Proceedings of

DiSS 2021

The 10th Workshop on Disfluency in Spontaneous Speech

Université Paris VIII Vincennes
Saint-Denis, France
25–26 August, 2021

ISBN: xxx-xxx-xxx-xxx-x

Edited by
Ralph L. Rose & Robert Eklund

<This page intentionally left blank>

Proceedings of

DiSS 2021

The 10th Workshop on Disfluency in Spontaneous Speech

Université Paris VIII Vincennes
Saint-Denis, France
25–26 August, 2021

ISBN: xxx-xxx-xxx-xxx-x

Edited by
Ralph L. Rose & Robert Eklund

Conference website: http://diss2021.fr/

Cover design by Ralph L. Rose & Robert Eklund
Graphics and photographs by Robert Eklund (front cover) and Université Paris 8 Vincennes – Saint-Denis (back cover) (except ISCA and Paris 8 logotypes)
Proceedings of DiSS 2021, The 10th Workshop on Disfluency in Spontaneous Speech
Workshop held at Université Paris VIII Vincennes (online), Saint-Denis, France, 25–26 August, 2021
Editors: Ralph L. Rose & Robert Eklund
Université Paris VIII Vincennes – Saint-Denis
2 rue de la Liberté - 93526, Saint-Denis, France
ISBN: xxx-xxx-xxx-xxx-x
DOI: https://doi.org/10.18463/DISS-2021-001

Table of contents

Second language acquisition and proficiency

Tasks and levels

Special day on (dis)fluency in speech and language disorders

Committees

Organizing Committee

Ivana Didirková, Chair
Université Paris 8 Vincennes – Saint-Denis, France

Robert Eklund
Linköping University, Sweden

Pierre-Olivier Gaumin
Université Paul-Valéry Montpellier 3, France

Fabrice Hirsch
Université Paul-Valéry Montpellier 3, France

Takeki Kamiyama
Université Paris 8 Vincennes – Saint-Denis, France

Sébastien Le Maguer
ADAPT Centre / Trinity College Dublin, Ireland

Ralph L. Rose
Waseda University, Japan

Sabina Tabacaru
Université Paris 8 Vincennes – Saint-Denis, France

Proceedings

Ralph L. Rose
Waseda University, Japan

Robert Eklund
Linköping University, Sweden

International Scientific Committee

Jens Allwood
University of Götenburg, Sweden

Judit Bóna
Eötvös Loránd University, Hungary

Ludivine Crible
The University of Edinburgh, Scotland

Liesbeth Degand
Université catholique de Louvain, Belgium

Andrea Deme
Eötvös Loránd University, Hungary

Christelle Dodane
Université Paul-Valéry Montpellier 3, France

Robert Eklund
Linköping University, Sweden

Camille Fauth
Université de Strasbourg, France

Dodji Gbedahou
Université Paul-Valéry Montpellier 3, France

Mária Gósy
Eötvös Loránd University, Hungary

Robert Hartsuiker
Ghent University, Belgium

Fabrice Hirsch
Université Paul-Valéry Montpellier 3, France

Peter Howell
University College London, United Kingdom

Robin Lickley
Queen Margaret University, United Kingdom

Takeki Kamiyama
Université Paris 8, France

Loulou Kosmala
Université Sorbonne Nouvelle, France

Kikuo Maekawa
The National Institute for Japanese Language and Linguistics, Japan

Alexandra Markó
Eötvös Loránd University (ELTE), Hungary

Marie-Claude Monfrais-Pfauwadel
Université Paul-Valéry Montpellier 3, France

Marine Pendeliau-Verdurand
Université Paul-Valéry Montpellier 3, France

Claire Pillot-Loiseau
Université Sorbonne Nouvelle, France

Laurent Prévot
Aix-Marseille University, France

Typhanie Prince
Université Paul-Valéry Montpellier 3, France

Ralph Rose
Waseda University, Japan

Vered Silber-Varod
The Open University of Israel, Israel

Sabina Tabacaru
Université Paris 8, France

Shu-Chuan Tseng
Institute of Linguistics, Taiwan

Preface

Organized for the first time in Berkeley in 1999, then successively in Edinburgh (2001), Göteborg (2003), Aix-en-Provence (2005), Tokyo (2010), Stockholm (2013), Edinburgh (2015), Stockholm (2017), and Budapest (2019), the Disfluency in Spontaneous Speech (DiSS) workshops are a privileged place for specialists working on questions related to speech fluency.

In 2021, 16 years after its first French edition, DiSS was planned to return to France. Due to the particular circumstances related to the pandemics, however, the 2021 edition is a fully virtual event. Despite the situation, we are pleased to see, again, many quality contributions to the field. During this edition, communications on disfluency in discourse and second language acquisition, filled pauses, and attitudes related to disfluent speech will be presented. Other papers bring new insights on the importance of the task on studies in speech fluency.

In addition, just like two years ago, DiSS 2021 features a co-located special day on (dis)fluency in pathological speech, including presentations addressing different disorders, such as dysarthria, multiple sclerosis, Alzheimer's disease, and developmental language disorders.

DiSS 2021 also welcomes three keynote speakers: Liesbeth Degand (Université catholique de Louvain, Belgium), Vered Silber-Varod (Open University of Israel), and Bridget Welsh (Michigan State University, US).

Our thanks go to everyone who helped organize this event: the whole organizing committee to help with the logistics, the scientific committee for their insightful comments, all the contributors, and all the participants to make DiSS an exciting gathering. Special thanks to the Université Paris 8 Vincennes – Saint-Denis and Université Paul-Valéry 3 Montpellier, and the French ANR (Agence Nationale de la Recherche) funding agency.

Saint-Denis, August 2021

Ivana Didirková

Discourse markers as markers of (dis)fluency: The role of peripheral position

Liesbeth Degand
Université catholique de Louvain, Louvain-la-Neuve, Belgium

Abstract

Discourse markers (DM) can be seen as enabling fluent speech though also as markers of disfluent moments in speech production. The present work seeks to resolve this apparent contradiction by examining the syntagmatic distribution of DMs. The Louvain Corpus of Annotated Speech—French was analyzed to look at DMs in peripheral and non-peripheral positions. Results show that speakers tend to start clauses and turns with sequential DMs, while ending with interpersonal DMs. In contrast, speakers tend to use rhetorical DMs at the start of intonation units, while ending with sequential DMs. In conclusion, peripheral clausal DMs can be seen as markers of fluency, placing them at the discourse-grammar interface.

Introduction

Studies on the relationship between discourse markers (DMs) and (dis)fluency have a Janus-headed face. On the one hand, DMs are described as structuring devices key to the local and global organization of discourse. As such, they contribute to its overall fluency. On the other hand, they have been described as traces of impediments in the speech production process, thus signalling disfluency. In other words, DMs are characterized by "functional ambivalence", a notion reflecting their effects as *symptoms* of production difficulties and as *signals* of inferences to be made (Crible, 2018, 3). The research question that follows from this observation is: How can we (reliably) disentangle fluent from disfluent use of Discourse Markers? The preliminary and incomplete answer to this question is: By looking at their syntagmatic distribution.

The syntagmatic distribution of Discourse Markers

Two main lines of research directly come to mind when addressing the relationship between DMs and their distribution in the flow of speech. The first line of research concerns studies investigating how a DM's position may influence its particular (contextual) meaning or function, with a focus on DMs in peripheral position. According to the *Subjectivity, Intersubjectivity and Peripheries Hypothesis* (SIPH, Jiménez, Arguedas, & Bordería, 2018), subjective meanings of DMs tend to be associated with the left periphery (initial position), while intersubjective meanings tend to be associated with the right periphery (final position) (e.g. Beeching & Detges, 2014), even if this hypothesis has been nuanced by quite some authors (see e.g. Haselow, 2012, Heim, 2019, Traugott, 2012, inter alia). The second line of research on the syntagmatic distribution of DMs concerns their role as segmentation or boundary markers (Horne et al., 1999). It is observed that "[p]articipants … employ discourse markers at conversational action (…) boundaries, in order to construct the frame shifts taking place throughout their interaction (…), often by projecting (…) the nature of these shifts (…)." (Maschler & Schiffrin, 2015, 194). It is this second line of research that I aim to connect to the (dis)fluency account of DMs through the hypothesis that DMs in peripheral position have a fluent boundary marking function, while DMs in non-peripheral position would be symptomatic of disfluent use. We will show that this hypothesis needs to be fine-tuned considering the type of host unit under study.

What type of units of talk do Discourse Markers bracket?

On the basis of previous work investigating the relationship between DM function, DM position and the linguistic type of host unit, we know that DMs do not pattern similarly at the boundaries of syntactic clauses, intonation units or speech turns (Degand & Crible, 2021). In other words, in order to fully grasp the boundary function of DMs, we need to consider the nature of the segmentation unit at stake.

The data of this study consists of a sample of (semi)-interactive and spontaneous speech of LOCAS-F (Louvain Corpus of Annotated Speech—French; Degand, Martin, & Simon, 2014) corresponding to 15 interactions, 20,086 words, 72 minutes. The data was sound-aligned and annotated under Praat and Exmaralda. Different levels of segmentation were defined independently from one another by means of "surface" elements only, i.e. distinct operational criteria belonging to a single level of linguistic analysis.

https://doi.org/10.18463/diss-2021-020-degand

- Dependency syntax for clausal segmentation, with peripheral DMs defined as immediately preceding or following the clause;
- Prosodic boundaries and intonation contours for prosodic segmentation, with peripheral DM defined as first DM or final DM of the unit;
- Speaker changes for turns, with periphery defined as first or last DM uttered by speaker in one turn.

Discourse Markers were manually identified on the basis of syntactic optionality, high degree of grammaticalization, procedural meaning, discourse-level scope, there was no closed list (Crible, 2018), resulting in a set of 853 tokens for the present study. These DMs were manually annotated in terms of domains and functions (Crible & Degand, 2019). Where the ideational domain covers DM uses related to states of affairs in the world, semantic relations between events; the rhetorical domain considers speaker's meta-commenting, reasoning and attitude; the sequential domain accounts for DMs used as local and global structuring devices, expressing progressing steps of speech flow; and the interpersonal domain concerns speaker-hearer management and addressee-oriented uses.

Peripheral distribution of Discourse Markers

Table 1 presents the peripheral distribution of DMs in clauses, intonation units and interactional turns. Strikingly, DMs seem to work as clausal boundary markers (with a strong preference for initial boundary). This observation leads us to consider syntactic (dependency) clauses as potentially "better" units of discourse segmentation. This is in contrast with intonation units which do not seem to be fit as units of discourse segmentation.

*Table 1: Peripheral distribution of DMs (*Intonationally autonomous DMs are not represented (9.3%))*

	Initial	Final	Medial
Clause	653	120	80
	773 (90,6%)		9.4%
Intonation unit*	185	171	418
	356 (41,7%)		49%
Turn	202	56	595
	258 (30.2%)		69,8%

To further investigate these initial results, I investigated the domains in which peripheral DMs work at the different levels of segmentation (see Tables 2 and 3).

General results suggest that within segments, initial and final position fulfil different speaker needs, in line with both SIPH and the processual view of grammar, but intonation units are less in line with the predictions. More in particular, clauses and turns both share a tendency to start with sequential DMs and another tendency to end with interpersonal DMs, although these uses are not exclusive. Intonation units stand apart with a larger proportion of rhetorical uses in initial position and of sequential uses in final position. Across segment types, peripheral use is better accounted for in clausal segmentation than at intonational or turn level.

Table 2: Domain distribution of initial DMs across segments ($X^2(6) = 85.22$; $p < .0001$)

	Clause-initial	Intonation-initial	Turn-initial
Ideational	109	47	32
Rhetorical	231	91	37
Sequential	282	33	110
Interpersonal	49	14	23

Table 3: Domains of final DMs across segments ($X^2(4) = 29.36$; $p < .0001$; ideational left out)

	Clause-initial	Intonation-initial	Turn-initial
Ideational	1	11	2
Rhetorical	28	50	18
Sequential	26	68	10
Interpersonal	65	42	26

Conclusions

Syntax is where DMs act most frequently as boundary markers, i.e. clauses are Schiffrin's units of talk that DMs are bracketing, rather than intonation units or turns. Peripheral clausal DMs can therefore be considered as markers of fluency with functionally-motivated uses both in initial position and final position, this is also the case for turns but to a lesser extent. DMs in medial position do not fulfil their bracketing function. These general results are in line with Thompson and Couper-Kuhlen (2005, 481) stating that: "the clause is a locus of interaction, in the sense that it is one of the most frequent grammatical formats which speakers orient to in projecting what actions are being done by others' utterances and in acting on these projections".

Thus, DMs are placed more firmly at the discourse-grammar interface, as elements that are out of the scope of micro-syntax but still play an important role at the macro-syntactic or macro-grammatical level, thus accounting for linguistic regularities that are grounded functionally.

References

Beeching, K. & U. Detges (eds.). 2014. *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change*. Studies in Pragmatics, Volume 12. Leiden, Netherlands: Brill. https://doi.org/10.1163/9789004274822

Crible, L. & L. Degand. 2019. Domains and Functions: A Two-Dimensional Account of Discourse Markers. Discours. *Revue de Linguistique, Psycholinguistique et Informatique. A Journal of Linguistics, Psycholinguistics and Computational Linguistics* 24. https://doi.org/10.4000/discours.9997

Crible, L. 2018. *Discourse Markers and (Dis)fluency. Forms and functions across languages and registers*. Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/pbns.286

Degand, L. & L. Crible. 2021. Discourse markers at the peripheries of syntax, intonation and turns. Towards a cognitive-functional unit of segmentation. In: D. Van Olmen & J. Šinkūnienė (eds.), *Pragmatic Markers and Peripheries*, 19–48. Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/pbns.325.01deg

Degand, L., L. Martin & A.-C. Simon. 2014. Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté [Basic discourse units and their left periphery in LOCAS-F, an annotated multigenre oral corpus]. In: F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer & S. Prévost (eds.), *CMLF 2014 -4ème Congrès Mondial de Linguistique Française*, July 19–23, 2014, Berlin, Germany, 2613–2626. https://doi.org/10.1051/shsconf/20140801211

Haselow, A. 2012. Subjectivity, intersubjectivity and the negotiation of common ground in spoken discourse: Final particles in English. *Language & Communication* 32(3), 182–204. https://doi.org/10.1016/j.langcom.2012.04.008

Heim, J. 2019. Turn-peripheral Management of Common Ground: A Study of Swabian *gell*. *Journal of Pragmatics* 141, 130–146. https://doi.org/10.1016/j.pragma.2018.12.007

Horne, M., P. Hansson, G. Bruce, J. Frid, & M. Filipsson. 1999. Discourse Markers and the Segmentation of Spontaneous Speech: The Case of Swedish Men 'but/and/so'. *Working Papers, Lund University, Dept. of Linguistics* 47, 123–139.

Jiménez, S. S., M. E. Arguedas, & S. P. Bordería. 2018. Beyond the notion of *periphery*: An account of polyfunctional discourse markers within the Val.Es.Co. model of discourse segmentation. In: K. Beeching, C. Gezzin, & P. Molinelli (eds.), *Positioning the Self and Others: Linguistic Perspectives*, 105–125. Pragmatics and Beyond New Series, 292. Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/pbns.292.05sal

Maschler, Y. & D. Schiffrin. 2015. Discourse markers: Language, meaning, and context. In: D. Tannen, H. E. Hamilton, & D. Schiffrin (eds.), *The Handbook of Discourse Analysis*, 2nd Edition, 189–221. Hoboken, NJ, USA: John Wiley & Sons. https://doi.org/10.1002/9781118584194.ch9

Thompson, S. A. & E. Couper-Kuhlen. 2005. The Clause as a Locus of Grammar and Interaction. *Discourse Studies* 7(4-5), 481–505. https://doi.org/10.1177/1461445605054403

Traugott, E. C. 2012. Intersubjectification and clause periphery. *English Text Construction* 5(1), 7–28. https://doi.org/10.1075/etc.5.1.02trau

DiSStory: A computational analysis of 9 editions of Disfluency in Spontaneous Speech workshop

Vered Silber-Varod
The Open University of Israel, Israel

Abstract

What are the most prominent research topics of the DISS workshops? Do we see any shift over the years? Can we identify the specific terms used in the research of disfluency? At the 10th workshop of DiSS, I will present some answers I have come up with using a data-driven approach on the database of abstracts published in the proceedings of DiSS workshops from 1999 to 2021. In this talk I call the participant to "Trust the text", as Sinclair and Carter (2004) entitled their book, and to join the journey into the DiSS story.

Introduction

The first Workshop on Disfluency in Spontaneous Speech (DiSS) was held as a one-day satellite meeting of the International Congress for Phonetic Sciences on July 30, 1999 at the University of California Berkeley. In the 12 papers that were presented back then, the authors discussed issues such as "Which speakers are most disfluent in conversation, and when?" (Bortfeld et al., 1999), "Uhs and interrupted words: The information available to listeners" (Brennan & Schober, 1999), "Speech Repairs: A Parsing Perspective" (Core & Schubert, 1999), "Between-Turn Pauses and Ums" (Fox Tree, 1999). It seems that most of the studies in that first workshop were carried out on English speech databases, but other languages were represented as well, such as Swedish in the paper "A Comparative Analysis of Disfluencies in Four Swedish Travel Dialogue Corpora" (Eklund, 1999) and Japanese in the paper "Detecting and Correcting Speech Repairs in Japanese" (Heeman & Loken-Kim, 1999). Like in future DiSS workshops, the organizers assigned each paper to three thematic sessions: Production issues, Perception, and automatic speech recognition (ASR)/computational linguistics (CL) approaches. In the second workshop in 2001 (Lickley & Shriberg, 2001), these themes indeed repeated and new ones were added: Annotation and disfluency types, prosody and phonetics, and disfluency as a general cognitive phenomenon.

In the third workshop, an introduction (Preambulum) was written to the proceedings (Eklund, 2003, 3–4), mentioning a list of disfluency types "… pauses, hesitations, 'err' words, truncated words, repetitions, prolonged sounds, repairs, etc." and the varied fields of research within which these phenomena were studied: stuttering research, general linguistics, and psychotherapy. Eklund (2003, 4) further acknowledge the rich terminology that was developed in the field and concluded that "… these proceedings cover several different disciplines and are thus illustrative of the interdisciplinary character of this area."

Following this rich and diverse field of research, I have decided to look for the terms, words, and concepts that researchers used along the years in the ten editions (Table 1) and to try to identify patterns or peaks of interest in their occurrences, using a data-driven approach.

Methodology

Already in 2004 Sinclair and Carter (2004) entitled their book "Trust the text: Language, Corpus and discourse". For many years researchers are using a corpus-based analysis to describe the features of a particular language or genre. The accelerated use of the corpus linguistics approach was the result of the development of sophisticated algorithms for natural language processing. The use of advanced statistical

Table 1. A summary of contributions in each of the 10 workshops.

Year	Location	Papers presented
1999	University of California Berkeley, USA	12
2001	University of Edinburgh; Edinburgh, Scotland, UK	26
2003	Göteborg University; Göteborg, Sweden	19
2005	Université de Provence; Aix-en-Provence, France	34
2010	University of Tokyo; Tokyo, Japan	31
2013	KTH Royal Institute of Technology; Stockholm, Sweden	19
2015	University of Edinburgh; Edinburgh, Scotland, UK	24
2017	KTH Royal Institute of Technology; (Stockholm, Sweden	14
2019	ELTE Eötvös Loránd University; Budapest, Hungary	21
2021	Paris 8 University (Paris Saint-Denis), France (online conference)	20
Total		220

https://doi.org/10.18463/diss-2021-021-silber_varod

methods to analyze the data of large-scale textual databases makes it possible to identify patterns of change and to encode essential qualities expressed in the texts (Silber-Varod, Eshet-Alkalai, & Geri, 2016; 2019).

One of the spin-offs of corpus linguistics is called 'Culturomics' a term used by Michel and colleagues (2011). Culturomics quantitatively investigates massive digital arrays of written text and spoken language to examine cultural patterns in various disciplines (Bohannon, 2011). This type of investigation is also termed "distance reading" (Moretti, 2014).

Although data-driven approach for discourse analysis is dated, I cannot avoid relating it to the emergence of data science field of research. Many developments in recent years have contributed to the leap in data science, including the huge use of social networks, developments in the field of big data, computer vision and natural language processing. Today our tour includes only few NLP tools.

Material

The ten reference documents in a bibtext format were taken from the Filled Pause Research Center website (Rose, filledpause.org). All metadata were manually removed leaving the titles and the abstracts.

Preprocessing

The textual data underwent:

1. Lemmatization for English texts using online lemmatizer (Trudove, SEOHorseSense.com)
2. Applying stop list
   a. English stop list using those embedded in AntConc (Anthony, 2020) and in Voyant tools (Sinclair & Rockwell, 2021).
   b. "DiSS stop list" was manually created to remove further academic writing vocabulary, such as *study*, *we*, *paper*, *results*, *found*, and more.

Corpus size after the preprocessing consists of 17,375 word-tokens and 3,403 unique word forms (types).

Findings

What are the most frequent words in each workshop?

Looking at the word clouds in Figure 1, we do see differences in the most frequent words in each year. Recognition and word are most frequent in the first workshop; pause and error in the second; word and repair in 2003; pause and prosodic in the fourth. Although pause is most frequent in several years, I did not want to include it in the DiSS stop-list, beacause it is interesting to see that it is not that prominent in every single year.

What are the languages researchers talked about during the past years at DiSS workshops?

The frequency does not say too much but the authors phrasing choices in the abstracts. The range values however indicate in how many years these languages are mentioned. Notice that the range data were automatically extracted from AntConc (Anthony, 2020), based only on the abstracts' content. Table 2 presents the frequency and distribution among the 10 editions (Range column) of the 25 languages found in the abstracts and titles. Only Japanese was mentioned in all ten abstract proceedings and English (whether, American, British, or Canadian) in nine. This was manually re-checked and I found that studies on English corpora were presented in 2003, however, not all authors bothered to mention the language in the abstract or title of their paper. Eight languages were mentioned only in a single workshop.

How unique is each workshop?

The last analysis that will be presented here is the unique contribution of each workshop. My goal was to compare the titles and abstracts of a single year to the documents of previous years and by this to achieve an evolution of the topics along the years. There are several methods/algorithms to achieve it and the terminology is also varied. In Voyant tools it is called Distinctive words. The idea behind it is to compare a certain portion of the corpus to the rest of the corpus, or to a different corpus. In AntConc this called Keyness. This tool, like others, allows to identify characteristic words in the corpus. The most frequent method is called Term Frequency-Inverse Document Frequency (TF-IDF), a numerical statistic that is intended to reflect how important a word is to

Figure 1. Ten word-clouds per each edition. The cloud shape 1 represents the 1999 edition; P represents the 2021 edition (P for Paris). © https://www.wordclouds.com/.

a document in a collection or corpus (Rajaraman & Ullman, 2011).

Since 1999 was the first year, no comparison was made and all words with frequency > 5 were listed as "unique". From 2001 and on I compared to the documents of previous years. The keyness lists were created automatically using AntConc tool (Anthony, 2020).

I then mapped each keyword into a category. For example, Japanese was categorized as Language, recognition was categorized as Technology, and so on. By the end of this manual annotation process, all the keywords were mapped into 10 categories: Dissfluency types, Technology, Subjects (speakers), Method terms, Structure-unit (linguistics), Prosody, Languages, Corpora, Acoustic terms, and Cognition. Figures 2-4 present their distribution and their trend lines (black dotted line).

Figure 2 presents those categories with a slope downward. Those are: Structure-unit (linguistics), Methods, Technology, and Dissfluency types. The meaning of this trend is that in future workshops, we can expect to find less and less divergence in the terminology of this category. This is not to claim that papers will not discuss those issues(!), but that the vocabulary has reached a certain saturation.

Table 2. Frequency and distribution (Range column) of the 25 languages found in the abstracts and titles of the 10 editions.

Language	Frequency	Range
Japanese	92	10
English	47	9
German	28	7
Mandarin	21	6
French	45	5
Hungarian	14	5
Swedish	10	5
Spanish	18	4
American English	5	4
Hebrew	13	3
Chinese	10	3
Portuguese	10	3
European Portuguese	7	3
British English	4	3
Italian	7	2
Estonian	6	2
Tok-Pisin	3	2
Thai	5	1
Canadian	3	1
Russian	2	1
Sri-Lankan English	2	1
Taiwanese	2	1
Arabic	1	1
Austronesian	1	1
Latin-American Spanish	1	1

Figure 3 presents two categories that had a level trend until 2019 and the trend changed in 2021. Those are: Acoustics and Prosody.

Figure 4 presents those categories with a slope upward. Those are: Cognition, Corpus, Subjects, and Language. This means that in future workshops, we can expect to find more divergence in the terminology of those categories.

The uniqueness analysis exemplifies how computational methods (such as the Keyness algorithm) benefits from qualitative methods (such as annotation and labeling) to achieve meaningful insights.

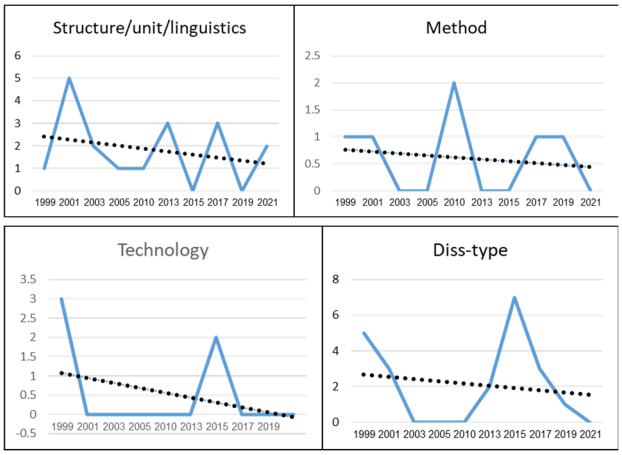

Figure 2. Downward slopes of four Key categories over the years.

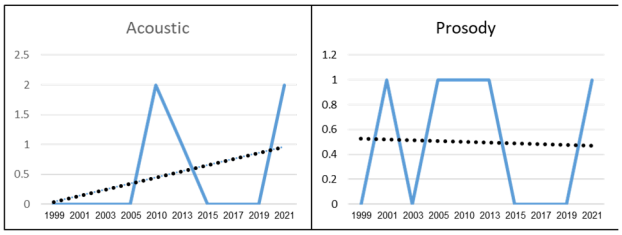

Figure 3. Trends of key terms in the acoustic category and the prosodic category over the years.

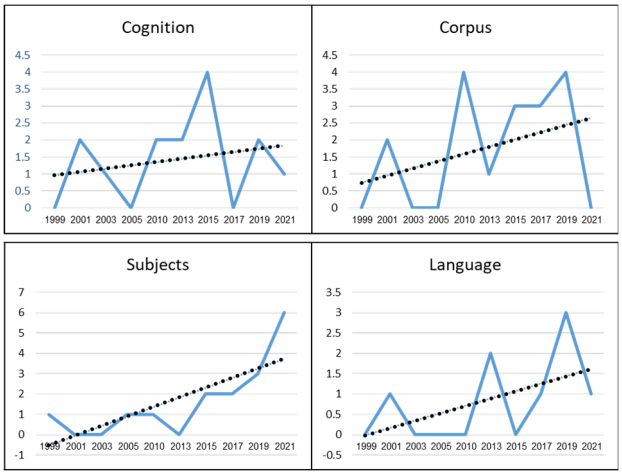

Figure 4. Upward slopes of four Key categories over the years.

Summary

The computational analysis showed that we can forecast trends in the field, like the rising interests in different types of speakers (subjects) and in more varied languages. On the other hands the applicative aspect of disfluencies becomes marginal.

I also showed several target words and topics that are still missing from the discourse of the workshop: Neuro terminology is marginal and linguistics as well.

Although my analysis is based on a small amount of data, I trust it to be "clean" and without "noises". I emphasized the difference between computational methods that unveil the use of different terms and the interpretations that are not part of the process but are required by scholars in the field on top of the findings.

References

Anthony, L. 2020. AntConc (Version 3.5.9). http://www.laurenceanthony.net/software/antconc/ (accessed 31 July, 2021).

Bohannon, J. 2011. Google Books, Wikipedia, and the future of culturomics. *Science 331*(6014), 135. https://doi.org/10.1126/science.331.6014.135

Bortfeld, H., S. D. Leon, J. E. Bloom, M. F. Schober, & S. E. Brennan. 1999. Which speakers are most disfluent in conversation, and when?. In: *Proceedings of Disfluency in Spontaneous Speech Workshop*, Berkeley, CA, USA, 30 July, 1999, 7–10.

Brennan, S. E. & M. F. Schober. 1999. Uhs and interrupted words: The information available to listeners. In: *Proceedings of Disfluency in Spontaneous Speech Workshop*, Berkeley, CA, USA, 30 July, 1999, 19–22.

Core, M. G. & L. K. Schubert. 1999. Speech Repairs: A Parsing Perspective. In: *Proceedings of Disfluency in Spontaneous Speech Workshop*, Berkeley, CA, USA, 30 July, 1999, 47–50.

Eklund, R. 2003. Preambulum. In: R. Eklund (ed.), *Proceedings of DiSS '03, Disfluency in Spontaneous Speech Workshop*, Gothenburg Papers in Theoretical Linguistics 90, Göteborg, Sweden, 5–8 September, 2003, 3–4.

Eklund, R. 1999. A Comparative Analysis of Disfluencies in Four Swedish Travel Dialogue Corpora. In: *Proceedings of Disfluency in Spontaneous Speech Workshop*, Berkeley, CA, USA, 30 July, 1999, 3–6.

Fox Tree, J. E. 1999. Between-Turn Pauses and Ums. In: *Proceedings of Disfluency in Spontaneous Speech Workshop*, Berkeley, CA, USA, 30 July, 1999, 15–17.

Heeman, P. A. & Loken-Kim, K.H. 1999. Detecting and Correcting Speech Repairs in Japanese. In: *Proceedings of Disfluency in Spontaneous Speech Workshop*, Berkeley, CA, USA, 30 July, 1999, 43–46.

Lickley, R. & E. Shriberg (eds.). 2001. *Proceedings of the second Workshop on Disfluency in Spontaneous Speech (DiSS '01)*, Edinburgh, Scotland, UK; 29–31 August, 2001.

Michel, J. B. et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science 331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Moretti, F. 2013. *Distant reading*. London/New York: Verso Books.

Rajaraman, A. & J. D. Ullman. 2011. *Data Mining*. In: A. Rajaraman & J. D. Ullman (eds.), *Mining of Massive Datasets,* Cambridge, UK: Cambridge University Press, 1–17. https://doi.org/10.1017/CBO9781139058452.002.

Rose, R. n.d. Filled Pause Research Center [website]. https://filledpause.org/diss/ (accessed 31 May, 2021).

Silber-Varod, V., Y. Eshet-Alkalai, & N. Geri. 2016. Analyzing the discourse of Chais Conferences for the study of innovation and learning technologies via a data-driven approach. *Interdisciplinary Journal of e-Skills and Life Long Learning 12*, 297–313. https://doi.org/10.28945/3604

Silber-Varod, V., Y. Eshet-Alkalai, & N. Geri. 2019. Tracing research trends of 21st-century learning skills. *British Journal of Educational Technology 50*(6), 3099–3118. https://doi.org/10.1111/bjet.12753

Sinclair, J. & R. Carter. 2004. *Trust the text: Language, corpus and discourse*. London/New York: Routledge.

Sinclair, S. & G. Rockwell. 2021. Voyant Tools [website]. https://voyant-tools.org/ (accessed 18 March 2021).

Trudov, A. n.d. SEO Horse Sense [website]. https://seohorsesense.com/ (accessed 31 July, 2021).

Attitudinal correlates of word-internal disfluencies in Japanese communication

Toshiyuki Sadanobu
Kyoto University, Kyoto, Japan

Abstract

Through a case observation and a questionnaire survey, this presentation seeks to elucidate the patterns of word-internal disfluency in Japanese communication and determine how speakers implement these patterns. Two conclusions can be drawn: (i) Four possible patterns of word-internal disfluency exist in Japanese communication. Some cases show that disfluency that superficially appears not to be prolonged may come under prolongation. (ii) Some deviations are observed in disfluency patterns in accordance with the speaker's attitude; all four patterns can be seen to occur in hesitant attitudes, whereas those expressed in the attitude of surprise primarily belong to the "suspending and restarting" pattern. However, where the degree of surprise is low or close to disgust, disfluency is more likely to be expressed as "prolonging and continuing."

Introduction

The objective of this research is to clarify the patterns of word-internal disfluency in Japanese communication and the attitudes in which they are expressed. Disfluency, as used in this paper, refers to deviations from smooth pronunciation in able-bodied person's speech in general (excluding pronunciation resulting from emphasis).

Word-internal disfluency can be observed from three perspectives: the position where disfluency occurs, the form of expression, and the treatment after getting stuck. Previous research has focused on prolonging, where the "Morphology Matters Hypothesis" has been proposed (Eklund, 2001, 6–7, 2004, 251).

This hypothesis states that the position in which prolongation occurs reflects the complex internal structures of words in the language. Specifically, Swedish and American English, which tend to have several consecutive consonants within a syllable (with up to three consonants before a vowel and up to eight consonants after a vowel in Swedish), are likely to have complex morphological structures, and therefore 30% of hesitant prolongation has been reported to occur at the beginning of words (the first segment), 20% in the medial part of the words (neither the beginning nor the end of the words), and 50% at the end of the words (the final segment). In other words, hesitant prolongation is not uncommon at the beginning of the words or even in the middle of words in these languages. Conversely, Tok Pisin, which lacks such complex sequences of consonants within a syllable (allowing up to two consonants before a vowel and up to one consonant after a vowel), is reported to have the corresponding distribution of 15% at the beginning of the words, 0% in the medial, and 85% at the end (Eklund, 2001, 6–7, 2004, 251), and is classified with Japanese (10-5-85%, Den, 2003) and Mandarin Chinese (4-1-95%, Lee et al., 2004; Eklund, 2004, 251). More recently, this hypothesis has been supported by reports that Hungarian, with its rich morphology, is more similar to the former group (18-19-63%, Gósy & Eklund, 2017), while German, which closely resembles Swedish, does not follow the hypothesis (7-15-78%, Betz, Eklund & Wagner, 2017), suggesting that the "Morphology Matters Hypothesis" alone is not enough to explain the positions of hesitant prolongation (Betz, Eklund & Wagner, 2017).

The aforementioned studies analyzed the prolongation type of disfluency from a macroscopic perspective based on a large corpus. However, there may be cases where the prolongation type requires a more detailed analysis. It should also be noted that, although the above research focused on disfluency expressed with an attitude of hesitation, the actual circumstances of the "hesitation" have yet to be clarified.

This presentation will focus on word-internal disfluency in Japanese communications, showing what types of disfluency patterns exist within words and illuminating the attitudes with which the speakers express these patterns.

Four possibilities for the patterns of word-internal disfluency

Two possible types of word-internal disfluency can be assumed: disfluency resulting from suspension in pronunciation (Suspending) and disfluency from prolonged pronunciation (Prolonging). In addition to these, there are two other possibilities for the speaker's treating after disfluency, other than rephrasing the utterance, inserting fillers, or giving up on the utterance and

https://doi.org/10.18463/diss-2021-002-sadanobu

stopping speaking: One is to go back to the beginning of the word after getting stuck (Restarting) and the other is to continue pronouncing the rest of the word without going back to the beginning (Continuing). Suppose, for example, that in the word "AB," disfluency occurs when the speaker says "A." In this case, there are a total of four possible patterns of disfluency: (i) Suspending + Restarting "A, AB"; (ii) Suspending + Continuing "A, B"; (iii) Prolonging + Restarting "A:AB"; and (iv) Prolonging + Continuing "A:B" (Table 1, Sadanobu et al., 2018).

Table 1. Four Patterns of Disfluency within the Word "AB."

Treatment Forms of expression	Restarting	Continuing
Suspending	"A, AB"	"A, B"
Prolonging	"A:AB"	"A:B"

We examined whether the above four possible patterns exist in real cases. The database used for the observation included 264 stories by native Japanese speakers among entries in the speech contest of funny stories, "My Funny Talk," (Sadanobu, 2018), which we have been organizing since 2010. These are available free of charge on the Internet with Japanese subtitles (some of them multilingual; http://www.speech-data.jp/chotto/tile/tile.cgi).

Given that these stories were contest entries, the speakers appeared to be somewhat prepared for what they were going to say. Nevertheless, all four of the above patterns were attested in these utterances (Sadanobu, 2021). The following are actual examples of them, one by one.

Suspending + Restarting ("A, AB")

In Op. 8 from 2011, the speaker utters [soɕi˳, soɕi˳te] when she should say /sosite/ [soɕi˳te] ("and," 00:21–00:22).

The utterance stops when [ɕi˳], the second mora of the conjunction /sosite/, is pronounced, and 0.24 second later, the speaker goes back to the beginning of the word and restarts it as /sosite/ (Figure 1 and other figures are based on Praat; Boersma & Weenink, 2006).

Suspending + Continuing ("A, B")

In Op. 10 from 2011, the speaker states [okɑ, ɕi:ndesɯjo] when she should say /okasiiɴdesuyo/ [okɑɕi:ndesɯjo] ("it is strange," 02:30–02:31). In the part [okɑ, ɕi:], the utterance stops when /ka/, the second mora of the adjective /okasii/ ("strange," the stem: *okashi*), is pronounced, and the pronunciation of the rest of the word /sii/ is continued 0.08 seconds later (Figure 2).

Prolonging + Restarting ("A:AB")

In Op. 13 from 2013, the speaker states [kjɯ:ʑɯ::kjɯ:ʑɯ::roku] when she should say /kju:zju:roku/ [kjɯ:ʑɯ:roku] ("ninety-six," 00:35–00:37). Here, the speaker returns to the beginning of the word after prolonging /zju:/ (0.29 s), the second syllable of /kju:zju:/ ("ninety") and restarts it as /kju:zju:roku/. The restarted /kju:zju:roku/ also becomes anchored in the "Prolonging + Continuing" manner described below, meaning that two patterns of word-internal disfluency can be observed consecutively here (Figure 3).

Prolonging + Continuing ("A:B")

In Op. 7 from 2013, the speaker states [ke:sɑ] when she should say /kesa/ [kesɑ] ("this morning," 00:59–01:01). Here, the speaker prolongs /e/

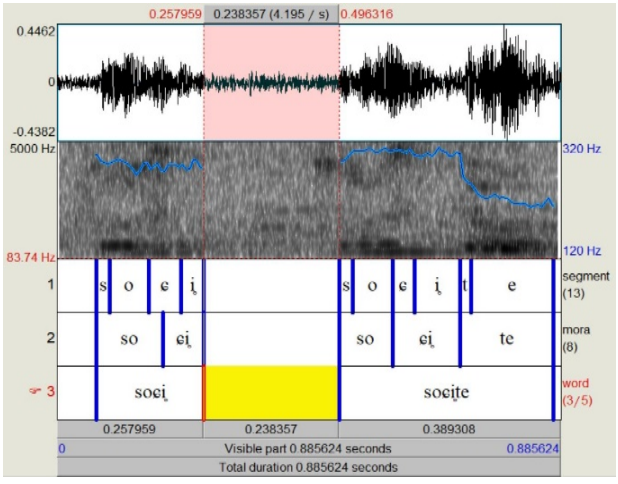

Figure 1. Sound wave, spectrogram, and F0 contour for [soɕi˳, soɕi˳te]

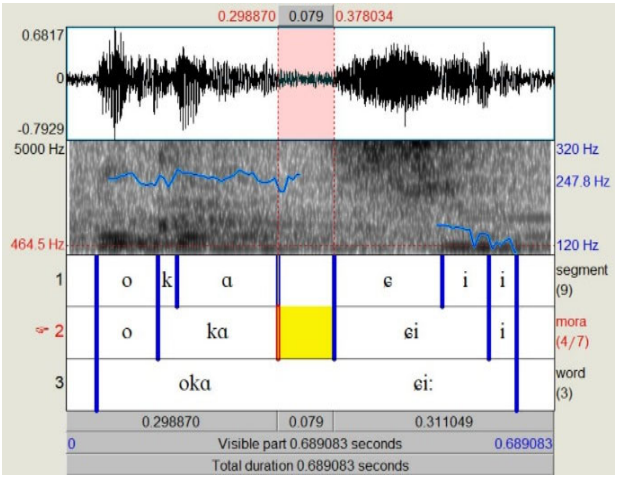

Figure 2. Sound wave, spectrogram, and F0 contour for [oka, ɕi:]

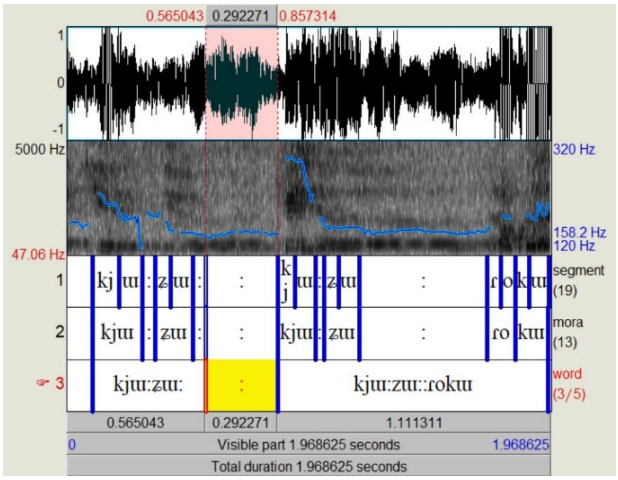

Figure 3. Sound wave, spectrogram, and F0 contour for [kjɯ:zɯ::kjɯ:zɯ::rokɯ]

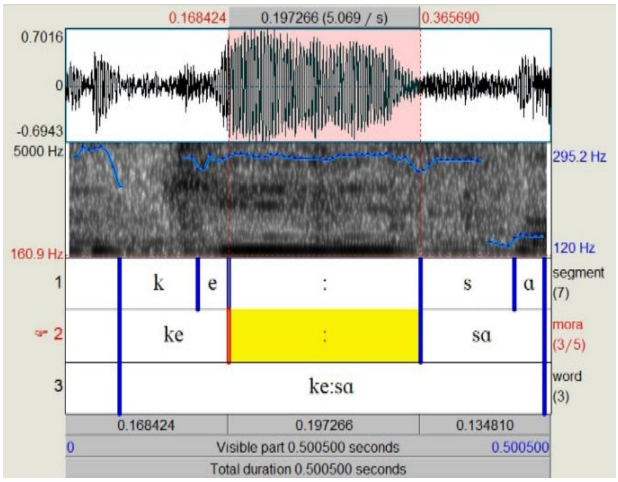

Figure 4. Sound wave, spectrogram, and F0 contour for [ke:sɑ]

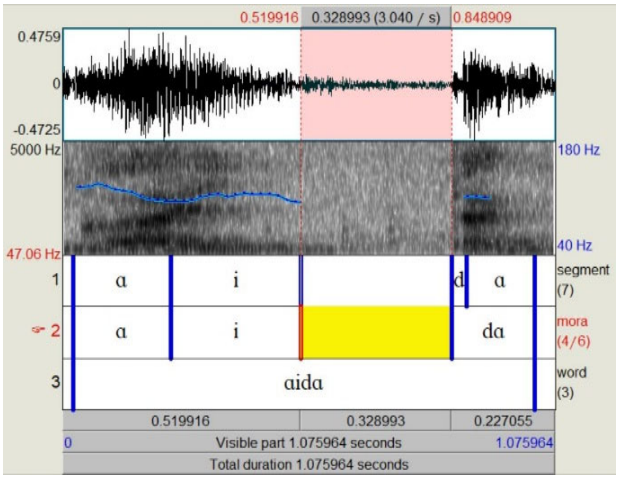

Figure 5. Sound wave, spectrogram, and F0 contour for [ai..da]

(0.20 s), the voice of the first mora of the simplex noun /kesa/, and pronounces the rest of the word /sa/ continuously (Figure 4).

Ambiguity between patterns

Of the four patterns described above, Suspending + Continuing ("A, B") and Prolonging + Continuing ("A:B") may not be distinguishable by superficial observation alone.

The two patterns are indistinguishable when the B immediately following the disfluency begins with a plosive or affricate. These sounds originally cause the air passage to close at the beginning. Therefore, when the B begins with a plosive or affricate, it is not possible to determine whether the silent segment is simply the result of a stop in pronunciation (Suspending), or whether the closure of the airflow at the beginning of B is extended as a result of prolonging the pronunciation of B (Prolonging).

Here is an example that begins with a plosive sound: In Op. 70 in 2011, the speaker says [ɑi..dɑ] (0:08–0:09) when he should say /aida/ [ɑidɑ] ("while"). A silent segment follows immediately after the second mora /i/ of the noun /aida/ is pronounced. This is natural because the plosive sound [d], which is expected immediately after *ai* is pronounced, begins with a stoppage of the airflow. However, as the silent segment has a significantly longer duration (0.33 s) than usual, according to a native Japanese speaker's intuition, it can be concluded that disfluency occurs in this part (Figure 5).

The disfluency shown here, involving a longer-than-usual silent segment, is ambiguous because this can be considered either Suspending, where the pronunciation simply stops immediately after [ɑi], or Prolonging, where the airflow closure at the beginning of the plosive [d], the consonant of /da/, is extended. A similar ambiguity may arise for disfluency, accompanied by affricates.

In addition to the four patterns above, another conceivable pattern is Prolonging + Suspending + Continuing ("A:, B"), but since this pattern is uncommon and can be considered the result of a change in pattern from Prolonging + Continuing ("A:B") to Suspending + Continuing ("A, B"), it will not be addressed here.

Since it was determined that it can be difficult at times to differentiate between the "Suspension + Continuation" pattern and the "Prolongation + Continuation" pattern, further preparatory study is required before quantitatively comparing these patterns. On the other hand, it is possible at this time to investigate the correspondence between disfluency patterns and attitude

Attitude during speech

The examples provided here are those that appear in calm speech. Conversely, in novels and comics, it

is not uncommon to see characters speaking disfluently in an attitude of being surprised or shaken up. Word-internal disfluency is traditionally observed here in the manner of Suspending + Restarting as follows: (i) *Ya, yattana!* (*Matsubayashi Hen'ya*, Shūgorō Yamamoto, 1938); (ii) *Ma, makoto de gozari masuru ka* (*Shinshi Taikōki*, Ryōtarō Shiba, 1968); (iii) *Na, nante koto o!!* (*Doraemon*, vol. 25, Fujiko F. Fujio, 1982). We used a questionnaire survey to identify the patterns of word-internal disfluency that occur in different types of attitudes in actual daily communications.

