < *0.008:*,* p < *0.001:**,* p < *0.0001:****

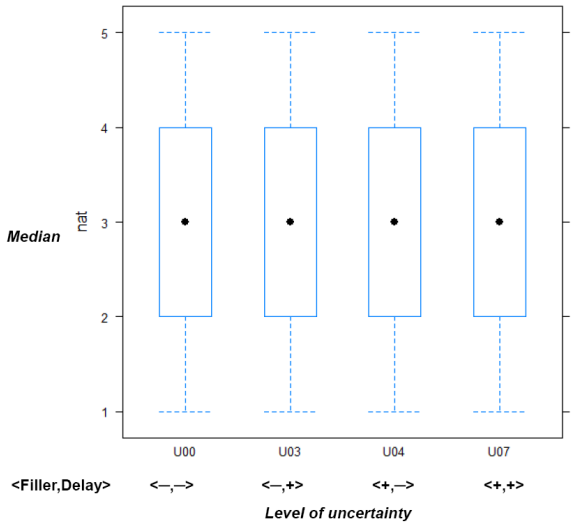


Figure 2: *Study I: Naturalness judgments*

6.4. Discussion

Our results suggest an additive principle of the uncertainty cues. If both cues, i.e. *delay* and *filler*, are activated the perceived uncertainty is significantly higher than for the deactivation of one cue or of both cues. With respect to the relative contribution of the cues, our data show no significant difference between the effect of *delay* vs. *filler*. Moreover, no significancies occur regarding differences in naturalness ratings. Also, no correlation between uncertainty and natural jugments is found.

7. Perception study II

The goal of the second experiment is to investigate the effect of *intonation* and/or *delay* on the perception of uncertainty and naturalness.

7.1. Material and hypothesis

Like in the first study, we use four different levels of intended uncertainty (cf. Table 1, right side). For explaining the structure of the stimuli we use the example *bananas* again:

U0: 'Was siehst Du?'/ [delay 1000 ms] 'Bananen.'

U3: 'Was siehst Du?' [delay 1000 ms] [delay 4000 ms] 'Bananen.'

U8: 'Was siehst Du?' [delay 1000 ms] 'Bananen?'

U11: 'Was siehst Du?' [delay 1000 ms] [delay 4000 ms] 'Bananen?'

As in the previous experiment the stimuli were divided into four sets, each containing eight stimuli (4 items, 4 distractors). Overall, each item stimulus occurred exactly once.

We assume that U0 yields to the lowest level of perceived uncertainty, whereas U11 leads to the highest level of perceived uncertainty. Based on our previous studies [14, 15] the following hierarchy is hypothesized: U0<U3<U8<U11.

7.2. Procedure

Seventy-nine undergraduate students (62 f, 12 m) from the University of Duisburg-Essen took part in the experiment, all of them native speakers of German. They were tested in four groups (g1: $N = 21$, g2: $N = 27$, g3: $N = 15$, g4: $N = 16$). The procedure was the same as in the previous study. For each dialogue, subjects had to judge the answer of the system on a questionnaire with respect to uncertainty and naturalness.

For the statistical analysis we again used the Kruskal-Wallis Rank Sum Test, the Wilcoxon Signed Rank Test with Bonferroni correction, and the Spearman's Rho Test.

7.3. Results

Regarding the perception of **uncertainty** our data reveal (cf. Figure 3) an overall difference between judgments ($p < 0.0001$, Kruskal-Wallis Rank Sum Test). Pairwise comparisons (Wilcoxon Signed Rank Test with Bonferroni correction) show significant differences between judgments for all comparisons ($p < 0.0001$ each time). There is only one comparison with $p < 0.008$ (U3 vs. U8).

Results for the perveiced **naturalness** are illustrated in Figure 4. There are no significant differences between judgments (Kruskal-Wallis Rank Sum Test: $p > 0.05$, Wilcoxon Signed Rank Tests with Bonferroni correction: each time $p > 0.008$). Spearman's Rho Test yields a coefficient of -0.04, indicating no correlation between the uncertainty and naturalness judgments.