Method

The survey was conducted online in April 2021. The respondents were 109 native Japanese speakers, who were paid to answer questions, of whom 33 were male and 73 were female, while 3 did not respond. In terms of age, there were 2 teens, 20 people in their 20s, 31 in their 30s, 31 in their 40s, 12 in their 50s, 6 in their 60s, and 4 over 70; 3 people did not respond to this question. The respondents' hometown areas were not examined. Since no significant deviations could be identified in relation to gender or age, the genders and ages of the respondents are treated without distinction. The respondents were asked to rate how natural each given speech sounded on a 5-point scale (1 point: *very unnatural*, 5 points: *very natural*). They could play the utterances as many times as they liked, with no time limit for responding.

A total of six questions were presented. In the first three of the six questions, the four aforementioned word-internal disfluency patterns were presented as four speech sounds. The respondents were asked how natural each of the speech sounds was. These three questions are as follows:

Question 1: In a conversation with an acquaintance, speaker 1 says, "My child is going to study abroad in California," and speaker 2 is surprised to hear this. Judge the naturalness of each of the four utterances of speaker 2 as a statement by the speaker:

(i) [kɑ, kɑɾifoɾɯniɑ=desɯ̥ (COP)-kɑ (Q)]
(ii) [kɑ, ɾifoɾɯniɑ=desɯ̥-kɑ]
(iii) [kɑːkɑɾifoɾɯniɑ=desɯ̥kɑ]
(iv) [kɑːɾifoɾɯniɑ=desɯ̥kɑ] ("Oh California")

Question 2: A person, speaker 1, who has moved into the house next door comes to say hello. When speaker 2 asks speaker 1 where they have been, speaker 1 unexpectedly replies with the name of a distant country, "Brazil." Speaker 2 is startled and shouts "Brazil" with disfluency. Judge how natural each of speaker 2's four utterance is as follows.

(i) [bɯ, bɯɾɑʑiɾɯ]
(ii) [bɯ, ɾɑʑiɾɯ]
(iii) [bɯːbɯɾɑʑiɾɯ]
(v) [bɯːɾɑʑiɾɯ]

Question 3: A quiz is being held in which contestants guess the name of a country by looking at its flag. The speaker thinks the flag might be Brazil's, but is unsure and hesitantly answers "Brazil" with disfluency. Judge how natural each of the four utterances of the speaker is as follows.

(i) [bɯ, bɯɾɑʑiɾɯ]
(ii) [bɯ, ɾɑʑiɾɯ]
(iii) [bɯːbɯɾɑʑiɾɯ]
(iv) [bɯːɾɑʑiɾɯ]

The latter three questions are specially designed to compare Suspending + Restarting with its "opposite" pattern, Prolonging + Continuing.

Question 4: A boss, speaker 1, who is watching the news during his lunch break at work advised an employee, speaker 2, to go home. When speaker 2 asks why, the boss, speaker 1, says, "Your house has been broken into and your wife is being held hostage." Judge how natural each of the two utterances of speaker 2's responses is as follows.

(i) [so, soɾe (that) =ɯɑ(TOP) tɑihen(terrible)=dɑ(COP) dʑɑ kɑeɾ-ɑse-te-itɑdɑki-mɑsu]
(ii) [soːɾe=ɯɑ tɑihen=dɑ dʑɑ kɑeɾ-ɑse-te-itɑdɑki-mɑsu] ("Oh no! I'll go home now")

Question 5: At an important business meeting, one of the other party's employees is absent. When speaker 1 asks why the employee is absent, another employee, speaker 2, replies, "He's gone home because his house has been broken into and his wife is being held hostage." Judge how natural each of the two utterances as a response to the answer is as follows.

(i) [so, soɾe=ɯɑ tɑihen=dɑ nɑɲigoto=mo nɑ-kereba ii=desɯ-ne]
(ii) [soːɾe=ɯɑ tɑihen=dɑ nɑɲigoto=mo nɑ-kereba ii=desɯ-ne] ("That sounds tough. I hope that nothing will happen")

Question 6: Judge how natural the responses of a surgeon are in each of two utterances after being told "You've successfully performed many difficult surgeries. You must have some magical powers in your hands."

(i) [so, soɾe=ɯɑ do(how)=deɕjoː(might be) -kɑ]
(ii) [soːɾe=ɯɑ doː=deɕjoː-kɑ] ("I'm not sure about that")

Results

The results for Questions 1–3 show a deviation toward Suspending + Restarting in the evaluations of Questions 1 and 2, this is mixed with an attitude of greater surprise, compared to the evaluation of Question 3, which reflects an attitude close to pure hesitation. As for Question 1, the medians of (i)–(iv) were 5, 2, 2, and 2, respectively. A Wilcoxon Signed-rank test shows a significant difference of naturalness between utterances (i) and (ii) (iii) (iv) ($p < 0.01$). The same can be said for Question 2 (medians: (i) 4, (ii) 1, (iii) 1, (iv) 1. $p < 0.01$), whereas this special preference for (i) is less clear in the case of Question 3 (medians: (i) 4, (ii) 3, (iii) 4, (iv) 3). The naturalness of (ii)–(iv) is significantly higher for Question 3 than those in Question 2 ($p < 0.01$), and the difference between (i) and (iii) is not significant ($p = 0.912$), although the difference between (i) and (ii) (iv) is significant ($p < 0.01$, $p < 0.05$).

Regarding the results for Questions 4–6, there is a preference for Suspending + Restarting for Question 4, where the response is assumed to be mixed with an agitated attitude since the matter is strongly related to the speaker themselves. The medians of (i) and (ii) were 4 and 2, respectively, and the difference of their naturalness is significant ($p < 0.01$). The same is true for Question 5 where the matter is strongly related to others (medians: (i) 4, (ii) 3), although the naturalness of (ii) is significantly higher than for Question 4 ($p < 0.01$). In the case of Question 6 (medians: (i) 3, (ii) 4), the naturalness of (ii) is significantly higher than for Question 4 ($p < 0.01$), and the difference in naturalness between (i) and (ii) is not found ($p = 0.594$), where the word-internal disfluency is with an attitude of disgust rather than surprise.

Concluding Remarks

As seen above, there are some deviations observed in word-internal disfluency patterns depending on the speaker's attitude.

All four patterns can appear in speech with a hesitant attitude, while in speech with an attitude of surprise, only Suspending + Restarting appears. However, if the degree of surprise is low or close to disgust, word-internal disfluency is more likely to be expressed in the manner of Prolonging + Continuing.

Acknowledgments

We thank our colleagues, especially Akiko Tabata for all their help. This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research, (S) 20H05630, and by NINJAL-Project "Cross-linguistic studies of Japanese prosody and grammar" and "Multiple approaches to analyzing the communication of Japanese language learners.".

References

Betz, S., R. Eklund & P. Wagner. 2017. Prolongation in German. In: R. Eklund & R. Rose (eds.), *Proceedings of DiSS 2017, The 8th Workshop on Disfluency in Spontaneous Speech*, 18–19 August, 2017, Stockholm, Sweden, 13–16.

Boersma, P. & D. Weenink. 2006. Praat: Doing phonetics by computer (version 6.1.48). https://www.praat.org/ (accessed 28 March 2021).

Den, Y. 2003. Some strategies in prolonging speech segments in spontaneous Japanese. In: R. Eklund (ed.), *Proceedings of DiSS 2003, Disfluency in Spontaneous Speech Workshop*, 5–8 September, 2003, Gōteborg, Sweden, 87–90.

Eklund, R. 2001. Prolongations: A Dark Horse in the Disfluency Stable. In: *Proceedings of DiSS '01 Disfluency in Spontaneous Speech*. 29–31 August, 2001, Edinburgh, Scotland, UK, 5–8.

Eklund, R. 2004. *Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues*. Ph.D. dissertation, Linkōping University, Sweden.

Gósy, M. & R. Eklund. 2017. Segment prolongation in Hungarian. In: R. Eklund & R. Rose (eds.), *Proceedings of DiSS 2017, The 8th Workshop on Disfluency in Spontaneous Speech*, 18–19 August, 2017, Stockholm, Sweden, 29–32.

Lee, T., Y. He, Y. Huang, S. Tseng, & R. Eklund. 2004. Prolongation in Spontaneous Mandarin. In: *Proceedings of Interspeech (ICSLP) 2004*, 4–8 October, 2004, Jeju Island, Korea, vol. III, 2181–2184.

Sadanobu, T. 2018. The "My Funny Talk" Corpus and speaking style variation in Japanese. In: D. G. Hebert (ed.), *International Perspectives on Translation, Education and Innovation in Japanese and Korean Societies*, Cham: Springer International Publishing, 133–147.

Sadanobu, T. 2021. Disfluency and taking the floor in Japanese-language conversation. Presentation at *The 17th International Pragmatics Conference*, 1 July, 2021, Zurich University of Applied Sciences (online).

Sadanobu, T., J. Somodi, J. Hidasi, V. Echbach-Szabo, A. N. Tekmen, D. Jayathilake, D. Dilshara-Jayasuriya, J. Arai, T. Shochi, M. Luo, A. Susairaj, K. Ryu & Y. Park. 2018. Gengo Ruikei kara mita Hiryūchousei: Kouchakugo to Enshingata Zokkouhoushiki no Tsukkae [Disfluency from a typological perspective: With special reference to "prolongation + continuation"]. *The Japanese Journal of Language in Society* 21(1), 113–128.

Why are some speech errors detected by self-monitoring "early" and others "late"?

Sieb Nooteboom and Hugo Quené
Utrecht University, Utrecht, The Netherlands

Abstract

In this paper we attempt to answer the question why in self-monitoring some segmental speech errors are detected in internal, some in external speech, and others not at all. This was done by re-analyzing data obtained in two earlier published SLIP experiments. It is hypothesized that detection of errors that are similar to the correct target takes longer than detection of errors that are dissimilar. It is also hypothesized that the time available for error detection in internal speech and for detection at all is limited. Results show that indeed a major factor is the strength of phonetic contrast between two competing response candidates.

Introduction

Errors of speech can be detected by self-monitoring internal (early) or external speech (late; cf. Levelt, Roelofs & Meyer, 1999; Hartsuiker, Kolk & Martensen, 2005). It has been demonstrated that this leads to a bimodal distribution of error-to-interruption times for repaired errors, the two peaks being separated by some 450 or 500 ms (Nooteboom & Quené, 2017), confirming that detection of segmental speech errors is a two-stage process. The bimodal distribution of log interruption times can be described as two overlapping gaussians, by applying an uninformed gaussian mixture model (Fraley & Raftery, 2002; Fraley et al., 2012). This allows us to fit a separation value for the two gaussians. All interruption times below this separation value are assigned to "early detections", all longer interruption times are assigned to "late detections". An example of an early detection is the repaired error K.. PAF KIEP, an example of a late detection is the repaired error KAF PIEP.. PAF KIEP. It is assumed that early detected repaired errors are detected in internal speech, i.e. before speech initiation, and that late detected errors are detected in external speech, i.e. after speech initiation (cf. Nooteboom & Quené, 2017). In this paper we attempt to find out why some errors are detected early (i.e. internally) and others late (i.e. externally). Speech errors may also remain unrepaired, assumedly because they were not detected, for example KAF PIEP instead of PAF KIEP.

We hypothesize (1) that detection of segmental speech errors depends on comparing the sound forms of competing simultaneously active response candidates from onset to offset (KAF as an error for PAF is detected by comparing the planned sound form KAF with the simultaneously active form PAF), (2) that detection of errors similar to the correct target takes more time than detection of errors that are more dissimilar, (3) that the time available for detection in internal speech is limited, (4) that, if this time is exceeded, the error will be passed on to self-monitoring overt speech, and (5) that if the speech error also exceeds the time available for detection in overt speech, it will remain undetected. We thus distinguish between early detected, late detected and undetected speech errors.

From this account of self-monitoring internal and overt speech for segmental speech errors we derive the following predictions:

1) There are relatively more dissimilar speech errors than similar speech errors detected internally.
2) There are relatively more similar speech errors than dissimilar speech errors detected externally.
3) There are relatively more similar than dissimilar errors that remain undetected and therefore unrepaired.

An interesting question is how similarity is to be assessed. In the literature on speech errors we find examples of assessing similarity by counting distinctive features (Nooteboom, 1967; Dell, 1986). However, Guenther (2016, chapter 1) proposed that during speech production, specification of speech sounds may be different at different levels of representation. He suggested that at least the following specifications are involved in speech planning: (1) abstract phonemes (2) targets in auditory perceptual space (involved in early planning of articulation) and (3) speech motor commands (involved in specifying articulatory gestures).

So now we are confronted with two questions to be answered in this paper: (1) Is it correct that the strength of the contrast between error and correct response candidates determines whether an error is detected early or late (or not all)? (2) If so, is the contrast more phonological, to be assessed by counting distinctive features, or more phonetic, to be assessed by determining the relative strength of the contrast?

The first question will be answered by comparing frequencies and repair frequencies of errors

https://doi.org/10.18463/diss-2021-003-noteboom-quene

involving a single distinctive feature (similar), viz. place or mode of articulation, with those of errors involving at least two distinctive features (dissimilar), viz. place plus mode of articulation. This will be done in Experiment 1. The second question will be answered by comparing error frequencies and repair frequencies of segmental speech errors involving (a) voicing errors in word initial stop consonants, (b) similar errors as defined for Experiment 1, (c) dissimilar errors as defined for Experiment 1, and (d) vowel errors. Voicing of consonants in word initial position is a relatively weak contrast in Dutch (Van Alphen & McQueen, 2006). Voiced and corresponding unvoiced initial stop consonant are distinguished in Dutch by the length of prevoicing. Unvoiced consonants are not aspirated. Phonetically voicing contrast is strengthened by an additional contrast in force of articulation, traditionally captured by fortis or tense for voiceless consonants and lenis or lax for voiced consonants. Vowel oppositions are supposed to provide a relatively strong contrast in Dutch.

Experiment 1

Experiment 1 has been reported as Experiment 1 in Nooteboom and Quené (2017). It was originally set up to investigate temporal aspects of detecting and repairing segmental speech errors in a SLIP (Spoonerisms of Laboratory Induced Predisposition, cf. Baars, Motley & MacKay, 1975) experiment. Here we limit description of Experiment 1 to those aspects that are relevant to the current task.

Method of Experiment 1

There were 106 participants. Interactive segmental speech errors were elicited by having CVC CVC Dutch word pairs (stimulus items), each preceded by 5 CVC CVC word pairs, the last three of which triggered a reversal of the two word initial consonants, as in BOUW JOOL, LIJF DEED, KEN PIT, KOET POP, KAS PIET, preceding the stimulus word pair PAF KIEP. There were two stimulus lists. In each list there were 32 stimuli, 16 with the two initial consonants differing in a single distinctive feature (place or mode of articulation; similar), and 16 with the two initial consonants differing in two distinctive features (place plus mode of articulation; dissimilar).

There were 23 filler stimuli, preceded by 0, 1, 2 3 or 4 CVC CVC word pairs not triggering a segmental reversal. After each test stimulus and each filler stimulus a sequence of "?????" was presented, as a cue to speak aloud the last word pair seen. After the "?????" there followed a presentation of the Dutch word for "repair?", to elicit sufficient repairs.

Each speaker was tested individually in a sound-treated booth. Presentation of precursors, stimuli, and "repair" cues always lasted 900 ms followed by a blank interval of 100 ms. Responses were categorized using Praat (Boersma & Weenink, 2016) for the current purpose as (0) fluent and correct, (1) hesitations and omissions, (2) completed and interrupted single elicited segmental errors, (3) completed and interrupted single other errors, (4) completed and interrupted multiple other errors. The current analysis focuses mainly on category 2.

Results of Experiment 1

A first breakdown of the observed speech errors is given in Table 1.

Table 1. Numbers of responses, broken down by response category and repair status.

response category	repair status		total
subjects	not repaired	repaired	
fluent & correct	5821	0	5821
hesitations & omissions	64	40	104
single elicited errors	298	115	413
single other errors	187	31	218
multiple errors	192	36	228
total	6562	222	6784

In our further analyses we will mainly focus on the 413 single elicited, errors, i.e. those errors the SLIP technique was meant to elicit. Of these 298 were not repaired, presumably because they were not detected by self-monitoring either in internal speech or after speech initiation, and 115 were repaired before or after speech initiation.

Speech errors were either "detected early", or "detected late", or else remained unrepaired, supposedly "undetected". The categorization as to whether a specific repaired error was detected "early" or "late" was made on the basis of an uninformed gaussian mixture model applied to "error-to-interruption times" (cf. Fraley & Raftery, 2002; Fraley et al., 2012). The resulting bimodal distribution is shown in Figure 1.

A relevant independent variable is "similar" vs "dissimilar" interacting consonants in the error; a relevant dependent variable is whether the error was detected "early" or "late" or "not detected" at all. This breakdown is given in Table 2.

Here we see that "similar" errors are far more frequent than "dissimilar" errors. This difference was found to be significant in a Bayesian binomial logistic mixed-effects regression (GLMM; Bürkner, 2017, 2018), with errors as hits and participants as random intercepts, and similarity as fixed effect; the response category "not detected" was used as

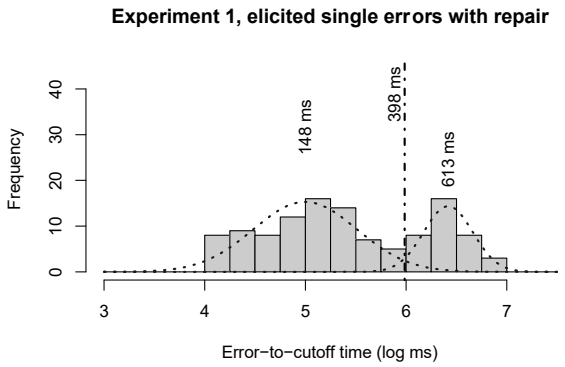

Figure 1. Histogram of log-transformed durations of error-to-interruption intervals, for N = *114 repaired errors (the error-to-interruption interval of 1 repaired error was missing). Dotted lines indicate the estimated distributions from an uninformed gaussian mixture model (see text). The vertical dashed line indicated the interpolated boundary value.*

Table 2. Numbers of single elicited segmental errors, broken down by "similar" vs "dissimilar" and by "early detected" vs "late detected" vs "not detected".

	Detection			
error category	early	late	not	Total
similar	26	18	192	236
dissimilar	54	17	166	177
total	80	35	298	413

baseline. The data reported in this paper, and full details of all statistical analyses, are available at https://osf.io/gxjnm/. Log odds are on average −2.79 for items eliciting interaction among "similar" consonants (baseline) and they are lower by −0.32 for items involving "dissimilar" consonants (with 95% highest posterior density interval [−0.52, −0.12]), suggesting a significant difference between similar and dissimilar.

The three-way classification of errors as "early detected", "late detected" and "not detected" (Table 2) was further analyzed with another Bayesian multinomial mixed effects regression model, with participants as random effect, contrast as fixed predictor (with similar interacting consonants as baseline). The odds of an error being detected early are indeed far lower for similar items (posterior mean −2.06) than for dissimilar items (mean −0.71, with non-overlapping posterior density intervals. The difference in (very low) odds of late detections between "similar" and "dissimilar" was not found to be significant.

Discussion of Experiment 1

The results of Experiment 1 demonstrate that strength of contrast determines whether segmental speech errors are detected in internal speech or not. Of course, in the current analysis, strength of contrast was expressed by counting distinctive features. In Experiment 2 we attempt to find out whether strength of contrast should rather be expressed phonetically.

Experiment 2

Experiment 2 has been reported as Experiment 2 in Nooteboom and Quené (2017). Experiment 2 is largely identical to Experiment 1, but in addition to stimuli eliciting interactions between "similar" and "dissimilar" consonants, we also added a category of stimuli eliciting interactions between "voiced" and "unvoiced" consonants and a category of stimuli eliciting interactions between the vowels of the two CVC words. Here we limit description of Experiment 2 to aspects that are relevant to the current task.

Method of Experiment 2

There were 124 participants. There were two stimulus lists. In each list there were 32 stimuli eliciting interactions between word initial consonants differing in place and/or mode of articulation, of which 16 differing in a single distinctive feature (place or mode of articulation; similar), and 16 with the two initial consonants differing in two distinctive features (place plus mode of articulation; dissimilar). There were also 16 stimuli eliciting interactions between voiced and unvoiced word initial consonants and 16 stimuli eliciting interactions between the vowels of the two CVC words. There were also 46 filler stimuli, with a number of precursors varying between 0 and 4. The precursors of the fillers did not prime interactions. Further details of the materials, the procedure and the scoring were the same as in Experiment 1. The current analysis focuses mainly on category 2.

Results of Experiment 2

A first breakdown of the numbers of single elicited errors is given in Table 3.

Speech errors were either "detected early", or "detected late", or else remained unrepaired, supposedly "undetected". The categorization as to whether a specific repaired error was detected "early" or "late" was made on the basis of an uninformed gaussian mixture model applied to "error-to-interruption times" (cf. Fraley & Raftery, 2002; Fraley et al., 2012). The resulting bimodal distribution is shown in Figure 2.

As is clear from Table 4, the numbers of total errors are conspicuously different for the four classes of stimuli. By far the most errors are made against "voicing", which confirms that the voicing contrast is relatively weak in Dutch. By far the fewest errors

are made against "vowels", which confirms that the contrast between vowels is relatively strong in Dutch. These differences were again analyzed in a Bayesian GLMM (Bürkner, 2017, 2018), with errors as hits and participants as random intercepts, and similarity as fixed effect (see https://osf.io/gxjnm/ for details). Log odds are on average −2.72 for items eliciting interaction among "similar" consonants (baseline) and they are lower by −0.71 logits (95% HPDI [−1.02, −0.43]) for items involving "dissimilar" consonants, again suggesting a significant difference between similar and dissimilar consonants. Moreover, the log odds of errors involving consonantal voicing contrast are higher than the baseline by +0.91 logits (95% HPDI [0.72, 1.10]), and the log odds of errors involving vowels are lower by −0.96 logits (95% HPDI [−1.28, −0.66]).

The three-way classification of errors as "early detected", "late detected" and "not detected" (Table 4) was further analyzed with another Bayesian multinomial mixed effects regression model, with participants as random effect, contrast as fixed predictor (with similar interacting consonants as baseline). The only significant effect was that the log odds for early detection were significantly lower for voicing errors (30:532) than for similar errors (45:214; the difference being −1.36 logits, 95% HPDI [−1.95, −0.78]).

Discussion of Experiment 2

Experiment 2 did not replicate the significant difference in error detection between errors involving similar and dissimilar consonants. Apparently, results were somewhat noisier than in Experiment 1. However, the significant difference between voicing errors and similar errors, both error categories involving a contrast of a single distinctive feature, suggests that speech errors are detected on the basis of more phonetic than phonological contrast (see general discussion below). We had also predicted a significant difference in frequency of early detection between dissimilar errors and vowel errors. That this effect did not show up possibly is due to the circumstance that detection of vowel errors in Dutch takes considerably more time than detection of consonant errors. This is so because the first part of Dutch long vowels and Dutch diphthongs sound as a corresponding Dutch short vowel.

General discussion

The main question we have attempted to answer in the present investigation is: "Why are some segmental speech errors detected by self-monitoring in internal speech, others in external speech, and others not at all?" The results of Experiment 1 demonstrate that a major factor is the strength of contrast between two competing response candidates, as assessed by the relative frequency of error commitment: Detection of segmental speech errors involving a weak contrast takes more time than detection of segmental speech errors involving a stronger contrast. The time available for detection in internal speech, before speech initiation, is limited. If this time is exceeded for a particular speech error, detection is likely to be postponed to a later stage of speech preparation, where articulation is initiated. In case also the time needed for detection of an error at this later stage of the speaking process is exceeded, the error remains undetected and unrepaired.

Table 3. Numbers of responses, broken down by response category and repair status.

response category	repair status		total
	not repaired	repaired	
fluent & correct	13069	0	13069
hesitations & omissions	228	67	295
single elicited errors	956	184	1140
single other errors	570	35	605
multiple errors	473	34	507
total	15296	320	15616

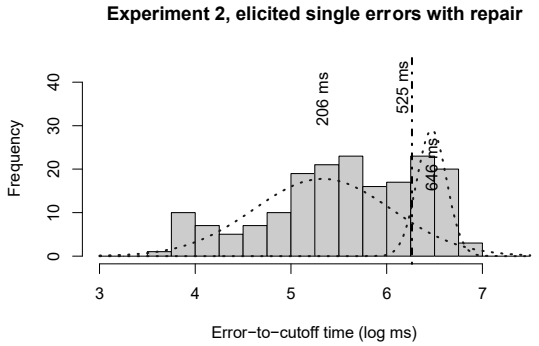

Figure 2. Histogram of log-transformed durations of error-to-interruption intervals, for N = 182 repaired errors (error-to-interruption intervals of 2 repaired errors were missing). Dotted lined indicate the estimated distributions from an uninformed gaussian mixture model (see text). The vertical dashed line indicated the interpolated boundary value.

Table 4. Numbers of single elicited segmental errors, broken down by error category (see text) and by detection status: detected "early" or "late" or "not detected".

	detection			
error category	early	late	not	total
voicing	30	16	532	579
similar	45	16	214	275
dissimilar	38	7	125	171
vowels	23	7	85	115

We have also attempted to find out whether contrast between competing segments involved in error detection by self-monitoring is phonological, i.e. in terms of number of distinctive features, or rather phonetic. Results of Experiment 2 suggest that contrast on the levels of representation where segmental errors are detected by self-monitoring is phonetic, to be specified in terms of more gradient segmental properties such as auditory perceptual contrast or articulatory contrast. The evidence for this conclusion in Experiment 2 stems from the comparison in frequency of "early" detection between errors involving the weak contrast in voicing and errors involving the stronger contrast in place or mode of articulation.

We wish to point out that if we apply the feature system proposed by Chomsky and Halle (1968), there sometimes is a major difference in terms of distinctive features between our errors against similar consonants in Experiment 1 and voicing errors in Experiment 2. For example, place of articulation of /t/ is specified by two distinctive features, viz. +coronal and +anterior (to distinguish labiodental /t/ from palatal /c/ that is +coronal and −anterior), whereas voicing is always specified by a single distinctive feature. This feature system brings phonology somewhat closer to phonetics. However, traditionally voiceless and voiced consonants were also assigned the features fortis and lenis or tense and lax, over and above the presence or absence of voicing. The strength of contrast between similar consonants differing in place or mode of articulation on the one hand and voicing on the other cannot be captured by counting features in some reified abstract phonological feature system, as done for example in Ulicheva et al. (2021).

The current results fit into an account of speech planning and self-monitoring for speech errors with different stages of planning. We follow Guenther (2016) in assuming that articulatory movements are planned internally in terms of sequences of targets in auditory perceptual space. During this stage segmental errors are detected "early". About 450 or 500 ms later, these segments are transformed into articulatory gestures, specified in terms of temporally coordinated motor commands. During this stage "late" error detection occurs.

References

Baars, B. J., M. T. Motley & D. G, MacKay. 1975. Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* 14(4), 382–391. https://doi.org/10.1016/S0022-5371(75)80017-X

Boersma, P. & D. Weenink. 2016. Praat: Doing phonetics by computer (version 6.0.19). https://www.praat.org/ (accessed 24 January 2021).

Bürkner, P.C. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.C. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395–411.

Chomsky, N. & M. Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Dell, G. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Fraley, C. & A. E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631. https://doi.org/10.1198/016214502760047131

Fraley, C., A. E. Raftery, T. B. Murphy & L. Scrucca. 2012. mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Report No. 597, Department of Statistics, University of Washington.

Guenther, F. H. 2016. *Neural Control of Speech*. Cambridge, MA: The MIT Press.

Hartsuiker, R. J., H. H. J. Kolk & H. Martensen. 2005. Division of labor between internal and external speech monitoring. In: R. Hartsuiker, Y. Bastiaanse, A. Postma & F. Wijnen (eds.), *Phonological Encoding and Minitoring in Normal and Pathological Speech*, Hove: Psychology Press, 187–205.

Levelt, W. J. M., A. Roelofs & A. S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Nooteboom, S. G. 1967. The tongue slips into patterns. In: A. Sciarone, A. J. van Essen, A. A. van Raad (eds.), *Nomen, Leyden Studies in Linguistics and Phonetics*, The Hague: Mouton, 114–132.

Nooteboom, S. G. & H. Quené. 2017. Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *Journal of Memory and Language* 95, 19–35. https://doi.org/10.1016/j.jml.2017.01.007

Ulicheva, A., K. D. Roon, Z. Cherkasova & P. Mousikou. 2021. Effects of phonological features on reading-aloud latencies: A cross-linguistic comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. http://dx.doi.org/10.1037/xlm0000893.

Van Alphen, P. M. & J. M. McQueen. 2006. The effect of voice onset time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance* 32(1), 178–196. https://psycnet.apa.org/doi/10.1037/0096-1523.32.1.178

Speech disfluencies as actual and believed cues to deception: Individuality of liars and the collective of listeners

Nette Vandenhouwe and Robert J. Hartsuiker
Ghent University, Ghent, Belgium

Abstract

There is no consensus about the relationship between disfluencies and deception in speech production. However, it is well established that listeners believe deceptive speech to contain more disfluencies than truthful speech. Here, we used an interactive game to collect the speech of liars and the veracity decisions of listeners. Using Multivariate Pattern Analysis (MVPA), we determined the predictive value of speech disfluencies as both actual and believed cues to deception. We found that patterns of disfluencies can indeed be used to predict both an utterance's veracity and a listener's decision about that veracity better than chance. However, there was much individual variation in how lies altered speech, whereas listeners were consistent in how they thought the speech of others indicates lying.

Introduction

It has been estimated that people lie once or twice a day on average (DePaulo et al., 1996). Because it can be very important to tell whether somebody is telling the truth or not, researchers have looked for ways of sorting the liars from the truthtellers. Studies on deceit have often asked whether there are discernible differences in a person's behavior when telling the truth vs. lies, and whether listeners can exploit those differences to evaluate the truthfulness of speakers (DePaulo et al., 2003; Loy, Rohde, & Corley, 2018).

Such studies usually assume that while a person is lying their behavior displays signals that can warn others that they are being deceived. These particular signals are called *actual cues to deception*. On the other hand, for the listener we identify *believed cues to deception*: the changes in behavior that listeners believe to be associated with lying (Levine, 2018; Vrij & Semin, 1996). Here, we ask whether the disfluencies in the speech of potential liars are such actual and believed cues.

Deception and disfluencies in speech

Research on actual cues to deception has not reached consensus about the exact relationship between disfluencies and deception. Some studies found that people utter *more* speech disfluencies while lying, whereas others observed *fewer* disfluencies during deception. As seen below, explanations of these opposite findings differ in their assumptions about the causes and functions of disfluencies in speech.

Many studies reported evidence for an increased prevalence of disfluencies during lying: their positive relationship with deception was found for different types of disfluencies, in different types of lies, and in different situations (DePaulo et al., 1982; Vrij, Edward, & Bull, 2001; Vrij & Winkel, 1991; Whelan, Wagstaff, & Wheatcroft, 2014). This positive relationship can be understood in terms of the *Cognitive Hypothesis*, which states that lying is cognitively more demanding than telling the truth. Lying is a complex process that involves additional tasks compared to truth-telling. Because liars are already spending so much of their cognitive resources on the deception, speech disfluencies would occur more frequently (Loy et al., 2018). This explanation of the positive relationship assumes that disfluencies occur more often in situations of high cognitive load, which is compatible with accounts that view disfluencies as diagnostic of difficulties in conceptualizing, planning, or executing speech (Fox Tree, 2001; Levelt, 1989).

However, other studies reported the opposite pattern: humans produce fewer disfluencies during lying. These studies typically focus on pauses and differentiate between silent pauses, ums, and uhs. In addition to disfluency proportion, measures like duration, pitch, and intensity are considered. Specifically, during lying participants produce fewer silent pauses, ums, and uhs (Arciuli, Mallard, & Villar, 2010; Benus et al., 2006; Villar, Arciuli, & Mallard, 2012; Villar & Castillo, 2017). Further, uhs were longer but ums were shorter and louder when uttered in deceptive speech (Arciuli et al., 2010; Benus et al., 2006). Additionally, deception seems more strongly related with the use of um than uh (Benus et al., 2006). These studies illustrate the importance of treating um as a standalone variable, separate from uh, speech errors, and hesitations (Arciuli et al., 2010). An account of the negative relationship between deceit and disfluency is the *Attempted Control Hypothesis*. It states that liars attempt to sound believable by controlling behaviors that may signal their deception to others (Loy et al., 2018). Liars would plan their speech in an attempt to

https://doi.org/10.18463/diss-2021-004-vandenhouwe-hartsuiker

sound more fluent, as they believe that fluent speech signals truthfulness to listeners (e.g. King, Loy, & Corley, 2017). This account assumes an alternative role of speech disfluencies in language. Following Clark and Fox Tree (2002), it assumes that um and uh, rather than being mere symptoms of speech difficulties, are conventional words that serve specific purposes. One such purpose would be to announce an upcoming delay in speech, with um and uh announcing a major and minor delay respectively. Um and uh are thus fundamentally different from each other and from other signals of delay. It is feasible that liars would suppress uttering um and uh as they know that these pauses signal speech production difficulty (delays), and in turn also deception, to listeners (Arciuli & Villar, 2009; Clark & Fox Tree, 2002).

Findings on what listeners believe to be signs of deception are clear-cut: listeners consistently believe that deceptive speech contains more errors, pauses, and hesitations (Global Deception Reasearch Team, 2006; King et al., 2017; Loy et al., 2018).

The current study

Our aim was to investigate further to what extent patterns of speech disfluencies are consistent and reliable actual and believed cues to deception. To do so, we adopted the game paradigm of Loy et al. (2018): in each experimental session, a speaker lied or told the truth about the location of a treasure to a listener who was trying to assess the veracity of this speaker's messages. In that way, we could examine the speech of liars and veracity decisions of listeners while they were engaged in a meaningful social interaction.

Rather than using typical analysis approaches which treat disfluency measures as dependent variables by testing them individually to determine whether they vary between conditions, we used Multivariate Pattern Analysis (MVPA). The MVPA classifier finds patterns in all measures available at once and for each participant individually. Subsequently, it attempts to use these personalized patterns to classify data into conditions, here true vs. false (Pistono & Hartsuiker, 2021).

More specifically, using MVPA, we examined whether the veracity of speakers' utterances and the decisions of listeners could be predicted based on patterns contained in several speech disfluency measures. We considered the proportions in every disfluency category but also different duration and intensity measures. The Cognitive Hypothesis predicts a positive relationship between deceit and disfluencies, whereas the Attempted Control Hypothesis predicts a negative relationship. We further expected listeners to associate disfluency with deceit.

Method

Participants

We tested 48 native Dutch speaking bachelor students (24 dyads) from the Faculty of Psychology and Educational sciences at Ghent University (46 female, M_{age} = 18.67, SD_{age} = 2.47). Dyad members always had the same sex and did not know each other before the experiment.

Material and design

The stimuli were identical to those of Loy et al. (2018) and consisted of 96 black-and-white line drawings of objects. Specifically, 48 images were from the dataset of Snodgrass and Vanderwart (1980) and 48 were manipulated versions of the original images. This led to 48 visually similar image pairs (Figure 1). This particular way of pairing images was done with the aim of eliciting relatively long and complex utterances from the speakers. The images on the speaker's screen were accompanied by images of a pile of dirt and a pile of coins (Figure 1).

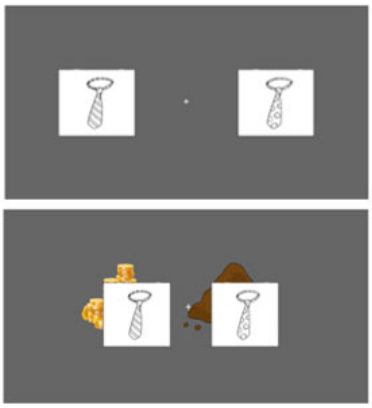

Figure 1. Example of a trial. Screen of the listener at the top and screen of the speaker at the bottom.

Each image pair occurred in four unique trials in which the original images were counterbalanced according to location (left or right) and type of pile associated with the image (treasure or dirt). The manipulated pair mate always appeared in the opposite position with the opposite type of pile behind it. We created four lists. Across the lists, each of the 48 original images appeared once in every unique trial and each version of a trial was included equally often. Each dyad saw two lists in two blocks of one list (48 trials) each. Across dyads, the four lists and each of their six possible combinations were presented equally often.

Procedure

In the experiment, participants sat opposite each other, enabling them to see each other's face but not each other's stimuli. In each dyad, one participant received the role of speaker and the other the role of listener, which they maintained throughout the whole experiment. Speakers were instructed to explain to listeners on each trial behind which image the treasure was hidden. The image pairs were visible to both the speaker and listener. Treasures and dirt were only visible to the speaker. Speakers were free to try to mislead listeners into looking for the treasure in the wrong location. They could do this by lying about the location of the treasure or by telling the truth, hoping that the listener would not trust what they are saying. For both participants, the goal was to gain as many coins as possible. The listener gained coins by correctly guessing where the treasure was hidden. The speaker gained coins when the listener indicated the wrong location. Listeners could make their choice by pressing the F- or the J-key when choosing the image on the left or right respectively. After this, a feedback message appeared that informed them about the winner of the trial and the current number of coins of each player, which was cumulative over the course of the experiment. The winner of the game received one euro as a reward.

Data processing and analysis

The speech of the 24 speakers was transcribed and annotated. Further, speakers' descriptions on each trial were coded as false or true, with false trials defined as trials on which the speaker uttered a statement about the location of the treasure that was factually incorrect and with true trials as trials with a statement that was factually correct. Second, the decisions of the 24 listeners were coded as false or true, with false trials defined as trials on which the listener chose the image not described by the speaker as concealing the treasure and with true trials as trials with the listener choosing the image described as concealing the treasure. For the data of the speakers, the mean percentage of true trials was 54.17% (SE = 1.80, min = 31.25%, max = 75.53%), and for the data of the listeners it was 54.98% (SE = 1.64, min = 32.29%, max = 64.84%). These match with a general bias towards both telling and expecting the truth (Loy et al., 2018).

Our coding system consisted of seven disfluency categories: filled pauses (ums, uhs, and mms), ums, uhs, silent pauses, repetitions, repairs (restarts, substitutions, and additions), and prolongations (Hartsuiker & Notebaert, 2010; Shriberg, 1996). Both the category filled pauses and ums and uhs were included to examine whether treating ums and uhs as separate categories influences results. The first and second coder of this study both coded 21% of the speech data and agreed on 72.98% of all labeled disfluencies. Table 1 presents summary statistics of all disfluency measures that were collected from the coded speech data: the raw counts of each disfluency category; the durations of the whole trial utterance, the onset of the utterance, the fluent and disfluent part of the utterance, and each and every coded disfluency in the speech data; and finally, the peak and standard deviation of the intensity of each coded disfluency.

Table 1. Descriptive statistics of all speech disfluency measures collected from speech data.

Measure		Raw count	Mean (SD)
Filled pause		1223	-
	Um	530	-
	Uh	662	-
Silent pause		1310	-
Repetition		280	-
Repair		304	-
Prolongation		1273	-
Utterance Dur (s)		-	4.01(2.12)
Onset Dur (s)		-	2.20(0.68)
Fluent Dur (s)		-	2.95(1.14)
Disfluent Dur (s)		-	1.07(1.39)
Disfluency Dur (s)		-	0.55(0.33)
Intensity Peak (dB)		-	52.87(12.83)
Intensity SD (dB)		-	6.17(3.14)

We used MVPA to investigate whether speech veracity and veracity decisions could be classified based on information contained in speech disfluency patterns. Linear discriminant analysis classifiers were trained for each participant individually. The classifications were performed in a leave-one-out cross-validation approach (15 folds). The accuracy measure was the proportion of correctly classified trials or disfluencies. Accuracies were compared to chance level, which is 50% for a two-class problem. Further, we determined which disfluency features played a significant role at the group level by testing whether their mean weight (i.e. their contribution in the classification) was significantly different from zero (Pistono & Hartsuiker, 2021).

Results

Speaker data—trial level

The first MVPA classified the speaker data on the trial level. This meant that we tried to classify the veracity of the utterances (i.e. trials) of speakers and that we based the classification on the disfluency features that were collected for each trial

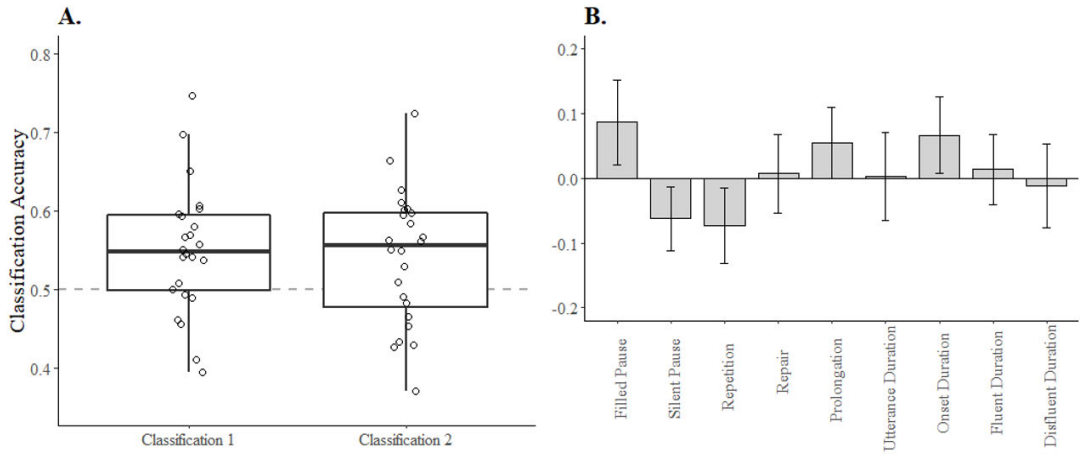

Figure 2. A. Participants' trial classification accuracies. The dashed line indicates chance level. Each dot represents accuracy for a single participant. Classification 1 is based on features including the filled pause category and classification 2 is based on features including the um and uh categories. B. Mean weights of features in classification 1. Error bars represent the SE of the mean.

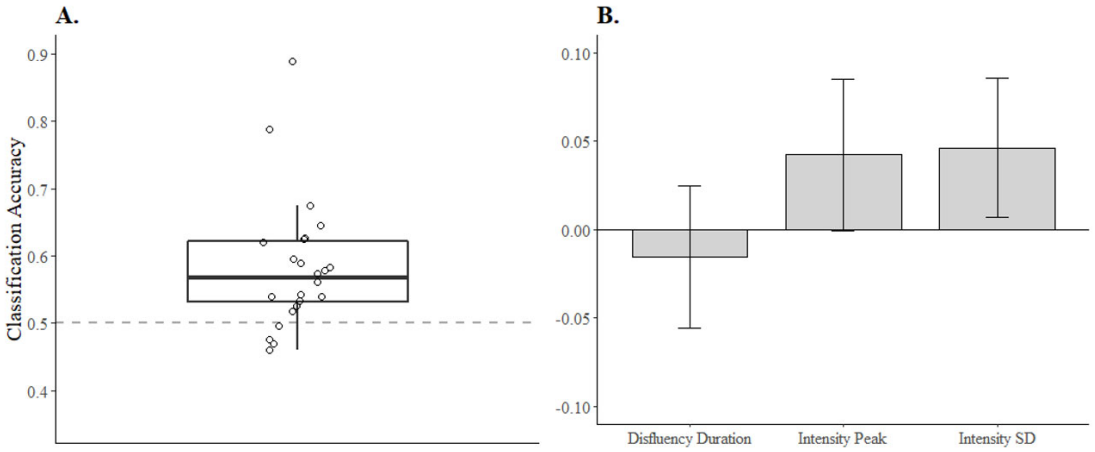

Figure 3. A. Participants' disfluency classification accuracy. The dashed line indicates chance level. Each dot represents accuracy for a single participant. B. Weights of features. Error bars represent the SE of the mean.

(Figure 2 A, classification 1; Figure 2 B). Mean classification accuracy was 54.99%, which was significantly above chance ($t(23) = 3.00$, $p = 0.003$). Tests on the contribution of features found that none of them was significant at the group level. Thus, although these disfluency patterns could predict the veracities of utterances better than chance, these patterns were not consistent from one participant to another. An analysis that treated um and uh as separate categories showed similar results: the classifier could predict veracity above chance, but the patterns were not consistent across participants (Figure 2 A, classification 2).

Speaker data—disfluency level

The second MVPA classified the speaker data on disfluency level. This meant that we tried to classify the veracity of the sentences in which certain disfluencies were uttered and this using disfluency features that were collected for each and every disfluency that was extracted from the speech data (Figure 3 A and 3 B). Mean classification accuracy was 58.24%, which was significantly above chance ($t(23) = 4.15$, $p < 0.001$). None of the features was significant at the group level. We conclude that disfluency patterns can predict the veracities of the utterances from which the disfluencies were extracted but that these patterns are not consistent from one participant to another.

Listener data

A final MVPA classified listener data (trial level). This meant that we tried to classify the decisions of listeners about speech veracity based on disfluency features that potentially influenced these listener decisions (Figure 4 A, classification 1; Figure 4 B). Mean classification accuracy was 56.67%, which was significantly above chance ($t(23) = 4.42$, $p < 0.001$). Three features were significant at the group level. We conclude that disfluency patterns

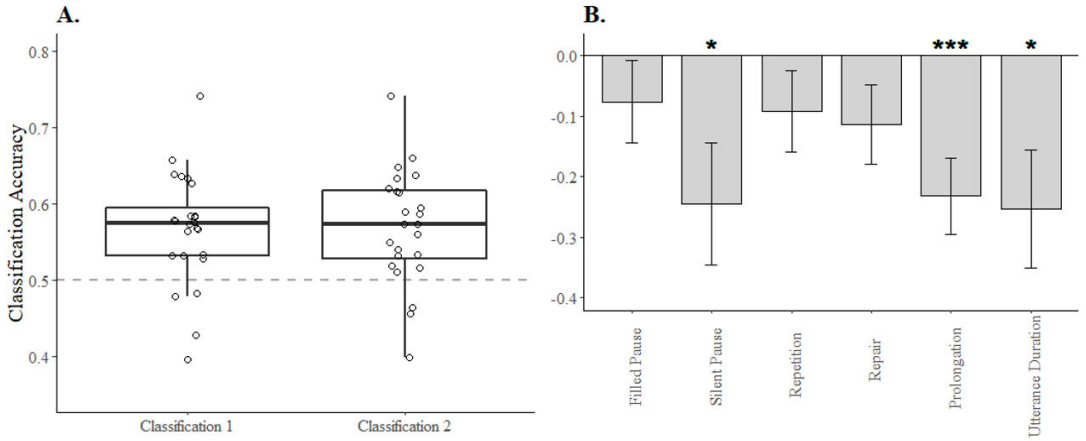

Figure 4. A. Participants' trial classification accuracy. The dashed line indicates chance level. Each dot represents accuracy for a single participant. Classification 1 is based on features including the filled pause category and classification 2 is based on features including the um and uh categories. B. Weights of features in classification 1. Error bars represent the SE of the mean. Asterisks indicate significance at the group level.

predict veracities of listeners' decisions and are consistent across participants. Specifically, listeners believe that many silent pauses and prolongations and utterances with longer durations are indicative of deceit. Conclusions are the same for the classification with um and uh as separate features (Figure 4 A, classification 2).

Discussion

The purpose of this study was to investigate the validity of several disfluency measures as cues to deception. We used a social game paradigm to collect speech of liars and decisions of listeners. We asked whether the veracity of speakers' utterances and listeners' decisions could be predicted based on patterns of disfluency measures.

With MVPA, we used all disfluency measures at once to classify veracity data for each participant individually. Our analyses demonstrated that patterns of disfluency measures contain information that allow for better-than-chance classifications of the veracity of utterances. However, none of the disfluency features reached significance at the group level. Thus, although disfluency patterns did contain valuable information for each participant individually, they were not consistent from one participant to another. Note that both the Cognitive and Attempted Control hypothesis predict that disfluencies consistently increase or decrease during lying. However, our results provide support for neither. These MVPAs illustrate the importance of taking individual differences into account when investigating cues to deception. A major conclusion to be drawn from the literature is that it is impossible to identify reliable cues to deception: no cues have been found that can in every situation and for every speaker correctly discriminate between truth and lie (e.g. Levine, 2018). This conclusion is not surprising when taking our results into account: humans differ in how lies alter their behavior. Therefore, which exact cues are valid varies across individuals.

Second, the information in patterns of disfluency measures allowed us to classify the veracity decisions of listeners better than chance. Importantly, three disfluency features reached significance at the group level: listeners consistently believe that utterances with more silent pauses, more prolongations, and longer durations are indicative of deceit. This is in line with previous findings (e.g. Global Deception Reasearch Team, 2006). Comparing the MVPAs on speaker vs. listener data suggests that beliefs about valid lie detection strategies are better generalizable across participants than the real changes in speaker behavior between lying and truth-telling.

Finally, we note a possibility that there were individual differences in how much speakers and listeners learned from the feedback after each trial and the way in which they adapted their deception or deception detection strategies as a result.

In conclusion, MVPA, which analyzes data at the individual level, showed that disfluency patterns can be used to predict the veracities of utterances and listeners' decisions. However, a discrepancy is observed in the generalizability of these patterns: participants differ in how lies alter their speech but agree on how the speech of others exposes lies. These findings have implications for how we search for valid cues to deception and develop lie detectors that, one day, may be used to catch real-life liars.

References

Arciuli, J., D. Mallard & G. Villar. 2010. "Um, i can tell you're lying": Linguistic markers of deception versus truth-telling in speech. *Applied Psycholinguistics 31*(3), 397–411. https://doi.org/10.1017/S0142716410000044

Arciuli, J. & G. Villar. 2009. Lies, Lies and More Lies. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*, July 29–Aug 1, 2009, Amsterdam, Netherlands, 2329–2334.

Benus, S., F. Enos, J. Hirschberg & E. Shriberg. 2006. Pauses in deceptive speech. In: R. Hoffmann and H. Mixdorff (eds.), *Proceedings of the International Conference on Speech Prosody*, May 2–5, 2006, Dresden, Germany, paper 212.

Clark, H. H., & J. E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

DePaulo, B. M., S. E. Kirkendol, D. A. Kashy, M. M. Wyer & J. A. Epstein. 1996. Lying in Everyday Life. *Journal of Personality and Social Psychology* 70(5), 979–995. https://doi.org/10.1037/0022-3514.70.5.979

DePaulo, B. M., B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton & H. Cooper. 2003. Cues to deception. *Psychological Bulletin* 129(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

DePaulo, B. M., R. Rosenthal, J. Rosenkrantz & C. Rieder Green. 1982. Actual and Perceived Cues to Deception: A Closer Look at Speech. *Basic and Applied Social Psychology*, 3(4), 291–312. https://doi.org/10.1207/s15324834basp0304_6

Fox Tree, J. E. 2001. Listeners' uses of um and uh in speech comprehension. *Memory & Cognition* 29(2), 320–326. https://doi.org/10.3758/BF03194926

Global Deception Reasearch Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology* 37(1), 60–74. https://doi.org/10.1177/0022022105282295

Hartsuiker, R. J. & L. Notebaert. 2010. Lexical access problems lead to disfluencies in speech. *Experimental Psychology* 57(3), 169–177. https://doi.org/https://doi.org/10.1027/1618-3169/a000021

King, J. P. J., J. E. Loy & M. Corley. 2017. Contextual Effects on Online Pragmatic Inferences of Deception. *Discourse Processes* 55(2), 123–135. https://doi.org/10.1080/0163853X.2017.1330041

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

Levine, T. R. 2018. Scientific Evidence and Cue Theories in Deception Research: Reconciling Findings From Meta-Analyses and Primary Experiments. *International Journal of Communication* 12, 2461–2479.

Loy, J. E., H. Rohde & M. Corley. 2018. Cues to Lying May be Deceptive: Speaker and Listener Behaviour in an Interactive Game of Deception. *Journal of Cognition* 1(1), 1–21. https://doi.org/10.5334/joc.46

Pistono, A. & R. Hartsuiker. 2021. Eye-movements can help disentangle mechanisms underlying disfluency. *Language, Cognition and Neuroscience*, 1–18. https://doi.org/10.1080/23273798.2021.1905166

Shriberg, E. 1996. Disfluencies in Switchboard. In: *The 4th International Conference on Spoken Language Processing (Addendum)*, October 3–6, 1996, Philadelphia, PA, USA, 11–14.

Snodgrass, J. G. & M. Vanderwart. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* 6(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Villar, G., J. Arciuli & D. Mallard. 2012. Use of "um" in the deceptive speech of a convicted murderer. *Applied Psycholinguistics* 33(1), 83–95. https://doi.org/10.1017/S0142716411000117

Villar, G. & P. Castillo. 2017. The Presence of 'Um' as a Marker of Truthfulness in the Speech of TV Personalities. *Psychiatry, Psychology and Law* 24(4), 549–560. https://doi.org/10.1080/13218719.2016.1256018

Vrij, A., K. Edward & R. Bull. 2001. Stereotypical Verbal and Nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin* 27(7), 899–909. https://psycnet.apa.org/doi/10.1177/0146167201277012

Vrij, A. & G. R. Semin. 1996. Lie experts' beliefs about nonverbal indicators of deception. *Journal of Nonverbal Behavior* 20(1), 65–80. https://doi.org/10.1007/BF02248715

Vrij, A. & F. W. Winkel. 1991. Cultural patterns in Dutch and Surinam nonverbal behavior: An analysis of simulated police/citizen encounters. *Journal of Nonverbal Behavior* 15(3), 169–184. https://doi.org/10.1007/BF01672219

Whelan, C. W., G. F. Wagstaff & J. M. Wheatcroft. 2014. High-Stakes Lies: Verbal and Nonverbal Cues to Deception in Public Appeals for Help with Missing or Murdered Relatives. *Psychiatry, Psychology and Law* 21(4), 523–537. https://doi.org/10.1080/13218719.2013.839931

Fine phonetic details for DM disambiguation in French: A corpus-based investigation

Yaru Wu[1,2], Mathilde Hutin[1], Ioana Vasilescu[1], Lori Lamel[1], Martine Adda-Decker[1,2] and Liesbeth Degand[3]
[1]Université Paris-Saclay (CNRS, LISN), Orsay, France
[2]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), Paris, France
[3]Université catholique de Louvain, Louvain-la-Neuve, Belgium

Abstract

In this study we examine phonetic variation of discourse markers in French, using for this purpose the 4-hour richly annotated LOCAS-F corpus. Both linguistic factors and stylistic variables are considered: speech style, part-of-speech category, mean phone duration and vowel formant distributions with respect to the word status. The results show that the use of discourse markers increases with the degree of spontaneity of the speech. Coordinating conjunctions are the part-of-speech which is most frequently used as discourse markers. Moreover, the mean phone duration tends to be shorter and the vowel space more centralized when words are employed as discourse markers, suggesting that discourse markers undergo hypoarticulation and, more generally, reduction.

Introduction

Languages are known to be ambiguous, and the primary sources of ambiguity in the lexicon are polysemy and homophony, both resulting in similar phonological forms referring to different objects. While polysemy refers to *similar* phonological forms with different, yet related meanings (e.g. French /kafe/ refers to the plant, the drink and the place where drinks are consumed), homophony refers to two *identical* phonological forms that have different meanings, and sometimes belong to different grammatical categories (e.g. French /sɑ̃/ refers to the numeral *cent*, "hundred", the noun *sang*, "blood", or the conjugated verb *sens/sent*, "(I, you, he/she/it) smell(s)"). Homophony and polysemy are challenging phenomena for both humans and automatic systems. For instance, it has been highlighted that the acquisition of a novel word is harder when the new entry has a homophone in the existing lexicon of the learner (Swingley & Aslin, 2007), especially if the new entry and the existing homophone belong to the same grammatical category (Dautriche, Swingley & Christophe, 2015). As for automatic processing, "word sense disambiguation" is a well-known task in NLP (natural language processing) that refers to "the ability to computationally determine which sense of a word is activated by its use in a particular context" (Navigli, 2009). Homophone disambiguation is also challenging for automatic speech recognition as it has been pointed out by studies that compare humans and ASR systems in transcribing ambiguous spoken samples, mainly due to the reduction phenomena that may increase the rate of near-homophone forms among function words (Nemoto, Vasilescu, & Adda-Decker, 2008).

This paper proposes an addition to the linguistic and speech processing literature on disambiguation by investigating a question that seems to have been less studied: that of achieving disambiguation by modeling fine-grained phonetic details that can be automatically extracted and statistically modeled. We know that phonetic variation depends on word frequency (Pierrehumbert, 2008; Phillips, 1984). For instance, it has been shown that word-frequency influences word-duration, allowing homophonous nouns and verbs in English to be distinguished (Lohmann and Conwell, 2020), and that more frequent words appear to have more centralized vowels due to shortening (Dinkin, 2008). Phonetic variation also depends on grammatical function in that function words are shorter, acoustically poorer and more prone to reduction (Adda-Decker & Snoeren, 2011; Ernestus & Warner, 2011). Finally, we know that phonetic variation depends on pragmatic usage, since fine prosodic cues can help disambiguate the function of words such as discourse markers (Didirková and colleagues, 2018, 2019; Lee et al., 2020).

Our working hypothesis is that, since fine phonetic features are word-specific and permit the distinction of meaning and/or function, they can be exploited to disambiguate homophones or polysemic words. To test this hypothesis, we take into account a well-known case of ambiguity: discourse markers (henceforth DMs). DMs are words or expressions such as *well*, *you know*, *I mean*, that are highly frequent in language use and have been shown to play an essential role as fluency devices (Crible, 2018) and in discourse planning and communication management (Levelt, 1993; Hasselgren, 2002). They usually emerge from other parts of speech with which they then co-exist in the language (Degand &

https://doi.org/10.18463/diss-2021-005-wu-etal

Fagard, 2011), and they are key items (Vasilescu, Rosset, & Adda-Decker, 2010) although problematic for various speech processing domains (Adda-Decker et al., 2003). Even when they fulfill the function of DMs, they often retain a high degree of polysemy (da Silva, 2006), whereas their high frequency in conversational speech raises the issue of contextual homophony due to reduction processes (Adda-Decker & Lamel, 2017).

In the present study, we first estimate the proportion of words in each POS category which also occur as DMs. These polysemous items, feeding the DM class, will be analyzed as a function of speech style. We then statistically compare the behavior of polysemous word types that behave as DMs (henceforth "DM uses", e.g. opening boundary *alors*) and as other POS (henceforth "Non-DM uses"; e.g. temporal adverb *alors*), as well as the behavior of DM word types and all other words in the corpus (henceforth "Others", e.g., *table*, "table" or *manger*, "eat"). Secondly, we focus on two phonetic features, duration and acoustic realization of vowels, both indicators of reduction (Meunier & Espesser, 2011), that we correlate with the classes defined above ("DM uses", "Non-DM uses", "Others") and with stylistic variables (speech style, gender). To that extent, we use a 4-hour, manually segmented and annotated French corpus to investigate whether different functions of word tokens that fulfill different POS as well as DM functions, bring about different phonetic behaviors. The final aim of the proposed approach is to take advantage of such smaller size, enriched data to model patterns that can be extrapolated to larger scale, more heterogeneous and less annotated corpora.

In the remainder of this paper, we first describe our corpus and methodology before presenting the results. The relation between the DMs, POS categories and speech styles, and the effect of the pragmatic use as DMs of some word tokens on phonetic features (mean phone duration, vocalic realization) are detailed in the "Results" section.

Corpus

For the present study, we chose to study the LOCAS-F (Degand, Martin, & Simon, 2014) corpus because of its already fine-grained manual annotations for parts-of-speech and DMs. It is composed of 42 sound tracks of 3-to-5-minute audio files, from a total 48 speakers. Multiple social practices were included in this corpus of primarily Belgian and metropolitan French, such that several speech styles are represented. The corpus is richly annotated in POS, DMs and other metadata.

Methodology

POS categories were identified manually by specialists, all of which are considered in our analyses. In the section dedicated to POS, only the nine categories that are polysemous with respect to the DM class are presented (see Table 1). The recordings were categorized into three speech styles according to the degree of preparation: (formal) prepared speech, (less formal) semi-prepared speech and (informal) unprepared speech. More details on the speech styles can be found in Degand et al. (2014). Concerning the vocalic dispersion, measurements of the first and second formants were extracted using PRAAT (Boersma & Weenink, 2006). As /o/ and /ɔ/ were manually annotated with similar standards in the LOCAS-F corpus, we decided to group the two vowels ("o-ɔ") in the analyses. Vowel spaces are illustrated for discourse markers ("DM uses") *vs* all other words ("Others") and then for the subset of word types fulfilling the function of DMs ("DM uses") *vs* the same word types employed with other POS functions ("Non-DM uses"). Due to limited space, we cannot present vowel spaces as a function of speech style. Outliers were mostly observed at the bottom left corner of the vowel spaces and were excluded from the analyses.

Table 1. Polysemous POS categories with respect to DM class.

Abbreviation	Complete category name
ADJ	Adjective
ADV	Adverb
CON	Coordinating conjunction
DET	Determiner
ITJ	Interjection
NOM	Noun
PRO	Pronoun
PRP	Preposition
VER	Verb

For the statistical analyses on POS and speech style, a generalized linear model (GLM) in R (R Core Team, 2013) was carried out on whether or not a word token can fulfill the DM function. The model was used to test the effect of POS categories and speech style. ADJ, PRO and DET were grouped together, given that we observe almost no "DM uses" for these POS. This grouping left us 7 categories of POS for the statistical analyses. The fixed effects considered were: part-of-speech (reference: CON) and speech style (reference: prepared speech).

Results

In this section, we present the DMs' ("DM uses") distribution and phonetic properties (duration,

vowels realization) compared to all the word tokens available in LOCAS-F ("Others") and to the same word tokens fulfilling other POS functions ("Non-DM uses").

DMs and Speech style

Table 2 gives the occurrences and percentages of word tokens annotated as DMs ("DM uses") *vs* all other word tokens ("Others") for each speech style. The word forms most often used as DMs are *et* ("and", 479 occurrences), *mais* ("but", 290 occurrences), *donc* ("so", 147 occurrences), *alors* ("then", 103 occurrences), *puis* ("then", 80 occurrences). Observed rates underline that DMs ("DM uses") are more frequent in unprepared speech (7%) than in prepared speech (2%). This trend is consistent with the literature as more spontaneous settings entail the use of such items for dialog management or planning purposes.

Table 2. Occurrences and rates of DM uses vs *Others as a function of speech style.*

	Prepared	**Semi-prepared**	**Unprepared**
DM uses	252 (2%)	440 (5%)	1321 (7%)
Others	12009 (98%)	8270 (95%)	16472 (93%)
Total	12261	8710	17793

Table 3 restricts the comparison to the word types corresponding to "DM uses" *vs* "Non-DM uses", for each speech style. Among word types that can fulfill a DM function, 8% of occurrences are used as discourse markers in prepared speech and more than 20% of the occurrences (24%) are used as discourse markers in unprepared speech. These results suggest that the unprepared settings increase the local ambiguity as some similar forms are more likely to fulfill the DM function.

Table 3. Occurrences and rates of DM uses vs *Non-DM uses as a function of speech style.*

	Prepared	**Semi-prepared**	**Unprepared**
DM uses	252 (8%)	440 (18%)	1321 (24%)
Non-DM uses	2729 (92%)	1981 (82%)	4073 (76%)
Total	2981	2421	5394

POS and Speech style

Figure 1 shows the rate of "DM uses" that is of word tokens having the meaning/function of DMs, as a function of the POS category from which they emerge. It is worth mentioning that we focus here only on similar word types that can be used as DMs (e.g. *alors*) and not on word clusters (e.g. *bein alors*). Separated results are displayed for each speech style. Selected POS correspond to the main categories likely to behave as DMs, as specified in Table 1. The figure shows the propensity of some POS to be more polysemous and fulfill the DM function in particular in unprepared speech. The categories that are most often used as DMs are Coordinating conjunctions (CON; e.g. *et*, "and"), followed by adverbs (ADV; e.g. *alors*, "so"), interjections (ITJ; e.g. *euh*, "uh") and prepositions (PRP; e.g. *pour*, "for"). Interestingly, pronouns (*moi*, "me") and determiners (DET; *le*, "the" or *mon*, "my") are almost never used as DMs in our corpus. The GLM results show that it is less likely to observe DMs in any of the other category than in CON ($p < 0.001$) and that it is more likely to observe DMs in less prepared speech than in more prepared speech ($p < 0.001$ for all pairwise comparison).

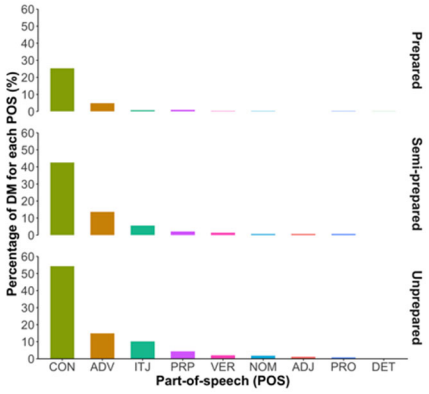

Figure 1. Rates of word tokens which occur as DMs displayed as a function of POS and speech style.

In the following two sections we focus on the phonetic properties of the class of words behaving as DMs ("DM uses") compared to the two other classes as identified above ("Non-DM uses", "Others").

Mean phone duration per word

Figure 2 gives an overview of mean phone duration per word for "DM uses" *vs* "Others". The highest bar is located slightly more to the left for "DM uses" than for "Others". This suggests that the local speech rate during the production of DM is slightly higher than for other words.

While the results displayed in Figure 2 could be affected by word frequency (given that word types corresponding to DMs are in general frequent words, thus produced with higher local speech rate), the analyses in Figure 3 support the hypothesis of phonetic features specific to DMs. They show that

for polysemous word forms that occur both as DM and other POS, the local speech rate is higher when they are employed as DMs. However, we may notice in the figures that the slope on the right side is decaying more slowly for DMs than for "Others". This reflects a poorly centered speech rate for DMs and suggests that hesitations which tend to increase segment durations might also be embedded in DM words.

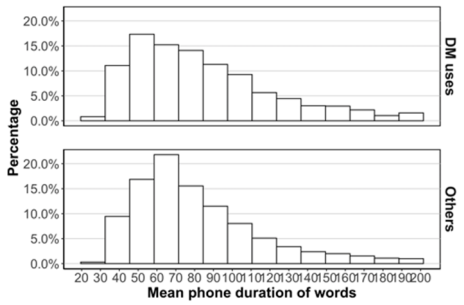

Figure 2. Mean phone duration per word for DMs vs other words.

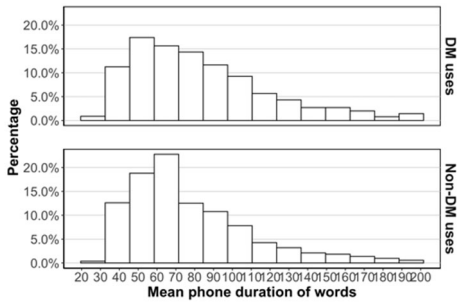

Figure 3. Mean phone duration per word for polysemous words employed as DMs vs Non-DM uses.

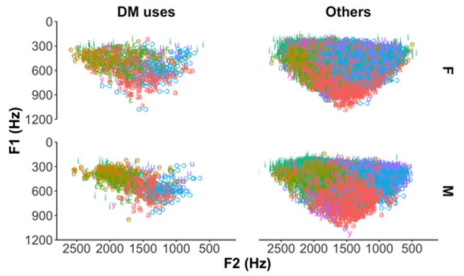

Figure 4. Vowel space for DMs (left) and other words (right) for female (F) and male speakers (M).

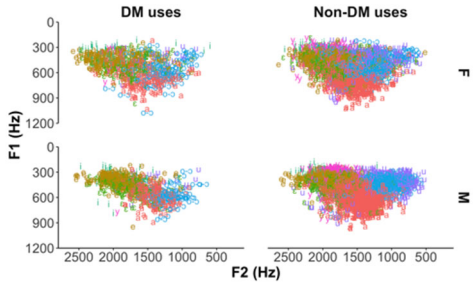

Figure 5. Vowel space for DMs (left) and DM words that are not used as DMs (right) for female (F) and male speakers (M).

Vowel space

This section focuses on the acoustic analysis of vowels: 1661 vowel segments are included in the "DM uses" class, 5636 in "Non-DM uses" and 38613 in "Others". Figure 4 illustrates the vowel space for DMs ("DM uses", left panel) *vs* other words ("Others", right panel) for female (F, top panel) and male speakers (M, bottom panel). The vowel space for other words is much larger than that for DMs, suggesting that vowels tend to be hypo-articulated when word forms are used as DMs. It is worth noting that DMs comprise only limited word types, and thus limited vowel identities (i.e., /i/, /y/, /e/, /ɛ /, /a/, /u/, /o/-/ɔ/). Moreover, the larger vowel space for the other words could also be due to the large range of word frequencies for words engaged in this category.

Figure 5 compares more closely the vocalic space of ambiguous words that can have both the role of DMs and other POS functions ("DM uses", left *vs* "Non-DM uses", right) for female (F, top) and male speakers (M, bottom). This comparison also allows us to control for word-frequency related variation. Similar to what is demonstrated in Figure 4, the group "DM uses" shows a smaller vowel space, suggesting that the acoustical realization of "DM uses" is more prone to reduction phenomena than in other situations ("Non-DM uses").

Conclusions

This study aims to investigate phonetic properties of discourse markers ("DM uses") compared to both similar phonological forms that can carry other POS functions ("Non-DM uses") and to all the available word forms ("Others") in the 4-hour richly annotated LOCAS-F corpus in French.

An account of "DM use" across data highlights that some POS are more polysemous and thus more likely to feed the DM class, in particular in more spontaneous settings. The POS that frequently fulfills the function of discourse marker is coordinating conjunction, followed by adverbs, interjections and prepositions. Then two phonetic parameters (mean phone duration per word and vowel space) are considered and quantified as function of classes of words (DM use, Non-DM use, Others). The distribution of mean phone duration per word, suggesting local speech rate, shows that mean phone duration, tends to be shorter for "DM uses" than for "Others" and for "Non-DM uses". Vowel space is smaller for "DM uses" than for "Others" and for "Non-DM uses", suggesting that discourse markers undergo hypoarticulation, and thus reduction, compared to other usages. Overall, our results encourage further investigations of patterns of

phonetic variation as cues for disambiguation in connected speech.

Acknowledgements

This research was supported by DATAIA/MSH Paris-Saclay "Excellence" grant and by IdEX U. de Paris, ANR-18-IDEX-0001, "Emergence" grant.

References

Adda-Decker, M., B. Habert, C. Barras, G. Adda, P. B. D. Mareuil & P. Paroubek. 2003. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. In: R. Eklund (ed.), *Proceedings of DiSS '03: Disfluency in Spontaneous Speech*, September 5–8, 2003, Göteborg, Sweden, 67–70.

Adda-Decker, M. & L. Lamel. 2017. Discovering speech reductions across speaking styles and languages. In: F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler & M. Zellers (eds.), *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*. Berlin, Germany: De Gruyter Mouton, 101–128. https://doi.org/10.1515/9783110524178-004

Adda-Decker, M. & N. D. Snoeren. 2011. Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics* 39(3), 261–270.

Boersma, P. & D. Weenink. 2006. Praat: Doing phonetics by computer (version 6.1.38). https://www.praat.org/ (accessed 24 January 2021).

Crible, L. 2018. *Discourse Markers and (Dis)fluency: Forms and functions across languages and registers*. Amsterdam, Netherlands: John Benjamins Publishing Company. https://doi.org/10.1075/pbns.286

da Silva, A. S. 2006. The polysemy of discourse markers: The case of *pronto* in Portuguese. *Journal of Pragmatics* 38(12), 2188–2205. https://doi.org/10.1016/j.pragma.2006.03.009

Dautriche, I., D. Swingley & A. Christophe. 2015. Learning novel phonological neighbors: Syntactic category matters. *Cognition* 143, 77–86. https://doi.org/10.1016/j.cognition.2015.06.003

Degand, L. & B. Fagard. 2011. Alors between discourse and grammar: the role of syntactic position. *Functions of language* 18(1), 29–56.

Degand, L., L. Martin & A.-C. Simon. 2014. Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté [Basic discourse units and their left periphery in LOCAS-F, an annotated multigenre oral corpus]. In: F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer & S. Prévost (eds.), *CMLF 2014 -4ème Congrès Mondial de Linguistique Française*, July 19–23, 2014, Berlin, Germany, 2613–2626. https://doi.org/10.1051/shsconf/20140801211

Didirková, I., G. Christodoulides, L. Crible & A. C. Simon. 2018. Naïve annotations of French *et* and *alors*: comparison with experts and effect of implicitation. In: L.-M. Ho-Dac & P. Muller (eds.), *Cross-Linguistic Discourse Annotation Applications & Perspectives (Proceedings of TextLink)*, March 19–21, 2018, Toulouse, France, 38–44.

Didirková, I., L. Crible & A. C. Simon. 2019. Impact of prosody on the perception and interpretation of discourse relations: studies on "et" and "alors" in spoken French. *Discourse Processes* 56(8), 619–642.

Dinkin, A. J. 2008. The real effect of word frequency on phonetic variation. *University of Pennsylvania Working Papers in Linguistics* 14(1), 97–106.

Ernestus, M. & N. Warner. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics* 39(SI), 253–260.

Hasselgren, A. 2002. Learner corpora and language testing: Small words as markers of learner fluency. In: S. Granger, J. Hung, & S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam, Netherlands: John Benjamins, 143–174. https://doi.org/10.1075/lllt.6.11has

Lee, L., D. Jouvet, K. Bartkova, Y. Keromnes & M. Dargnat. 2020. Étude comparative de corrélats prosodiques de marqueurs discursifs français et anglais selon leur fonction pragmatique [Comparative study of prosodic correlates of French and English discourse markers according to their pragmatic function]. In: C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, S. Schneider (eds.), *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1: Journées d'Études sur la Parole*, June 8–19, 2020, Nancy, France, 335–343.

Levelt, W. J. 1993. *Speaking: From intention to articulation* (Vol. 1). Cambridge, MA, USA: MIT Press.

Lohmann, A. & E. Conwell. 2020. Phonetic effects of grammatical category: How category-specific prosodic phrasing and lexical frequency impact the duration of nouns and verbs. *Journal of Phonetics* 78, Article 100939. https://doi.org/10.1016/j.wocn.2019.100939

Meunier, C. & R. Espesser. 2011. Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics* 39(3), 271–278.

Navigli, R. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)* 41(2), 1–69.Nemoto, R., I. Vasilescu & M. Adda-Decker. 2008. Speech Errors on Frequently Observed Homophones in French: Perceptual Evaluation vs Automatic Classification. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, May 26–June 1, 2008, Marrakech, Morocco, 2189–2195

Phillips, B. S. 1984. Word frequency and the actuation of sound change. *Language* 60(2), 320–342. https://doi.org/10.2307/413643

Pierrehumbert, J. B. 2008. Word-specific phonetics. In: C. Gussenhoven and N. Warner (eds.), *Laboratory Phonology 7*, Berlin, Germany: De Gruyter Mouton, 101–140.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Swingley, D. & R. N. Aslin. 2007. Lexical competition in young children's word learning. *Cognitive psychology* 54(2), 99–132.

Vasilescu, I., S. Rosset & M. Adda-Decker. 2010. On the functions of the vocalic hesitation euh in interactive man-machine question answering dialogs in French. In: *Proceedings of DiSS-LPSS Joint Workshop 2010*, September 25–26, 2010, Tokyo, Japan, 111–114.

Hesitations distribution in Italian discourse

Loredana Schettino[1], Simon Betz[2] and Petra Wagner[2]
[1] University of Salerno, Salerno, Italy
[2] Bielefeld University, Bielefeld, Germany

Abstract

The acknowledgment of the functional role of hesitations in speech has increased the research interest in investigating and modeling their occurrence in discourse. This study explores hesitation combinations and distribution in Italian discourse. Though clusters represent less frequent occurrences than standalone hesitations, it is still worth examining their composition, distribution, and context of occurrence for a better understanding of hesitations' role in discourse. Also, the emerging patterns may provide interesting findings for technological applications, such as integrating hesitations models in conversational agents' production to improve their communicative efficiency and naturalness.

Introduction

This study is part of a project that aims at modeling the occurrence of disfluencies, and more specifically, hesitation phenomena in Italian semi-spontaneous discourse for technological applications. It was developed from the CHROME project which includes among its general goals modeling multimodal data for the design of virtual agents serving in museums.

In the last forty years, prompted by Chafe's (1980) influential study, a positive view has been established, considering disfluency not as the opposite of fluency, as the term might suggest, but rather a component of fluency. In this light, Götz (2013) proposes the more neutral term *fluencemes* and Voghera (2017) identifies disfluencies among the most pervasive *speech functional linguistic correlates*. On the one hand, speakers can correct their uttered sequences through the deletion, insertion, or substitution of speech material, i.e., phenomena commonly gathered as *repairs* or *Backward-Looking Disfluencies*. On the other hand, speakers can gain some time when the planning process needs it, also providing time for listeners to process information, producing *hesitations* or *Forward-Looking Disfluencies,* under which pauses, fillers, lengthenings are commonly subsumed. (Levelt, 1989; Ginzburg, Fernández, & Schlangen, 2014). So, the role of hesitations in reducing the temporal pressure due to the dynamic simultaneity of speech planning, production, and reception processes is by now recognized. These considerations are crucial for researchers' raising interest in investigating and modeling hesitations composition, distribution, and function in discourse.

Hesitations have been observed to occur in simple and complex nested structures (Shriberg, 1994). In more complex compositions, strong positive correlations among different types of disfluencies are found (Merlo & Mansur, 2004, 499). In their contrastive study on the clustering of discourse markers and pauses in French and English, Degand and Gilquin (2013) observe the high frequency of the sequence of discourse markers followed by a filled pause. Building on these findings, Crible, Degand, and Gilquin (2017, 87) find that in French this pattern frequently occurs in clause-final position with "punctuating" function. Also, in French, Kosmala and Morgenstern (2017) find the recurrent occurrence of two combinations of hesitations—filled pauses preceding silences and filled pauses following lengthenings—and observe their structuring role in speech. The idea of hesitations as a tool for discourse structuring is further supported by findings in Schettino et al. (to appear). In their Italian data about a third of hesitant silent pauses and fillers occur at clause or topic-comment boundaries. As a matter of fact, as claimed by Tottie (2016, 100), speech planning can be identified as hesitations' basic function, but, according to their context, these phenomena may carry out other possible functions, like structuring discourse or highlighting key elements in speech (Kjellmer, 2003; Schegloff, 2010).

Moreover, Betz et al. (2015) highlight the importance of investigating micro-structure and the "phonotactics of disfluencies" (Betz et al. 2015, 2222) for the definition of a model of hesitation insertion to improve incremental dialogue synthesis systems. In their German data, clustered configurations are less frequent than standalone hesitations and mostly include a filled pause followed by a silence.

Despite the relevance of this topic, little is still known about hesitations combinations and distribution in Italian discourse. Our research intends to contribute to filling this gap by addressing the following questions:

1. In which combinations do hesitations occur in Italian discourse?
2. Where do they occur in discourse?
3. Does this correlate with their function in context?

https://doi.org/10.18463/diss-2021-006-schettino-etal

In this study, hesitations are defined as a temporary delay in speech delivery marked by phenomena like silent pauses, fillers, and lengthenings which do not have a propositional content but carry procedural meaning.

Corpus and Methods

We performed a corpus-based analysis on a dataset from the C.H.R.O.M.E. corpus (Origlia et al., 2018). It consists of about 80 minutes of semi-spontaneous speech by three female expert guides leading visits at San Martino's Charterhouse. Per each guide, 2 visits were considered, where they show the same point of interest.

Disfluency phenomena were annotated using the ELAN software (2020; Sloetjes & Wittenburg, 2008) according to a three-level annotation scheme (described in Schettino et al., to appear). On the first level, disfluencies' macro-structure is labeled, namely: the *reparandum,* the region to be repaired; the *reparans*, the repaired one; the *interregnum*, the one where the delay occurs (Shriberg, 1994). The second level is for the microstructure, that is the types of disfluency phenomena composing the disfluent event. Each of these phenomena is identified as *Backward-Looking* or *Forward-Looking* (Ginzburg et al., 2014) on the third level.

Among the types of disfluency phenomena, the following hesitations were considered:

- Silent Pauses (SP) perceived as a hesitant pause in context (Lickley, 2015);
- Lengthenings (LEN), marked prolongation of segmental material (Betz, 2020, 14);
- Filled Pauses (FP), non-verbal filler, vocalizations ("eeh", "ehm");
- Lexicalized Filled Pauses (LFP), strongly semantically bleached verbal fillers.

As the identification of hesitant pauses does not depend on absolute measures but is related to the context of occurrence, this annotation relies on subjective perceptual judgment, so the interrater reliability was tested measuring Cohen's *K* for the annotations by two expert raters ($K = 0.91$, high agreement, Landis & Koch, 1977).

Each of these hesitation types was associated with possible function/s according to their co-text.

- *Word Searching* (WS), items involved in the search for a target word (Tottie, 2020).
  «*potete intuire <ehm> <sp> la<aa> la bellezza*»
  «you can grasp <uhm> <sp> the<ee> the beauty»
- *Structuring* (STR), items structuring discourse on syntax (clause) and information structure levels (topic-comment). For example:
  «*la Certosa di San Martino qui a Napoli<ii> <ehm> ha almeno due anime <sp> <eeh> una<aa> <ehm> racconta…*»
  «the Charterhouse here in Neaples<vv> <ehm> has two souls <sp> <eeh> one<ee> <ehm> tells the story…»
- *Focusing* (FOC), items preceding semantically heavy and often emphasized elements (Kjellmer, 2003). For example:
  «*quindi la<aa> Certosa ha un'origine <sp> trecentesca*»
  «so the<ee> Charterhouse has a <sp> 14th century origin»
- *Hesitative* (HES), hesitations' basic function of speech planning (Tottie, 2016). This label is assigned to items for which no other function can be identified. For example:
  «non possiamo vedere<ee> molto bene»
  «we can't see<ee> it properly».

For this level, the interrater agreement was substantial ($K = 0.73$).

The analysis focused on the content of the *Interregna*, namely standalone hesitations and clusters, and their position within tonal units (Crocco, 2005). Two metrics were considered to assess the position of hesitation phenomena regarding the tonal units. One relies on percentage values based on the unit length, so lower values correspond to phenomena that occur towards the beginning of the tonal unit, conversely, higher values stand for phenomena that occur towards the end of the tonal unit (100%). The second metrics is based on the unit constituents, hesitation phenomena were classified as follows: *Initial*, when occurring at the beginning of a unit or within the first syntactic constituent when the unit consists of several ones; *Medial*, when occurring between different constituents; *Final*, when occurring at the end of a unit or within the last syntactic constituent when the unit consists of several ones; *Unit*, when the item corresponds to the tonal unit; *Cross-unit*, when the item belongs to an *interregnum* that covers the final part of a unit and the initial part of the following one.

The statistical analysis was conducted using Generalized and Linear Mixed Models, including Speaker and Item as random variables (GLMM, LMM using "lme4" package, Bates et al., 2015).

Results

Hesitation Combinations

In the data, 940 *Interregnum* intervals were detected. They were comprised of standalone hesitations (79%), clusters (18%), and other phenomena like breath-noises and tongue-clicks (3%).

As shown in Figure 1, most hesitations occur in a standalone fashion (66%), whereas fewer occur in clusters (34%). Considering type and function of phenomena as predictors of standalone vs. clustered hesitation (as binomial categorical outcome), lengthenings are less likely to occur in clusters ($E = -0.90$, $SE = 0.43$, $z = -2.09$, $p = 0.03$), a similar tendency is found for lexical fillers ($E = -0.71$, $SE = 0.37$, $z = -1.91$, $p = 0.05$). Further, the function of Word Searching is a significant predictor of hesitation clusters ($E = 1.33$, $SE = 0.46$, $z = 2.87$, $p = 0.004$).

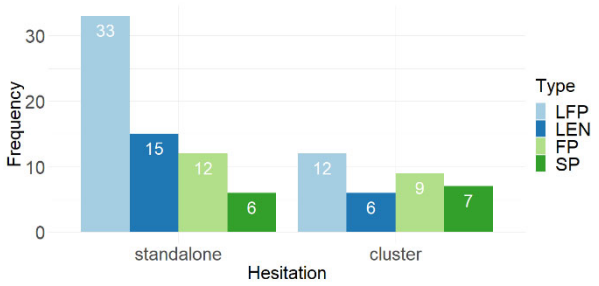

Figure 1. Frequency (%) of standalone hesitations and hesitation clusters per type

As for combinations, they mostly consist of two items (80%), followed by combinations of three items (17%), four items (3%), up to the rare maximum of five items (1%). Figure 2 shows the frequency of occurrence of the combinations found for the first couple of items. Lexical fillers are mostly followed by another lexical filler, the next most frequent combination is with a lengthening or a silence; lengthenings are mostly followed by a filled pause; filled pauses by a lexical filler or a silence; silent pauses are mostly followed by a filled pause.

Hesitation Position

The model featuring the position (%) within the tonal unit as outcome and hesitations type, function, and combination as predictors yielded the following results. No significant distinction is found for the position in which standalone items (44%) and clusters (37%) are most likely to occur. Still, Figure 3A illustrates that among hesitation types, filled pauses (29%, $E = -0.17$, $SE = 0.05$, $t = -3.49$) and silent pauses (31%, $E = -0.14$, $SE = 0.05$, $t = -2.85$) occur significantly more toward the beginning of the tonal unit than lexical fillers (48%) and lengthenings (49%). As for hesitation functions, as depicted in Figure 4A, items with a Structuring function occur significantly more toward the beginning of the tonal unit than the others ($E = -0.08$, $SE = 0.02$, $t = -4.38$).

These results are consistent with the outcome of the second analysis, where each one of the position categories was processed as a binomial dependent variable and modeled as a function of hesitation type and function. As shown in Figure 3B, filled and silent pauses are significantly more likely to occur in *Initial* position than lengthenings and lexical fillers ($p < 0.001$). Also, the significant effect of hesitation Structuring function is confirmed ($p < 0.0001$, see Figure 4B).

Finally, it is worth noticing that 11% of hesitation clusters stretch across the boundaries between tonal units, and these mostly include hesitations with a Word Searching function (69%).

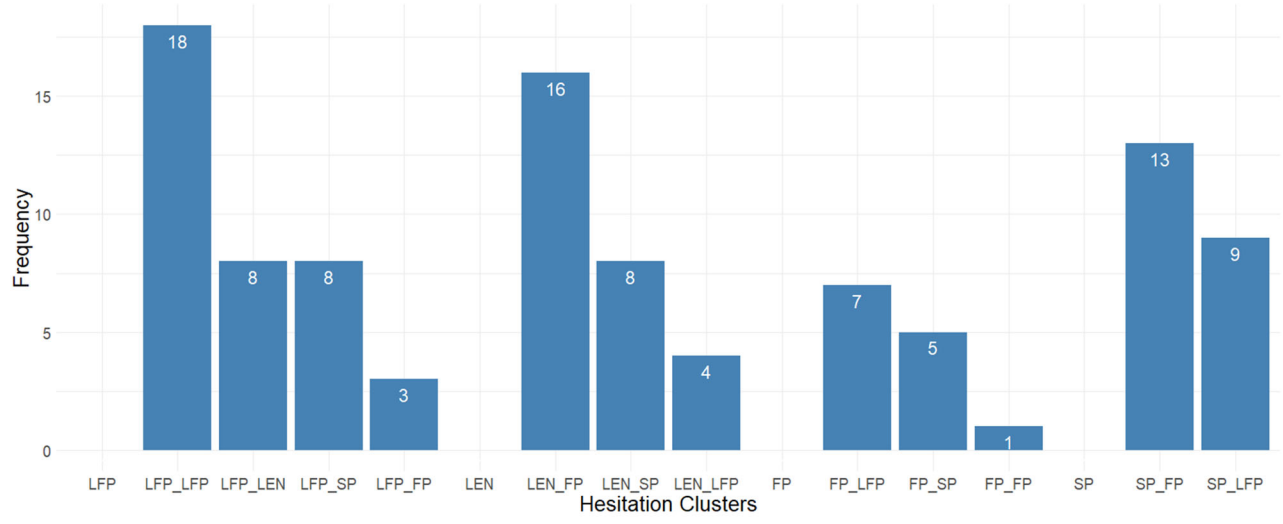

Figure 2. Frequency (%) of the combinations in hesitation clusters.

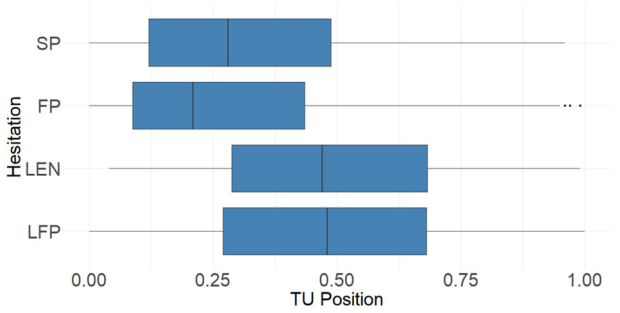

Figure 3A. Position (%) of hesitations within the tonal units per type.

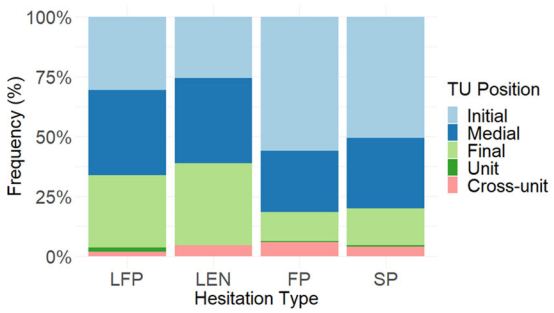

Figure 3B. Frequency of hesitaton positions within the tonal units per type.

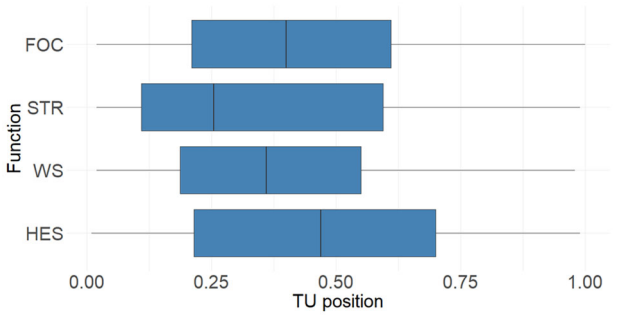

Figure 4A. Position (%) of hesitations within the tonal units per function.

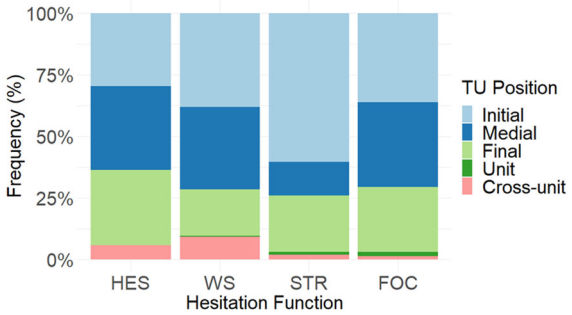

Figure 4B. Frequency of hesitation positions within the tonal units per function.

Discussion

Results confirm the observation that hesitation phenomena mostly occur alone rather than in clusters (Betz et al., 2015). Silent pauses and filled pauses most likely occur in clusters together or with lengthenings and lexical fillers. Further, most frequent sequences include a lexical filler followed by a lengthening or a silent pause, which are both most often followed by a filled pause, the latter then being most often followed by a lexical filler. So, based on these recurring combinations, the following sequence could be proposed:

"original utterance [LFP – LEN – SP – FP – LFP] + continuation"

Hence, much like in the model defined by Betz and his colleagues (2018, 7), a hesitation insertion model could start with less intrusive items like lexical fillers and lengthenings, then include the more salient silent and filled pauses, and eventually, a lexical filler could be introduced in order to take extra time and at the same time compensate for the disfluent event emphasizing upcoming speech. As a matter of fact, as discussed in Betz et al. (to appear), disfluency salience seems to be language-specific and in Italian, silences might be even more salient phenomena than fillers. Moreover, Schettino et al. (to appear) finds lexical fillers to strongly correlate with the focusing and structuring functions.

As for hesitation position relative to the prosodic unit, it was found that early positions in the unit are most likely occupied by silent pauses, filled pauses, and phenomena structuring discourse at the clause or information structure level, whereas later positions are preferred by the less intrusive lengthenings and lexical items, and by hesitations dealing with the search of a target word and those highlighting key concepts.

Finally, a number of clusters stretches across tonal units, mostly due to word retrieval problems.

Conclusion

This study focused on hesitation combination patterns. Though they are rarer occurrences than standalone hesitations, it is still worth examining their composition, distribution, and context of occurrence for a better understanding of hesitations' role in discourse.