7.4. Discussion

In line with the findings of study I, we observe an additivity of the uncertainty cues, but this time for *intonation* and *delay*. Our data also suggest that *rising intonation* alone contributes more strongly to the perception of uncertainty than *delay* alone. With respect to the perception of naturalness our data do not provide evidence for a difference in ratings. In a similar way, no correlation between uncertainty perception and naturalness perception can be observed.

8. Conclusion

We presented two perception studies on the influence of disfluencies in uncertainty perception. The utterances were characterized by different combinations of uncertainty cues and were generated by an articulatory synthesizer. Results show in general significant differences of perceived uncertainty. Our data provide evidence for an additivity of the cues with respect to uncertainty perception. However, we cannot observe an effect of prosodic cues of uncertainty on the perception of naturalness of the synthetic utterances. In addition, no significant correlation could be observed between the judgments of perceived uncertainty and perceived naturalness. In previous studies [8, 9], it was also found that *fillers* and *filled pauses* do not significantly decrease the naturalness of synthetic speech.

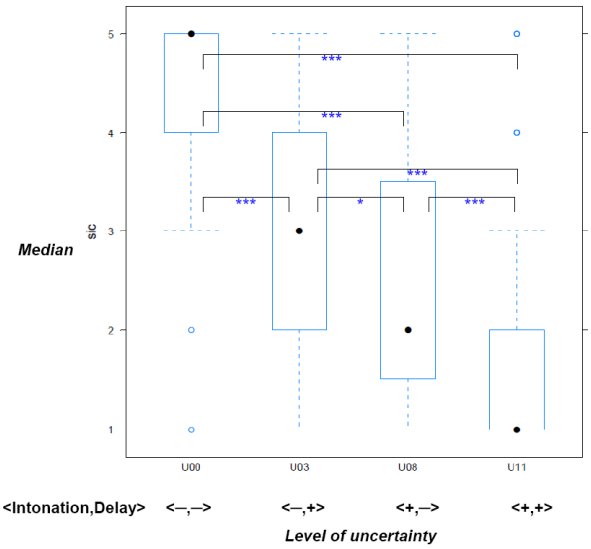


Figure 3: *Study II: Uncertainty judgments;* p < *0.008:*,* p < *0.001:**,* p < *0.0001:****

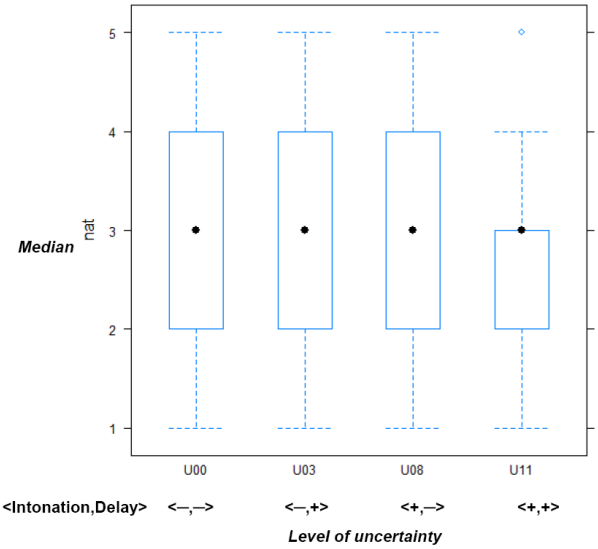


Figure 4: *Study II: Naturalness judgments*

For future work, we deem it necessary to further investigate the interplay between uncertainty perception and naturalness perception. Also, different scenarios need to be considered in order to test the role of disfluencies for the expression of uncertainty and its benefit for human–machine communication.

Many studies have shown that prosody is not only conveyed but also perceived in the visual channel (for synthetic speech e.g. [17, 18]) and the role of visual prosody and uncertainty has been studied for instance in [6, 7, 19]. In our future work we would also like to further investigate *audiovisual* prosody of uncertainty and its interplay with naturalness perception.

9. Acknowledgements

We thank Bernhard Fisseni and Denis Arnold for helpful comments on this paper.