Moreover, these results may provide interesting findings for technological applications, such as the improvement of conversational agents' communicative efficiency and naturalness through the integration of hesitations based on a linguistic model. Further studies may involve investigations on dialogic datasets and the interplay of hesitations in clusters, also including backward-looking disfluencies.

Acknowledgments

Work funded by the Italian National Project PRIN "Cultural Heritage Resources Orienting Multimodal Experiences (CHROME)" (B52F15000450001).

Notes

[1] This work results from the collaboration between the authors. Loredana Schettino: data wrangling, analysis, writing. Simon Betz: advice and discussion. Petra Wagner: supervision.

References

Bates D., M. Mächler, B. Bolker & S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Betz, S., N. Bryhadyr, L. Kosmala & L. Schettino. To appear. A crosslinguistic study on the interplay of fillers and silences. Submitted to Disfluency in Spontaneous Speech 2021, 25–26 August 2021, Paris 8 University, France.

Betz S., P. Wagner & D. Schlangen. 2015. Micro-structure of disfluencies: Basics for conversational speech synthesis. In: *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*, 610 September, 2015, Dresden, Germany, 2222–2226.

Betz, S., B. Carlmeyer, P. Wagner & B. Wrede. 2018. Interactive hesitation synthesis: modelling and evaluation. *Multimodal Technologies and Interaction*, 2(1), 9. https://doi.org/10.3390/mti2010009

Betz, S. 2020. *Hesitations in Spoken Dialogue Systems*. Ph.D. dissertation, Universität Bielefeld.

Chafe, W. 1980. Some reasons for hesitating. In: H. W. Dechert & M. Raupach. (eds), *Temporal variables in speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton, 169–180. https://doi.org/10.1515/9783110816570.169

Crible, L., L. Degand & G. Gilquin. 2017. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis) fluency. *Languages in Contrast*, 17(1), 69–95. https://doi.org/10.1075/lic.17.1.04cri

Crocco, C. 2005. Etichettatura prosodica [Prosodic tagging]. In: *Spoken Italian – Multilevel Database*. http://www.parlaritaliano.it (accessed 25 June 2021)

Degand, L. & G. Gilquin. 2013. The clustering of 'fluencemes' in French and English. Presentation at *7th International Contrastive Linguistics Conference (ICLC 7) – 3rd Conference on Using Corpora in Contrastive and Translation Studies (UCCTS)*, July 10–13, 2013, Ghent, Belgium.

ELAN (Version 6.0) [Computer software]. 2020. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan (accessed 25 June 2021)

Ginzburg, J., R. Fernández & D. Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics* 7(9), 1–64. https://doi.org/10.3765/sp.7.9

Götz, S. 2013. *Fluency in native and nonnative English speech*. Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/scl.53

Kjellmer, G. 2003. Hesitation. In Defence of Er and Erm. *English Studies* 84(2), 170–198. https://doi.org/10.1076/enst.84.2.170.14903

Kosmala, L. & A. Morgenstern. 2017. A Preliminary Study of Hesitation Phenomena in L1 and L2 Productions: A Multimodal Approach. In: R. Eklund & R. Rose (eds.), *Proceedings of DiSS 2017: The 8th Workshop on Disfluency in Spontaneous Speech*, August 18–19, 2017, Stockholm, Sweden, 37–40.

Landis, J. & G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–74. https://doi.org/10.2307/2529310

Levelt, W. J. 1989. *Speaking: From intention to articulation* (Vol. 1). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Lickley, R. J. 2015. Fluency and Disfluency. In: A. M. Redford, *The Handbook of Speech Production*, Hoboken, NJ: Wiley Blackell, 445–474. https://doi.org/10.1002/9781118584156.ch20

Merlo, S. & L. L. Mansur. 2004. Descriptive Discourse: Topic Familiarity and Disfluencies. *Journal of Communication Disorders* 37(6), 489–503. https://doi.org/10.1016/j.jcomdis.2004.03.002

Origlia, A., R. Savy, I. Poggi, F. Cutugno, I. Alfano, F. D'Errico, L. Vincze & V. Cataldo. 2018. An Audiovisual Corpus of Guided Tours in Cultural Sites: Data Collection Protocols in the CHROME Project. In: B. N. De Carolis, C. Gena, T. Kuflik, A. Origlia & G. E. Raptis (eds.), *Proceedings of the 2018 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage, Castiglione della Pescaia*, Article 8.

Schegloff, E. A. 2010. Some other "uh (m)" s. *Discourse Processes*, 47(2), 130–174. https://doi.org/10.1080/01638530903223380

Schettino L., S. Betz, F. Cutugno & P. Wagner. To appear. Hesitations and Individual Variability in Italian Tourist Guides' Speech. In: *Proceedings of the XVII Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV)*. Zürich, Switzerland.

Shriberg, E. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. dissertation, University of California, Berkeley.

Sloetjes, H. & P. Wittenburg. 2008. Annotation by category – ELAN and ISO DCR. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, May 26–June 1, 2008, Marrakech, Morocco, 816–820

Tottie, G. 2016. Planning what to say: Uh and um among the pragmatic markers. In: G. Kaltenbock, E. Keizer & A. Lohmann (eds.), *Outside the Clause: Form and Function of Extra-Clausal Constituents*, Amsterdam, Netherlands: John Benjamins, 97–122. https://doi.org/10.1075/slcs.178.04tot

Tottie, G. 2020. Word-search as word-formation? The case of uh and um. In: P. Núñez-Pertejo, M. J. López-Couso, B. Méndez-Naya & J. Pérez-Guerra (eds.), *Crossing linguistic boundaries: systemic, synchronic and diachronic variation in English.* London: Bloomsbury Academic, 29–42. https://doi.org/10.5040/9781350053885.ch-002

Voghera, M. 2017. *Dal parlato alla grammatica: Costruzione e forma dei testi spontanei*. [From Speech to Grammar: Construction and form of spontaneous texts.] Roma: Carocci.

Investigating disfluencies contribution to discourse-prosody mismatches in French conversations

Laurent Prévot [1,2] Roxane Bertrand [1] and Stéphane Rauzy [1]
[1]Aix Marseille Université & CNRS, Aix-en-Provence, France
[2]Institut Universitaire de France, Paris, France

Abstract

In conversation, discourse and prosodic units association can be articulated through an interesting range of configurations. The situation in which these units are mismatching is the least studied and understood of these configurations. We make the hypothesis in this paper that disfluencies are a major cause for such mismatches. Our quantitative analysis based on a 8 hour corpus of French conversations manually annotated with disfluencies, discourse units (DU) and prosodic units (PU), confirms that disfluencies do play a major role in PU-DU mismatch but also that other sources should be considered. In the analysis, we also provide some insight about the different types of disfluencies and their frequency in the different DU-PU configurations.

Introduction

Discourse, Prosody and Syntax interplay is a crucial aspect of linguistic analysis of conversation. Previous literature had described many aspects of the association between these three levels in terms of boundary alignment (Degand & Simon, 2009; Prévot et al., 2015; Lacheret-Dujour & Kahane, 2020) and explored them in relation to *discourse genre* and *speaking style* (Degand & Simon, 2009b). The configuration least understood remains the case in which major prosodic unit and discourse unit boundaries do not match. In this paper, we hypothesize that a large number of those mismatches are related to disfluency (Shriberg, 1994). We analyse an 8 hours corpus of French conversations (Bertrand et al., 2008) that had been manually annotated with prosodic units, discourse units and disfluencies. After introducing previous work, we present our annotated data and a set of quantitative analyses aiming at better understanding the impact of disfluencies on mismatches between these units.

Related Work

While much work has been done on the link between prosody and syntax, and more particularly intonation and syntax, much remains to be done on the link between prosody and discourse. Some studies have shown that syntax and prosody play a role in the construction and identification of TCU (Turn-Constructional Units) (Ford & Thompson, 1996; Selting, 1996). In French, some studies attempt to model such a unit at the interface of syntax, prosody and discourse (see Degand & Simon, 2009; Lefeuvre & Moline, 2011 for a review of different approaches; Lacheret-Dujour & Kahane, 2020).

Following Lacheret-Dujour and Kahane (2020) or Degand and Simon (2009) we consider the different levels as autonomous. The basic discursive unit (BDU) in Degand and Simon refers to the "segments that speakers use to build a representation (interpretation) of the discourse. BDUs have a cognitive function since they correspond to steps of production and discourse processing. BDUs require syntax and prosody and their different matching give rise to several types of BDU corresponding to different discursive strategies.

The syntactic and discourse units of Lacheret-Dujour & Kahane (2020) is based on macro syntactic approach (Deulofeu, 2016) taking into account the illocutionary force (Austin, 1962) of the unit.

However, some difficulties remain in segmenting these units due to the specific phenomena frequents natural conversations. Among them, we consider that disfluencies represent a source of confusion for analyzing these levels.

Disfluencies are very frequent in spontaneous speech (about 1 every 15 words in the CID, Pallaud et al., 2019) and can occur at phonetic or morphosyntactic level anywhere in the utterance. They consist in an interruption of the flow that can be repaired or abandoned (Shriberg, 1994).

Concerning more precisely PU-DU mismatches, called *mixed-BDU* in Degand and Simon (2009b), they are not considered crucial in their analysis but they state that this 'unexpected' category deserves more attention, at least to understand why it occurs at significant rates. Lacheret-Dujour and Kahane (2020) called them *asynchronous* (12% of their prosody-syntax units) and relate it to difficulties for the speaker to produce and plan the utterance, which indeed suggest to look with more attention at their relationship with disfluencies.

https://doi.org/10.18463/diss-2021-007-prevot-etal

Data

This work is performed on the whole Corpus of Interactional Data (8 conversations of 1 hour each). In this corpus participants have a chat about "unusual situations" or "conflicts at work". See Bertrand et al. (2008) and Blache et al. (2017) for details on the corpus. The annotations used in this study are coming from three independent annotation campaigns. Overall discourse and prosodic segmentation have been performed through independent annotation campaigns realized by naive annotators trained and equipped with guidelines. Disfluency annotations have also been annotated in this way and an expert (one of the authors of the present paper) manually corrected and enriched the whole dataset.

Compared to earlier work, the amount of units annotated is much larger since the study deals with 17,102 discourse units and 30,970 prosodic units.

Disfluencies annotation

Disfluency phenomena were manually annotated following the guidelines presented in Pallaud et al. (2019). The disfluencies are defined as interruptions of the verbal fluency of the utterance at the morphosyntactic level. Some of these interruptions are characterized by utterances that are simply given up (referenced hereafter as *DISI*), some others correspond to a suspension of the verbal fluency but which continues without any impact on the syntactic structure *(DISS),* and a last kind implies the repair of the morphosyntactic sequence with the presence of a truncated word *(DIST*) and / or of a *break (DISB*), for which the annotation scheme of Shriberg (1995) is applied. This scheme proposes a three terms structure composed of the *Reparandum* (the term to be repaired), the *Interregnum* (*Break point*, which can be empty) and the *Reparans* (the repairing term). In case of multiple repairs, the disfluency annotation follows a tree structure which traces the paradigmatic pile. The annotation task does not present any major difficulty except the ambiguity in deciding whether an utterance is marked as abandoned or marked as the repaired term of the next utterance. The categories introduced here are illustrated in examples (1) and (2) below.

Prosodic Units

Prosodic units *(PUs)* are based on the two main consensual units in French (Di Cristo, 1998; Jun & Fougeron, 2000; etc). The Accentual Phrase (AP) is the lowest tonal unit which is the domain of primary and secondary stress. The right boundary of AP is demarcated by a final rise (LH) and the lengthening of the final syllable. The Intonation Phrase (IP) is higher than AP. It is marked by a major f0 movement on the last or two last syllables of the IP, a large final lengthening and often followed by a pause (Di Cristo, 1998; Fougeron & Jun, 1998). We will only consider the latter here.

The guidelines were simplified to be used by naive annotators (2 annotators for each file). The annotation was conducted manually and the annotators did not have strict instructions regarding silent pauses or hesitations. Thus, as long as disfluency items did not interfere with the prosodic phrasing, the annotators were free to annotate them either independently of the rest of the utterance or by integrating them. The prosodic units then obtained reflect how annotators have treated disfluencies. This first step of non-expert annotation was partly aimed at focusing on true sources of difficulty and then enabled us to better disentangle between the problematic items (Portes & Bertrand, 2011). Also, we hypothesize that the presence of disfluencies could have an impact on the mismatch between discourse and prosodic units.

Manual prosodic segmentation with our guidelines has proven to be relatively reliable with κ-scores (Cohen, 1960) ranging between 0.5 and 0.65 for naive coders and 0.75–0.85 for expert coders.

Discourse Units

Our discourse unit segmentation was inspired by Muller et al. (2012) and corresponds to Elementary Discourse Units used in Afantenos et al. (2012) but adapted to our interactional spoken data and simplified to be used by naive annotators. The guidelines combined semantic (eventualities identification), discourse (discourse markers) and pragmatic (speech acts) instructions. Such a mixture of levels has been made necessary by the nature of the data featuring both rather monologic narrative sequences and highly interactional ones. The annotation was performed on the transcript alone without access to audio files (but including pause and timing information). Manual discourse segmentation with our guidelines has proven to be reliable with κ-scores ranging between 0.8 and 0.85. In this approach *DUs* are semantic counterparts of independent syntactic clauses, at discourse level. They are also closely related to the macro-syntactic *Illocutionary Units* (Lacheret-Dujour & Kahane, 2020) as well as to the Discourse Units of Degand & Simon (2009a).

In the annotation we distinguished between *Discourse Units (DU)* and *Abandoned Discourse Units (ADU)* (illustrated in (1) below) that correspond to false starts that cannot be easily related

to the material coming after (as illustrated below in example (1)). As a consequence, *ADU* are disfluencies in which there is no *reparans;* and disfluencies should not interfere with *DUs.*

Illustration

Example (1) below illustrates the *ADU* vs. *DU*, as well as interrupted units *(DISI)* at disfluency level. Example (2) illustrates the different disfluency categories: suspensive *(DISS)*, with break *(DISB)*, Truncated words *(DIST) as well as the reparandum (REP)* along with PU and DU structures. Finally, (3) shows a case of PU crossing DU boundaries.

(1) *<[que j'avais envie (d-)$_{DIST}$ enfin bref]$_{PU}$ >$_{ADU}$ (#)$_{DISI}$ <[et (#)$_{DISS}$ on l'a accueillie (b-)$_{DIST}$ (a-)$_{DIST}$ on lui a rien demandé]$_{PU}$>$_{DU}$*

(2) *<[(ou des)$_{REP}$ (euh non)$_{DISB}$]$_{PU}$ [(pas des)$_{REP}$ (f-)$_{DISB}$ pas des frustrations]$_{PU}$>$_{DU}$ <[(des (#)$_{DISS}$ espèces de)$_{REP}$ (euh)$_{DISB}$]$_{PU}$] (# mhm #)$_{DISB}$ [[(ouais)$_{DISB}$]$_{PU}$ [(des)$_{REP}$ des vues]$_{PU}$[différentes]$_{PU}$ [sur le boulot]$_{PU}$[quoi]$_{PU}$ >$_{DU}$*

(3) *< ... [tu as un décalage]$_{PU}$ [quand même par rapport à l' âge]$_{PU}$ [c' est normal >$_{DU}$ <**surtout**]$_{PU}$ [à cet âge -là]$_{PU}$ ...>$_{DU}$*

Disfluencies and PU-DU congruence

We approach the relationship between disfluencies and PU-DU congruence by scrutinizing what happens at PU and DU boundaries in disfluent vs. fluent sequences. More precisely we start by comparing PU-DU matching within ADUs and within other DUs. Our prosodic units being overall much smaller than our discourse units, we then explore disfluencies when DU-PU mismatches, excluding the ADU case before looking in detail the different disfluency categories in this context. Finally explore other potential sources of mismatches.

Abandoned Discourse Units

By definition, ADUs are disfluent speech. Figure 1 illustrates the difference between *DU* and *ADU,* the latter hosting a much larger proportion of some mismatches between *PU* and *DU*.

PU crossing DU boundaries

We then consider with Figure 2 the DU case, excluding ADU by looking at whether a given *PU* includes or not a disfluency when it is either internal or matching a DU or crossing a DU boundary. Given our fine-grained *PUs,* the majority of DU-PU relationships are either 1-to-1 mapping or one *DU*

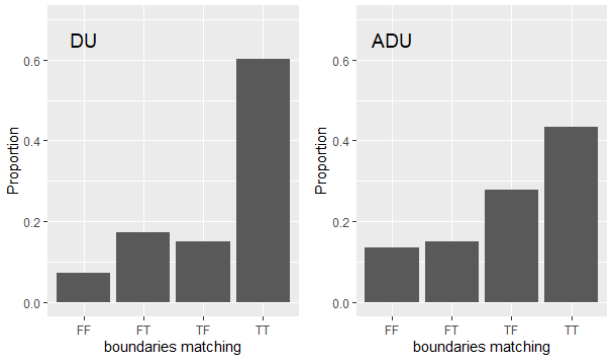

Figure 1. For the two groups DU and ADU, the proportions of mismatches encoded as TT (no mismatch), TF (mismatch on the right boundary), FT (mismatch on the left boundary and FF (mismatch on both boundaries).

including several *PU* (while matching *PU* left and right boundaries). However, *10.8%* of our *PUs* are crossing *DUs* boundaries *(as in example (3) above)*. Figure 2 illustrates that disfluencies are more frequent in those mismatch situations than in matching boundaries cases. This explains a major source of mismatches between *PUs* and *DUs,* putting aside *ADU* cases.

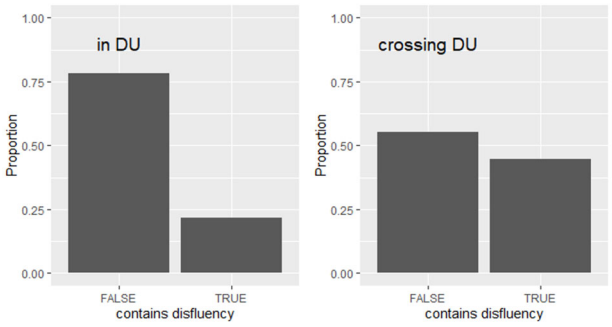

Figure 2. Comparison of PU hosting disfluency or not depending on its relation with DUs

Disfluency type

Among the PUs in DUs, we compare PUs ending *DUs* vs. non-ending *DUs* according to the type of disfluency. The Figure 3 illustrates that *breaks* tend to terminate a PU (but not a DU which is not surprising given *DU* definition), and a new PU starts with the *reparans*. Suspensive disfluency, that does not alter the syntactic flow but is likely to impact the prosodic flow with empty or filled pauses, is also a phenomena that tends to close *PU*.

Other sources of DU-PU mismatch

In order to figure out better what happens in the mismatch zone, we extracted the tokens distribution of such zones and normalized these raw frequencies based on the distribution of DU-final and DU-initial tokens (which seems to be the best candidate for such

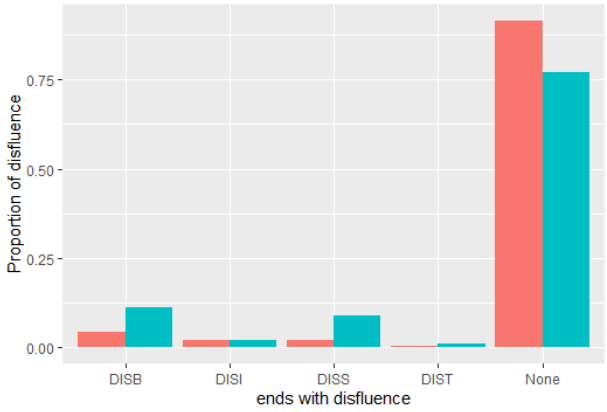

Figure 3. PUs ending DUs (red) vs. non-ending DUs (blue) according to the type of disfluency (with break (DISB), abandoned (DISI), without repair (DISS), truncation (DIST) or not ending with a disfluency (None).

a normalisation). The tokens over-represented in mismatch areas are Reported Speech (RS) verb introducer *(dire – say),* spoken particles *(quoi / hein / bon / tu_vois / enfin / tu_sais) ;* filled pauses *(euh)*, and to a less extent, first person pronouns *(je / moi).*

The presence of filled pauses in this short list simply confirms the impact of disfluency on DU-PU alignment. First person pronouns can also be a confirmation in that direction.

When Reported Speech (RS) verb introduces are the lexical items the most associated with PU-DU mismatches. (Lacheret-Dujour & Kahane, 2020) also mentioned RS as a source of asynchronous units. The main cause is that changing perspective through reporting speech clearly initiates a new discourse unit starting right after the verb introducing the RS, but sometimes the initial element of RS is prosodically grouped with the introducer.

Spoken particles create two challenges. At prosodic level, even if they are extremely short, their phrasing can vary a lot from one example to another leading to very different PU segmentations. At discourse level, some of them can be both DU-initial or DU-final (*enfin / bon / tu_vois*). It makes it difficult to decide whether they should be included in the DUs they follow or in the one after.

On the side of the spectrum some lexical items are associated PU-MU matches. This is the case of clearly initial discourse markers such as *et (and) / parce que (because) / donc (so) / mais (but) / alors (then) / ben (well)*. Those items with their clear signal of initiating a new unit could be used as some kind of synchronisation place for the different levels. The second part of the french negation *pas* falls also in this category, but in this case in final position. There are also some other items for which we do not have a clear explanation: *là / ça / y / ils*

Conclusion

This study allowed us to refine our understanding of the impact of disfluencies on discourse-prosody interfaces. Results largely confirm what is known on this matter, namely that disfluency strongly impacts prosodic units but less discourse ones once false starts are put aside. Disfluencies explain a sizable proportion of such mismatches that are not easy to analyse from discourse-prosodic interface viewpoint. Some other sources of mismatches (such as direct reported speech) could be further investigated in order to cover the whole range of phenomena generating those DU-PU mismatches. In this paper, we pushed the analysis both in terms of scale (8 hours of conversational speech) as well as in terms of granularity specifically with regards to the different types of disfluencies involved.

References

Afantenos, S., N. Asher, F. Benamara, M. Bras, C. Fabre, M. Ho-dac, A. Le Draoulec, P. Muller, M.-P. Péry-Woodley, L. Prévot, J. Rebeyrolles, L. Tanguy, M. Vergez-Couret & L. Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, May 21–27, 2012, Istanbul, Turkey, 2727–2734.

Austin, J. L. 1962. *How to do things with words*. Oxford: Clarendon Press.

Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde & S. Rauzy. 2008. Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, *49*(3), 105–134.

Blache, P., R. Bertrand, G. Ferré, B. Pallaud, L. Prévot & S. Rauzy. 2017. The corpus of interactional data: A large multimodal annotated resource. In: N. Ide & J. Pustejovsky (eds.), *Handbook of linguistic annotation*, Dordrecht: Springer, 1323–1356. https://doi.org/10.1007/978-94-024-0881-2_51

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Degand, L. & A.-C. Simon. 2009a. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours*, 4. https://doi.org/10.4000/discours.5852

Degand, L. & A.-C. Simon. 2009b. Mapping prosody and syntax as discourse strategies: How Basic Discourse Units vary across genres. In: D. Barth-Weingarten, N. Dehé & A. Wichmann (eds.), *Where prosody meets pragmatics*, Leiden, Netherlands: Brill, 79–105. https://doi.org/10.1163/9789004253223_005

Deulofeu, J. 2016. La macrosyntaxe comme moyen de tracer la limite entre organisation grammaticale et organisation du discours [Macrosyntax as a means of drawing the line between grammatical organization and the organization of discourse]. *Modèles linguistiques*, *38*(74), 135–166. https://doi.org/10.4000/ml.2040

Di Cristo, A. 1998. Intonation in French. In: D. Hirst & A. Di Cristo (eds), *Intonation systems: A survey of Twenty Languages*. Cambridge, UK: Cambridge University Press, 195–218.

Ford, C. E. & S. A. Thompson. 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In: E. Ochs, E. A. Schegloff & S. A. Thompson (eds.), *Interaction and grammar*, Cambridge, UK: Cambridge University Press, 134–184. https://doi.org/10.1017/CBO9780511620874.003

Fougeron, C. & S-A. Jun. 1998. Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics* 26(1), 45-69. https://doi.org/10.1006/jpho.1997.0062

Jun, S. & C. Fougeron. 2000. A Phonological model of French intonation. In: A. Botinis (ed.) *Intonation: Analysis, Modeling and Technology*, Dordrecht: Kluwer Academic Publisher, 209–242. https://doi.org/10.1007/978-94-011-4317-2_10

Lacheret-Dujour, A. & S. Kahane. 2020. Unités syntaxiques et unités intonatives majeures en français parlé: inclusion, fragmentation, chevauchement [Syntactic units and major intonation units in spoken French: inclusion, fragmentation, overlap]. In: F. Neveu, B. Harmegnies, L. Hriba, S. Prévost & A. Steuckardt (eds.), 7e Congrès Mondial de Linguistique Française, July 6–10, 2020, Montpellier, France, Article 14005. https://doi.org/10.1051/shsconf/20207814005

Lefeuvre, F. & E. Moline. 2011. Unités syntaxiques et unités prosodiques: Bilan des recherches actuelles [Syntactic and prosodic units: Review of current research]. *Langue française*, 170, 143–157. https://doi.org/10.3917/lf.170.0143

Muller, P., M. Vergez-Couret, L. Prévot, N. Asher, B. Farah, M. Bras & L. Vieu. 2012. Manuel d'annotation en relations de discours du projet ANNODIS [Annotation manual in discourse relations of the ANNODIS project]. *Carnets de Grammaire* 21, 1–34.

Pallaud, B., R. Bertrand, L. Prevot, P. Blache & S. Rauzy. 2019. Suspensive and Disfluent Self Interruptions in French Language Interactions. In: L. Degand, G. Gilquin, L. Meurant, A. C. Simon (eds.), *Fluency and Disfluency across Languages and Language Varieties*, Louvain-la-Neuve, Belgium: Presses universitaires de Louvain, 109–138.

Portes, C. & R. Bertrand. 2011. Permanence et variation des unités prosodiques dans le discours et l'interaction [Permanence and variation of prosodic units in speech and interaction]. *Journal of French Language Studies* 21(1), 97–110. https://doi.org/10.1017/S0959269510000499

Prévot, L., S.-C. Tseng, K. Peshkov & A. C. H. Chen. 2015. Processing units in conversation: A comparative study of French and Mandarin data. *Language and Linguistics*, *16*(1), 69–92. https://doi.org/10.1177/1606822X14556605

Selting, M. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, *6*(3), 371–388. https://doi.org/10.1075/prag.6.3.06sel

Shriberg, E. E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California at Berkeley.

Shriberg, E. E. 1995. Acoustic properties of disfluent repetitions. In: K. Elenius & P. Branderud (eds.), *Proceedings of the 13th International Congress of Phonetic Sciences*, August 13–19, 1995, Stockholm, Sweden, 4, 384–387.

Filled pauses in university lectures

Jessica Di Napoli
RWTH Aachen University, Aachen, Germany

Abstract

Previous studies have shown that filled pauses such as uh and um may provide cues to listeners to discourse structure and information structure. The present study employs a corpus-based approach to investigate to what extent filled pauses occur in this function in eight undergraduate lectures in American English. Results show that filled pauses occur most frequently in initial (i.e., post-pausal) position, and that they often cluster together following topic changes. Filled pauses are also shown to occur before important words in the corpus. Together, the results suggest that filled pauses in lectures may highlight important information and mark discourse structure at various levels. The findings contribute to gaining a better understanding of filled pause use across different registers and provide support of filled pauses as signals which benefit listeners.

Introduction

Filled pauses (such as English *uh* and *um*) are characteristic of online speech production across a number of different spoken language registers. Previous studies have shown that filled pause frequency and distribution vary not only across languages (de Leeuw, 2007; Crible, Degand, & Gilquin, 2017), language varieties (Tottie, 2014), and social factors (Fruehwald, 2016; Wieling et al., 2016), but also across speech situations (Schachter et al., 1991; Shriberg, 1994; Tottie, 2014). The present study explores the characteristics of filled pauses in undergraduate university lectures.

A study by Schachter et al. (1991) suggests that filled pause frequency in lectures is, to some extent, dependent on the academic discipline and the degree of choice implied in its respective working vocabulary. This is consistent with filled pauses as a *symptom* of cognitive processing; however, filled pauses can also serve as a *signal* to the listener (see de Leeuw, 2007 for an overview). It is the latter which is the primary focus of this paper.

Of particular interest to the present study, filled pauses have been shown to serve as potential cues to listeners of discourse and information structure. For example, in Dutch, filled pauses are more frequent at major discourse boundaries than minor discourse boundaries (Swerts, 1998). In English, filled pauses are more likely to occur in sentence-initial or utterance-initial position than in medial position (Beattie, 1979; Shriberg, 1994; Eklund & Shriberg, 1998; O'Connell & Kowal, 2005; de Leeuw, 2007). The frequent occurrence of filled pauses here can be attributed to the higher cognitive load associated with planning larger syntactic and discourse units (Beattie, 1979; Shriberg, 1994; Eklund & Shriberg, 1998). Listeners, in turn, can exploit these observed regularities (see Finlayson & Corley, 2012; see also Corley & Stewart, 2008) to extract information about the structure of ongoing discourse. Filled pauses can therefore be posited as cues to discourse units of various sizes.

In fact, listeners positively evaluate speakers who produce filled pauses at clause and discourse boundaries (Fischer & Schümchen, 2019). In addition, filled pauses aligned with clause boundaries have been shown to have an influence on listeners' grammaticality judgements of sentences and on their interpretation of the syntactic structure of sentences (Bailey & Ferreiera, 2003; Rose, 2019). This suggests that filled pauses as markers of discourse structure are beneficial to listeners.

Other studies on perception have shown that filled pauses can lead listeners to expect discourse-new, rather than given referents (Arnold, Fagnano, & Tanenhaus, 2003) and low frequency, rather than high frequency words (Corley, MacGregor, & Donaldson, 2007; see also Beattie & Butterworth, 1979 for production). Words preceded by filled pauses have also been shown to be more likely remembered by listeners (Corley et al., 2007). Additionally, filled pauses can mark important words and information, helping listeners to respond correctly to comprehension questions (Fischer & Schümchen, 2019). Above the word level, filled pauses have been shown to facilitate recall of discourse (Fraundorf & Watson, 2011).

Given these potential benefits to listeners, the present study investigates the frequency and positioning of filled pauses in a pedagogical context. Specifically, using a corpus-based approach, the study investigates whether undergraduate lectures in American English display evidence of filled pauses signaling discourse structure and marking important words. The analysis focuses on the co-occurrence of filled pauses with silent pauses and surrounding words, as well as the relative distributions of *uh* and *um*.

https://doi.org/10.18463/diss-2021-008-dinapoli

Method

The speech material for the present study was obtained from *The Michigan Corpus of Academic Speech* (MICASE), a corpus of academic speech recorded at the University of Michigan in the late 1990's and early 2000's (Simpson et al., 1999). In particular, I examined a subset of the lecture part of the corpus featuring large (>40 students), highly monologic, undergraduate lectures given by native speakers of American English. Eight lectures, across four academic divisions, matched these criteria. Details regarding the lectures contained in the corpus for this study are presented in Table 1. In total, the corpus contains approximately 505 minutes of speech and 86,730 words.

Corpus analysis methods were adopted to investigate the frequency and positioning of filled pauses across the corpus. Queries were performed in CQPweb (Hardie, 2012) to search for instances of *uh* and *um* and to calculate their respective raw and relative frequencies (per 1000 words and per minute of lecture) for each lecture. I also queried for the co-occurrence of *uh* and *um* with silent pauses, as perceived and transcribed by the transcribers of the corpus. The MICASE transcriptions distinguish between four types of silent pause (see Simpson, Lee, & Leicher, 2002: 1) a brief mid-utterance pause accompanied by a non-phrase-final intonation contour; 2) a brief pause accompanied by an utterance-final intonation contour; 3) a longer pause of two to three seconds; and 4) a long pause (four seconds or longer). All query results were verified manually.

I subsequently categorized each occurrence of *uh* and *um* as occurring in one of four positions (cf. O'Connell & Kowal, 2005; de Leeuw, 2007): 1) *initial* (preceded by a silent pause, followed by speech); 2) *medial* (preceded and followed by speech); 3) *final* (preceded by speech, followed by a silent pause); and 4) *isolated* (preceded and followed by a silent pause). Silent pauses preceding and following filled pauses were then classified according to type of pause (brief mid-utterance pause, brief utterance-final pause, and long pause).

Sentences and clauses are not transcribed in MICASE; punctuation marks indicate prosodic structure as signaled by pauses and intonation. In order to evaluate the position of filled pauses with respect to syntactic structure, I performed, as a proxy measure, queries searching for the frequency of the different parts of speech of words immediately preceding and following *uh* and *um*. For each query, a distinction was made between sequences with no intervening silent pause, those with a brief, mid-utterance silent pause, and those with a brief utterance-final or a long silent pause.

Next, I performed a cluster analysis to examine which specific words frequently occurred following *uh* and *um*, with respect to which words were most frequent across the corpus as a whole, and for texts individually. Filled pauses in both initial and medial position, where they are immediately followed by speech, and sequences containing a brief, mid-utterance pause directly after the filled pause were included in the analysis. Sequences containing long pauses and utterance-final pauses were excluded. For each individual text, I examined how often the six most frequent content words (tokens), taken to represent important words in the lectures after observing a clear link between the words and the topics of the lectures (see Table 1), were produced with a preceding filled pause.

Table 1. Relative frequency of filled pauses and percentage of um by lecture, together with key lecture characteristics including academic discipline (BS = Biological and Health Sciences, HA = Humanities and Arts, PS = Physical Sciences, SS = Social Sciences), topic, text ID, gender and age of speaker, and the six most frequent content words.

Discipline/Topic	Text ID	Gender	Age	Six most frequent content words (tokens) *Words in italics were preceded by a filled pause*	FPs/1000 wds	FPs/min	% *um*
BS / Cancer	LEL175SU106	M	51+	cells, *cancer*, cell, metastasize, immune, metastasis	5.4	0.97	8.8
BS / Drugs	LEL500SU088	M	31–50	*serotonin*, receptor, receptors, effects, hallucinatory, give	7.7	1.39	5.3
BS / Evolution	LEL175JU154	F	31–50	things, population, time, natural, selection, organisms	11.2	1.52	98.0
HA / Ancient Rome	LEL215SU150	M	31–50	*Roman, Caesar, people, Augustus, Rome, Sulla*	31.4	6.37	27.4
HA / Art history	LEL320JU143	F	31–50	*people*, painting, modern, *Manet*, artist, time	19.4	3.46	95.3
PS / Chemistry	LEL200JU105	F	51+	*family*, get, ion, ions, reaction, charge	8.1	1.23	1.6
PS / Physics	LEL485JU097	M	51+	light, time, moving, hundred, meters, frame	17.6	3.08	17.0
SS / Psychology	LEL500JU034	M	31–50	evolution, evolutionary, right, *Darwin*, *behavior*, see	11.2	2.07	30.6
mean across lectures					14.0	2.51	35.5

Additionally, I classified all content words which occurred in a cluster with a preceding filled pause according to their overall frequencies in each text. Relative frequencies were determined as a percentage based on the ranking of the word in the word list for the text, divided by the total number of word types in the text. Words were then classified into one of three categories: 1) high frequency (0–20%); 2) medium frequency (20–50%); and 3) low frequency (>50%). For the purposes of the present study, these categories were determined on the basis of absolute word frequencies across the texts. High frequency words with a relative frequency of up to 20 percent tended to occur at least five times in a text. Medium frequency words occurred, on average, between two and four times in a text, while low frequency words tended to occur only once.

Finally, using AntConc (Anthony, 2019), I also looked for evidence of filled pauses clustering around major discourse boundaries through the use of concordance plots. This relied on first identifying clusters of filled pauses in the concordance plot for each text, visible as dark bands in the plot, and then analyzing the discourse structure of the surrounding text. If a shift in topic could be identified, this was marked as a major discourse boundary, and the position and frequency of filled pauses with respect to the boundary were then annotated.

Results

The eight university lectures investigated contain a total of 1251 filled pauses, with 747 occurrences of *uh* and 504 occurrences of *um*. The relative frequency of filled pauses per lecture, together with defining characteristics of the lectures (speaker age and gender, academic discipline, etc.), is presented in Table 1. As can be seen in the table, there is a high degree of variation between lectures, with the lecture on cancer displaying the lowest frequency of filled pauses (0.97 FPs/min), and the lecture on ancient Rome by far the highest (6.37 FPs/min). There is some variation across academic disciplines, with lectures in the humanities displaying the highest rates of occurrence of filled pauses, and lectures in the biological and health sciences the lowest. Across lectures, filled pauses occurred at a rate of 14 FPs/1000 words or 2.51 FPs/min.

Table 1 also shows the percentage of *um* produced by lecturers (with respect to the total number of filled pauses they produced). Overall, *uh* was much more frequent than *um* in the corpus, however, individual lectures varied regarding the relative frequencies of the two filled pause forms, in particular with respect to the age and gender of the speaker. As is clear in the table, while the two younger female speakers use *um* almost exclusively (on average, 97% of the time), older speakers, and male speakers, use *uh* much more frequently than *um*. The predominance of male speakers in the corpus (5M, 3F) could account for the higher frequency of *uh* across lectures.

Turning now to the co-occurrence of filled pauses with silent pauses in the corpus, results (see Table 2) show that both *uh* and *um* occur frequently with silent pauses. Both filled pause forms, and especially *uh*, occur most frequently in initial position, where they are preceded by a silent pause and followed by speech. The second most frequent position overall, and in particular for *um*, is isolated, where the filled pause is preceded and followed by a silent pause. Overall, 87 percent of all filled pauses in the corpus occur together with a silent pause, most often a *preceding* pause.

The frequency of the different pause types co-occurring with filled pauses varied according to whether the silent pause preceded or followed the filled pause, and according to filled pause form. Results are summarized in Table 3. Overall, filled pauses occur most frequently with brief, mid-utterance and utterance-final pauses. Co-occurrence with longer pauses is relatively rare, both preceding and following the filled pause. Pauses preceding *uh* are primarily brief, utterance-medial pauses, followed by brief, utterance-final pauses, while for *um*, the opposite tendency appears. Silent pauses preceding *um* are primarily brief, utterance-final pauses. Silent pauses following *uh* and *um* are overwhelmingly brief, mid-utterance pauses.

Table 2. Filled pauses according to position with respect to co-occurring silent pauses. Percentages are given for filled pause types separately and pooled across the corpus (overall).

		FP Position (%)			
FP	N	Initial	Medial	Final	Isolated
uh	747	59.2	14.7	11.4	14.7
um	504	47.3	10.1	9.3	33.3
all	1251	54.4	12.9	10.5	22.2

*Table 3. Silent pauses preceding vs. following co-occurring filled pauses according to type (**brief** mid-utterance, brief **utterance-final**, and **long**). Percentages are given for filled pause types separately and pooled across the corpus (overall).*

	Preceding FP (%)			**Following FP (%)**		
FP	brief	utt-final	long	brief	utt-final	long
uh	64.5	35.0	0.5	98.0	0.0	2.1
um	37.1	60.4	2.5	98.6	0.0	1.4
all	52.9	45.8	1.4	98.3	0.0	1.7

With respect to the most frequent parts of speech of words immediately preceding and following *uh* and *um*, possibly with an intervening silent pause, the results (see Table 4) show different patterns, depending on the position of the word with respect to the filled pause. For words *preceding* filled pauses, nouns are the most frequent word class, followed by verbs and adverbs. Silent pauses occur frequently between these word classes and the following filled pause, between 70 and 86 percent of the time. These pauses include both brief, mid-utterance pauses, and longer and utterance-final pauses. In contrast, words *following* filled pauses are most frequently conjunctions, followed by pronouns and nouns. Where these word classes follow a filled pause, they occur less frequently with an intervening perceived pause (only approximately 30% of the time), realized almost exclusively as a brief, mid-utterance pause.

Table 4. Most frequent parts of speech of words immediately preceding and following filled pauses according to their relative frequency (percentage of total), together with percentage of time they co-occur with an intervening silent pause.

	POS	% FPs	% no pause	% brief pause	% utt-fin / long pause
Before FP (*the uh*)	Noun	46.2	12.2	38.9	49.0
	Verb	12.7	29.6	48.4	22.0
	Adv	12.2	18.4	45.4	36.2
After FP (*uh the*)	Conj	24.7	70.8	29.2	0.0
	Pro	12.9	67.1	32.3	0.6
	Noun	11.8	70.1	29.9	0.0

Continuing to focus on the context of words following filled pauses, the cluster analysis showed that, for the most part, individual words which were very frequent in the corpus overall were also frequent after filled pauses. The most frequently occurring clusters were: *uh/m and*, *uh/m the*, and *uh/m he*. Some words, however, ranked much higher in terms of frequency following a filled pause than across the corpus generally. This includes nouns such as *Darwin*, *Caesar*, *tryptophan*, and *photography*, and adjectives such as *nonrandom*. These words appear to be directly related to the specific topics covered in the lectures (see Table 1).

The additional investigation by text found that of the six most frequent content words in each text (see Table 1), lecturers produced filled pauses before an average of 1.6 of these words (range = 0–6, median = 1). Six of the eight lecturers produced at least one of the most frequent words together with a preceding filled pause at least once. Interestingly, several lecturers produced filled pauses before multiple occurrences of these frequent words, up to a maximum of four times.

There was a clear tendency for content words occurring in a two-word cluster with a preceding filled pause to be high or medium frequency words. Across lectures, an average of 40.0 percent of content words preceded by filled pauses were high frequency words (mean relative frequency = 8.4%), and an average of 38.3 percent were medium frequency words (mean rel. freq. = 33.0%). Filled pauses occurred somewhat less frequently preceding low frequency content words, with low frequency words occurring in, on average, 21.9 percent of clusters (mean rel. freq. = 66.1%). Individual lecturers predominantly displayed the same tendency, producing filled pauses primarily before high and medium frequency content words. There was, however, one lecture which displayed the opposite tendency, with filled pauses occurring slightly more often before low frequency words.

Examination of concordance plots revealed additional patterns regarding filled pause use over the course of a lecture. For all lectures, regardless of whether or not the lecturer produced filled pauses frequently, there is evidence of clustering of filled pauses around changes in topic. In these portions of the texts, filled pauses occur at a rate much higher than in surrounding portions, typically with at least one filled pause occurring in each utterance just after the topic change. An example from the science lecture on evolution is presented in Figure 1. Here there are dark bands visible where filled pauses cluster together in proximity to the topic changes marked and described in the figure.

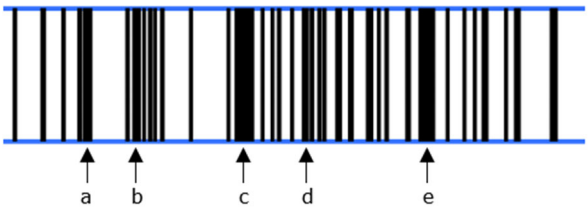

Figure 1. Excerpt of concordance plot from lecture on evolution (LEL175JU154) showing clusters of filled pauses at discourse boundaries. Arrows mark topic shifts co-occurring with filled pauses: a) ways of thinking before Darwin (introducing Thales); b) essentialism and creationism combine to form natural theology; c) illustrative example from Lamarck; d) introducing Cuvier; e) Darwin after his return from Galapagos.

Discussion and conclusion

The present study investigated the characteristics of filled pauses in American English undergraduate lectures. The main goal of the study was to determine whether lectures show evidence of filled pauses potentially functioning as signals to the listener (in

this case, undergraduate students) of discourse structure and important words (see Swerts, 1998; Fischer & Schümchen, 2019). Although the study is based on a relatively small sample of lectures, initial results offer some important insights and directions for future research.

Filled pauses occurred across lectures at a rate of 14 FPs/1000 words or 2.51 FPs/min, which is consistent with previous studies on filled pause frequency in university lectures (see Schachter et al., 1991) and classroom lessons (see Crible et al., 2017) in English. Also consistent with previous findings (see Schachter et al., 1991), some variation in overall filled pause frequency was observed across academic disciplines. Additionally, the relative frequency of *uh* vs. *um* was found to vary according to the age and gender of the speaker. These results are consistent with a language change in progress in Germanic languages, including English, led by younger female speakers, where *um* is increasingly replacing *uh* (Fruehwald, 2016; Wieling et al., 2016).

Filled pauses were found to co-occur frequently with silent pauses (in 87 percent of cases). The majority of co-occurring pauses *precede* the filled pauses, and both *uh* and *um* occur most frequently in initial position. This is consistent with de Leeuw (2007) and O'Connell and Kowal (2005), but stands in contrast to Clark and Fox Tree (2002), perhaps because of the focus on conversational speech in the latter study. There is no evidence from the present study that suggests that *uh* and *um* signal upcoming delays; rather, their frequent occurrence in initial position suggests that they signal prosodic, and possibly also syntactic, structure.

Interestingly, the present study does suggest a potential difference in the relative positioning of *uh* and *um*. *Um* was more frequent in isolated position and occurred more often at utterance boundaries, preceded by an utterance-final pause, while *uh* occurred more often at utterance-internal boundaries, preceded by a mid-utterance pause. This is in line with Shriberg (1994) who proposed that *um* is more characteristic of the global planning of larger units, while *uh* may be linked to more local lexical choice. Analysis of more lectures is needed to confirm this observed tendency.

Further support for filled pauses as signaling prosodic and syntactic structure in lectures comes from the analysis of the parts of speech of words surrounding filled pauses. The results show that filled pauses are most frequently preceded by nouns, but that in this context, there is a frequent occurrence of an intervening silent pause. This suggests that a prosodic and/or syntactic boundary often intervenes between the noun and the following filled pause (see Grice & Baumann, 2007). In contrast, filled pauses are most often followed by conjunctions, most typically without an intervening silent pause. This in line with filled pauses occurring in clause-initial (see Shriberg, 1994) and prosodic phrase-initial position. Current ongoing syntactic and prosodic analysis aims to confirm this.

The study also found evidence of filled pauses signaling higher-level discourse structure in the lectures. Filled pauses were observed to occur more frequently and in closer proximity to one another in correspondence to a change in topic, consistent with previous findings for Dutch (see Swerts, 1998).

Additionally, filled pauses were found to occur before important, high and medium frequency words in the lectures (see Fischer & Schümchen, 2019). This appears to contrast with previous research that has shown a frequent co-occurrence of filled pauses with *low* frequency words (Beattie & Butterworth, 1979; Corley et al., 2007). One potential reason for this could be the system used to classify words in terms of frequency in the present study. High frequency words in this study are not necessarily high frequency words in English overall, but only in the context of the individual lectures investigated.

In conclusion, the present study has implications for research on filled pause use in different communicative situations (Tottie, 2014), in particular pedagogical situations, and provides initial evidence of filled pauses as marking discourse structure and important information in university lectures.

Acknowledgments

I would like to thank two anonymous reviewers for their helpful comments on an earlier version of this paper.

References

Anthony, L. 2019. AntConc (version 3.5.8). https://www.laurenceanthony.net/software. (accessed 18 February 2019).

Arnold, J. E., M. Fagnano, & M. K. Tanenhaus. 2003. Disfluencies Signal Theee, Um, New Information. *Journal of Psycholinguistic Research* 32(1), 25–36. https://doi.org/10.1023/A:1021980931292

Bailey, K. G. D. & F. Ferreira. 2003. Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language* 49, 183–200. https://doi.org/10.1016/S0749-596X(03)00027-5

Beattie, G. W. 1979. Planning units in spontaneous speech: some evidence from hesitation in speech and speaker gaze direction in conversation. *Linguistics* 17, 61–78. https://doi.org/10.1515/ling.1979.17.1-2.61

Beattie, G. W. & B. L. Butterworth. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22(3), 201–211. https://doi.org/10.1177/002383097902200301

Clark, H. H. & J. E. Fox Tree. 2002. Using Uh and Um in Spontaneous Speaking. *Cognition* 84(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Corley, M., L. J. MacGregor, & D. I. Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105(3), 658–668. https://doi.org/10.1016/j.cognition.2006.10.010

Corley, M. & O. W. Stewart. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*. *Language and Linguistics Compass* 2(4), 589–602. https://doi.org/10.1111/j.1749-818X.2008.00068.x

Crible, L., L. Degand, & G. Gilquin. 2017. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast* 17(1), 69–95. https://doi.org/10.1075/lic.17.1.04cri

de Leeuw, E. 2007. Hesitation Markers in English, German, and Dutch. *Journal of Germanic Linguistics* 19(2), 85–114. https://doi.org/10.1017/S1470542707000049

Eklund, R. & E. Shriberg. 1998. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogs. In: R. H. Mannell & J. Robert-Ribes (eds.), *Proceedings of the 5th International Conference on Spoken Language Processing*, 30 November – 4 December, 1998, Sydney, Australia, vol. 6, 2627–2630.

Finlayson, I. R. & M. Corley. 2012. Disfluency in dialogue: an intentional signal from the speaker? *Psychonomic Bulletin & Review* 19, 921–928. https://doi.org/10.3758/s13423-012-0279-x

Fischer, K. & N. Schümchen. 2019. Hesitation markers and audience design: Position matters. In: *Proceedings of the 1st International Seminar on the Foundations of Speech – Breathing, Pausing and the Voice*, 1–3 December, 2019, Sønderborg, Denmark, 19–20.

Fraundorf, S. H. & D. G. Watson. 2011. The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language* 65, 161–175. https://doi.org/10.1016/j.jml.2011.03.004

Fruehwald, J. 2016. Filled Pause Choice as a Sociolinguistic Variable. *Penn Working Papers in Linguistics* 22(2), 41–49.

Grice, M. & S. Baumann. 2007. An introduction to intonation – functions and models. In: J. Trouvain & U. Gut (eds.), *Non-Native Prosody: Phonetic Description and Teaching Practice*, Berlin: De Gruyter Mouton, 25–51. https://doi.org/10.1515/9783110198751.1.25

Hardie, A. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

O'Connell, D. C. & S. Kowal. 2005. *Uh* and *Um* Revisited: Are They Interjections for Signaling Delay? *Journal of Psycholinguistic Research* 34(6), 555–576. https://doi.org/10.1007/s10936-005-9164-3

Rose, R. 2019. The structural signaling effect of silent and filled pauses. In: R. L. Rose & R. Eklund (eds.), *Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech*, 12–13 September, 2019, Budapest, Hungary, 19–22. https://doi.org/10.21862/diss-09-006-rose

Schachter, S., N. Christenfeld, B. Ravina, & F. Bilous. 1991. Speech Disfluency and the Structure of Knowledge. *Journal of Personality and Social Psychology* 60(3), 362–367. https://doi.org/10.1037/0022-3514.60.3.362

Shriberg, E. E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California at Berkeley.

Simpson, R. C., S. L. Briggs, J. Ovens, & J. M. Swales. 1999. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI, USA: The Regents of the University of Michigan.

Simpson, R. C., D. Y. W. Lee, & S. Leicher. 2002. *MICASE Manual: The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI, USA: The Regents of the University of Michigan.

Swerts, M. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics* 30(4), 485–496. https://doi.org/10.1016/S0378-2166(98)00014-9

Tottie, G. 2014. On the use of *uh* and *um* in American English. *Functions of Language* 21(1), 6–29. https://doi.org/10.1075/fol.21.1.02tot

Wieling, M., J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, & M. Liberman. 2016. Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change* 6(2), 199–234. https://doi.org/10.1163/22105832-00602001

A crosslinguistic study on the interplay of fillers and silences

Simon Betz[1], Nataliya Bryhadyr[1], Loulou Kosmala[2] and Loredana Schettino[3]
[1] Faculty for Linguistics and Literary Studies & CITEC, Bielefeld University, Bielefeld, Germany
[2] Institut du Monde Anglophone, Sorbonne Nouvelle University – Paris 3, Paris, France
[3] Department of Humanities, University of Salerno, Salerno, Italy

Abstract

We present a crosslinguistic study on the interplay of hesitation silences and fillers in conversation. The research questions have been addressed for English in a previous DiSS workshop paper (Betz & Kosmala, 2019) and this study extends the analysis to German, Italian and French. The research questions are: 1) Does the type of the filler influence following silence duration 2) Does the duration of the filler correlate with silence duration 3) Does silence duration vary depending on its distance from filler. The analysis shows cross-linguistic similarities and differences, thus highlighting the role and the language- and culture-specific nature of disfluencies.

Introduction

Silences and fillers are a very common form of hesitation and disfluency and have been studied extensively in different languages, such as English (e.g. Clark & Fox Tree, 2002; Shriberg, 2001), German (e.g. Trouvain, Fauth, & Möbius, 2016), Italian, (e.g. Esposito et al., 2007) and French (e.g. Candea, 2000; Grosjean & Deschamps, 1972). Given their high frequency in spontaneous speech, these markers have also been subject to several investigations in multilingual studies, e.g. Campione and Véronis' (2002) large-scale study on silent pauses duration in English, French, German, and Italian, or de Leeuw's (2007) comparative study of filled pauses in German, English, and Dutch. These studies have demonstrated cross-linguistic differences, which gives support to Clark and Fox Tree's (2002) argument that fillers are language-specific. In line with this body of research, the present study aims to examine the duration and co-occurrence of fillers and silences in German, Italian, and French, based on a previous study on English (Betz & Kosmala, 2019).

Less is known about the clustering of silences and fillers specifically, and the way it may influence their duration. However, we can still find several studies that have investigated their co-occurrence in detail. For instance, Degand and Gilquin (2013) conducted a study on the clustering of fillers, discourse markers, silences, and other disfluency markers in English and in French. Their study showed that frequent clusters often included fillers and silences, with for instance "euh" followed by a short pause or a long pause in French, and "uhm" followed or preceded by a short pause in English. More recently, another study conducted by Grosman, Simon, and Degand (2018) examined the impact of syntax and speech genre on the frequency and duration of silences, based on a multi-genre corpus in French. More evidence further suggests that fillers and silences are often found in combination, both in English and in French (see Grosjean & Deschamps, 1972), and that the form of the filler in English (either produced with a central vowel *uh* or a nasal consonant *um*) may affect the duration of silences. For instance, Smith and Clark (1993) claimed that um-type fillers were typically followed by longer pauses than uh-type fillers because speakers intentionally chose between *uh* and *um* to signal word retrieving difficulties. This led to Clark and Fox Tree's (2002) assumption that *uh* signals a minor delay in speech, while *um* signals a major one. However, this hypothesis has also been challenged (see Finlayson & Corley, 2012).

Similar results are reported in Betz and Kosmala (2019) on semi-spontaneous English dialogues. The duration of silences was found to be longer when they occurred in an utterance with an um-type filler, as opposed to a uh-type. In addition, silences were found to be longer when they immediately co-occurred with *um*, but only in medial position (as opposed to initial position). Overall, their results corroborated Clark and Fox Tree (2002) and showed that longer fillers were associated with longer silences in the utterance. The distance between the two hesitations (either immediately adjacent, or further away in the utterance) was also found to influence the duration of silences, in the sense that silences tend to be longer in vicinity of fillers and longer following fillers, which sparked the assumption that fillers might ground hesitation in dialogue, after which longer silences are tolerable.

Following Betz and Kosmala (2019), we aim to extend our analysis of fillers and silences to other languages, mainly French, German, and Italian. We address the following research questions:

https://doi.org/10.18463/diss-2021-009-betz-etal

1. Does the type of the filler influence the duration of the following silence?
2. Does the duration of the filler correlate with silence duration?
3. Does silence duration vary depending on its distance from filler?

Corpus and Methods

As mentioned above, this study is based on earlier work conducted on English data (Betz & Kosmala, 2019), and the materials for the present work are taken from different corpora in three languages, German, French, and Italian.

The German data consists of 9 dyadic dialogues by 18 speakers of different gender, ca. 30 minutes each (DUEL corpus, Hough et al., 2016). The speakers had a task to furnish an imaginary apartment of 200 square meters with a fictional budget of 500,000€. Only dialogue was allowed as a tool to achieve this.

The French data is taken from the DisReg Corpus (Kosmala, 2020) which includes semi-spontaneous interactions of 12 native speakers of French, all students from the same university. They were asked to freely discuss various topics (e.g. funny anecdotes at university, last film seen on TV) in pairs. The duration of the selected sample is 30 minutes approximately (5 minutes per dialogue).

The Italian data consists of two task-oriented dialogues by four native speakers of Neapolitan Italian, ca. 15 minutes per couple (CLIPS Corpus, Savy & Cutugno, 2009). The interlocutors were given similar pictures and asked to perform a "spot the difference" task, during the interaction they were not allowed to see each other, so they could only rely on the verbal channel.

Following Betz and Kosmala (2019), we investigated the clustering of fillers and silences in the data, and we coded the form of fillers (uh/um), and the distance in words between silences and fillers: 0 for the first position after a filler (*filler* + *silence*), positive values for subsequent positions (e.g. *filler* + *word* + *silence*), −1 for the last position before the filler (*silence* + *filler*), and negative values for greater distance before a filler (e.g. *silence* + *word* + *filler*).

To answer research questions 1 and 3, we used Wilcox tests to compare mean durations as the data is not normally distributed. The correlations for question 2 are analyzed using the Spearman correlation test, which outputs a correlation coefficient *rho* and a significance value *p*. *Rho* denotes how strong the correlation is and in which direction. Note that dealing with spontaneous speech data frequently does not yield high correlation coefficients due to noise, variation etc., weak correlations can still be significant.

Results

German Data

962 silent pauses were identified: 76% without fillers in the vicinity, and 24% co-occurring with fillers. Among the latter, 8% precede and 30% follow *um*s (38% in total), whereas 22% precede and 38% follow *uhs* (60% in total).

The mean duration of silences following *um*s (692 ms) is significantly higher than the mean duration of silences following *uhs* (561 ms, $W = 4769.5$, $p = 0.004538$).

Silences' duration correlates significantly, though weakly with both *uh* duration (*rho* = 0.24, $p = 0.0003$) and *um* duration (*rho* = 0.21, $p = 0.04$).

In general, as reported in Figure 1, the post-filler silence duration is significantly l onger than the pre-filler silence duration ($W = 3864$, $p = 0.00075$). Further, Figure 1 shows that the peak in silence duration is at 0, standing for silences directly following the filler. In position −1 there is the secondary peak, hence also directly preceding fillers, silences are relatively long. The remaining silences are quite scattered, but with a higher range in the post-filler region.

French Data

339 silent pauses were extracted from the data, 64% without fillers in the vicinity, and 36% co-occurring with fillers. 1% of silences precede and 10% follow *um*s (11% in total), whereas 32% precede and 57% follow *uh*s (89% in total).

In French data, silences after *um*s (583 ms) are shorter than those after *uh*s (652 ms), though not significantly ($W = 771$, $p = 0.6069$).

As for the silences-filler correlation, silences are strongly and negatively correlated with ums (*rho* = -0.68, $p = 0.01$), but not significantly with *uh*s (*rho* = -0.08, $p = 0.43$). As Figure 2 shows, post-filler silences are on average shorter than the pre-filler ones. This difference is not significant ($W = 1856$, $p = 0.2399$). In effect, post-fillers silences show a higher range and outliers compared to the pre-filler ones. Looking at the plot of silence duration as a function of distance from filler (see Figure 2), we can see a peak at position 0. However, besides silences directly after fillers, no other post-filler silences show higher duration ranges.

Italian Data

Out of a total of 311 instances, just 23% of silent pauses co-occur with fillers. The latter were quite

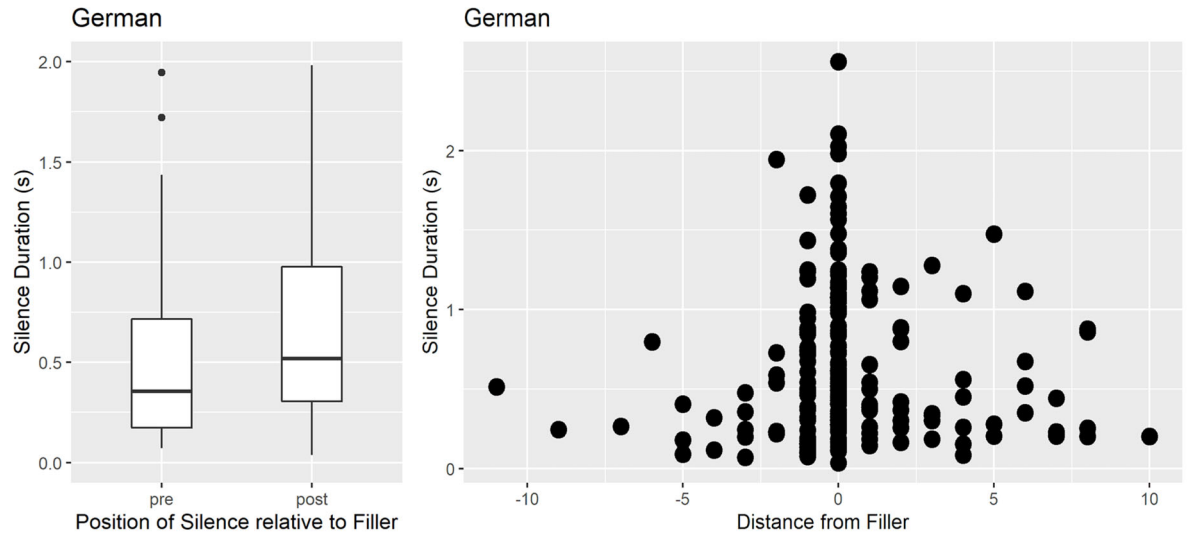

Figure 1. For German, silence duration pre-filler and post-filler (left) and silence duration for each distance-from-filler position (right).

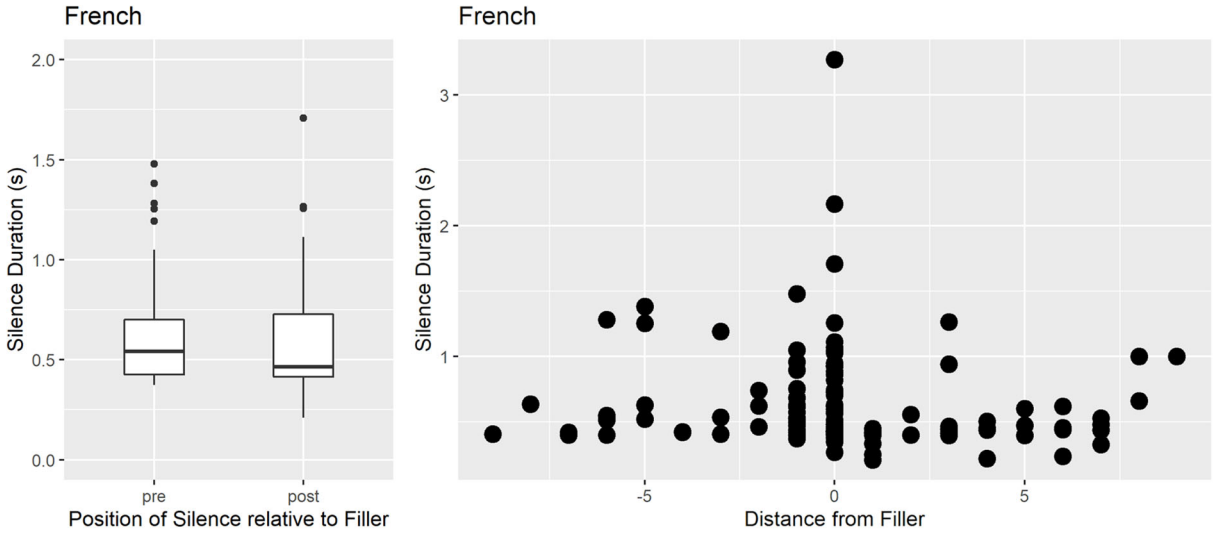

Figure 2. For French, silence duration pre-filler and post-filler (left) and silence duration for each distance-from-filler position (right).

evenly split between *uh*s and *um*s: 31% preceding and 18% following *um*s (49% in total); 29% preceding and 22% following *uh*s (51%) in total.

Silences occurring after *um* (405 ms) are on average longer than those after *uh* (388 ms). Though, not significantly ($W = 670.5$, $p = 0.7999$).

As for the correlation between filler duration and silence duration, it was found to be significant for *uh* (*rho* = 0.35, $p = 0.03$), though not significant for *um* (*rho* = 0.25, $p = 0.13$) co-occurrences.

In general, post-filler silences are longer than pre-filler ones (respectively, the mean duration is 479 ms and 340 ms, Figure 3). This difference is not significant by a very small margin ($W = 467.5$, $p = 0.0742$). Further, Figure 3 reveals that Italian lacks the peak at position 0, but it exhibits peaks at −1 and +1.

Clusters

The analyses exhibited cross-linguistic differences and visual inspection of the plots suggested a tendency for high silence duration in clusters with fillers. We thus conducted an exploratory post-hoc analysis to examine whether (1) silences in position 0 are longer than silences in other positions and (2) whether silences on positions 0 and −1 combined are longer than silences in other positions. Wilcox tests were again used to compare the means and Table 1 summarizes the results.

Discussion

From our data emerge different tolerances for silence duration in each language, Italian showing on average shorter silent pauses than German and

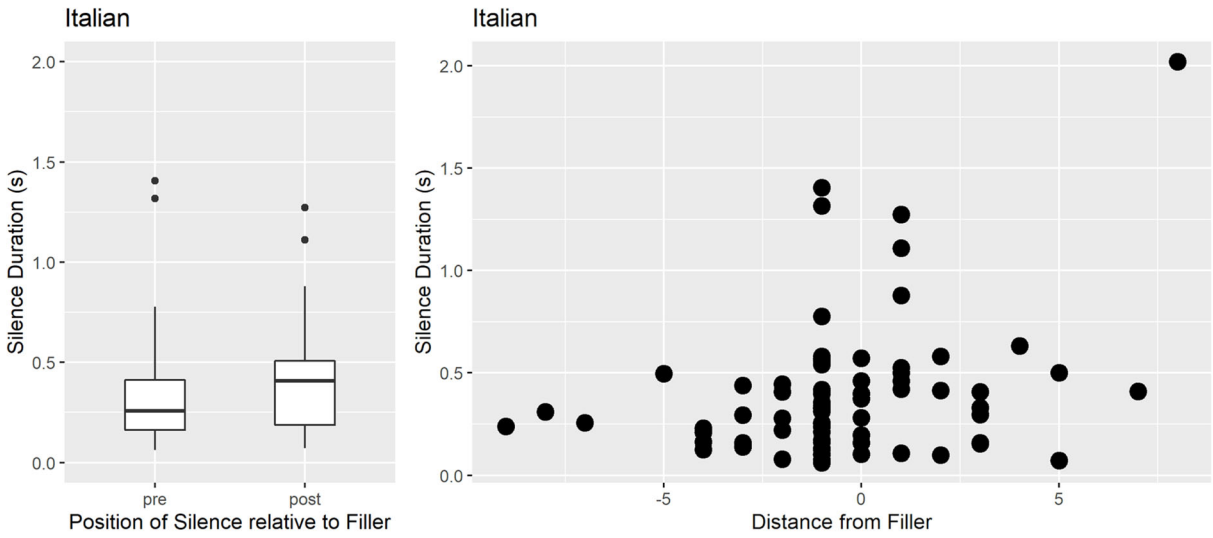

Figure 3. For Italian, silence duration pre-filler and post-filler (left) and silence duration for each distance-from-filler position (right).

Table 1. P *values from the comparison of silences in position 0 (I col.) and in both position 0 and −1 (II col.) with the other silences per each language*

	Pos 0 vs rest	**Pos 0,−1 vs rest**
German	0.0001 ***	<0.0001 ***
English	<0.0001 ***	0.04981 *
French	0.0993 .	0.007417 *
Italian	0.7881	0.07672 .

French. Besides this relative tolerance, German results seem to confirm the trends attested for English (Betz & Kosmala, 2019), whereas French and Italian present different tendencies.

Firstly, in German, silence duration varies due to the preceding filler type, being systematically longer after fillers with nasalization. So, as observed in English, *um*s could be interpreted as signals for major delay in speech as opposed to *uh*s, signaling a minor one (Clark & Fox Tree, 2002). The same cannot be claimed for the Italian and French data, showing non-systematic tendencies.

Similarly, in German, longer fillers are regularly followed by longer silences, whereas in Italian and French only a partially systematic correlation could be observed.

Lastly, it is worth noticing the duration of silences with directly neighboring fillers. As Table 1 shows, there is a slight trend for durations to increase in such clusters across languages, but this picture is only clear in German and English.

Overall, in German, there seems to be the same increased tolerance for silence duration after fillers as observed for English. For French and Italian, this cannot be safely stated. It is up for future research to explore this issue further. The dataset we had at hand for French and Italian was rather small, so we do not want to over-interpret these tendencies. However, it might be the case that the idea that fillers ground hesitations in dialogue which has been formulated in Betz and Kosmala (2019), and which relates to Jefferson's (1988) concept of standard maximum silence, has to be reconsidered. It is possible that the idea of fillers as the most salient hesitation is based on the Germanic languages perspective. It might well be that in French or Italian, silences, and not fillers, are the most salient disfluency and that the grounding is inverse compared to German and English. The idea of "fillers increase silence tolerance" might be formulated more universally as "salient hesitations increase tolerance for following less salient hesitations".

Conclusion

The present contrastive study investigates the clustering of silences and fillers and its effect on silences' duration. It sheds light on the cross-linguistic "proximity effect" and the different tolerance for silences. The results confirm the language-specific character of disfluencies (see Clark & Fox Tree, 2002) and highlight that hesitation markers and their interplay respond not only to phonological, syntactic, and semantic constraints (Ginzburg, Fernández, & Schlangen, 2014) but also to pragmatic culture-specific dynamics regarding individual languages, which is relevant to consider when modelling the occurrence of disfluencies in conversation, e.g. for technological applications.

Future investigations could aim to extend observation to a larger dataset for Italian and French

and consider silences and fillers interplay with other hesitation types they could cluster with, such as lengthenings and lexical fillers.

Notes

[1] Authors appear in alphabetical order. Responsibilities: Simon Betz—German data, data analysis, *Results* section; Loulou Kosmala—French data, *Introduction* and *Corpus and Methods* sections; Loredana Schettino—Italian data, *Discussion* and *Conclusion* sections. Nataliya Bryhadyr—data preparation and assistance.

References

Betz, S. & L. Kosmala. 2019. Fill the silence! Basics for modeling hesitation. In: R. L. Rose & R. Eklund (eds.), *Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech*, 12–13 September, 2019, Budapest, Hungary, 11–14. https://doi.org/10.21862/diss-09-004-betz-kosm

Campione, E. & J. Véronis. 2002. A large-scale multilingual study of silent pause duration. In: *Speech Prosody 2002*, 11–13 April, 2002, Aix-en-Provence, France, 199–202.

Candea, M. 2000. Contribution à l'étude des pauses silencieuses et des phénomènes dits “d'hésitation” en français oral spontané. Etude sur un corpus de récits en classe de français. [Contribution to a study of silent pauses and “hesitation” phenomena in spontaneous French speech. A corpus study of French oral narratives in class]. PhD dissertation, Université de la Sorbonne nouvelle – Paris III.

Clark, H. H. & J. E. Fox Tree. 2002. Using Uh and Um in Spontaneous Speaking. *Cognition* 84 (1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Degand, L. & G. Gilquin. 2013. The clustering of ‘fluencemes’ in French and English. Presentation at *7th International Contrastive Linguistics Conference (ICLC 7) – 3rd Conference on Using Corpora in Contrastive and Translation Studies (UCCTS)*, July 10–13, 2013, Ghent, Belgium.

de Leeuw, E. 2007. Hesitation markers in English, German, and Dutch. *Journal of Germanic Linguistics*, 19(2), 85–114. http://dx.doi.org/10.1017/S1470542707000049

Esposito, A., V. Stejskal, Z. Smékal, & N. Bourbakis. 2007. The significance of empty speech pauses: Cognitive and algorithmic issues. In: F. Mele, G. Ramella, S. Santillo, & F. Ventriglia (eds.), *International Symposium on Brain, Vision, and Artificial Intelligence: Advances in Brain, Vision, and Artificial Intelligence,* Berlin, Germany: Springer, 542–554. https://doi.org/10.1007/978-3-540-75555-5_52

Finlayson, I. R. & M. Corley. 2012. Disfluency in Dialogue: An Intentional Signal from the Speaker? *Psychonomic Bulletin & Review* 19(5), 921–28. https://doi.org/10.3758/s13423-012-0279-x

Ginzburg, J., R. Fernández & D. Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics* 7(9), 1–64. https://doi.org/10.3765/sp.7.9

Grosjean, F. & A. Deschamps. 1972. Analyse Des Variables Temporelles Du Français Spontané. [Analysis of temporal variables in spontaneous French]. *Phonetica* 26(3), 129–56. https://doi.org/10.1159/000259407

Grosman, I., A. C. Simon, & L. Degand. 2018. Variation de la durée des pauses silencieuses: impact de la syntaxe, du style de parole et des disfluences. [Variation in the duration of silent pauses: impact of syntax, speech style and disfluencies]. *Langages* 211, 13–40. https://doi.org/10.3917/lang.211.0013

Hough, J., Y. Tian, L. de Ruiter, S. Betz, D. Schlangen, & J. Ginzburg. 2016. DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. In: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (eds.), *Proceedings of LREC 2016, 10th International Conference on Language Resources and Evaluation*, 23–28 May, 2016, Portorož, Slovenia, 1784–1788.

Jefferson, G. 1988. Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In: D. Roger & P. Bull (eds.) *Conversation: An interdisciplinary perspective*, Clevedon, UK: Multilingual Matters, 166–196.

Kosmala, L. 2020. Euh le saviez vous? Le rôle des (dis)fluences en contexte interactionnel: étude exploratoire et qualitative. [Uh did you know? The role of (dis)fluencies in interactional contexts.] In: F. Neveu, B. Harmegnies, L. Hriba, S. Prévost & A. Steuckardt (eds.), *7e Congrès Mondial de Linguistique Française*, July 6–10, 2020, Montpellier, France, Article 01018. https://doi.org/10.1051/shsconf/20207801018

Savy, R. & F. Cutugno. 2009. CLIPS: Diatopic, diamesic and diaphasic variations in spoken italian. In: M. Mahlberg, V. González-Díaz, C. Smith (eds.), *Proceedings of the Fifth Corpus Linguistics Conference*, July 20–23, 2009, Liverpool, UK, Article 213.

Shriberg, E. 2001. To 'errrr 'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1), 153–169. https://doi.org/10.1017/S0025100301001128

Smith, V. L. & H. H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language* 32, 25–38. https://doi.org/10.1006/jmla.1993.1002

Trouvain, J., C. Fauth, & B. Möbius. 2016. Breath and non-breath pauses in fluent and disfluent phases of German and French L1 and L2 read speech. In: J. Barnes, A. Brugos, S. Shattuck-Hufnagel & N. Veilleux (eds.), *Proceedings of Speech Prosody 2016 (SP2016)*, 31 May–3 June, 2016, Boston, MA, USA, 31–35. https://doi.org/10.21437/speechprosody.2016-7

The acoustic characteristics of *um* and *uh* in spontaneous Canadian English

Gabrielle Morin and Benjamin V. Tucker
University of Alberta, Edmonton, Alberta, Canada

Abstract

The present study investigates and compares the acoustic characteristics of uh *[ə] and* um *[əm] spontaneous speech. The data comes from a corpus of Western Canadian conversational spontaneous speech. Measures of duration, fundamental frequency, F1 and F2 were extracted from 1,048 instances of* um *and* uh*. Results indicate that longer durations occurred when markers preceded silent pauses.* Um *was found to have higher F1 and lower F2 than* uh*. F0 was overall lower for* um *in comparison to* uh*. These results provide a preliminary understanding of* um *and* uh *as markers in spontaneous Canadian English. Canadian English shows a similar proportion of* um *over* uh *usage in comparison to American and British English. Findings on vowel duration show no significant difference between* um *and* uh*. Differences in f0, F1 and F2 provide additional indication of how* um *and* uh *are different.*

Introduction

Um [əm] and *uh* [ə] have been reported to be among the most frequently observed disfluencies in spontaneous speech (Shriberg, 2001). We follow Le Grézause (2017) and classify *um* and *uh* as markers as opposed to fillers and filled pauses. In the present study, we investigate the acoustic characteristics of *um* and *uh* in Canadian English.

There are considerable differences across languages with regard to the frequency of occurrence of these markers. Over the past five decades, there has been an increase in *um* occurrences across British and American English dialects while *uh* has significantly decreased (Wieling et al., 2016). In their analysis of multiple spoken language corpora, Wieling et al. (2016) found that the proportion of *um* over *uh* increased from 0.3 to around 0.5 for female speakers of American English and British English as of 2013. They also found that the frequency of *um* occurrence relative to all other words has been consistently increasing in American English. Their results show a significant relationship between age and frequency of occurrence of *um* in all four American and British English corpora, demonstrating the tendency for younger generations to use *um* more than older generations (Wieling et al., 2016). However, Horváth (2010) found a greater usage of *uh* in comparison to *um* in Hungarian spontaneous speech. There is very little research with regard to Canadian English and the occurrence of *um* and *uh* as markers. Part of this may be because Canadian English is often combined with American and/or British English dialects rather than being examined individually. Canadian English is also interesting because it has strong historical influences from British English and currently remains in close contact with American English (Boberg, 2010).

Previous work has shown that the duration of *um* is consistently greater than *uh*, likely because it is composed of two phonemes rather than one (Clark & Fox Tree, 2002; Swerts, 1998). However, data analyzing the vowel duration alone has found that *um* is shorter than *uh* (Hughes, Wood, & Foulkes, 2016). The duration of these markers plays an important role in the surrounding environments. Markers have been categorized into major (*um*) or minor (*uh*) delays depending on their following silence, where *um* tends to precede longer pauses than *uh* (Clark & Fox Tree, 2002; Swerts, 1998).

Fundamental frequency (f0) is another phonetic property that has the potential to differentiate *um* and *uh* (Shriberg, 2001). While the f0 patterns can vary depending on the surrounding environments, there is evidence that the f0 of markers is generally lower than the speaker's relative f0 levels (Gabrea & O'Shaughnessy, 2000), with *uh* having a lower f0 than *um* in Dutch (Swerts et al., 1998). Analyzing the formants and intensity of the vowel segments in each marker can also signal differences in the production of *um* and *uh.* Work by Hughes et al. (2016) did not find major differences between F1 and F2 for the vocalic midpoints of *uh* and *um*.

The present study investigates two main questions of interest. First, is *uh* or *um* the most common form of marker found in Canadian English speech? Second, what are the acoustic characteristics of *uh* and *um*? In order to address the second question, measures of duration, fundamental frequency, F1 and F2 were extracted for each marker. Following previous research, we hypothesize that:

1. *Um* and *uh* will have an equal occurrence frequency across speakers (Wieling et al., 2016).
2. *Uh* will have a longer vowel duration than *um* (Hughes et al., 2016).

https://doi.org/10.18463/diss-2021-010-morin-tucker

3. *Uh* will have a lower f0 than *um* (Swerts, 1998).
4. *Um* and *uh* will have similar F1 and F2 values (Hughes et al., 2016).

Method

Corpus

The conversational speech data used in this analysis is from the *Corpus of Spontaneous Multimodal-Interactive Language* (CoSMIL) (Järvikivi & Tucker, 2015). Sixteen native Canadian English speakers (14 female and 2 male; 18–23 years old) participated in the recording sessions. Participants were undergraduate students enrolled in an introductory linguistics course at the University of Alberta, each receiving credit for their participation. Participants signed up as pairs and came to do the experiment together.

The recordings were made in an observation studio, which was set up to use two high quality head-mounted microphones and two opposing ceiling mounted video cameras. The researcher controlled data acquisition from a control room and could observe the interaction via a one-way mirror. Participant pairs engaged in a 45-minute conversation while sitting across from each other in the observation room. Each participant was fitted with an over the ear omnidirectional head-mounted microphone (Countryman E6) with a flat frequency response cap. Each speaker was recorded on one channel of a stereo recording, which were subsequently separated into individual files for each speaker for later analysis. Topics were provided to help initiate conversation, however the conversation portion of the experiment was not controlled and participants were encouraged to talk about whatever topic they wanted. As a result conversational topics varied widely between participants. All sixteen recordings from CoSMIL were time aligned with orthographic transcription via ELAN (2020) and phonetically aligned for Praat (Boersma & Weenink, 2020) using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008). These alignments were used in the analysis for this study.

Data Extraction

A custom script was written to extract our acoustic measures of interest via Praat. We extracted vowel duration, mid-point formant values (F1 and F2), and mean f0 of each vowel. As a control, we also extracted speech rate, which was defined as the number of syllables produced in the surrounding 6 seconds (3 seconds preceding and 3 seconds following). Finally, we extracted the preceding and following word along with their duration. Data was extracted from individual speakers so that f0 and formant extraction values could be appropriately tailored to each speaker. Markers for this study were defined as *um, uh,* and *er*. For comparison purposes, we also counted the instances of *like* produced by each participant. *Like* can also be used as a marker with a range of grammatical functions in spontaneous speech of Western Canadian English (Podlubny, Geeraert, & Tucker, 2015).

Statistical Analysis

The majority of the data was modeled using Linear Mixed Effects Regression (Bates et al., 2015) with subject as a random effect. We investigated vowel duration, F1, F2 and fundamental frequency as dependent variables for the *um* and *uh* markers. We used the identity of the Marker (*um* or *uh*), the Following or Preceding context (word vs silent period (sp)), and Speech Rate as our independent variables. We used a backward stepwise model fitting procedure testing individual predictor effects along with possible two-way interactions. Non-significant effects were removed until a final best-fit model was achieved. Effects were considered significant if the *t* value exceeded an absolute value of 2. All possible random slopes were explored after the stepwise modeling procedure and any random slopes which improved the models fit and did not result in an error or warning were retained.

Results

A total of 1,055 markers were extracted from the eight conversations in the CoSMIL dataset, or about 66 markers per speaker with their rate of production ranging from 20 to 129 markers per conversation. Of the markers there were 7 instances of *er*, 498 of *uh*, and 550 *um*. As a result of so few instances of *er*, these were excluded from the statistical analyses leaving 1,048 instances of *um* and *uh* markers. We also counted a total of 5,513 instances of *like* in the corpus or about 344 instances of *like* per speaker, with individual speakers ranging between 151 to 585 productions of *like* during their conversations.

Duration

As an initial model, we performed a *t*-test to compare the duration of the Marker. In this analysis, *um* (mean = 428 ms) is significantly longer ($t(918.15) = -2.836$, $p < 0.005$) than *uh* (mean = 243 ms) which is likely due to the fact that *um* is made up of two segments.

We then investigated the duration of the vowels in the marker, as illustrated in Figure 1. We investigated all two-way interactions in an attempt to

find the most parsimonious model. We report only those predictors and interactions that were significant in the best fitting model as described in the Statistical Analysis section. Marker by Subject as a random slope improved the model fit and was retained in the final model. There is a significant interaction between Marker and Following context. The interaction illustrates that when the Following context is held constant the Markers are not significantly different from each other (sp: $t = -0.065$; word: $t = -1.825$). When the Marker is held constant there is a significant difference as a result of the Following context. For both *uh* and *um* the vowel is shorter when the Following context is a word (*uh*: $\beta = -0.0842$, *se (standard error)* = 0.0112, $t = -7.487$; *um*: $\beta = -0.0457$, $se = 0.011$, $t = -4.142$). When a word follows the Marker the duration of the vowel is shorter. ($\beta = -0.0805$, $se = 0.0122$, $t = -6.598$). We also find that the faster the speech rate the shorter the vowel ($\beta = -0.009$, $se = 0.003$, $t = -2.934$).

Fundamental Frequency

In our f0 data there were instances where the pitch tracking algorithm failed to extract a valid measure and these items were excluded from the analysis, leaving 1029 items for the analysis. No random slopes were found to improve model fit. We have chosen not to transform our f0 values in this model as most of our speakers are female and it is hoped that the speaker random effect will account for some of the speaker variability. We found that f0 is lower when the segment is shorter (effect size: 40 Hz, $\beta = -43.987$, $se = 12.292$, $t = -3.578$) and the f0 is lower when the speech rate is faster (effect size: 62 Hz, $\beta = -8.745$ $se = 1.292$, $t = -6.768$). The f0 is slightly higher for the *uh* markers (8 Hz, $\beta = 8.349$, $se = 3.612$, $t = 2.312$). The f0 is higher when there is a following word as opposed to when there is following silence (11 Hz, $\beta = 11.738$, $se = 3.256$, $t = 3.605$).

Formants

We also analyzed the formant characteristics of the vowels in *um* and *uh*. This comparison is illustrated in the vowel plot in Figure 2. In this analysis we transformed the formant values using the $\log_{10}$ function and also included Segment Duration as a covariate in the model. In the model of F1 no random slopes were found to improve the model fit and in the F2 model Marker and Previous context by subject significantly improved the model fit.

We found that *um* has a significantly higher F1 compared to *uh* ($\beta = -0.078$, $se = 0.014$, $t = -5.517$) and that F1 is higher when the following item is a

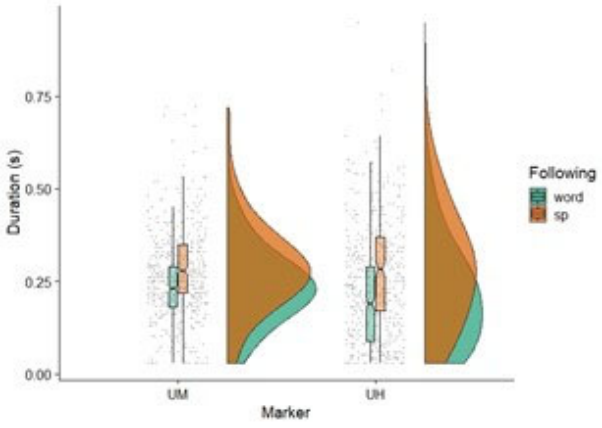

Figure 1. Raincloud plot (Allen et al., 2021) of the vowel durations of the markers um and uh split by the following content, silent period (sp) is in brown and lexical content (word) is in green.

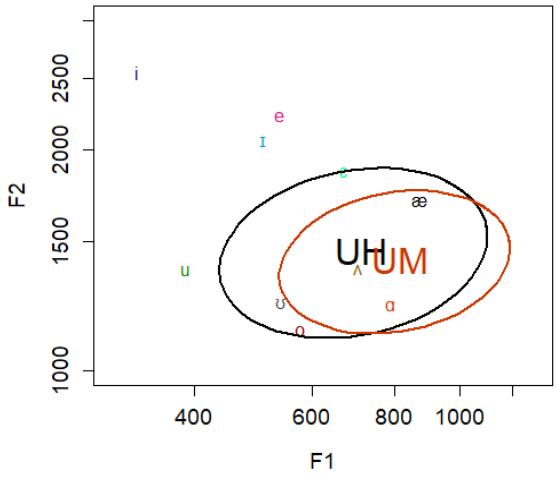

Figure 2. F1 by F2 plot of formant measures for the markers um and uh using phonTools (Barreda, 2015). The label indicates the average formant value and ellipses are plotted at 1.96 standard deviations. Average values from Hillenbrand et al. (1995) are plotted to provide some context.

word ($\beta = -0.028$, $se = 0.012$, $t = 2.203$). There was also a significant interaction between Segment Duration and the Previous context. When the Previous context is a silent period there is a slight increase (effect size: 66 Hz, $\beta = -0.17$, $se = 0.0602$, $t = 2.822$) in F1 as Segment Duration increases but when the preceding context is a word, we see that as segment duration increases the frequency of F1 also increases (effect size: 289 Hz, $\beta = 0.374$, $se = 0.094$, $t = 3.981$). As speech rate increases so does F1 frequency ($\beta = 0.015$, $se = 0.005$, $t = 2.99$).

In the model analyzing F2 we find that *uh* has higher F2 than *um* ($\beta = 0.054$, $se = 0.0145$, $t = 3.726$). Speech rate was not significant in this model. Previous and Following context significantly interacted with Segment Duration: the effect is the

same for both. There is not effect when the preceding context is a silent period but the effect is significant when there is a word preceding ($\beta = -0.128$, $se = 0.052$, $t = -2.453$) or following ($\beta = -0.12$, $se = 0.051$, $t = -2.336$). In both cases, when a word is present, longer segment duration decreases the F2.

Discussion

In summary, the results from the present analysis indicate that there is a fairly equal but small bias toward the occurrence of *um* (550) as opposed to *uh* (498) in the 1,048 extracted markers. In testing our first research hypothesis, we find that *um* is the more common form of marker found in Canadian English speech, though only slightly more common. These results also confirm our original hypothesis that Canadian English would reflect the usage of *um* and *uh* of other English varieties, showing a similar proportion of *um* over *uh* instances as those found most recently in 2013 (Wieling et al., 2016). Following Wieling et al., (2016), we suspect that the similarity in marker proportion is likely due to cross-linguistic changes within native English speaking countries that are often influenced by societal extralinguistic forces. Interestingly, our data indicates a relatively low occurrence of *um* and *uh* markers when compared to the occurrences of *like* in the corpus. Our counts indicate that *like* occurs 5 times as often. We have not seen previous comparisons of these markers and believe that the high frequency of *like* is potentially due to the increased functional role it plays in speech (Podlubny et al., 2015).

Our findings on overall marker duration confirm that *um* has a longer duration than *uh*, likely due to the phonemic difference between the two markers (*um* /əm/ has two phonemes while *uh* /ə/ has one, Clark & Fox Tree, 2002; Swerts, 1998). These overall marker durations are consistent with previous findings. Contrary to Hughes et al. (2016) and our second research hypothesis, we do not find a significant difference in the duration of the vowels in *um* and *uh*. However, most of our participants are female, while all of the participants from Hughes et al. (2016) were male. It is possible that there are gender differences in the usage of the two markers. We do note that the reported vowel durations for both markers in our study in comparison to Hughes et al. (2016) might suggest that the vowel duration of *um* is longer in Canadian English than in other dialects. The duration of both the *uh* and *um* vowel segments are longer when followed by a silent pause than when followed by a word, suggesting that Canadian English aligns with previous claims that these prolonged vowels are used by speakers to signal an upcoming delay (Clark & Fox Tree, 2002).

The results show other acoustic phonetic differences between *um* and *uh* as well. Specifically, our third research hypothesis investigates fundamental frequency. We find that fundamental frequency is slightly lower for *um* in comparison to *uh*, disconfirming our original hypothesis (Swerts, 1998). We believe this may be due to the following voiced nasal contributing to a lower f0 in the *um* vowel, however results concerning the effect of following consonants on vowels is variable (Hanson, 2009). While the present findings generally agree with the literature, we are cautious in our interpretations as the sample size is fairly limited.

For our fourth and last research hypothesis we found that *um* has a higher F1 and lower F2 than *uh*, contradicting our original hypothesis (Hughes et al., 2016). We suspect this difference is due to the high between-speaker variability and stylistic differences that are often reported in acoustic analyses of filled pauses (Hughes et al., 2016; Clark & Fox Tree, 2002) as well as the gender differences noted previously.

We believe that additional research of Canadian English spontaneous speech datasets is necessary and recommend two possible directions. First, additional investigation of *like* as a marker in spontaneous speech is necessary. *Like* is an increasingly common marker that fills many functions in conversational speech (e.g. Fox Tree & Tomlinson, 2007; Podlubny et al., 2015). Acoustic characteristics of *like* have been shown to signal its usage as a marker in comparison to its other functions (Podlubny et al., 2015). Second, further investigation of the functional role of the *um* and *uh* as stance markers (Le Grézause, 2017) in Canadian English is important. Following Swerts (1998), investigation of an interaction between phrase position, fundamental frequency, and duration for *um* and *uh* would be beneficial. The current results are an important first step to our preliminary understanding of the acoustic characteristics and differences of *um* and *uh* in spontaneous Western Canadian English.

References

Allen, M., D. Poggiali, K. Whitaker, T. R. Marshall, J. van Langen, & R. A. Kievit. 2021. Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, 4, 63. https://doi.org/10.12688/wellcomeopenres.15191.2

Barreda, S. 2015. phonTools: Functions for phonetics in R. R package version 0.2-2.1.

Bates, D., M. Mächler, B. Bolker, & S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boberg, C. 2010. *The English Language in Canada: Status, History and Comparative Analysis*. Cambridge, UK: Cambridge University Press.

Boersma, P. & D. Weenink. 2020. Praat: doing phonetics by computer (version 6.1.32). http://www.praat.org/ (accessed 12 November 2020).

Clark, H. H., & J. E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3 %

Fox Tree, J. & J. M. Tomlinson Jr. 2007. The Rise of *Like* in Spontaneous Quotations. *Discourse Processes*, 45(1), 85–102. https://doi.org/10.1080/01638530701739280

Gabrea, M., & D. O'Shaughnessy. 2000. Detection of filled pauses in conversational speech. In: *Proceedings, Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, 16–20 October, 2000, Beijing, China, 517–520.