10. References

[1] P. Rozin and A.B. Cohen, "High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans". In: *Emotion* 3(1), pp. 68–75, 2003.

[2] D. Keltner and M.N. Shiota, "New Displays and New Emotions: A Commentary on Rozin and Cohen." In: *Emotion* 3(1), pp. 86–91, 2003.

[3] C.C. Kuhlthau, *Seeking Meaning: A Process Approach to Library and Information Services*, Norwood, NJ: Ablex, 1993.

[4] V.L. Smith and HH. Clark, "On the Course of Answering Questions". In: *Journal of Memory and Language* 32, pp. 25–38, 1993.

[5] S.E. Brennan and M. Williams, "The feeling of anothers knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers". In: *Journal of Memory and Language* 34, pp. 383–398, 1995.

[6] M. Swerts and E. Krahmer, "Audiovisual prosody and feeling of knowing". In: *Journal of Memory and Language* 53, pp. 81–94, 2005.

[7] E. Marsi and F. van Rooden, "Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System". In: *Proceedings of the Workshop on Multimodal Output Generation*. Enschede, Netherlands, 105–116, 2007.

[8] J. Adell, A. Bonafonte and D. Escudero-Mancebo, "Modelling Filled Pauses Prosody to Synthesise Disfluent Speech". In: *Proceedings of Speech Prosody 2010*, Chicago, IL, 2010.

[9] S. Andersson, L. Georgila, D. Traum, M. Aylett, R.A.J. Clark, "Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech". In: *Proceedings of Speech Prosody 2010*, Chicago, IL, 2010.

[10] R. Eklund, *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Linköping University, Sweden, 2004.

[11] I.R. Murray and J.L. Arnott, "Synthesizing Emotions in Speech: Is it Time to Get Excited?" In: *Proceedings of the International Conference on Spoken Language Processing* 1996, Philadelphia, PA, pp. 1816–1819, 1996.

[12] P. Birkholz, *3D-Artikulatorische Sprachsynthese*, Berlin: Logos, 2006.

[13] P. Birkholz, B.J. Kröger and C.J. Neuschaefer-Rube, "Model-Based Reproduction of Articulatory Trajectories for Consonant-Vowel-Sequence". In: *EEE Transactions on Audio, Speech, and Language Processing*, 19(5), pp. 1422–1433, 2011.

[14] C. Wollermann and E. Lasarcyk, "Modeling and Perceiving of (Un)Certainty in Articulatory Speech Synthesis". In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis*. Bonn, Germany, pp. 40–45, 2007.

[15] E. Lasarcyk and C. Wollermann, "Do prosodic cues influence uncertainty perception in articulatory speech synthesis?" In: *Proceedings of the 7th ISCA Workshop on Speech Synthesis*. Kyoto, Japan, pp. 230–235, 2010.

[16] A. Batliner, A. Kießling, S. Burger and E. Nöth, "Filled Pauses In Spontaneous Speech". In: *Proceedings of 13th Intl. Congress of Phonetic Sciences* 3, pp. 472–475, 1995.

[17] E. Krahmer, Z. Ruttkay, M. Swerts and W. Wesselink, "Pitch, Eyebrows and the Perception of Focus". In: *Proceedings of Speech Prosody* 2002. Aix-en-Provence, France, pp. 443–446, 2002.

[18] B. Granström and D. House, "Inside out – acoustic and visual aspects of verbal and non-verbal communication". In: *Proceedings of the 16th International Congress of Phonetic Sciences* 2007, Saarbrücken, Germany, pp. 11–18, 2007.

[19] I. Oh, *Modeling Believable Human–Computer Interaction with an Embodied Conversational Agent: Face-to-Face Communication of Uncertainty*. PhD thesis, Rutgers The State University of New Jersey, NJ, 2006.

Author index

< This page intentionally left blank >

ISBN 978-91-981276-0-7
eISBN 978-91-981276-1-4
ISSN 1104-5787
ISRN KTH/CSC/TMH--13/01-SE
TRITA TMH 2013:1