Hanson, H. M. 2009. Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America*, 125(1), 425–441. https://doi.org/10.1121/1.3021306

Hillenbrand, J., L. A. Getty, M. J. Clark, & K. Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5), 3099–3111. https://doi.org/10.1121/1.411872

Horváth, V. 2010. Filled pauses in Hungarian: Their phonetic form and function. *Acta Linguistica Hungarica,* 57(2–3), 288–306.

Hughes, V., S. Wood, & P. Foulkes. 2016. Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech Language and the Law*, 23(1), 99–132. https://doi.org/10.1558/ijsll.v23i1.29874

Järvikivi, J. & B. V. Tucker. 2015. Corpus of Spontaneous Multimodal Interactive Language (CoSMIL). University of Alberta.

Le Grézause, E. 2017. *Um and Uh, and the Expression of Stance in Conversational Speech*. Ph.D. dissertation, University of Washington.

Podlubny, R. G., K. Geeraert, & B. V. Tucker. 2015. It's All About, *Like*, Acoustics. In: The Scottish Consortium for ICPhS 2015 (eds.), *Proceedings of the 18th International Congress of Phonetic Sciences*, 10–14 August, 2015, Glasgow, Scotland, UK, Paper 0477.

Shriberg, E. 2001. To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1), 153–169. https://doi.org/10.1017/S0025100301001128

Swerts, M. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4), 485–496. https://doi.org/10.1016/S0378-2166(98)00014-9

Wieling, M., J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, & M. Liberman. 2016. Variation and Change in the Use of Hesitation Markers in Germanic Languages. *Language Dynamics and Change*, 6(2), 199–234. https://doi.org/10.1163/22105832-00602001

Yuan, J. & M. Liberman. 2008. Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America*, 123(5), 3878–3878. https://doi.org/10.1121/1.2935783

Variation in jitter, shimmer, and intensity of filled pauses and their contexts in native and nonnative speech

Ralph L. Rose
Faculty of Science and Engineering, Waseda University, Tokyo, Japan

Abstract

Various acoustic parameters of filled pauses (e.g., uh/um in English, e-(to) in Japanese) have been investigated including duration, pitch, and formants. Less investigated have been jitter, shimmer, and intensity. The present work looks at systematic variation in these properties of filled pauses and their immediate contexts in a crosslinguistic speech corpus. Filled pauses were examined within the five token (word) window centered on the filled pause, exploring variation with respect to first (L1 Japanese) and second language (L2 English) speech as well as L2 proficiency. Results show that relative to the central filled pause, higher jitter and shimmer occur before the filled pause and higher intensity afterward. Proficiency group differences are weak, but suggest that jitter differences are greater in high proficiency speakers and shimmer differences greater in low proficiency speakers. Results vary somewhat from earlier work, but suggest jitter and shimmer may be advance indicators of upcoming disfluency.

Introduction

Study of the acoustic properties of filled pauses (e.g., *uh*/*um* in English, *e*-(*to*) in Japanese) has looked at many features in the past two or more decades of pausological research. Although often containing rather wide variation across and sometimes within individuals, systematic observations have been found with respect to their duration (cf. Jehoul, 2019; Watanabe et al 2015), the duration of adjacent silent pauses (Clark and Fox Tree, 2002; Rose, 2015), their pitch contours (Shriberg and Lickley, 1993; Tseng 1999), their pitch contexts (Maekawa, 2013), their formant values in native, bilingual, and nonnative speech (Lo, 2020; Rose, 2017) among other trends (see Lickley, 2017 for a recent review). Accounts of these various trends have built on pragmatic principles (e.g., interactional constraints; Clark and Fox Tree 2002), phonetics (e.g., language-based prosodic effects; Maekawa and Mori, 2017), or cognitive limitations (e.g., working memory capacity; Don and Lickley 2015).

There remain many acoustic features of filled pauses that have not been studied much. Three of these include jitter, shimmer, and intensity. While these properties are used productively as features in machine learning applications and speech technologies (see e.g., Farrús et al 2007) and to the extent that training datasets contain naturally produced filled pauses, these datasets may lead to better and more realistic performance (e.g., López-de-Ipiña et al, 2020), little is known about how systematically these features emerge in filled pauses and their contexts (though see Maekawa and Mori, 2017, discussed further below).

The present study is thus intended as an exploration of jitter, shimmer, and intensity in filled pauses in a crosslinguistic speech corpus. The exploration considers how these features might vary between languages, relative to word tokens preceding and following a filled pause, and across proficiency levels of nonnative speech. The paper is organized as follows. After reviewing some relevant background of these three acoustic properties in the next section, the paper describes the corpus used. The methodology section describes what kind of variation is expected and how it can be observed. Afterward follows the results and analysis and the paper concludes with the discussion.

Background

Filled pauses

Filled pauses occur commonly in everyday human speech and are known to occur in a wide variety of forms. However, most languages have clearly dominant forms. In English, these are [əː] and [əm], typically rendered as *uh* and *um*, respectively in North American orthography (Maclay and Osgood, 1959; Kasl and Mahl, 1965; Shriberg, 1994; Clark and Fox Tree, 2002). In Japanese, every vowel sound is attested as a filled pause, but by far the most common filled pause form is [ɛː(to)] and a nasal [ŋː] is a common variant (Maekawa et al 2003), often transcribed as *e*-(*to*) and *N*-, respectively. These variants in English and Japanese will be the focus of the present study.

Jitter, shimmer, and intensity

Jitter is a measure of the variation in the periodicity of speech, while shimmer is a measure of the variation in the amplitude of speech. There are many ways of calculating each of these, but in the

https://doi.org/10.18463/diss-2021-011-rose

jitter(local) and *shimmer(local)* methods available in Praat (Boersma and Weenink, 2006), they are calculated only over portions of the signal where voicing pulses are detected and are a ratio of the mean absolute value of sequential deviation in the period/amplitude to the mean period/amplitude. Hence, a voice with precisely no variation in pitch and no variation in intensity would have jitter/shimmer values of 0, while a so-called "gravelly" voice would have much higher jitter and shimmer values. Theoretically, the maximum jitter value would be 1, but practically, values are much lower. In fact, jitter ≤ 1.04% and shimmer ≤ 3.81% have been suggested ranges for normal non-pathologic speech when producing long sustained vowels in a controlled laboratory setting (Williamson, 2014).

Jitter and shimmer are known to vary with such things as age (Goy and Pichora-Fuller, 2016), emotion (Erickson et al, 2008), language background (Cantor-Cutiva et al, 2021), loudness and gender (Brockmann et al, 2008), as well as pathological conditions (e.g., López-de-Ipiña et al, 2020).

Maekawa and Mori (2017) compared filled pause vowels to lexical vowels in Japanese spontaneous speech and observed that they differ on a number of acoustic parameters. Duration was the greatest contributor to the difference, while jitter made a significant, though minor contribution. Specifically, filled pause vowels in Japanese were longer and exhibited greater jitter than lexical vowels. They suggest that this effect might be partly explained by the use of breathy phonation during filled pauses, as breathiness may be a cue of politeness in Japanese.

Intensity, measured as the energy in the speech signal, has been studied more than jitter and shimmer, but is quite difficult to measure reliably due to many intervening factors including distance from the microphone, orientation of microphone relative to vocal tract, and environment. All of these factors may vary during a single recording session. Maekawa and Mori (2017) also looked at intensity in their comparison of filled pause and lexical vowels and found that it made the second largest contribution to the difference after duration, with filled pause vowels showing lower intensity.

Nonnative speech

Numerous studies of nonnative speech have observed various acoustic differences in speakers' first (L1) and second language (L2) speech production, particularly as they relate to speakers' proficiency in their L2. For example, speakers tend to speak more slowly and pause longer in their L2 than in their L1 and this difference is greater for lower proficiency speakers (Rose, 2013). With respect to filled pauses, Rose (2017) observed that native speakers of Japanese use longer filled pauses in their L2 than in their L1. However, this was consistent across proficiency levels and did not seem to be modulated by articulation rate.

With respect to jitter, shimmer, and intensity (JSI), I can find almost no work on the comparison of these measures in filled pauses in L1 and L2 speech, let alone across L2 proficiency levels. Yet, there are several reasons why differences might be expected. If Maekawa and Mori's (2017) account is correct, low proficiency L2 speakers might exhibit more frequent production difficulties and accounting for these difficulties—as a matter of politeness—may lead to greater breathiness and thus different JSI measures than those for higher proficiency speakers. Furthermore, the appearance of this acoustic information might not be on the filled pause alone but could appear earlier in context (as the speaker realizes their production difficulty), or perhaps even immediately afterward as a spillover effect. The present work seeks to bring data to bear on these conjectures by answering the following research questions.

- How do filled pause JSI vary between languages?
- How do filled pause JSI vary across L2 proficiency levels?
- How do filled pause JSI vary with respect to their contexts?

Method

Corpus

As an exploratory study, the present work makes use of a speech corpus to answer the main research questions. The Crosslinguistic Corpus of Hesitation Phenomena (CCHP: Rose, 2013) is a corpus of speech recordings in which 35 university-aged native speakers of Japanese were asked to speak in response to three elicitation tasks: reading aloud, picture description, and topic narrative. Participants spoke for 2–3 minutes in response to each task and completed the tasks in both Japanese (L1) and English (L2). The speech recordings have already been transcribed and annotated for various hesitation phenomena (e.g., filled pauses, silent pauses, repair sequences). Time alignment information is available for all filled pauses and their immediate context (preceding and following word tokens). Also available is meta information including an estimate of the participant's L2 proficiency based on standardized test performance, living abroad experience, and their own self-assessment.

For the present study, the reading aloud recordings were excluded because filled pauses are sparse in this task, with many speakers using none. Hence, the remainder of the study is focused on the filled pauses used in the other spontaneous speech tasks.

Given the difficulties with intensity measurement as noted in the Background section above, some explanation about intensity in the CCHP recordings is warranted. Before speaking, participants were asked to maintain a consistent distance from the microphone while speaking. However, no further actions were taken to control the speaking configuration and thus intensity. Indeed, it is likely that each speaker varied somewhat from recording to recording as well as within each recording. However, main part of the analysis below will focus on only a narrow window surrounding each filled pause and normalize values relative to the filled pause. It is believed that only intensity variation due to configurational variation would be rare and essentially negligible within these narrow windows.

Procedure

The data for the present study comes from an analysis of the filled pauses in spontaneous speech recordings. For the context investigation, the data set is limited to filled pauses in continuous contexts: 5-token windows with a filled pause at its center and where the filled pause does not have an adjacent silent pause . In other words, a 5-word sequence with the filled pause in position 3 and two words before and after (e.g., "the father uh had a"). Words were used as the unit of measure in part because the corpus provides alignment information only at the word—and not the syllable or even phoneme—level. Further, if the politeness account is valid, then it is likely that speakers would be recognizing their upcoming difficulty at the level of word and phrase formulation rather than at articulation. Nonetheless, this does mean that there is no control of word length in the present analysis.

Measurements of the jitter, shimmer, and intensity of all of these tokens were taken using Praat (Boersma and Weenink, 2006). Periodicity was analyzed using the *PointProcess(periodic, cc)* procedure and then for jitter, the *jitter(local)* procedure was used; for shimmer, the *shimmer(local)* procedure was used; and for intensity, the *Get intensity (db)* procedure was used.

Measurements were taken across each whole token. As noted in the background section, Praat's algorithms compute jitter and shimmer only where voicing pulses are detected. Hence, this means that voiced consonants might also be included. However, consonants in general have much shorter duration than vowels and therefore it is believed they would have only a small influence in overall jitter and shimmer values for each word. Alternatively, devoiced vowels would be excluded. This would likely affect the Japanese data more where high vowel devoicing is a well-known phenomenon (cf. Hasegawa, 1999). However, it is believed that this would not have a systematic effect, but at most only decrease the robustness of the Japanese speech data.

For the proficiency level information, the participants were separated into two groups, high and low, based on the proficiency level estimate provided in the corpus. In the statistical analyses below, mixed effects modeling was used with **language** and **proficiency level** as fixed effects and **participants** as a random effect. Statistics were computed using the *nlme* (version 3.1-149) package in *R* (version 4.0.3).

Results

Shown in Table 1 are the number of filled pauses in the spontaneous speech recordings in the corpus by 34 speakers (one speaker who lacked proficiency information was removed).

Table 1. Number of filled pauses in spontaneous speech recordings of CCHP in the corpus as a whole and the subset of those which have no adjacent silent pause

Proficiency Group	Whole corpus		No pause	
	Ja (L1)	En (L2)	Ja (L1)	En (L2)
High	943	641	176	65
Low	850	661	151	36

Results (see Table 2) show that participants produce filled pauses in Japanese with higher jitter [$t(2932) = 3.38$, $p < 0.001$], higher shimmer [$t(2866) = 5.50$, $p < 0.001$], and lower intensity [$t(3059) = 5.85$, $p < 0.001$] than those produced in English. While jitter patterns are similar across proficiency levels, the shimmer difference is largely driven by the high proficiency speakers while low proficiency speakers actually show the same shimmer for both L1 and L2 [$t(2866) = 3.61$, $p < 0.001$]. A similar pattern appears for intensity where low proficiency speakers show no difference between L1 and L2 in contrast to high proficiency speakers [$t(3059) = 3.41$, $p < 0.001$].

The context analysis focuses on the "no pause" subset of the data described in Table 1. Jitter results (see Figure 1) show that both high and low group participants' speech has higher jitter before the filled pause [$t(1723) = 2.05$, $p < 0.05$], the low proficiency

Table 2. Jitter, shimmer, and intensity measurements with 95% confidence intervals for 3,095 filled pauses in CCHP.

		Ja (L1)	En (L2)
Jitter	High	0.038±0.002	0.034±0.003
	Low	0.038±0.003	0.034±0.002
Shimmer	High	0.146±0.006	0.130±0.007
	Low	0.140±0.007	0.139±0.007
Intensity (db)	High	66.94±0.48	68.29±0.50
	Low	67.39±0.46	67.61±0.50

speakers' jitter in English is marginally more contrastive [$t(1723) = 1.89$, $p = 0.058$].

Shimmer results (see Figure 2) show that while shimmer overall is higher in the context than in the filled pause [$t(1677) = 2.70$, $p < 0.01$], this is mainly due to the speech of the high proficiency speakers in English [$t(1677) = 2.10$, $p < 0.05$].

Intensity results (see Figure 3) show that the intensity is generally higher in the context of the filled pause than in the filled pause itself [$t(1814) = 2.11$, $p < 0.05$], but that this is mostly expressed in English more than in Japanese in the tokens adjacent to the filled pause [$t(1814) = 2.25$, $p < 0.05$].

Discussion

This study has sought to answer the questions of how the acoustic features of jitter, shimmer, and intensity vary in filled pauses and their contexts with respect to first and second language speech as well as second language proficiency. Overall, the results suggest that filled pauses in native Japanese speakers' L2 speech is lower than that in the surrounding context while in L1 speech, the filled pauses are consistent with their context. There is some slight difference between the three acoustic characteristics with jitter and shimmer differences showing up mostly before the filled pause and intensity differences showing up mostly after. This suggests that if the account of Maekawa and Mori (2017) is correct, the politeness cues are actually appearing in an anticipatory manner in advance of the actual overt disfluency.

Proficiency group differences are mild. If anything, low proficiency speakers show greater jitter before the filled pause while high proficiency speakers show greater shimmer. This is a curious result as it is difficult to suppose that the processing difficulties that high versus low proficiency speakers experience would generate such a subtle acoustic difference. Perhaps this observation—which is already weak to begin with—is just an artifact of individual variation.

One question that might arise is whether the observed language differences might be due to discourse processing. That is, could jitter, shimmer, and intensity be affected by whether the filled pause

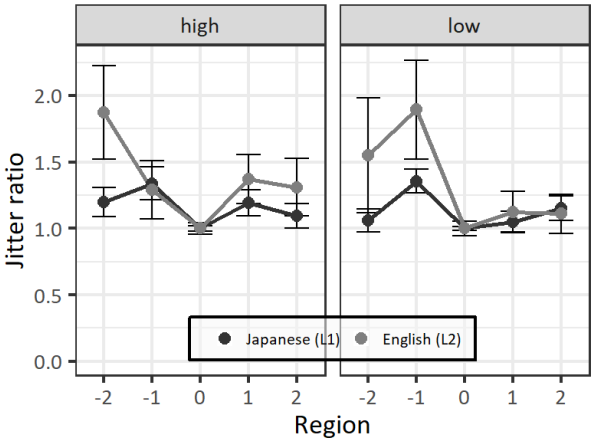

Figure 1. Jitter values in the context 5-token context surrounding a filled pause (0). Values are expressed as ratios to that of the center filled pause. Error bars represent standard error of measurement.

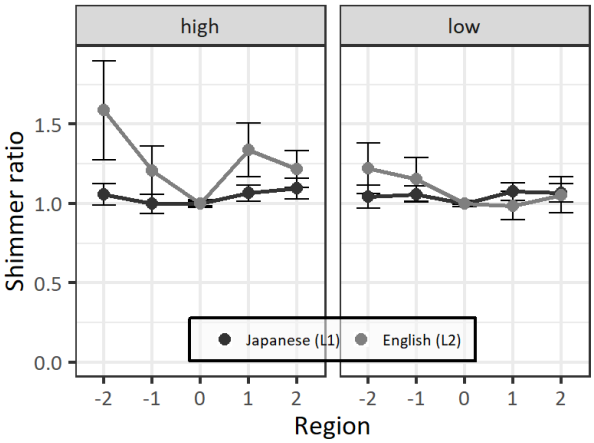

Figure 2. Shimmer values in the context 5-token context surrounding a filled pause (0). Values are expressed as ratios to that of the center filled pause. Error bars represent standard error of measurement.

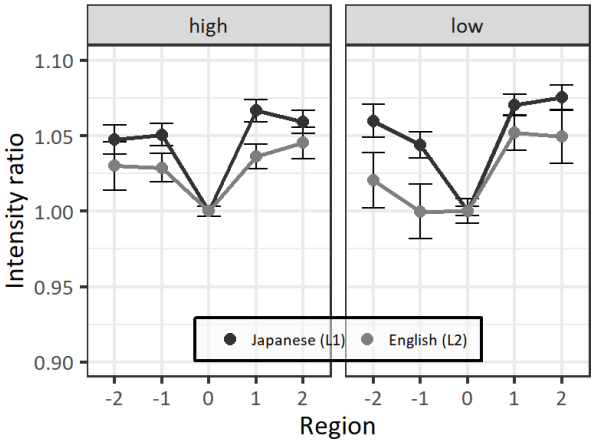

Figure 3. Intensity values in the context 5-token context surrounding a filled pause (0). Values are expressed as ratios to that of the center filled pause. Error bars represent standard error of measurement.

is at a major discourse boundary or not. A post-hoc test of this hypothesis evaluated this by including utterance boundary (i.e., whether the filled pause is at an utterance boundary) as a factor in the model. However, this analysis shows no effect of boundary.

At first glance, the main results might appear to be at odds with those of Maekawa and Mori (2017) who observed higher jitter for filled pauses than lexical vowels. But it could be the case that in the immediate context of filled pauses, speakers have heightened jitter (and shimmer) overall, with filled pauses representing local minima even if higher than average for the lexical vowels in the whole speech sample. Unfortunately, the corpus annotation does not make it possible to test this hypothesis precisely, but a preliminary test—comparing the filled pause jitter to the jitter overall for the speech sample which contains it—supports this conjecture with participants' mean filled pause jitter 0.015 higher than the overall jitter [$t(33) = 5.69$, $p < 0.001$].

Despite the politeness account of the use of filled pauses, it may actually be possible that speakers are not consciously using these subtle acoustic cues to communicate mental states to their interlocutors, given doubts about speaker design in the use of disfluencies (cf. Finlayson and Corley, 2012). Nevertheless, these acoustic differences may still be systematically reflecting certain cognitive processes or difficulties and based on the present data, might even suggest their anticipatory use. From the perceptual side, early evidence suggests that listeners might not actually be very sensitive to jitter and shimmer as cues (Kreiman and Gerratt, 2003), but some recent evidence counters this (Erickson et al 2008). Clearly, more research is warranted on patterns of production and perception of jitter, shimmer, and intensity in filled pauses.

References

Boersma, P. & D. Weenink. 2006. Praat: Doing phonetics by computer (version 6.1.38). https://www.praat.org/ (accessed 24 January 2021).

Brockmann, M., C. Storck, P. N. Carding, & M. J. Drinnan. 2008. Voice Loudness and Gender Effects on Jitter and Shimmer in Healthy Adults. *Journal of Speech, Language, and Hearing Research* 51(5), 1152–1160. https://doi.org/10.1044/1092-4388(2008/06-0208)

Cantor-Cutiva, L. C., P. Bottalico, C. Nudelman, J. Webster & E. J. Hunter. 2021. Do Voice Acoustic Parameters Differ Between Bilingual English-Spanish Speakers and Monolingual English Speakers During English Productions? *Journal of Voice* 35(2), 194–202. https://doi.org/10.1016/j.jvoice.2019.08.009.

Clark, H. H. & J. E. Fox Tree. 2002. Using Uh and Um in Spontaneous Speaking. *Cognition* 84(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Don, S. & R. Lickley. 2015. Uh I forgot what I was going to say: How memory affects fluency. In: *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015)*, 8–9 August 2015, Edinburgh, Scotland.

Erickson, D., A. Rilliard, T. Shochi, J. Han, H. Kawahara & K. Sakakibara. 2008. A cross-linguistic comparison of perception to formant frequency cues in emotional speech. In: *Proceedings of the 11th Oriental COCOSDA Workshop*, 25–27 November 2008, Nara, Japan, 209–214.

Farrús, M., J. Hernando, & P. Ejarque. 2007. Jitter and Shimmer Measurements for Speaker Recognition. In: *8th Annual Conference of the International Speech Communication Association (Interspeech)*, 27–31 August 2007, Antwerp, Belgium, 778–781.

Finlayson, I. R. & M. Corley. 2012. Disfluency in dialogue: an intentional signal from the speaker? *Psychonomic Bulletin & Review* 19(5), 921-928. https://doi.org/10.3758/s13423-012-0279-x.

Goy, H. & M. K. Pichora-Fuller. 2016. Effects of age on speech and voice quality ratings. *The Journal of the Acoustical Society of America* 139(4), 1648–1659. https://doi.org/10.1121/1.4945094

Hasegawa, Y. 1999. Pitch accent and vowel devoicing in Japanese. In: J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (eds.), *14th International Congress of Phonetic Sciences (ICPhS)*, 1–7 August 1999, San Francisco CA, USA, 523–526.

Jehoul, A. 2019. *Filled pauses from a multimodal perspective. On the interplay of speech and eye gaze*. Ph.D. Dissertation, Katholieke Universiteit Leuven.

Kasl, S. V. & G. F. Mahl. 1965. Relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology* 1(5), 425–433. https://doi.org/10.1037/h0021918

Kreiman, J. & B. R. Gerratt. 2003. Jitter, shimmer, and noise in pathological voice quality perception. In: C. d'Allessandro & K. R. Scherer (eds.), *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, 27–29 August 2003, Geneva Switzerland, 57–62.

Lickley, R. 2017. Disfluency in typical and stuttered speech. In: C. Bertini, C. Celata, G. Lenoci, C. Meluzzi, I. Ricci, (eds.), *Fattori Socali e Biologici Nella Variazione Fonetica* [Social and Biological Factors in Speech Variation] (Studi AISV), Milano, Italy: Associazione Italiana Scienze della Voce, 373–387. https://doi.org/10.17469/O2103AISV000019

Lo, J. J. H. 2020. Between Äh(m) and Euh(m): The Distribution and Realization of Filled Pauses in the Speech of German-French Simultaneous Bilinguals. *Language and Speech* 63(4), 746–768. https://doi.org/10.1177/0023830919890068

López-de-Ipiña, K., U. Martinez-de-Lizarduy, P.M. Calvo, B. Beitia, J. García-Melero, E. Fernández, M. Ecay-Torres, M. Faundez-Zanuy & P. Sanz. 2020. On the analysis of speech and disfluencies for automatic detection of Mild Cognitive Impairment. *Neural Computing and Applications* 32(20), 15761–15769. https://doi.org/10.1007/s00521-018-3494-1

Maclay, H. & C. E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15(1), 19–44. https://doi.org/10.1080/00437956.1959.11659682

Maekawa, K. 2003. Corpus of spontaneous Japanese: Its design and evaluation. In: *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, 12–16 April 2003, Tokyo, Japan.

Maekawa, K. 2013. Prediction of F0 height of filled pauses in spontaneous Japanese: a preliminary report. In: R. Eklund (ed.), *The 6th Workshop on Disfluency in Spontaneous Speech (DiSS 2013)*, 21–23 August 2013, Stockholm, Sweden, 41–44.

Maekawa, K. & H. Mori. 2017. Comparison of Voice Quality between the Vowels in Filled Pauses and Ordinary Lexical Items. *Journal of the Phonetic Society of Japan* 21(3), 53–62. https://doi.org/10.24467/onseikenkyu.21.3_53

Rose, R. L. 2013. Crosslinguistic Corpus of Hesitation Phenomena: A Corpus for Investigating First and Second Language Speech Performance. In: F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier (eds.), *Proceedings of 14th Annual Conference of the International Speech Communication Association (Interspeech)*, 25–29 August 2013, Lyon, France, 992–996.

Rose, R. L. 2015. Um and Uh as Differential Delay Markers: The Role of Contextual Factors. In: *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015)*, 8–9 August 2015, Edinburgh, Scotland.

Rose, R. L. 2017. A Comparison of Form and Temporal Characteristics of Filled Pauses in L1 Japanese and L2 English. *Journal of the Phonetic Society of Japan* 21(3), 33–40. https://doi.org/10.24467/onseikenkyu.21.3_33

Shriberg, E. E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California at Berkeley.

Shriberg, E. E. & R. J. Lickley. 1993. Intonation of clause-internal filled pauses. *Phonetica* 50(3), 172–179. https://doi.org/10.1159/000261937

Tseng, S-C. 1999. *Grammar, prosody and speech disfluencies in spoken dialogues*. Master's Thesis, Bielefeld University.

Watanabe, M., Y. Kashiwagi, & K. Maekawa. 2015. The relationship between preceding clause type, subsequent clause length and duration of silent and filled pauses at clause boundaries in Japanese monologues. In: *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015)*, 8–9 August 2015, Edinburgh, Scotland.

Williamson, G. 2014. Acoustic Measures (Norms). Speech and Language Therapy Information (web site). Retrieved from https://www.sltinfo.com/acoustic-measures-norms/ on 30 June 2021

EGG analysis of filled pauses in Japanese spontaneous speech: Differences in Japanese native speakers and Chinese learners

Xinyue Li [1], Carlos Toshinori Ishi [1,2] and Ryoko Hayashi [3]
[1] ATR Hiroshi Ishiguro Labs, Kyoto, Japan
[2] RIKEN Guardian Robot Project, Saitama, Japan
[3] Kobe University, Kobe, Japan

Abstract

Previous studies on L2 learners of Japanese have shown that the appropriate use of filled pauses is a crucial skill in communication with native speakers. However, there is limited acoustic investigations on filled pauses produced by L2 learners of Japanese. The present study examines the production of filled pauses in Japanese native speakers and L1-Chinese L2 learners of Japanese, using open quotient features extracted from Electroglottography (EGG) signals. The results show that open quotient values of filled pauses were lower than those in ordinary lexical items for Chinese learners of L2 Japanese, suggesting that they may be using vocal tension as one cue to distinguish filled pauses from ordinary lexical items. However, no similar differences for open quotient were observed for the Japanese native speakers. Furthermore, open quotient-valued voice range profiles reveal that Chinese learners of L2 Japanese transfer their native glottal source cues when they produce filled pauses in Japanese.

Introduction

It is well known that spontaneous speech contains a high rate of disfluencies, such as repairs and filled pauses (Shriberg, 2001). Filled pauses like *um* and *uh* in English, *eeto* and *ano* in Japanese, *na4* (tone type is represented by a number) and *zhe4* in Chinese, are considered transmitting a variety of pragmatic information. Filled pauses are also associated with various phonetic characteristics that differentiate them from ordinary lexical items, such as duration patterns (Eklund & Shriberg, 1998), voice quality features (Maekawa & Mori, 2017) and laryngealization (Ogden, 2001). For instance, Maekawa and Mori (2017) conducted acoustic analyses of vowels in filled pauses and ordinary lexical items of Japanese speech. The results showed that voice quality features like spectral tilt-related indices, jitter and shimmer and prosodic features like F0 and intensity are closely linked to the encoding of filled pauses.

The glottal open quotient, defined as the duration of the glottal open phase normalized by the local glottal period (Timcke, von Leden, & Moore, 1958), is one glottal source measurement of voice quality that is useful for discriminating tense vs. lax voice (Henrich, d'Alessandro, & Doval, 2001), which is crucial to the production of filled pauses (Shriberg, 2001). For example, Shriberg (2001) revealed that filled pauses tend to end in creaky voice. Moreover, open quotient can be directly derived from an electroglottography (EGG) signal (Henrich et al., 2004). In this study, we use the EGG signals for calculating the open quotient values in filled pauses.

Filled pauses produced by native speakers were discussed so far, however, relatively little is known about the production of filled pauses speech by second language (L2) learners. Do they show a different pattern from native speakers in the encoding of filled pauses? Is it possible to acquire how to use filled pauses of the target language? Because whether or not the L2 learners can use filled pauses well is an important evaluation criterion for whether their speech is easy to understand (Takamura, 2012), so that acquisition of filled pauses in the target language is inevitable. Hence, the present study aims to clarify whether glottal source measurements differentiate the native speakers from L2 learners in production, using L1-Chinese L2 Japanese as a case study.

Method

The dataset used in the present study includes four sessions of free-topic casual conversations among three interlocutors, collected in our research institute (Ishi, Minato, & Ishiguro, 2019). Each session lasts about 20 to 30 minutes. Two sessions among Japanese native speakers (including 3 males and 3 females), and two sessions among one Japanese native speaker and two L1-Chinese learners of L2 Japanese (including 2 males and 2 females) were selected for analysis. All conversations are in Japanese. It is worth mentioning that we considered that beginners and intermediate learners of L2 Japanese can hardly use Japanese filled pauses

https://doi.org/10.18463/diss-2021-012-li-etal

appropriately in conversation (Konishi, 2017), so that in the present study, we selected Chinese speakers who had passed the highest level ("N1") of the standardized Japanese Language Proficiency Test (JLPT), living in Japan over two years. All Chinese learners speak Mandarin Chinese as their native language. Chinese learners also produced utterances with the same literal meaning in Mandarin Chinese.

Filled pauses are annotated in the dataset, in terms of the transcription texts and the speech act labels, by Japanese native speakers with professional annotation experience. For the ordinary lexical items, only the word-initial vowels were chosen, such as /aHmoNdo/ ('almond') and /eHsu/ ('ace'). In total, 271 filled pauses and 199 ordinary lexical items produced by Japanese native speakers, 338 filled pauses and 220 ordinary lexical items produced by L1-Chinese learners of L2 Japanese were analyzed. When recording the speech, VoceVista Electroglottograph portable devices were used to record the EGG signal from all speakers.

Results of Open Quotient (OQ)

In this study, we divided the data by speaker gender, since it is known that the voice quality of female and male may differ, especially for parameters related to the open quotient (Klatt & Klatt, 1990). The EGG signal is derived into the DEGG (differentiated EGG) signal to calculate the open quotient (Henrich et al., 2004). The glottal open/closed phases were semi-automatically obtained from the peaks and valleys in the DEGG signal (Ishi & Arai, 2018). For example, in Figures 1 and 2, "o" represents "glottal open phase", and "c" represents "glottal close phase". Average OQ values were then obtained for each target vowel segment.

The average open quotient values across each item were calculated and grouped by filled pauses/ ordinary lexical items and the speaker's L1 background. Figure 3 represents the results of this analysis (JN: Japanese native speaker, JFL: Chinese learners of L2 Japanese).

A two-way ANOVA for male speakers (filled pauses/ordinary lexical items × speaker's L1 background) revealed significant main effects of filled pauses/ordinary lexical items and L1 background ($ps < .001$) as well as a significant two-way interaction ($p < .01$). Post-hoc tests on the Chinese male speakers, corrected for multiple comparisons, indicated open quotient values of filled pauses were lower than ordinary lexical items ($p < .001$). In contrast, no similar significant differences for open quotient were observed for the Japanese male speakers. Then, a two-way ANOVA

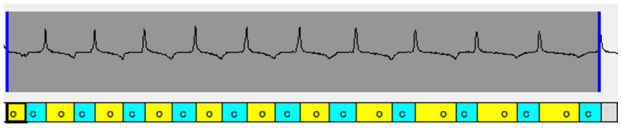

Figure 1. DEGG signal of vowel /a/ produced by Japanese male speaker.

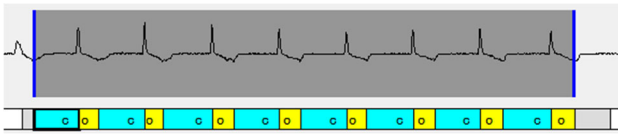

Figure 2. DEGG signal of vowel /a/ produced by Chinese male speaker.

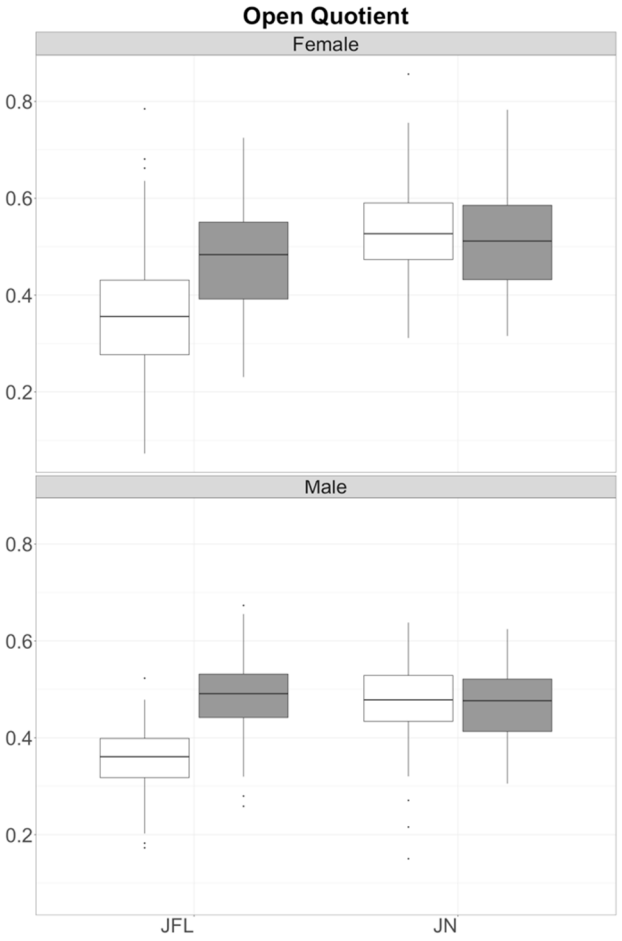

Figure 3. Distributions of Open Quotient for filled pauses (white) and ordinary lexical items (grey).

for female speakers also revealed significant main effects of filled pauses/ordinary lexical items and L1 background as well as a significant two-way interaction (p-values $< .01$). Post-hoc tests on the Chinese female speakers indicated open quotient values of filled pauses were lower than those in ordinary lexical items ($p < .001$). But no similar significant differences for open quotient were observed for the Japanese female speakers.

Results of OQ-valued VRP

Based on the open quotient, OQ-valued voice range profile (VRP) (Wakasa et al., 2018) of the filled pauses is utilized to clarify the dynamic properties of voice quality among Japanese speech produced by two Japanese native speaker and two Chinese learners of L2 Japanese, as well as Chinese speech produced by two Chinese native speakers (ten items for each speaker), shown in Figures 4 to 6. X-axis and y-axis represent F0 (in semitone intervals, with 100Hz as reference; Ishi et al., 2008) and power (in 2 dB intervals), color represents open quotient (the smaller open quotient, the deeper red color, indicating the tenser voice; the larger open quotient, the lighter green color, indicating the laxer voice).

Overall, open quotient decreases when F0 and power become higher in all groups. For Japanese filled pauses produced by native speakers, lax voice has been found (the minimum open quotient value is above .50, presented in green and light yellow color), contrary to that of Chinese learners of L2 Japanese and Chinese filled pauses by native speakers, which have lower open quotient (presented in wider orange and red color zone), suggesting a tenser voice. Moreover, Figures 5 and 6 showed a similar tendency of production in filled pauses, implying that the expression pattern of Japanese filled pauses of Chinese learners of L2 Japanese is influenced by their native language.

Discussion

For the results of open quotient, values in filled pauses were significantly lower than in ordinary lexical items for Chinese learners of L2 Japanese, but not observed for Japanese native speakers. These suggest that Chinese learners may be using vocal tension as one cue to distinguish filled pauses and ordinary lexical items, whereas the same does not happen for Japanese native speakers. According to Maekawa and Mori (2017), where Japanese in monologue speech produced by native speakers were argued, production of filled pauses showed a breathy and aperiodic phonation. The results in the present study may stem from the fact that Japanese native speakers use a lax voice to express filled pauses, different from Chinese learners of L2 Japanese.

For the analysis of OQ-valued VRP, open quotient decreases when F0 and power become higher in all groups, consistent with the results of Henrich, d'Alessandro, Doval, and Castellengo (2005), where open quotient is reported to be strongly related to F0 and power. And the results of OQ-valued VRP indicated that Chinese learners tend to produce filled pauses with tenser voice in both

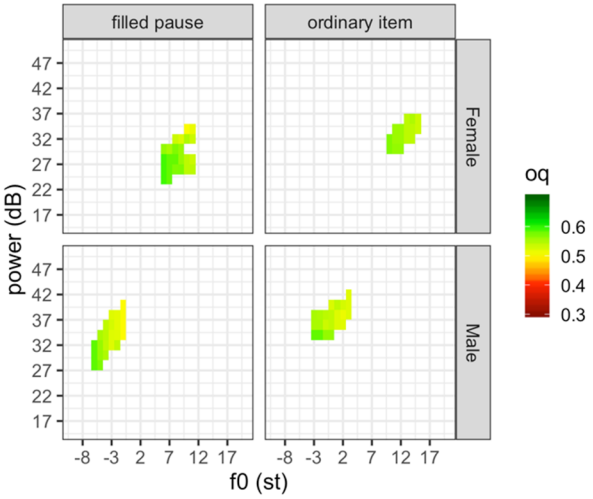

Figure 4. Japanese filled pauses by Japanese native speakers: OQ-valued voice range profile.

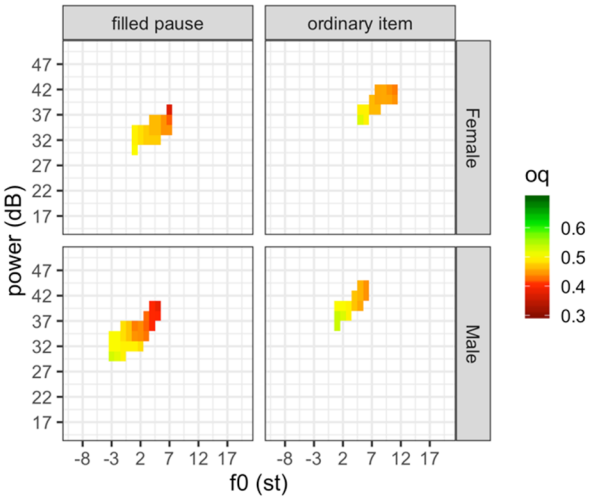

Figure 5. Japanese filled pauses by Chinese learners of L2 Japanese: OQ-valued voice range profile.

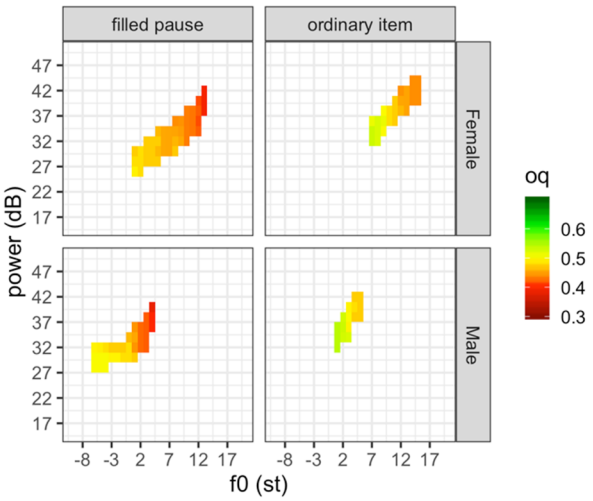

Figure 6. Chinese filled pauses by Chinese native speakers: OQ-valued voice range profile.

Japanese (i.e., the target language) and Chinese (i.e., the native language), whereas the same cannot be

said for Japanese speakers. These results imply that Chinese learners transfer their native glottal source cues when they produce filled pauses in Japanese. Furthermore, Chinese is considered as a tonal language with lexical tone-types (Fu et al., 1998), filled pauses in Chinese such as *na4* (tone type is represented by a number) and *zhe4* are usually falling tones or weak stress, might result in tense phonation. This is a piece of evidence that suggests that L2 transfer exists in the production of filled pauses.

Conclusion

The present study documents that glottal source measurements of open quotient differentiate Japanese native speakers from L1-Chinese L2 learners of Japanese in the production of filled pauses. Specifically, open quotient of filled pauses is significantly lower than ordinary lexical items for Chinese learners of L2 Japanese, but not for Japanese native speakers, suggesting that Chinese learners of L2 Japanese may be using vocal tension as one cue to distinguish filled pauses and ordinary lexical items. Furthermore, the results of OQ-valued VRP indicate that Chinese learners of L2 Japanese transfer their native glottal source cues when they produce filled pauses in Japanese.

Considering the limited number of data used in the present study, further works include gathering a greater number of speakers and investigating the speech communication instruction of filled pauses for L2 learning will be done.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20H05630 and partly supported by JSPS KAKENHI Grant Number JP20H05576, Japan.

References

Eklund, R. & E. Shriberg. 1998. Crosslinguistic disfluency modeling: a comparative analysis of Swedish and American English human–human and human-machine dialogues. In: *Proceedings of the 5th International Conference on Spoken Language Processing*, 30 November – 4 December, 1998, Sydney, Australia, Paper 0805.

Fu, Q. J., F. G. Zeng, R. V. Shannon, & S. D. Soli. 1998. Importance of tonal envelope cues in Chinese speech recognition. *The Journal of the Acoustical Society of America* 104(1), 505–510. https://doi.org/10.1121/1.413004

Henrich, N., C. d'Alessandro, & B. Doval. 2001. Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data. In: P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan (eds.), *Proceedings of the 7th European Conference on Speech Communication and Technology*, 3–7 September, 2001, Aalborg, Denmark, 47–50.

Henrich, N., C. d'Alessandro, B. Doval, & M. Castellengo. 2004. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America* 115(3), 1321–1332. https://doi.org/10.1121/1.1646401

Henrich, N., C. d'Alessandro, B. Doval, & M. Castellengo. 2005. Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *The Journal of the Acoustical Society of America* 117(3), 1417–1430. https://doi.org/10.1121/1.1850031

Ishi, C. T. & J. Arai. 2018. Periodicity, spectral and electroglottographic analyses of pressed voice in expressive speech. *Acoustical Science and Technology* 39(2), 101–108. https://doi.org/10.1250/ast.39.101

Ishi, C.T., H. Ishiguro, & N. Hagita. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50(6), 531–543. https://doi.org/10.1016/j.specom.2008.03.009

Ishi, C.T., T. Minato. & H. Ishiguro. 2019. Analysis and generation of laughter motions, and evaluation in an android robot. *APSIPA Transactions on Signal and Information Processing*, 8, e6. https://doi.org/10.1017/ATSIP.2018.32

Klatt, D. H. & L. C. Klatt. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87(2), 820–857. https://doi.org/10.1121/1.398894

Konishi, M. 2017. Differences in words used by learners of Japanese and native speakers: I-JAS comparison using different tasks. *NINJAL Research Papers* 13, 79–106.

Maekawa, K. & H. Mori. 2017. Comparison of Voice Quality between the Vowels in Filled Pauses and Ordinary Lexical Items. *Journal of the Phonetic Society of Japan* 21(3), 53–62. https://doi.org/10.24467/onseikenkyu.21.3_53

Ogden, R. 2001. Turn-holding, turn-yielding and laryngeal activity in Finnish talk-in-interaction. *Journal of the International Phonetics Association* 31(1), 139–152. https://doi.org/10.1017/S0025100301001116

Shriberg, E. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 30 (1), 153–169. https://doi.org/10.1017/S0025100301001128

Takamura, M. 2012. Nihongo gakushuu-sha no inta-byuuno hatsuwa ni okeru po-zu to fira- no tokuchou [Characteristics of Pauses and Fillers in Utterances of Japanese-learning Interviewees]. *Hikaku Bunka Kenkyuu* [Studies in Comparative Culture] 102, 101–111.

Timcke, R., H. von Leden, & P. Moore. 1958. Laryngeal vibrations: Measurements of the glottis wave. *AMA Archives of Otolaryngology* 68 (1), 1–19. https://doi.org/10.1001/archotol.1958.00730020005001

Wakasa, K., H. Terasawa, H. Kawahara, & K. Sakakibara. 2018. Seili Ongkyoteki tokutyo-ryo-bunseki niyoru operaggasyo- to gassyo-kasyo- no hikakukento [Comparison between operatic singing and choral singing by physiological and acoustic feature quantity analysis]. In: *Nihon onkyou gakkai kennkyuu happyoukai kouenn ronnbunnshuu (shuuki)* [Proceedings of Acoustic Society of Japan (Fall)], 12–14 September, 2018, Oita City, Oita, Japan, 1121–1124.

Attached filled pauses: Occurrences and durations

Mária Gósy [1] and Vered Silber-Varod [2]
[1]Hungarian Research Centre for Linguistics and ELTE University, Budapest, Hungary
[2]The Open University of Israel, Israel

Abstract

Filled pauses may reveal speech planning or execution problems that result in various positional and temporal patterns in spontaneous utterances. The purpose of this study was to analyze the position of the vocalic FPs, with respect to an adjacent word, in terms of occurrences and their durations produced by young (mean age: 25 years) and elderly (mean age: 76 years) speakers of Hungarian (a total of 32 participants). Elderly speakers produced significantly less and longer vocalic FPs than young speakers did. Both the occurrences and durations were significantly influenced by position of FPs and by age. In this paper, we introduced the conception of a functional difference between FPs attached either to the preceding or to the following word. The findings indicated different ways of resolving speech planning or execution problems depending on age.

Introduction

The phenomenon identified by the term 'filled pause' (henceforward FP) has been known and studied for about 60 years (Maclay & Osgood, 1959; Shriberg, 2001; Fox Tree, 2002; Corley & Stewart, 2008, etc.). In Hungarian, [ø]-like or [ə]-like vowels are the most frequent types to fill pauses in spontaneous utterances (e.g. Gósy et al., 2014). FPs were found to have various functions. They provide extra time for the speaker to aid speech planning and execution, monitoring and repair, as well as they signal conversational turns or occurring as discourse markers, etc. (e.g. Smith & Clark, 1993; Fox Tree, 2002; Watanabe et al., 2008; Finlayson & Corley, 2012; Urizar & Samuel, 2014). There are, however, serious difficulties when one tries to identify, separate or categorize these functions (Cutler, 1988).

There is a specific property of FPs that they can occur either between two silent pauses or attached to (co-articulated with) a word. FPs may occur inserted before the first segment of the word (FPword position) or after the last segment of the word (wordFP position). Clark and Fox Tree (2002) termed these two positions as cliticization. There are a limited number of papers focusing on the positioning of FPs related to the neighboring words in the literature. Speakers were reported to attach a FP onto a previous word, but not onto a following in native (British) English speech (Clark & Fox Tree, 2002). However, FPs were more frequent between a lexical item and a silent pause than between two silent pauses in another study also in (British) English speech (de Leeuw, 2007). Dutch and German speakers seemed to behave differently in positioning FPs; attached FPs were found to be common in Dutch while there was no difference in the positions of FPs in German (de Leeuw, 2007). Silber-Varod and colleagues (2016) found that in Hebrew, attached FPs are more common than FPs between silent pauses, and enclitic FPs (wordFP) are more common than proclitic FPs (FPword). The articulation gesture of attaching a FP to a word is easy to perform and provides a kind of concealment of the FP (and the speaker's difficulty) since it is not flashy in these positions.

Findings about the occurrences and durations of FPs in various age groups seem to be controversial (e.g. Bortfeld et al., 2001; Searl, Gabel, & Fulks, 2002). Some studies reported that elderly people used a larger number of FPs as opposed to young adults (Bortfeld et al., 2001; Roggia, 2012). Kemper (1992) found that old-old speakers (ages between 75 and 90) produced more FPs than young-old speakers (ages between 60 and 74) did. In contrast, other studies did not find such differences (Leeper & Culatta, 1995; Bóna, 2014; Gósy et al., 2014). Bóna (2011) found that as soon as the topic of the narrative became more challenging, young subjects produced a higher number of FPs than old subjects did. Emotional stress seemed to influence elderly speakers' pausing more than those of young ones (Caruso, McClowry, & Ludo, 1997).

The durational range of FPs is wide, and there are a great many factors that influence the measured values (average and contextual speech rate, thinking speed, difficulty of the topic to be discussed, etc.). The mean durations of FPs are reported to range from about 100 ms to about 750 ms or even longer (e.g. Shriberg, 2001; Clark & Fox Tree, 2002; Merlo & Barbosa, 2010; de Jong & Bosker, 2013). FPs' durations were shown to increase by age in some studies (e.g. Pindzola, 1990; Kemper, 1992) while others did not support

https://doi.org/10.18463/diss-2021-013-gosy-silber_varod

significant differences between young and old speakers in this respect (e.g. Horton, Spieler, & Shriberg, 2011; Bóna, 2014; Gósy et al., 2014). Durations of FPs occurring between two silent pauses were significantly longer than those occurring between a lexical item and a silent pause in two middle-aged Hungarian-speaking speakers' spontaneous utterances (Gósy, 2015).

We are of the opinion that more clear patterns can be found on the (sometimes controversial) behavior of FPs if the factor of their immediate position is also considered. The main body of the present study addresses whether occurrences and durations of FPs show differences depending on age (young and old speakers) and positions with regards to adjacent words and silent pauses in Hungarian. We hypothesized that (i) the occurrences of FPs would show significantly different patterns depending on age, (ii) the proportions of FPs in various positions would show significant differences, (iii) the durations of FPs would show significant differences depending on age, and (iv) the durations of FPs would show significant differences depending on their positions.

Methodology

Thirty-two spontaneous narratives produced by native Hungarian speakers (half of them were females) were randomly selected (with the exception of age and gender criteria) from the BEA Hungarian speech database (Gósy, 2012). Two distinct age groups were formed: (i) young speakers (aged between 22 and 28 years; mean = 25 years) and (ii) old speakers (aged between 70 and 80 years; mean = 76 years). The participants were asked to speak about their life and about their opinions on topics of current interest provided by the interviewer (who was the same person across all recordings). The mean speech rate was 4.4 syllables/s in young speakers while 3.8 syllables/s in old speakers.

Recordings were made in the same sound-attenuated room, under identical technical conditions using an Audiotechnica AT4040 cardioid condenser microphone connected directly to a computer using GoldWave to record samples at 44.1 kHz, 16 bits, monaurally. For the present study, more than 4.5 hours of speech samples from the database were used. The duration of recording per subject was 8.5 minutes (*std. dev.* = 0.3 minutes).

The speech material was manually annotated focusing on vocalic FPs (variants of the [ø] vowel and the [ə] neutral vowel). FPs were marked by öö while silent pauses were marked by SIL in annotations. Silent periods may contain also breath noise (Trouvain, Fauth, & Möbius, 2016). The positions of vocalic FPs were also coded as between silent pauses or silence on one side and lexeme on the other. Transcription was done in Praat (Boersma & Weenink, 2021). Occurrences of all FPs were analyzed according to the three possible positions. There are two instances where FPs are attached to the lexical items. FP can be attached to the first segment of the word (this is the FPword position) after a silent pause. FP can be attached to the last segment of the word (this is the wordFP position) and is followed by a silent pause. The third position occurs when FP is surrounded by silent pauses on both sides (this is the silFPsil position). (The occurrences of FPs surrounded by two words were extremely rare, thus, they were excluded from the analysis.) Examples: (i) *le akartam fényképezni a* SIL *ööhegyeket* ('I wanted to take pictures /of/ SIL öömountains'); (ii) *tehátöö* SIL *furcsa helyzet van* ('wellöö SIL there is a peculiar situation'); *probléma hogy ez* silFPsil *nehéz feladat* ('the problem is that this /is/ silFPsil /a/ difficult task').

A total of 1,068 FPs (284 of them were produced by the old and 784 by the young speakers) were found in the speech material. The total number of silent pauses were 1271, out of 904 were found in young while 367 in old speakers. The items of silFPsil type consisted of 240 silent pauses in young and 166 silent pauses in old speakers. Figure 1 shows spectrograms of FPs in the attached positions.

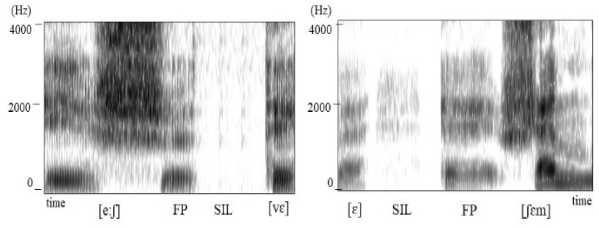

Figure 1. Speech fragments where FP is followed (left) and preceded (right) by a silent pause (containing some breath) and is attached to the first consonant [ʃ] of the following word sem *'neither' (right) and to the last consonant [ʃ] of the preceding word* és *'and' (left).*

The duration was measured as the interval (i) between the onset and offset of the second formants of the vocalic FP occurring between silent pauses, and (ii) between the onset/offset of the second formant of the vocalic FP and the onset/offset of the preceding and following segment based on traditional criteria. Durations were extracted using a specific Praat script. All inter-lexical pauses (Zellner, 1994) were considered. The shortest

silent pause in the vicinity of FPs was 40 ms. Prolongations and outlier data were excluded from (further) analysis.

To test statistical significance, MANOVA was performed on durations of vocalic FPs as dependent factors. As fixed effects, we entered 'age group', and 'position' into the model. Chi-Square and Mann–Whitney *U* tests were performed to analyze occurrences of FPs. In all cases, the confidence level was set at the conventional 95%.

Results

Occurrences

Considering all FPs, we found 5.7 incidents per minute in young speakers while 2.1 incidents per minute in old speakers. Statistical analysis revealed significant differences depending on both 'position' and 'age' (for position: *Chi-Square* = 131.511; $p < 0.001$; for age: *Chi-Square* = 234.082; $p < 0.001$). Young speakers produced the incidents of wordFP type in 3.1 per minute, while they produced the incidents of FPword type less frequently (1.8 incidents per minute). The incidents of the type silFPsil occurred the least frequently in their speech samples (0.9 incidents per minute).

Old speakers produced the incidents of FPword type in 0.8 per minute while there were no significant differences in occurrences of the incidents of wordFP and silFPsil types in their case (0.6 incidents per minute in both cases: Mann–Whitney *U* test: $Z = 0.957$, $p > 0.05$).

The distribution of FPs according to position, within an age group, showed significant differences in both young (*Chi-Square* = 179.429; $p < 0.001$) and elderly speakers (*Chi-Square* = 7.232; $p < 0.027$). Figure 2 demonstrates the different ratios of occurrences according to FP types expressed in percentages in both age groups.

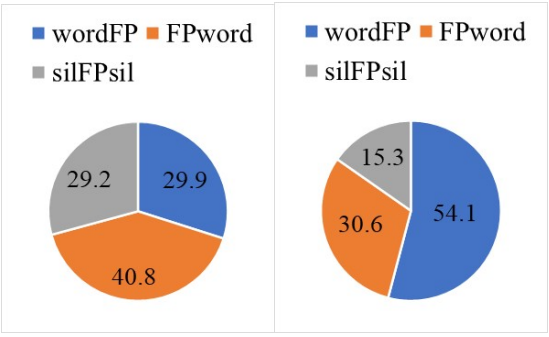

Figure 2. Occurrences of FPs (%) according to the three position types (left side: old speakers, right side: young speakers).

Durations

Young speakers produced significantly shorter FPs compared to old speakers (*mean* = 411 ms; *std. dev.* = 309 ms versus *mean* = 479 ms; *std. dev.* = 207 ms, respectively). Statistical analysis revealed significant differences in the durations of FPs depending on both factors of 'age' ($F(1, 1067) = 4.181$; $p = 0.041$; *partial* $\eta^2 = 0.004$) and 'position' ($F(1, 1067) = 37.905$; $p = 0.001$; *partial* $\eta^2 = 0.067$). In both groups, the shortest FPs were produced in FPword positions (for young speakers, *mean* = 341 ms, *std. dev.* = 353 ms; for old speakers, *mean* = 408 ms, *std. dev.* = 175 ms) while the longest ones were produced in silFPsil positions (for young speakers, *mean* = 606 ms, *std. dev.* = 286 ms; for old speakers, *mean* = 575 ms, *std. dev.* = 178 ms). The durations of FPs in wordFP positions were in between the other two types, both in young as well as in old speakers (for young speakers, *mean* = 395 ms, *std. dev.* = 264 ms; for old speakers, *mean* = 481 ms, *std. dev.* = 237 ms). Post hoc Tukey test revealed significant differences in durations of FPs, depending on positions in all comparisons ($p < 0.005$). The interaction of age and position was not statistically significant ($F(1, 1067) = 2.968$; $p = 0.052$; *partial* $\eta^2 = 0.006$). The durations of FPs in various positions seem to form the same patterns irrespective of age (Figure 3).

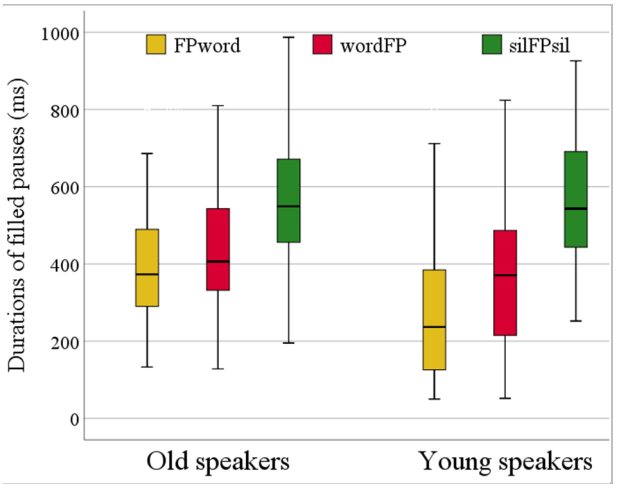

Figure 3. Durations of FPs in the three positions produced by young and old speakers (quartiles, medians).

The durations of the silent pauses produced by both young and old speakers preceding FPs were significantly different according to 'position' but not to 'age' ($F(2, 1067) = 249.78$, $p = 0.001$; *partial* $\eta^2 = 0.320$; $F(1, 1067) = 3.470$; $p = 0.063$; *partial* $\eta^2 = 0.002$, respectively). Post hoc Tukey tests revealed significant differences in all cases

($p < 0.05$). The interaction of 'position' and 'age' was not significant ($F(2, 1067) = 1.090$, $p = 0.337$). The durations of the silent pauses produced by both young and old speakers following FPs were significantly different according to both 'position' and 'age' ($F(2, 1067) = 79.770$, $p = 0.001$; *partial* $\eta^2 = 0.131$; $F(1, 1067) = 13.534$; $p = 0.001$; *partial* $\eta^2 = 0.02$, respectively). Post hoc Tukey tests revealed significant differences in all cases ($p < 0.05$). The interaction of position and age was significant ($F(2, 1067) = 4.201$, $p = 0.008$, *partial* $\eta^2 = 0.008$).

Silent pauses were shorter before FPword position than after wordFP position in young speakers; however, the opposite could be found in old speakers (Table 1). In silFPsil positions, the first silent pauses were significantly longer than the second ones only in old speakers ($F(1, 165) = 13.922$, $p = 0.001$; *partial* $\eta^2 = 0.078$; $F(1, 239) = 0.934$, $p = 0.335$; *partial* $\eta^2 = 0.004$, respectively).

Table 1. Mean durations of silent pauses according to positions (given in ms, std. dev. values are in brackets).

Age group	silFPword	wordFPsil	sil_1FP	$FPsil_2$
Young	484 (384)	537 (481)	725 (478)	665 (549)
Old	438 (295)	386 (327)	631 (505)	383 (330)

Conclusions

The goal of this study was to obtain information on occurrence of positions and durations of vocalic FPs produced by young and old Hungarian-speaking speakers. Young speakers produced close to twice as many FPs in their utterances than old speakers did. Our first hypothesis was confirmed. These findings add to the controversial data of the literature mentioned earlier (e.g. Bortfeld et al., 2001; Leeper & Culatta, 1995; Searl et al., 2002; Gósy et al., 2014). The attached FPs were all longer and less frequent in old speakers while they were shorter and more frequent in young speakers. However, practically no difference was found in the durations of FPs occurring between two silent pauses between young and old speakers. This finding suggests that these FPs might function as discourse markers with distinct traits from attached FPs. As such, both the occurrences and durations of FPs showed significantly different patterns, depending on age.

Our explanation for the findings comprises two aspects: Speech planning differences between young and old speakers as well as different speaking routines. The different speech rates of the young an old speakers may also contribute to temporal differences. Old people are supposed to activate their thoughts at a slower speed compared to young speakers. Thus, they might not often need extra time for selection of thoughts. The simultaneity of activation and selection of thoughts together with old speakers' simpler grammatical structures (see Horton et al., 2011) seem to be more transparent and more easily managed. Young speakers are supposed to activate numerous thoughts at the same time at a high speed that require continuous activation and selection of thoughts followed by transforming them into grammatical forms. These tasks require extra time. Obviously, old speakers have more routine in verbal communication including well-tried strategies as opposed to young speakers. We assumed that the proportions and durations of FPs in various positions would show significant differences which was confirmed. In FPword position, the speaker tries to solve the problem during the silent pause that precedes FP; however, this amount of time is not enough, therefore, the speaker starts producing a FP which increases the necessary time to continue (Maschler, 2001). The more frequent occurrence of this phenomenon in old speakers reflects that they are not able to solve the problem during speech planning, consequently they are in need for more time. In wordFP position, the speaker anticipates some problem with the continuation during the (last) word production. Young speakers' monitoring works better and faster than those of old speakers that provides an early identification of the problem. To gain extra time, the speaker lengthens the word by means of coarticulating a FP followed by a silent pause. This strategy is similar to that of segment prolongation phenomenon. If the silent pause+FP combination is not sufficient for problem solving, another silent pause is added (silFPsil position).

We suggest the conception of a functional difference between the FPword and wordFP positions. FPword position signals a speech planning problem while wordFP position signals the repair that is happens during that time. Speech planning problem means re-selection and/or re-organization of thoughts that needs longer time to perform as opposed to repairing obvious errors. This interpretation is supported by durations since FPs in the FPword positions were shorter than those in the wordFP positions. FPs surrounded by two silent pauses may signal both the planning problem (first silent pause and FP) as well as the momentary inability to repair it (FP and second silent pause) or some other processing strategy. The longer durations of FPs surrounded by silent pauses compared them to those of attached FP

types support this interpretation. Thus, we arrived at the conclusion that the FP position has a functional definiteness.

The data highlight some temporal equalization in the case of the attached FPs in both age groups. The mean durations of the silent pause and FP combinations are around 850 ms, indicating the inner control over the speech planning disharmony by the speaker. The silFPsil combinations show the tendency that the silent pauses following FPs are shorter than those preceding FPs. This tendency can be explained by the fact that FPs signal the (near-)resolution of the problem-solving process.

Our findings call attention to the interrelations of immediate positions and durations of FPs in spontaneous utterances that provide a better understanding of the surface effects of the speakers' speech planning difficulties.

Acknowledgements

This research has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development, and Innovation Fund, financed under the ELTE TKP2020-IKA-06 funding scheme.

References

Boersma, P. & D. Weenink. 2021. Praat: doing phonetics by computer (version 6.1.40). http://www.praat.org/ (accessed 27 February 2021).

Bóna, J. 2011. Disfluencies in the spontaneous speech of various age groups: Data from Hungarian. *Govor* 28, 95–115.

Bóna, J. 2014. Temporal characteristics of speech: The effect of age and speech style. *Journal of the Acoustical Society of America* 136(2), EL116–121. https://doi.org/10.1121/1.4885482

Bortfeld, H., S. D. Leon, J. E. Bloom, M. F. Schober, & S. E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech* 44(2), 123–147. https://doi.org/10.1177/00238309010440020101

Caruso, A. J., T. M. McClowry, & M. Ludo. 1997. Age-related effects on speech fluency. *Seminars in Speech and Language* 18, 171–180. https://doi.org/10.1055/s-2008-1064071

Clark, H. H. & J. E. Fox Tree, 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84(1), 73–111. https://doi.org/10.1016/s0010-0277(02)00017-3

Corley, M. & O. W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: the meaning of *um*. *Language and Linguistics Compass* 2(4), 589–602. https://doi.org/10.1111/j.1749-818x.2008.00068.x

Cutler, A. 1988. The perfect speech error. In: L. M. Hyman & C. N. Li (eds.): *Language, speech, and mind: Studies in honor of Victoria A. Fromkin*. London, UK: Routledge, 209–223.

Finlayson, I. R. & M. Corley. 2012. Disfluency in dialogue: An intentional signal from the speaker? *Psychonomic Bulletin & Review* 19(5), 921–928. https://doi.org/10.3758/s13423-012-0279-x

Fox Tree, J. E. 2002. Interpreting pauses and ums at turn exchanges. *Discourse Processes* 34(1), 37–55. https://doi.org/10.1207/s15326950dp3401_2

Gósy, M. 2012. BEA – A multifunctional Hungarian spoken language database. *The Phonetician* 105/106, 50–61.

Gósy, M. 2015. Beszédtervezési diszharmónia és a kitöltött szünetek összefüggései. [Interrelations of speech planning disharmony and FPs] *Magyar Nyelvőr* 139(4), 436–449.

Gósy, M., J. Bóna, A. Beke, & V. Horváth. 2014. Phonetic characteristics of filled pauses: the effects of speakers' age. *10th ISSP*, Cologne, 150–153.

Horton, W. S., D. H. Spieler, & E. Shriberg. 2011. A corpus analysis of patterns of age-related change in conversational speech. *Psychology of Aging* 25(3), 708–713. https://doi.org/10.1037/a0019424

de Jong, N. H. & H. R. Bosker, 2013. Choosing a threshold for silent pauses to measure second language fluency. In: Eklund, R. (ed.): *Proceedings of DiSS 2013: The 6th Workshop on Disfluency in Spontaneous Speech*, 21–23 August, 2013, Stockholm, Sweden, 17–20.

Kemper, S. 1992. Language and aging. In: F. I. M. Craik & T. A. Salthouse (eds.), *The Handbook of Aging and Cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 213–270. https://doi.org/10.1080/01924780903295796

Leeper, L. H. & R. Culatta. 1995. Speech fluency: Effect of age, gender and context. *Folia Phoniatrica et Logopedia* 47, 1–14. https://doi.org/10.1159/000266337

de Leeuw, E. 2007. Hesitation markers in English, German, and Dutch. *Journal of Germanic Linguistics* 19(2), 85–114. https://doi.org/10.1017/s1470542707000049

Maclay, H. & C. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15(1), 19–44. https://doi.org/10.1080/00437956.1959.11659682

Maschler, Y. 2009. *Metalanguage in interaction: Hebrew discourse markers*. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.181

Merlo, S. & P. A. Barbosa. 2010. Hesitation phenomena: a dynamical perspective. *Cognitive Processing* 11(3), 251–261. https://doi.org/10.1007/s10339-009-0348-x

Pindzola, R. H. 1990. Dysfluency characteristics of aged, normal-speaking black and white males. *Journal of Fluency Disorders* 15(4), 235–243. https://doi.org/10.1016/0094-730X(90)90004-C

Roggia, A. B. 2012. Eh as a polyfunctional discourse marker in Dominican Spanish. *Journal of Pragmatics* 44(13), 1783–1798. https://doi.org/10.1016/j.pragma.2012.08.010

Searl, J P., R. M. Gabel, & J. S. Fulks, 2002. Speech disfluency in centenarians. *Journal of Communication Disorders* 35(5), 383–392. https://doi.org/10.1016/S0021-9924(02)00084-9

Shriberg, E. 2001. To '*errrr*' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetics Association* 31(1), 153–169. https://doi.org/10.1017/s0025100301001128

Silber-Varod, V., H. Kreiner, R. Lovett, Y. Levi-Belz, & N. Amir. 2016. Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. In: J. Barnes, A. Brugos, S. Shattuck-Hufnagel & N. Veilleux (eds.), *Proceedings of Speech Prosody 2016 (SP2016)*, 31 May–3 June, 2016, Boston, MA, USA, 1211–1215. https://doi.org/10.21437/SpeechProsody.2016-249

Smith, V. L. & H. H. Clark, 1993. On the course of answering questions. *Journal of Memory and Language* 32(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Trouvain, J., C. Fauth, & B. Möbius. 2016. Breath and non-breath pauses in fluent and disfluent phases of German and French L1 and L2 read speech. In: J. Barnes, A. Brugos, S. Shattuck-Hufnagel & N. Veilleux (eds.), *Proceedings of Speech Prosody 2016 (SP2016)*, 31 May – 3 June, 2016, Boston, MA, USA, 31–35. https://doi.org/10.21437/speechprosody.2016-7

Urizar, X. & A. G. Samuel. 2014. A corpus-based study of fillers among native Basque speakers and the role of zera. *Language and Speech* 57(3), 338–366. https://doi.org/10.1177/0023830913506422

Watanabe, M., K. Hirose, Y. Den, & N. Minematsu. 2008. FPs as cues to the complexity of upcoming phrases for native and non-native listeners, *Speech Communication* 50(2), 81–94. https://doi.org/10.1016/j.specom. 2007. 06.002.

Zellner, B. 1994. Pauses and the temporal structure of speech. In: E. Keller (ed.), *Fundamentals of speech synthesis and speech recognition*. Chichester, UK: John Wiley, 41–62.

Gestures in fluent and disfluent cycles of speech: What they may tell us about the role of (dis)fluency in L2 discourse

Loulou Kosmala
Sorbonne Nouvelle University, Paris, France

Abstract

The present study looks at the production of gestures in fluent versus disfluent speech in L1-L2 interactions, following Graziano and Gullberg (2013, 2018). The aim of this paper is twofold: first to argue against the Lexical Retrieval Hypothesis (Krauss, Chen, & Gottesman, 2000) by comparing the distribution and function of gestures in fluent versus disfluent speech; second, to closely examine the unfolding of embodied (dis)fluencies, where vocal and visual-gestural actions are coordinated and situated within word searching sequences. The analyses are conducted on a video-recorded corpus of semi-spontaneous interactions between French and American speakers in tandem settings. Overall, our results support Graziano and Gullberg's (2018) findings, and show that gestures accompanying (dis)fluencies are not necessarily related to lexical difficulties. Additionally, the qualitative analyses highlight the interactional and multimodal role of (dis)fluencies, which offers a fresh perspective of these phenomena which have often been treated from an internal production perspective.

Background

In the field of Second Language Acquisition (SLA) one major question regarding gesture use is whether it can help learners resolve speech difficulties. When learners experience difficulties, their speech usually becomes filled with a number of disfluencies, which have often been viewed as indications of time out during verbal planning (Goldman-Eisler, 1958) and which are usually frequent during lexical searches (Stam, 2001). Additionally, it has been proposed that manual gestures arise when speakers experience lexical problems, and that these gestures can help facilitate word finding (Beattie & Butterworth, 1979; Krauss & Hadar, 1999). Gestures, which carry "the full expressive burden of a language" (Gullberg, 2011, 139) can thus compensate for lexical shortcomings. Indeed, it has been shown that L2 learners are likely to produce more gestures in their L2 than in their L1, to overcome "a lack of skill" in their target language (Gullberg, 1998), and according to the Lexical Retrieval Hypothesis, (henceforth LHR, Krauss, Chen, & Gottesman, 2000) word findings are more successful when accompanied by referential gestures, as they facilitate access to lexical memory. All this evidence seems to suggest that gestures help compensate for the lack of speech. However, for gestures to be truly compensatory, it would mean that they would have to occur *during* speech perturbations (i.e. disfluencies), which is rarely the case (Chui, 2005; Graziano & Gullberg, 2018; Kosmala, Candea, & Morgenstern, 2019; Yasinnik, Shattuck-Hufnagel, & Veilleux, 2005). Some studies have shown that gestures tend to be suspended prior to speech suspension (Seyfeddinipur & Kita, 2001) while others have stated that they were also likely to begin during pauses (Beattie & Butterworth, 1979). As Graziano and Gullberg (2018) point out, studies on gesture and disfluency production have led to contradictory findings, and the observation that gestures are more likely to occur with fluent rather than disfluent speech makes it difficult to assess theories such as the Lexical Retrieval Hypothesis. In order to address these issues, they conducted a study on the gestural behavior of different groups of speakers (competent L1 speakers, adult and child L2 learners), during fluent and disfluent speech in oral narratives in Dutch and Italian, and contrary to the LRH, their results showed that all groups produced not only referential gestures which can facilitate lexical search, but also pragmatic gestures that are not related to semantic content, but rather offer metalinguistic comments. Moreover, gestures were shown to occur significantly more in fluent than disfluent stretches of speech, and gestures tended to be held during disfluent speech. These results show a synchronicity between speech and gesture suspension, which suggests a very tight link between fluent speech and gesture production.

But gestures and disfluencies are used for so much more than simply to look for a word or to solve production difficulties. They can be used to elicit an answer from an interlocutor, or filling a sentence (Stam & Tellier, 2017; Tellier, Stam, & Bigi, 2013). Disfluencies are not only self-directed and production oriented as they can also positively contribute to the co-construction of meaning, for example through embodied completions. This practice can be defined as a completion of an action, previously initiated, through "gesture or embodied display" (Mori & Hayashi, 2006). In a study of

https://doi.org/10.18463/diss-2021-014-kosmala

interactions between L2 learners of Japanese, Mori and Hayashi (2006) have demonstrated the way L1 and L2 speakers coordinate their talk through gestures and embodied completions in the context of L2 use. As Rydell (2019) argues, searching for a word is not only an internal process resulting from language difficulties, it is also an embodied visible activity (Goodwin & Goodwin, 1986) which can be collaboratively negotiated by two or more speakers with the help of gaze and gesture. The same applies to disfluencies, and this paper will show that they are not only the result of internal processes such as verbal planning, but that they also actively participate in the unfolding of interactional sequences. Additionally, they can be used by speakers to display to one another whether they are engaging or disengaging in the interaction (Goodwin & Goodwin, 2004) by adjusting their body and talk. We thus adopt the view of disfluency as a fully multimodal (Kosmala et al., 2019) but also highly contextualized and situated phenomenon. Therefore, we will use the term "(dis)fluency" in this paper (Crible et al., 2019; Götz, 2013; Kosmala, 2021) in order to stress the fact that (dis)fluencies are not necessarily disruptive and associated with internal production difficulties, but that they can also embody more communicative and fluent actions. However, the interactional dimension of (dis)fluencies has received little attention in SLA disfluency research (except for a few exceptions, e.g. McCarthy, 2009; Peltonen, 2019, among others) and this may be due to theoretical and methodological differences. Most (but not all) studies conducted on (dis)fluencies are quantitative and come from a psycholinguistics background (e.g. Levelt, 1989; Seyfeddinipur, 2006; Shriberg, 1994, among others) which do not present them in specific situated interactional sequences. We believe that a close examination of talk within situated human interaction is a key addition to the quantitative treatment of (dis)fluencies.

Therefore, the aim of this paper is twofold: first, to compare the distribution and function of gestures in L1 and L2 fluent and disfluent speech of French and English in order to examine gestural behavior and test the LHR (in line with Graziano & Gullberg, 2018) second, to illustrate the unfolding of embodied (dis)fluencies in situated interactional sequences and to highlight their multimodal and interactional dimension.

Data Analysis

The data and the excerpts of interactions to be examined in this paper are taken from recordings of the SITAF Corpus (Horgues & Scheuer, 2015) which comprises interactions between French learners of English and American learners of French studying at Sorbonne Nouvelle University in France. The students (undergraduate level) were part of a tandem exchange program and had the opportunity to meet once a month to interact in their target language with their tandem partner. They were video recorded twice in a three-month interval throughout the academic year. The recordings selected for the present study are taken from 12 subjects (6 American speakers and 6 French speakers) who were engaged in argumentative tasks (they had to debate on a topic written on a piece of paper) in tandem settings (alternating from their L1 to L2). The interactions lasted 3–5 minutes on average, and the total duration of the data sample is of approximately 53 minutes. In line with Graziano and Gullberg (2018), we looked at both fluent and disfluent stretches of speech. Disfluent stretches of speech include the following (dis)fluencies (see Kosmala, 2021): filled pauses (uh and um), silences, self-repairs, repetitions, restarts, truncated words, non-lexical sounds (e.g. inbreath and clicks), and word/syllable prolongations.

The methodology used for the quantitative analyses of this study is adapted from Graziano and Gullberg (2018) which coded gesture phrases (preparation, stroke, suspension, retraction) for the gestures co-occurring with (dis)fluencies, and the functions of all the gestures, mainly *pragmatic* (which are related to aspects of utterance structure, speech acts, turn-taking mechanisms) and *referential* (which convey semantic content). The qualitative analysis will be presented with reference to interactional patterns (i.e. turn taking mechanisms, sequential organization) and the visual-gestural modalities of the talk (gesture, gaze behavior, and body orientation).

Quantitative results

A total of 623 gestures and 1903 (dis)fluencies were coded in the data (for a closer look at the distribution of (dis)fluencies see Kosmala, 2021). The distribution of all the gestures in L1 and L2 (both fluent and disfluent speech) showed that speakers produced on average 7.2 gestures per hundred words in their L1 ($N = 300$), and 10.2 per hundred words in their L2 ($N = 323$), which was found to be statistically significant ($LL = 210.03$ $p < .0001$).

For the distribution of gesture strokes in fluent versus disfluent stretches of speech in L1 and L2, results (see Figure 1) showed that gestures occurred predominantly more during fluent stretches of speech (77%, $N = 231/300$ in L1, and 60%, $N = 194/323$, in L2) than during disfluent ones overall (23%, $N = 69/300$ in L1 and 40% $N = 129/323$ in L2), but more gestures were found

during disfluent speech in L2 ($N = 129/294$), than in L1 ($N = 69/300$; $z = 4.58$, $p < .0001$).

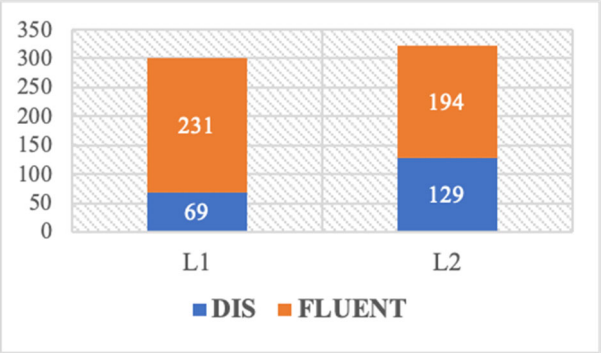

Figure 1. Proportion of gestures in fluent and disfluent cycles of speech (raw values)

These findings overall suggest a higher gestural activity in L2 than in L1, which supports previous work (e.g. Gullberg, 1998).

Results further showed that a majority of pragmatic gestures were found both in fluent and disfluent speech in L1 and L2 (Figure 2).

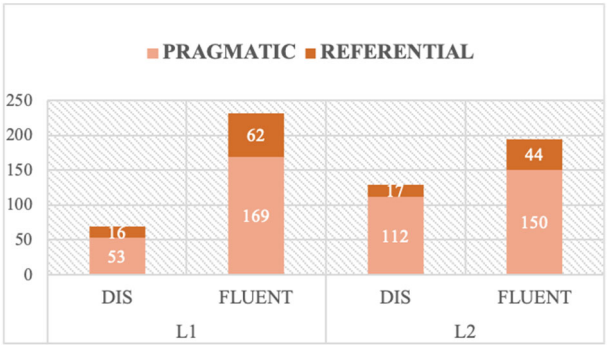

Figure 2. Proportion of pragmatic and referential gestures in L1 and L2 speech (raw values)

The proportion of referential gestures was slightly higher in fluent (26%, $N = 62/231$) than in disfluent stretches of speech (23%, $N = 16/69$) in L1 ($z = 0.6$, $p = 0.03$), as well as in L2 disfluent speech (13% $N = 17/129$) versus fluent speech (22% $N = 44/194$), but the latter did not reach significance ($z = 2{,}1$, $p > 0.05$). In addition, no significant differences were found between the distribution of referential gestures in disfluent L1 (23%) and L2 (13%) despite numerical evidence ($z = 1.80$, $p = 0.1$), but there were slightly more referential gestures during fluent speech in L1 (26%) than in L2 (22%) $z = 0.98$, $p > 0.05$. Taken together, these results do not support the view that speakers use more referential gestures to deal with lexical failures in their L2 (Stam, 2001), since they also use a great deal of pragmatic gestures to provide metalinguistic comments on various aspects of the interaction, in line with Graziano and Gullberg (2018). In fact, pragmatic gestures may be used by L2 speakers to co-create fluency (see the notion of "confluence" in McCarthy, 2009) and manage turn-taking in discourse, and are thus not necessarily associated with lexical difficulties. This is illustrated in the following section.

Qualitative analysis

We shall now turn to the micro analysis of a small excerpt taken from the data. In the previous section, we compared the production of gestures in L1 and L2 in fluent versus disfluent stretches of speech in order to investigate the gestural behavior of L1 and L2 speakers with regard to the LRH. We shall now move to another level of analysis involving the study of talk-in-interaction. This micro-analysis will allow us to get a full account of the visible bodily practices embodying (dis)fluencies within situated tandem interactions.

The following example is taken from participants F13 (French) and A13 (American) who alternated between speaking their L1 and their L2. In tandem interactions, speakers alternate between their "expert" and "non-expert" status, which establishes a reciprocal state of friendly mutual assistance, inviting the participants to help one another (Horgues & Scheuer, 2015). This may also invite speakers to enact certain visible bodily actions in order to display the current state of the talk to their partner. This is a reference to the *Participation Framework* (Goodwin & Goodwin, 2004) whereby coparticipants demonstrate their forms of involvement in the course of the talk. In the following excerpt in French, the L2 speaker (A13) is talking about the kinds of sensitive topics that friends can have during a conversation, and he does not exactly find the words for it.

1. *A13: um mais (…) en même temps
um but (...) at the same time
on peu:ut vraiment si si:i (…) dans une groupe euh
we really ca:an if i:if (...) in a group uh
((left hand held+ looks up e.))
qui:i qui discutons de:es des choses politiques
who:o who talk abou:ut about political stuff
((left hand rotating))
2. *F13: [mm mm
((head nod))
*A13: ou des choses euh (…) quoi tu tu]
or things uh (...) like you you]
((left hand held)) ((left open palm extended f.))
3. *F13: [religieuses politiques les les choses un peu:u un peu tabou

[religious things the the things that are a little:e taboo
((cyclic gesture+ looks at A13 g.))
4. *A13: [oui tout ça.
[*yes all that.*
((shoulder shrug g.))

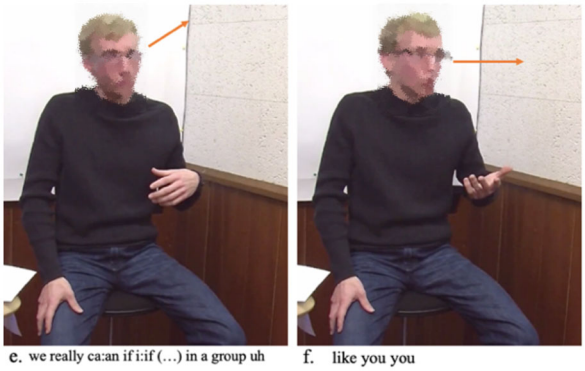

Figure 3. Left hand held then extended upwards towards F13 (turn 1)

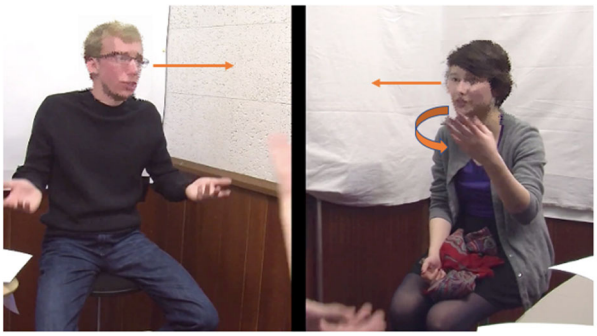

Figure 4. Shoulder shrug. State of mutual understanding

While A13 does not explicitly signal to his interlocutor that he is looking for a word, he still displays that his talk is currently being suspended, with the held gesture (Figure 3, e), and the different vocal (dis)fluencies (lengthening c*a:an i:i*f, silence and filled pause *uh*). After retrieving one noun phrase (*political stuff*) he initiates another one (*or things uh*) and finally shifts from his solitary word search activity to a joint one, by inviting his partner to take part in it. He does so by extending his left open palm (which was previously held) towards her (example of a *Palm Up Open Hand Gesture,* Müller, 2017 in Figure 2, f). This "offering" gesture (Streeck, 2009) appears to metaphorically hand over A13's current search to his partner, who joins it and offers a new lexical item (*religious things*). She also produces a cyclic gesture at the same time (Figure 4); these gestures can be used to express duration, continuity and process (Müller, 2017) and it appears here that she is producing it to ensure continuity between A13's previous utterance and her own. A state of mutual understanding is then accomplished when, A13, almost immediately after F13's subsequent turn, offers a positive assessment "yeah and all that", accompanied by a shoulder shrug which further displays his affiliation (Figure 4). In this case, the embodied (dis)fluencies emerged in a context of co-construction, which further supports the idea that word searches are not only internal activities associated with speech difficulties, but also collaborative ones that can be co-achieved (Rydell, 2019). This example has highlighted the multimodal and interactional dimension of (dis)fluencies which have often been viewed from a strictly verbal perspective. While L2 speakers tend not to gesture frequently during disfluent speech generally (41% of the time approximately), it is still essential to examine specific occurrences of embodied (dis)fluencies which can contribute to the building of interactional sequences. It further shows that gestural performance in L2 is not necessarily associated with lexical shortcomings, as speakers can make use of them, along with other bodily modalities (head movement, body orientation, gaze behavior), to offer metapragmatic comments on different aspects of the interaction.

Conclusion

The aim of this paper on gestures and (dis)fluencies was twofold: first to examine gestural distribution in fluent and disfluent cycles of speech in L1 and L2, following Graziano and Gullberg (2018) in order to test the Lexical Retrieval Hypothesis; second, to go beyond this level of analysis and analyze the occurrence of embodied (dis)fluencies captured in situated tandem interactions. We believe that a combination of quantitative and qualitative analyses enables us to paint a rich picture of these phenomena, by integrating different levels of analysis in different modalities (speech, gesture, and interaction level). Our quantitative findings showed similar results reported by Graziano and Gullberg (2018), mainly that gestures tend not to occur during disfluent speech, and that speakers produce a great number of pragmatic gestures and not only referential gestures during (dis)fluencies. Gestures can thus be seen as multimodal communication strategies (Gullberg, 2011) which consist in dealing with and solving lexical and interactional related difficulties encountered in speech. However, this does not necessarily imply that they are used to "compensate" lexical shortcomings, just like (dis)fluencies are not necessarily associated with production trouble. We believe that gestures can tell us a great deal about the role of (dis)fluencies in multimodal L2 discourse, uncovering interactional features that are not otherwise visible in speech only.

Acknowledgements

Many thanks to Aliyah Morgenstern and Maria Candea for re-reading a draft of this paper and for reviewing this work.

References

Beattie, G. W., & Butterworth, B. L. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, 22(3), 201–211. https://doi.org/10.1177%2F002383097902200301

Chui, K. 2005. Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, 37(6), 871–887. https://doi.org/10.1016/j.pragma.2004.10.016

Crible, L., A. Dumont, I. Grosman, & I. Notarrigo. 2019. (Dis)fluency across spoken and signed languages: Application of an interoperable annotation scheme. In: L. Degand, G. Gilquin, & A. C. Simon (eds.), *Fluency and Disfluency across Languages and Language Varieties*. Louvain-la-Neuve: Presses universitaires de Louvain, 17–39.

Goldman-Eisler, F. 1958. The predictability of words in context and the length of pauses in speech. *Language and Speech*, 1(3), 226–231. https://doi.org/10.1177%2F002383095800100308

Goodwin, C. & M. H. Goodwin. 2004. Participation. In: S. Duranti (Eds.), *A Companion to Linguistic Anthropology,* 222-224. https://doi.org/10.1002/9780470996522.ch10

Goodwin, M. & C. Goodwin. 1986. Gesture and coparticipation in the activity of searching for a word. *Semiotica*, 62(1–2), 51–76. https://doi.org/10.1515/semi.1986.62.1-2.51

Götz, S. 2013. *Fluency in native and nonnative English speech*, Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/scl.53

Graziano, M. & M. Gullberg. 2013. Gesture production and speech fluency in competent speakers and language learners. In: *Proceedings of the Tilburg Gesture Research Meeting (TiGeR)*, Tilburg, The Netherlands.

Graziano, M. & M. Gullberg. 2018. When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology* 9, 879. https://doi.org/10.3389/fpsyg.2018.00879

Gullberg, M. 1998. Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish. *Studies in Second Language Acquisition* 22(1), 122–123 https://doi.org/10.1017/S0272263100271052.

Gullberg, M. 2011. Multilingual multimodality: Communicative difficulties and their solutions in second-language use. In: J. Streeck, C. Goodwin, & C. LeBaron (eds.), *Embodied Interaction: Language and Body in the Material World*, Cambridge, UK: Cambridge University Press, 137–151.

Horgues, C. & S. Scheuer. 2015. Why some things are better done in tandem. In: J. A. Mompean & J. Fouz-González (eds.), *Investigating English Pronunciation*, New York, NY, USA: Springer, 47–82.

Kosmala, L. 2021. On the Specificities of L1 and L2 (Dis)fluencies and the Interactional Multimodal Strategies of L2 Speakers in Tandem Interactions. *Journal of Monolingual and Bilingual Speech*, 33(1), 69–101. https://doi.org/10.1558/jmbs.15676

Kosmala, L., M. Candea, & A. Morgenstern. 2019. Synchronization of (Dis)fluent Speech and Gesture: A Multimodal Approach to (Dis)fluency. In: A. Grimminger (ed.), *Proceedings of the 6th Gesture and Speech in Interaction (GESPIN)*, 11–13 September, 2019, Paderborn, Germany, 56–61.

Krauss, R. M., Y. Chen, & R. F. Gottesman. 2000. Lexical gestures and lexical access: A process. In: D. McNeill (ed.), *Language and Gesture*, Cambridge, UK: Cambridge University Press, 261–283. https://doi.org/10.1017/CBO9780511620850.017

Krauss, R. M., & U. Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. In: R. Campbell & L. Messing (eds.), *Gesture, speech, and sign*, Oxford, UK: Oxford University Press, 93–116. https://doi.org/10.1093/acprof:oso/9780198524519.003.0006

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

McCarthy, M. 2009. Rethinking spoken fluency. *Estudios de Lingüística Inglesa Aplicada (ELIA)*, 9, 11–29.

Mori, J. & M. Hayashi. 2006. The achievement of intersubjectivity through embodied completions: A study of interactions between first and second language speakers. *Applied Linguistics*, 27(2), 195–219. https://doi.org/10.1093/applin/aml014

Müller, C. 2017. How recurrent gestures mean: Conventionalized contexts-of-use and embodied motivation. *Gesture*, 16(2), 277–304. https://doi.org/10.1075/gest.16.2.05mul

Peltonen, P. 2019. Gestures as Fluency-enhancing Resources in L2 Interaction: A Case Study on Multimodal Fluency. In: P. Lintunen, M. Mutta & P. Peltonen (eds.), *Fluency in L2 Learning and Use*, Bristol, UK: Multilingual Matters, 111–128. https://doi.org/10.21832/9781788926317-010

Rydell, M. 2019. Negotiating co-participation: Embodied word searching sequences in paired L2 speaking tests. *Journal of Pragmatics*, 149, 60–77. https://doi.org/10.1016/j.pragma.2019.05.027

Seyfeddinipur, M. 2006. *Disfluency: Interrupting speech and gesture.* Ph.D. dissertation, Radboud University Nijmegen. https://doi.org/10.17617/2.59337

Seyfeddinipur, M., & S. Kita. 2001. Gesture as an indicator of early error detection in self-monitoring of speech. In: *Proceedings of DiSS '01: Disfluency in Spontaneous Speech*, 29–31 August, 2001, Edinburgh, Scotland, UK, 29–32.

Shriberg, E. E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California at Berkeley.Stam, G. 2001. Lexical failure and gesture in second language development. In: C. Cavé, I. Guaïtella, & S. Santi (eds.), *Oralité et gestualité: Interactions et comportements multimodaux dans la communication,* Paris, France: L'Harmattan, 271–275.

Stam, G., & M. Tellier. 2017. The sound of silence: The functions of gestures in pauses in native and non-native interaction. In: R. B. Church, M. W. Alibali, & S. D. Kelly (eds.), *Why Gesture?: How the Hands Function in Speaking, Thinking and Communicating*, Amsterdam, The Netherlands: John Benjamins, 353–377. https://doi.org/10.1075/gs.7.17sta

Streeck, J. 2009. *Gesturecraft: The manu-facture of meaning*, Amsterdam, The Netherlands: John Benjamins Publishing. https://doi.org/10.1075/gs.2

Tellier, M., G. Stam, & B. Bigi. 2013. Gesturing while pausing in conversation: Self-oriented or Partner-oriented?". In: *Proceedings of the Tilburg Gesture Research Meeting (TiGeR)*, Tilburg, The Netherlands.

Yasinnik, Y., S. Shattuck-Hufnagel, & N. Veilleux. 2005. Gesture marking of disfluencies in spontaneous speech. In: *Proceedings of DiSS '05: Disfluency in Spontaneous Speech Workshop*, 10–12 September, 2005, Aix-en-Provence, France, 173–178.

Categorical differences in the false starts of speakers of English as a second language: Further evidence for developmental disfluency

Simon A. Williams
University of Sussex, East Sussex, UK

Abstract

Although much is known about the formal properties of second language (L2) repair in general and error corrections in particular, less is known about other subtypes, here collectively referred to as false starts. Unlike L2 self-corrections, false starts are psycholinguistically more comparable with first language speaker equivalents and are of particular interest as possible sites of learner monitoring and modified output. Consistent with previous research on L2 repairs, this study found that lower-intermediate and advanced L2 speakers produced similar numbers of false starts. Their mapping by speaker proficiency level onto Levelt's (1989) model of speech production revealed that both groups were concerned with lexical and morphological false start repair but that lower-intermediate speakers produced more syntactic and advanced speakers more conceptual examples.

Motivation for the study

Of the various disfluencies in spontaneous speech, self-repairs are of particular interest as sites of language modification and possible intake. A number of distinctions and associated claims have been made depending on whether the repairs are self- or other-initiated (Shehadeh, 2001), correction or other imputed psycholinguistic function (Kormos, 1998), form or concept in nature (Zuniga & Simard, 2019). This study refers to Levelt's (1983) coding scheme for self-repairs, the basis for L2 adaptations (Kormos, 1998; van Hest, 1996), which describes four overt types: self-corrections ('error-repairs') and three others ('appropriacy', 'different', and 'rest'), which in the present study are collectively referred to as false starts. L2 self-corrections are defined here as the speaker's substitution of linguistic output regarded as non-standard with an alternative supposed to be standard.

Self-corrections are now relatively well understood as a distinct repair category. L1 research reports that they are most often 'instant' (51%) and 'anticipatory' (41%) and address only the trouble element, leaving the rest of the repairable element unchanged (Levelt, 1983, 86). L2 research reports that self-correction repair time is considerably shorter (Kormos, 2000b) and internal silent pauses shorter and less numerous than those in other types of repair (Williams & Korko, 2019). It can also be argued that L2 self-corrections differ psycholinguistically from those in L1 speech. They signify an attempt to redress the transgression of a shared social 'system' of the kind that Schegloff, Jefferson, and Sacks (1977) refer to as a form of 'socialization … for those who are still learning or being taught to operate with a system which requires … that they be adequate self-monitors and self-correctors' (Schegloff et al., 1977, 381).

Less is known about L2 false starts and how they might map onto the speech processing model of Levelt (1989) at different stages of language development. Levelt's (1989) model shows the output of the processing unit he calls the conceptualiser as a 'preverbal message' for the formulator (Levelt, 1989, 10). The formulator contains two sub-components, a grammatical encoder and a phonological encoder in an iterative relationship, which both call upon the lexicon. In L2 research, evidence for the primacy of the grammatical encoder comes from Dell, Oppenheim, and Kittredge's (2008) notion of a 'syntactic category constraint' (Dell et al., 2008, 2) associated with the formulator. And Hennecke (2013) provides further evidence that grammatical encoding is precedent to lexical selection and lemma formation. The present study follows the procedure in van Hest (1996), who found that lower-intermediate speakers produced more word-search and syntactic repairs and advanced speakers more conceptual repairs, and who analysed a combined category of non-error repairs as false starts.

The importance of understanding the production of false starts by fluency level may be summarised as follows:

1. The incidence and content of false starts may indicate fluency rather than disfluency
2. Developmental change can be inferred from the type of false start
3. The nature of false start elements in relation to speech models (Levelt, 1989) may indicate the processual level and the process involved, e.g. working memory processing of lexis and syntax, the pragmatic concerns of the speaker

https://doi.org/10.18463/diss-2021-015-williams

4. The identification of dissimilar elements may support arguments (Kormos, 1999; Kormos, 2000a; Kovac & Milatovic, 2012; O'Connor, 1988; van Hest, 1996) that advanced speakers' false starts are discoursal in nature.

The study sought to answer the following questions:

RQ1 Which of lexical/morphological, syntactic and conceptual categories of false starts of L2 learners of English are significantly different from each other?

RQ2 Which of lexical/morphological, syntactic and conceptual categories of false starts of L2 learners of English are associated with speaker level?

It was hypothesised that advanced-level speakers would produce more conceptual false starts and lower-intermediate speakers more lexical/ morphological and syntactic variants.

Method

Participants, settings and materials

The data for the study comes from a corpus of reformulations comprising the false starts and self-corrections of 56 speakers of English as an L2. Speech samples comprised a two-minute monologue on a familiar topic, e.g. 'Describe a business you would like to start' (Allen, Powell & Dolby, 2007; Hashemi & Thomas, 2011). Each participant was given a one-minute preparation time before starting to speak. On the basis of auditing the participants' speech samples, two EFL teachers had assigned them to lower-intermediate or advanced categories, with reference to the public version of the Speaking Band Descriptors of the IELTS exam. An independent samples *t*-test confirmed a significant difference, $M_{low} = 5.17$, $SD_{low} = 0.17$, $M_{high} = 7.05$, $SD_{high} = 0.64$, $t(54) = 14.2$ $p < .001$ between the proficiency levels ($N_{low} = 25$, $N_{high} = 31$). Two speakers who produced no false starts in the reformulations corpus were excluded. The number of false starts within the remaining two-minute samples ranged from one to eight ($M_{low} = 2.72$, $SD_{low} = 1.46$, $M_{high} = 3.34$, $SD_{high} = 1.95$). Participants in the lower-intermediate group ($N = 25$, $N_{females} = 13$, $M_{age} = 25.62$, 19–44 years) spoke 6 L1s; participants in the advanced group (N = 29, $N_{females} = 24$, $M_{age} = 26.84$, 20–43 years) spoke 12 L1s (Table 1). The original corpus and its collection are reported in more detail in Williams and Korko (2019).

Procedure

Following van Hest (1996) and Zuniga and Simard (2019), two judges (here, certified English language examiners) were asked to label the data, in the present case as (a) *lexical/morphological*, i.e. a word search or minor modification to word form; (b) *syntactic*, i.e. a revision of the phrase structure; or (c) *conceptual*, i.e. the expression of a completely fresh idea (Table 2). Because the raters were working with transcripts, false starts relying on phonological cues for identification, and any further examples occurring after the first in a compound sequence of false starts, were removed from the data set ($N = 176$), leaving ($N = 167$) false start exemplars. In addition, 50 words of the text surrounding the false start were supplied for context, as illustrated in the following samples.

#163: I will study something like project management and I want to run my own business [0.450] ahhahh and **if we** [0.390] like [0.264] **not boring** because I will do with my friend [0.377] so I.. I will do [0.470] what I want to [0.449] and what I love so it will not like I I I am working yeah uhhuhh uhhuhh [1.056] yeah and [0.427] er

#164: I have some advantages because my brother already has a restaurant I could have the name of the restaurant in my paella catering [0.600] ehr.. he has a really good reputation [0.315] and I think it would be pretty easy for me because I don't

Table 1. Reported first languages of participants

Speaker L1	Lower-Int	Advanced
Arabic	5	1
Bengali	-	2
Cantonese	-	1
Esan	-	1
Farsi	-	1
French	-	1
German	-	13
Japanese	3	-
Kurdish	3	-
Mandarin	10	2
Russian	-	1
Spanish	-	4
Thai	3	-
Turkish	1	-
Twi	-	1
Urdu	-	1
Total	**25**	**29**

Table 2. Example categories for the rating exercise.

False start category	**Example** False starts underlined
Lexical/ Morphological	we went er [0.522] we crossed France and Spain
Syntactic	and introduced our traditional Chi.. culture of China
Conceptual	people like me who.. who are into who wanna stand out

need like er much things to.. like **money for.. to invest** and this kind of things [0.571] so yeah

The exemplars were organised according to the theme of the prompt, e.g. all the false starts extracted from responses to the prompt that asked the speakers to describe a business they would like to start were grouped together The raters had no communication with each other and there was no discussion between them to achieve consensus.

An interrater reliability analysis using the Kappa statistic was performed to determine consistency between the raters; and, because the number of raters and the rating levels were small, and many participants were awarded the same rating, the percentage of specific agreement was also calculated. Finally, the association of the ratings with predetermined fluency level was confirmed by performing a chi-square test of independence to examine the relation between false start content and pre-established proficiency level.

Results

Interrater reliability was found to be *Kappa* = 0.797 ($p < 0.001$), 95% CI (0.719, 0.875). The average similarity rate was 87%. Items on which the raters disagreed were not included in the analysis, leaving 144 agreed categorisations. The relation between the variables was significant, χ^2 (2, N = 144) = 6.66, p = .036, Cramer's V = 0.215, which indicates a small to medium effect. Since the p-value is less than significance level α = 0.05, the null hypothesis can be rejected, and the conclusion reached that there is an association between false start content and proficiency level (Table 3). The majority of advanced learners' corrections were lexical/morphological in nature (44%), followed by conceptual corrections (33.3%), followed by syntactic modifications (22.6%). Among the lower intermediate learners, lexical/morphological corrections were also the most common (45%), but unlike the advanced learners their syntactic corrections (38.3%) outnumbered those conceptual in nature (16.7%). Adjusted residuals indicated that (1) advanced speakers were more likely to produce conceptual false starts and lower-intermediate speakers less likely (2.2, −2.2); and (2) lower-intermediate speakers were more likely to produce syntactic false starts and advanced speakers less likely (2.0, −2.0). (Table 3). χ^2 (2, N = 144) = 6.655, p = .036. Since the p-value is less than significance level α = 0.05, the null hypothesis can be rejected, and the conclusion reached that there is an association between false start content and proficiency level (Table 3). Findings with German and Mandarin speakers removed were not significant χ^2 (2, N = 75) = 4.757, p = .093; neither were findings from German and Mandarin speakers alone χ^2 (2, N = 69) = 2.590, p = .274.

Discussion

The interrater similarity rate (87%) compares favourably with (73%) in Levelt (1983). The largest category of false starts in both groups was lexical/morphological, i.e. speaker revisions of language forms of an idiosyncratic rather than of a systemic error kind, and consistent with the findings of Fathman (1980), Lennon (1984), Levelt's (1983) L1 study, and van Hest (1996). Conceptual was the smallest category. Lower-intermediate speakers reproduced the pattern (lexical/morphological 45%, syntactic 38.33%, conceptual 16.66%), but advanced speakers produced fewer than expected syntactic repairs and more than expected conceptual repairs, as suggested by adjusted residuals of −2.0 and 2.2 respectively.

Accordingly, in response to RQ1, all three false start categories are confirmed to be different from each other; and in response to RQ2, syntactic false starts are associated with lower-intermediate speakers, and conceptual false starts with advanced speakers. The hypothesis that advanced-level speakers would produce more conceptual false starts was confirmed, but that lower-intermediate speakers would produce more lexical/morphological and syntactic variants was rejected. The point of interest

Table 3. Type of false start by proficiency level.

Proficiency level		Type of false start			
		Lexical/ Morphological	Syntactic	Conceptual	Total
Lower-intermediate (N=25)	Count	27	23	10	60
	% within lower-intermediate level	45%	38.3%	16.7%	100%
	Adjusted residual	.1	2.0	−2.2	
Advanced (N=29)	Count	37	19	28	84
	% within advanced level	44%	23%	33%	100%
	Adjusted residual	−.1	−2.0	2.2	
Total	Count	64	42	38	144

is the lexical/morphological false start category, which is produced to much the same extent by both groups of speakers.

The fact that lower-intermediate speakers produced more lexical/morphological and syntactic repairs than conceptual revisions is consistent with van Hest's (1996) findings for proficiency levels and general reformulations, i.e. those not divided into corrections and false starts. Levelt's (1989) speech production model suggests that the conceptual revision of an already conceptualised utterance demands more cognitive work than the accessing of grammatical structure or vocabulary alone as it is further removed from the moment of articulation and entails all three stages—conceptualiser, formulator and articulator. Advanced speakers are more likely to possess greater automaticity and therefore to have more working memory with which to manage radical reformulations, such as those involved in conceptual false starts. Lower-intermediate speakers are able to handle syntactic revisions, which call for less cognitive reworking than conceptual false starts.

Implications

The study found no significant difference in the number of false starts by learner level (cf. Gilabert, 2007; Kormos, 1999). Contrary to the claims of some authors, the results show that lower-level learners are well able to produce false starts, but they focused on word and sentence form rather than conceptual content. That the majority of false starts are lexical followed by syntactic tends to confirm findings reported by Swain (1995) and Kovac and Milatovic (2012). The findings also bear out Kormos (1999), who suggests that, owing to a higher degree of automatization as they gain in proficiency, learners evolve the ability to focus on discourse level problems. It seems, however, that lexis/morphology continues to be an area of interest throughout L2 learner development and a future study might look for qualitative differences between the lexical/morphological false starts of different proficiency levels. To maximise the production of modified output implies minimising demands on processing (Mackey et al., 2010) and motivating conceptual revision. Suitable tasks to maximise learner production of false starts might therefore incorporate information transformation (Skehan & Foster, 2012) and be relatively unstructured, though with essential lexis supplied. Such tasks are likely to elicit numbers of false starts with accompanying benefits for acquisition and the development of learner automaticity.

A limitation of the study is the over-representation of two first languages in the data: German (45% of advanced speakers) and Mandarin (40% of lower-intermediate speakers). The literature already provides evidence of L1 influence on same-speaker L2 self-repairs (e.g. Derwing et al., 2009; Fox, Maschler, & Uhmann, 2009; Rieger, 2003; Riazantseva, 2001). As in Riggenbach (1991), if the *p* value required to reject the null hypothesis is set at .10 instead of the more conventional .05 or .001 because the number of tokens produced by the remaining participants is much lower, then the findings with German and Mandarin speakers removed again assume significance. This suggests that the study should be replicated with a more even spread of speaker first languages, or a single first language.

Acknowledgements

The author wishes to thank Patrick Bushell and Ian Cotton for rating the data; and Evan Hazenberg for helpful discussions and suggestions.

References

Allen, M., D. Powell, & D. Dolby. 2007. *IELTS Graduation: Student's Book*. Oxford: Macmillan.

Dell, G. S., G. M. Oppenheim, & A. K. Kittredge. 2008. Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language and Cognitive Processes*, *23*(4), 583–608. https://dx.doi.org/10.1080%2F01690960801920735

Derwing, T. M., M. J. Munro, R. I. Thomson, & M. J. Rossiter. 2009. The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533–557. https://doi.org/10.1017/S0272263109990015

Fathman, A. K. 1980. Repetition and correction as an indicator of speech planning and execution among second language learners. In: H. Dechert & M. Raupach (eds.), *Towards a cross linguistic assessment of speech production*, New York: Peter Lang, 77–85.

Fox, B. A., Y. Maschler, & S. Uhmann. 2009. Morpho-syntactic resources for the organization of same-turn self-repair: Cross-linguistic variation in English, German and Hebrew. *Gesprächsforschung–Online-Zeitschrift zur verbalen Interaktion* 10, 245–291.

Gilabert, R. 2007. Effects of manipulating task complexity on self-repairs during L2 oral production. *International review of applied linguistics in language teaching* 45(3), 215–240. https://doi.org/10.1515/iral.2007.010

Hashemi, L. & B. Thomas. 2011. *IELTS trainer: six practice tests with answers*. Cambridge: CUP.

Hennecke, I. 2013. Self-repair and language selection in bilingual speech processing. *Discours* 12, 1–20. https://doi.org/10.4000/discours.8789

Kormos, J. 1998. A new psycholinguistic taxonomy of self-repairs in L2: A qualitative analysis with retrospection. *Even Yearbook, ELTE SEAS Working papers in linguistics* 3, 43–68.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning* 49(2), 303–342. https://doi.org/10.1111/0023-8333.00090

Kormos, J. 2000a. The role of attention in monitoring second language speech production. *Language Learning* 50(2), 343–384. https://doi.org/10.1111/0023-8333.00120

Kormos, J. 2000b. The timing of self-repairs in L2 speech production. *Studies in L2 Acquisition* 22(2), 145–167. https://doi.org/10.1017/S0272263100002011

Kovac, M. M. & B. Milatovic. 2012. Analysis of repair distribution. Error correction rates, and repair successfulness in L2. *Studia Linguistica* 67, 225–255. https://doi.org/10.1111/stul.12000

Lennon, P. 1984. Retelling a story in English as a second language. In: H. W. Dechert, D. Möhle, & M. Raupach (eds.), *Second Language Productions*, Tübingen, Germany: Gunter Narr Verlag, 50–68.

Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition* 14(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

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

Mackey, A., R. Adams, C. Stafford, & P. Winke. 2010. Exploring the relationship between modified output and working memory capacity. *Language Learning* 60(3), 501–533. https://doi.org/10.1111/j.1467-9922.2010.00565.x

O'Connor, N. 1988. Repairs as indicative of interlanguage variation and change. In: T. J. Walsh (ed.), *Georgetown University Round Table in Languages and Linguistics 1988: Synchronic and diachronic approaches to linguistic variation and change*, Washington: Georgetown University Press, 251–259.

Riazantseva, A. 2001. Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition* 23(4), 497–526. https://doi.org/10.1017/S027226310100403X

Rieger, C. L. 2003. Repetitions as self-repair strategies in English and German conversations. *Journal of Pragmatics* 35(1), 47–69. https://doi.org/10.1016/S0378-2166(01)00060-1

Schegloff, E., G. Jefferson, & H. Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53(2), 361–382. https://doi.org/10.2307/413107

Shehadeh, A. 2001. Self-and other-initiated modified output during task-based interaction. *TESOL Quarterly* 35(3), 433–457. https://doi.org/10.2307/3588030

Skehan, P. & P. Foster. 2012. Complexity, accuracy, fluency and lexis in task-based performance. In: A. Housen, F. Kuiken, & I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA,* Amsterdam, The Netherlands: John Benjamins, 199–220. https://doi.org/10.1075/lllt.32.09fos

Swain, M. 1995. Three functions of output in L2 learning. In: G. Cook & B. Seidlhofer (eds.), *Principles and practice in applied linguistics: Studies in honour of H G Widdowson*, Oxford, UK: Oxford University Press, 125–144.

van Hest, E. 1996. *Self-repair in L1 and L2 production*. Tilburg, The Netherlands: Tilburg University Press.

Williams, S. & M. Korko. 2019. Pause behaviour within reformulations and the proficiency level of L2 learners of English. *Applied Psycholinguistics* 40(3), 723–742. https://doi.org/10.1017/S0142716418000802

Zuniga, M. & D. Simard. 2019. Factors influencing L2 self-repair behavior: The role of L2 proficiency, attentional control and L1 self-repair behavior. *Journal of Psycholinguistic Research* 48(1), 43–59. https://doi.org/10.1007/s10936-018-9587-2

Hesitation phenomena in first and second languages: Evidence from reading in Russian as L1 and Japanese as L2

Valeriya Prokaeva and Elena Riekhakaynen
Saint Petersburg State University, Saint Petersburg, Russia

Abstract

The studies of speech disfluencies rarely involve spontaneous reading data. The current study aims at the identification and the comparative analysis of the hesitation phenomena during unprepared reading of texts in the native (Russian) and non-native (Japanese) language. Three groups of disfluencies are differentiated: silent pauses, filled pauses (including lexical fillers, non-lexical fillers, lengthenings, syllable-by-syllable pronunciation and paralinguistic phenomena), and other hesitations (error-related disfluencies, repetitions, self-interruptions and within-word breaks). The results suggest that disfluency is more frequent in non-native reading and is prevalent in the lower Japanese proficiency group, whilst the higher text complexity defined by a text type does not necessarily induce more hesitations. The self-correction phenomena were equally widespread in both L2 proficiency groups, whereas the number of noticed but uncorrected errors was higher in the lower Japanese proficiency group.

Introduction

In any unprepared utterance, we can observe various verbal and non-verbal markers that reflect the speaker's thinking and speech planning processes such as hesitation pauses, self-corrections, repetitions, and further various speech disfluencies (Maclay & Osgood 1959; Goldman-Eisler, 1961; Levelt, 1983; etc.). The bilingual studies of speech disfluency and the results of the comparative studies of hesitation in speech in the native and foreign languages allow us to test the hypotheses both about the nature of hesitations and about the speech production in the first and second languages. Numerous papers show that there are certain differences in hesitation phenomena in first and second language speech (Temple, 2000; Watanabe & Rose, 2012; Rose, 2017, etc.). Comparative studies on the Russian language as L1 and other languages as L2, however, are quite few. The studies on Russian as L2 are more common: see Chen (2016) for the study of Chinese students learning Russian; the data of native English speakers in Gilquin (2008), Riazantseva (2001), and German speakers in Belz et al. (2017). We hope that the present study conducted on the material of Russian and Japanese speech of Russian learners of Japanese will contribute to the theory of second language acquisition and the process of speech production.

Whereas speech disfluencies may occur not only in spontaneous speech, but also in reading (Fant, Kruckenberg, & Ferreira, 2003; Krivokapić, 2007), such data is rarely used in disfluency research as reading aloud is usually less common than spontaneous speech among adult speakers. However, reading allows us to obtain homogeneous and easily controlled data for the comparison of speech production in different conditions. Therefore, we consider reading aloud to be an appropriate material in a comparative study of speech disfluencies in the first and second language acquisition because it allows us to control the content, the format, and the size of the data and, thus, to check the effects of language on various hesitation phenomena.

Disfluencies in Second Language research

Production of speech in a second language, being a more cognitively laborious task, entails a greater number of errors and self-corrections among speakers, which may be due to the lower level of speech planning skills automation in L2. Some studies show that, when a speech error occurs, native speakers tend to fill pauses with various hesitation phenomena, while non-native speakers frequently leave them unfilled, which leads to poorer fluency in a second language (Temple, 1992). Besides, non-native language speech may contain hesitation strategies that are involuntarily transferred by the speaker from their first language, which affects their qualitative characteristics and can lead to specific types of speech disfluencies due to the cross-linguistic interference (Tedlock, 1983).

The self-correction strategies in second language speakers at different levels of proficiency also differ. The speakers with a higher level often do not resort to self-corrections: corrections remain unpronounced and stay at the level of inner speech, being expressed through repetitions of words and various hesitations (O'Connor, 1988).

Some studies show that the students with higher language proficiency use a wider range of hesitation strategies in contrast to less proficient students

https://doi.org/10.18463/diss-2021-016-prokaeva-riekhakaynen

(Rieger, 2003). Others suggest that using hesitation phenomena in L2 learning may become closely connected to the hesitation patterns in L1 with the rise of language proficiency (Rose, 2013). Quite obviously, further research in L2 speech disfluency production is required for better understanding of the hesitation phenomena and speech production.

In this paper, we aim at the comparative description of disfluencies occurring while reading in Russian as first and Japanese as second languages in order to find the influence of the three main factors on the disfluency characteristics in reading in L1 and L2: *language* (native or non-native), *text type* (static / descriptive or dynamic / narrative), and *level of L2 proficiency*. We decided to consider the text type because previous studies on reading in Russian have shown that descriptions (static texts) are more difficult to process than narratives (dynamic texts) (Petrova & Dobrego, 2016). Therefore, we might expect more speech disfluencies in the static texts.

Data and method

Material

We used two fragments (one narrative / dynamic and one description / static) from the Japanese novel *Natsu no Niwa* by Yumoto Kazumi and their translations to Russian as the material. The readability level for the Russian texts was 5.15 for the dynamic one and 5.64 for the static one (according to http://ru.readability.io/); for the Japanese texts—3.07 and 3.36 respectively (Upper intermediate level; according to https://jreadability.net/sys). The texts in Russian contained 176 and 173 words, the Japanese ones—192 and 210 words, sentence count was 15 sentences for each text, average word count for the Russian texts was 11.73 and 11.53, for the Japanese ones—17.5 and 19.1 (in Japanese *bunsetsu*). The numbers are provided for the dynamic and static texts respectively. Each text was presented on the computer screen in black Georgia font (16-point size) on a white background, line spacing 1.5.

Participants

The participants of the study were 10 native speakers of Russian (19–28 years old) with some Japanese studying experience (N3–N1 Japanese Language Proficiency Level JLPT), without speech or reading disorders. The participants were divided into two groups (5 people in each) according to their Japanese studying and usage experience:

- group one (N3–N2 level, no more than 5.5 years of studying, not actively engaged with Japanese at the workplace)—lower level of proficiency;
- group two (N2–N1 level, 5–12 years of studying, currently working with the Japanese language (teaching, translating))—higher level of proficiency.

All participants from the higher L2 proficiency group had extensive reading experience in Japanese (read in Japanese regularly), whereas the lower group readers rarely did so, with the only one exception (Participant 2). Reading skills in the native language were beyond the factors considered in the current study.

Procedure

The experiment was conducted in the accordance with the Declaration of Helsinki and the existing Russian and international regulations concerning ethics in research. All participants provided written consent to take part in the experiment.

The procedure was held with an interval of 1.5–2 weeks to increase the degree of reading spontaneity, as that the participants were presented similar tests in Russian and Japanese. Each participant read the original text and its translation on different days in the randomized order. At each stage of the experiment, the participants were asked to read aloud one text in the Russian language and one in Japanese. The participants were asked to read thoroughly and carefully at their own pace.

20 speech recordings were provided with orthographic annotations in Praat (Boersma & Weenink, 2006). The mean reading time for the Russian texts was 79 s (dynamic) and 86 s (static); for the Japanese ones, it was 150 s and 186 s respectively.

The classification of disfluencies in the material

The principles of description were based on the classification of hesitation phenomena in spontaneous speech and reading for the Russian language (Bogdanova-Beglaryan et al., 2013) and the filled hesitation phenomena classification in Japanese (Maekawa, 2003).

Silent pauses

The silent hesitation pauses were selected manually with a lower boundary of 100 ms. We considered pauses of 100 ms and higher to be a marker of hesitation if they appeared within a word or disrupted the unity of a clause (e.g. appeared between a noun and a particle). Pauses on clause boundaries were considered to contain a hesitation component if they exceeded 700 ms (see Barik, 1968, 157).

Filled pauses

1) *Lexical fillers* (this category included all metacommunicative insertions observed in reading when the speakers addressed their difficulties or reacted to them);
2) *vocalizations* (non-lexical filled pauses) such as *a-a*, *m-m*, *h-m,* etc.;
3) *vowel and consonant lengthenings* (*aikawarazu-u*);
4) *syllable by syllable pronunciation* (*ke-re-do* «but»);
5) *paralinguistic phenomena* (laugh, sighs, aspiration, etc.).

Other hesitation phenomena

1) *Error-related hesitations*
   Self-corrections. This phenomenon is widely referenced in other studies on different languages; see (Shriberg, 1994; Eklund, 2004; Maruyama & Sano, 2006).
   Noticed and uncorrected errors. We identify this phenomenon as a discrepancy between the utterance produced by the informant and the content of the text, accompanied by a hesitation pause or other hesitation phenomenon being a shred of evidence that the participant spent additional time processing a mispronounced word or a phrase, yet could not pronounce it correctly.
   Unnoticed errors.

Other hesitation phenomena also included some cases of:

2) *repetitions* (discussed in Maclay & Osgood, 1959; Henderson, Goldman-Eisler, Skarbek, 1966);
3) *self-interruptions* (*tsukawa= tsukawarete* «being used»);
4) *within-word breaks* (pronouncing words "part by part", when an identifiable pause (100 ms or more) occurs between word fragments).

Results

General overview

Overall number of hesitations was higher in Japanese text reading (7 times more). The most frequent hesitation markers in Russian text reading were silent pauses (see Table 1). Filled pauses, such as vocalizations, vowel, consonant lengthenings, and paralinguistic phenomena prevailed in Japanese text reading.

Table 1. The percentage of different hesitation phenomena

	Russian	Japanese
Silent pauses	227 (44.6%)	1328 (35.8%)
Filled pauses	205 (40.3%)	1772 (47.8%)
Other phenomena	77 (15.1%)	609 (16.4%)
Overall	509	3709

Silent pauses and filled hesitations distribution

The Student's *t*-test and Wilcoxon signed-rank test showed significant differences in the total number of hesitations while reading texts of the same type in different languages. There were more disfluencies in the Japanese text than in the Russian one while reading both the static text ($t = 5.638$, $df = 18.00$, $p < 0.001$) and the dynamic one ($t = -7.638$, $df = 9$, $p < 0.001$). When reading the static text in Japanese the participants made significantly more disfluencies than when reading the dynamic text ($Z = 54.00$, $p = 0.004$). The differences between the hesitations in reading static and dynamic texts in Russian did not reach significance ($t = 1.509$, $df = 9$, $p = 0.166$).

We obtained similar results while analyzing separately the distribution of silent pauses and all filled hesitations (filled pauses and other hesitation phenomena together). For filled hesitations, static ($t = 5.851$, $df = 9$, $p < 0.001$) and dynamic ($t = 6.718$, $df = 9$, $p < 0.001$) text reading differed. There were more filled hesitations while reading the Japanese static text than the dynamic one ($t = 2.878$, $df = 9$, $p = 0.018$).

The number of silent pauses was statistically higher while reading in L2 than in L1: for the static text ($t = 7.722$, $df = 18.00$, $p < 0.001$) and the dynamic one ($t = -8.223$, $df = 9$, $p < 0.001$). The static text reading in Japanese induced more hesitations than the dynamic text reading ($t = 3.658$, $df = 9$, $p = 0.005$). The difference in number of hesitations while reading different types of texts in Russian was never significant: $t = 1.752$, $df = 9$, $p = 0.114$ (filled hesitations only), $t = 0.975$, $df = 9$, $p = 0.355$ (silent pauses only).

For the two Japanese language proficiency groups, we found the lower L2 proficiency group participants to make more hesitations in general ($t = -3.242$, $df = 18.00$, $p = 0.005$) and filled hesitations only ($t = -3.644$, $df = 18.00$, $p = 0.002$); the differences in silent pauses did not reach significance ($t = -2.006$. $df = 18.00$, $p = 0.060$).

Filled pauses

Lexical fillers Lexical fillers almost never appeared in L1 reading. Except for two cases (nani 'what', eto 'well'), all lexical fillers contained words or phrases in Russian, even though the participants were instructed to read in the language in which the text was written. Overall, there were more lexical fillers in the Japanese static text and there were only a few of them in Russian (see Table 2).

Vocalizations We found that the number of vocalizations was statistically greater in Japanese texts ($t = -3.959$, $df = 9$, $p = 0.003$ for the static text, $t = -3.038$, $df = 9$, $p = 0.014$ for the dynamic one). We also observed the influence of the text type in Japanese ($t = 2.517$, $df = 9$, $p = 0.033$). The vocalizations were more frequent in the lower Japanese proficiency group ($t = -7.402$, $df = 18.00$, $p < 0.001$). For the distribution of different vocalizations in our data see Table 3.

Lengthenings The number of vowel lengthenings was statistically higher for reading in Japanese than in Russian both for the static text ($t = -4.955$, $df = 9$, $p < 0.001$) and for the dynamic one ($Z = 0.000$, $p = 0.006$). We found no influence of the text type here ($t = 1.430$, $df = 9$, $p = 0.186$ for Russian; $t = 2.127$, $df = 9$, $p = 0.062$ for Japanese) and the proficiency level ($U = 30.5$, $p = 0.151$).

We also did find a significant language effect for consonant lengthenings—they were more frequent in the Japanese texts ($t = -4.478$, $df = 9$, $p = 0.002$ for the static text; $t = -5.011$, $df = 9$, $p < 0.001$ for the dynamic one).

Syllable by syllable pronunciation This phenomenon was more typical for the reading in Japanese: we found 6 examples in the Russian texts and 52 examples in the Japanese ones. In Japanese, the phenomenon was observed in both short (ex. *o-ku* 'inside, interior') and long words (*kinmo-ku-us-sei* 'fragrant olive'), but was more common for the words consisting of three or four morae (see Table 4).

Paralinguistic phenomena There were more paralinguistic hesitations while reading the texts in the second language ($t = -3.061$, $df = 9$, $p = 0.014$ for the static texts; $t = -5.219$, $df = 9$, $p < 0.001$ for the dynamic ones). Moreover, these phenomena were more widespread in a lower language proficiency group ($t = -2.701$, $df = 18.00$, $p = 0.015$).

Error-related hesitation phenomena

There were a few curious findings concerning error-related hesitations. For *self-corrections*, we observed the language effect ($t = -5.045$, $df = 9$, $p < 0.001$ for the static texts; $t = -3.000$, $df = 9$, $p = 0.015$ for the dynamic ones) and the text type effect for Japanese ($t = -3.899$, $df = 9$, $p = 0.004$). However, there were no differences in the number of such hesitations between the two language proficiency groups ($t = -1.281$, $df = 18.00$, $p = 0.216$).

The number of *noticed but uncorrected* errors in Russian was too small to draw any comparison with Japanese (one in the static text, three in the dynamic one). Yet, for reading in Japanese, the comparison of the two groups of participants showed a significant difference in the number of such errors: they emerged more frequently in a lower proficiency group ($t = -2.652$, $df = 18.00$, $p = 0.016$). There was no difference in the number of *unnoticed errors* between these two groups ($t = -0.732$, $df = 18.00$, $p = 0.473$).

Table 2. Lexical hesitation pause fillers

Russian		Japanese	
dynamic	static	dynamic	static
2	3	14	34

Table 3. Vocalization types

Vocalization type	Russian	Japanese
a-a	86.6%	79.2%
a-m	-	6.3%
m-m	6.7%	7.5%
e-e, e-m	6.7%	7%

Table 4. Syllable by syllable word pronunciation (words)

Russian			
1-2 syllable	3 syllables	4 syllables	5+ syllables
0	2	2	2
Japanese			
1-2 morae	3 morae	4 morae	5+ morae
6	18	18	8

Discussion

The data showed some significant differences between the number of disfluencies and their types in first and second languages. According to the results, more hesitations (including silent pauses of hesitation and the overall number of filled hesitations separately) occurred in the L2 reading. The language factor affected the number of lexical fillers, vowel and consonant lengthenings, paralinguistic fillers, and self-corrections suggesting that the disfluency appears more frequently in L2 reading due to higher cognitive load experienced by the speaker, which is consistent with the results obtained for other languages: for instance, English L1 and German L2 (Fehringer & Fry, 2007) and the learners with different backgrounds (Tavakoli, 2010), but disagrees with the data of French as L1 and English as L2 (Kosmala & Morgenstern, 2017). We do realize, however, that some linguistic features of the language itself (e.g. differences in writing systems in two compared languages) might affect the number and the qualitative characteristics of hesitations in speech (de Johnson, O'Connell, & Sabin, 1979). For this reason, further comparison of the native Japanese speakers' data with our findings is required to eliminate this factor.

In some cases, the static text reading entailed more hesitations (all hesitations, filled hesitations,

silent pauses, lexical fillers, and vocalizations in the reading of Japanese texts), but there was no such difference in L1 text reading, suggesting that the text type is a weak predictor of the hesitation frequency in L1.

Our results also provided evidence for a greater number of all hesitations, filled hesitations, vocalizations, and paralinguistic phenomena in the lower L2 proficiency group. Interestingly, we did not find such effect for sound lengthenings. Probably, this hesitation phenomenon is highly individual. Furthermore, the difference between the two L2 proficiency groups did not reach significance for self-corrections, which is consistent with O'Connor (1988). The data showed that L2 lower proficiency group readers more often left the noticed errors uncorrected, which might be due to the lack of particular Japanese word knowledge or the stronger focus on the better text understanding than on accurate word pronunciation, compared to the L2 higher proficiency group. As for the unnoticed errors, presumably, the readers remain unaware of them and thus this phenomenon might not be connected to the second language proficiency level.

Conclusions

In this paper, we analyzed the oral reading data of 10 native Russian learners of Japanese. The study is the first to describe disfluency patterns in the reading of Russian learners of Japanese with different Japanese language proficiency level, and its main findings contribute to the assumption that hesitation phenomena appear more often in L2 speech production and in the speakers with less knowledge or experience of L2. We found a text type (static / dynamic) influence on the hesitation frequency in L2 reading due to the possible higher processing difficulty of the static text, while the reading of a static text in L1 might not be more cognitively loaded: it did not induce more hesitations. According to the error-related hesitation phenomena analysis, the number of self-corrections is not affected by the L2 level, whereas the number of noticed but uncorrected mistakes might be.

References

Barik, H. C. 1968. On Defining Juncture Pauses: A Note On Boomer's «Hesitation and Grammatical Encoding». *Language and Speech* 11(3), 156–159. https://doi.org/10.1177/002383096801100302

Belz, M., S. Sauer, A. Lüdeling, & C. Mooshammer. 2017. Fluently disfluent? Pauses and repairs of advanced learners and native speakers of German. *International Journal of Learner Corpus Research* 3(2), 118–148. https://doi.org/10.1075/ijlcr.3.2.02bel

Boersma, P. & D. Weenink. 2006. Praat: Doing phonetics by computer (version 6.1.42). https://www.praat.org/ (accessed 26 April 2021).

Bogdanova-Beglaryan, N. V., E. M. Baeva (Sapunova), I. S. Brodt (Panarina), O. V. Pavlova (Ilyicheva). 2013. *Zvukovoj korpus kak material dlya analiza russkoj rechi. Kollektivnaya monografiya. Chast' 1. Chtenie. Pereskaz. Opisanie* [Sound corpus as a material for the analysis of Russian speech. Collective monograph. Part 1. Reading. Retelling. Description], Saint Petersburg: Saint Petersburg University Press.

Chen, C. 2016. Russkaya Spontannaya Rech' na Nerodnom Yazyke: Analiz Khezitatsiy (na materiale russkoy rechi kitaytsev) [Spontaneous Speech in Russian as a Foreign Language: Analysis of Hesitation (a Case Study of Chinese Students' Speech)]. *Vest-nik Permskogo Universiteta. Serija: Rossijskaja i zarubezhnaja filologija* [Perm University Herald. Russian and Foreign Philology] 1(33), 53–62.

de Johnson, T. H., D. C. O'Connell, & E. J. Sabin. 1979. Temporal analysis of English and Spanish narratives. *Bulletin of the Psychonomic Society* 13(6), 347–350. https://doi.org/10.3758/BF03336891

Eklund, R. 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. Ph.D. dissertation, Linköping University.

Fant, G., A. Kruckenberg, J. B. Ferreira. 2003. Individual variations in pausing. A study of read speech. In: M. Heldner (ed.), *Proceedings of Fonetik 2003, Reports in Phonetics*, 2–4 June 2003, Umeå, Sweden, 193–196.

Fehringer, C., C. Fry. 2007. Hesitation phenomena in the language production of bilingual speakers: The role of working memory. *Folia Linguistica: Acta Societatis Linguisticae Europaeae* 41(1–2), 37–72. https://doi.org/10.1515/flin.41.1-2.37

Gilquin, G. 2008. Hesitation markers among EFL learners: Pragmatic deficiency or difference? In: J. Romero-Trillo (ed.), *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Berlin: Mouton de Gruiter, 119–150. https://doi.org/10.1515/9783110199024.119

Goldman-Eisler, F. 1961. A comparative study of two hesitation phenomena. *Language and Speech* 4(1), 18–26. https://doi.org/10.1177/002383096100400102

Henderson, A., F. Goldman-Eisler, A. Skarbek. 1966. Sequential temporal patterns in spontaneous speech. *Language and Speech* 9(4), 207–216. https://doi.org/10.1075/aral.15.2.03tem

Kosmala L. & A. Morgenstern. 2017. A preliminary study of hesitation phenomena in L1 and L2 productions: a multimodal approach. In: R. Eklund & R. L. Rose (eds.), *Proceedings of DiSS 2017, the 8th Workshop on Disfluency in Spontaneous Speech*, 18–19 August, 2017, Stockholm, Sweden, 37–40.

Krivokapić, J. 2007. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics* 35(2), 162–179. https://doi.org/10.1016/j.wocn.2006.04.001

Levelt, W. J. 1983. Monitoring and self-repair in speech. *Cognition* 14(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Maclay, H. & C. E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15(1), 19–44. https://doi.org/10.1080/00437956.1959.11659682

Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In: *Proceedings of ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition, (SSPR 2003)*, 13–16 April, 2003, Tokyo, Japan, 7–12.

Maruyama, T. & S. Sano. 2006. Classification and Annotation of Self-Repairs in Japanese Spontaneous Monologues. In: *Proceedings of the international symposium on linguistic patterns in spontaneous speech (LPSS)*, 17–18 November, 2006, Taipei, Taiwan, 283–298.

O'Connor, N. 1988. Repairs as indicative of interlanguage variation and change. In: T. J. Walsh (ed.), *Georgetown University Round Table in Languages and Linguistics 1988: Synchronic and diachronic approaches to linguistic variation and change*, Washington: Georgetown University Press, 251–259.

Petrova T. & A. Dobrego. 2016. Processing of static and dynamic texts: an eye-tracking study of Russian. In: *Proceedings of the 3rd International Multidisciplinary Scientific Conference on Science and Arts, SGEM 2016*, 24–30 August, 2016, Albena, Bulgaria, 991–997.

Riazantseva, A. 2001. Second Language Proficiency and Pausing. A Study of Russian Speakers of English. *Studies in Second Language Acquisition* 23(4), 497–526. https://doi:10.1017/S027226310100403

Rieger, C. L. 2003. Disfluencies and hesitation strategies in oral L2 tests. In: R. Eklund (ed.), *Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop*, Gothenburg Papers in Theoretical Linguistics 90, 5–8 September 2003. Göteborg University, 41–44.

Rose, R. L. 2013. Crosslinguistic corpus of hesitation phenomena: A corpus for investigating first and second language speech performance. In: F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, & P. Perrier (eds.), *Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 25–29 August 2013, Lyon, France, 992–996.

Rose, R. L. 2017. A comparison of form and temporal characteristics of filled pauses in L1 Japanese and L2 English. *Journal of the Phonetic Society of Japan* 21(3), 33–40. https://doi:10.24467/onseikenkyu.21.3_33

Shriberg, E. E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California at Berkeley.

Tavakoli P. 2010. Pausing patterns: differences between L2 learners and native speakers. *ELT Journal* 65(1), 71–79. https://doi.org/10.1093/elt/ccq020

Tedlock, D. 1983. *The Spoken Word and the Work of Interpretation.* Philadelphia: University of Pennsylvania Press.

Temple, L. 1992. Disfluencies in learner speech. *Australian Review of Applied Linguistics* 15(2), 29–44. https://doi.org/10.1075/aral.15.2.03tem

Temple, L., 2000. Second language learner speech production. *Studia Linguistica* 54(2), 288–297. https://doi.org/10.1111/1467-9582.00068

Watanabe, M. & R. L. Rose. 2012. Pausology and Hesitation Phenomena in Second Language Acquisition. In: P. Robinson (ed.), *The Routledge Encyclopedia of Second Language Acquisition*, New York/London: Routledge, 480–483.

Word-form related disfluency versus lemma related disfluency: An exploratory analysis of disfluency patterns in connected-speech production

Aurélie Pistono and Robert J. Hartsuiker
Ghent University, Ghent, Belgium

Abstract

Several language production levels may be involved in the production of disfluencies. In the current study, we conducted network task experiments to tackle disfluencies related to conceptualization, which we operationalized by impeding visual object recognition (i.e. blurring). Contrary to what was expected, blurriness did not lead to more disfluency. However, disfluency type and disfluency location were closely related. This suggests a distinction in the underlying function of disfluencies, some reflecting word-form related difficulties, others reflecting lemma related difficulties.

Introduction

The term 'disfluency' includes various phenomena such as filled or silent pauses, repeated words, and self-corrections. Despite the high frequency of these phenomena (Fox Tree, 1995), the question remains as to why speakers are so often disfluent. Within the language production system, several levels may be involved in the production of disfluencies. Several authors attempted to relate the pattern of disfluencies to difficulties at specific levels of production, using a Network Task (Figure 1). In this paradigm, participants describe a route through a network of pictures so that a listener could fill in a blank network by listening to the description. This allows for the manipulation of the items to create difficulties at specific stages (e.g. conceptual generation) while holding others constant (e.g. lexical selection). It has been shown, for example, that hampering the verbal monitoring system (Oomen & Postma, 2001), the initial stage of lexical access (Hartsuiker & Notebaert, 2010), or the conceptual generation of a message (Schnadt & Corley, 2006) affected the rate of disfluencies. More particularly, the latter study showed that blurred pictures were preceded by more prolongations than clear pictures, and that prolongations were the most frequently occurring category of disfluency. This type of disfluency could belong to three categories: "a", "the", "to" (e.g. "tooo a hammer"; "to theee hammer"). In the current study, we first aim to test whether conceptualization difficulties (i.e., blurriness) increases disfluency production and prolongations in Dutch speakers as well, similarly to English speakers tested in Schnadt and Corley. Second, we aim to analyze the effect of this manipulation on disfluency location. Indeed, we predict that blurriness might induce a specific pattern of disfluency, since the upcoming difficulty can be anticipated. More precisely, because grammatical gender is marked on determiners in Dutch, we will differentiate "early disfluencies", occurring before any act of choice, from "late disfluencies", occurring on the determiner of after its production. Given that determiner selection occurs after noun selection (Dhooge, De Baene, & Hartsuiker, 2016), we predict that disfluency related to visual identification difficulty will occur more often before determiner production than disfluency associated with clear pictures.

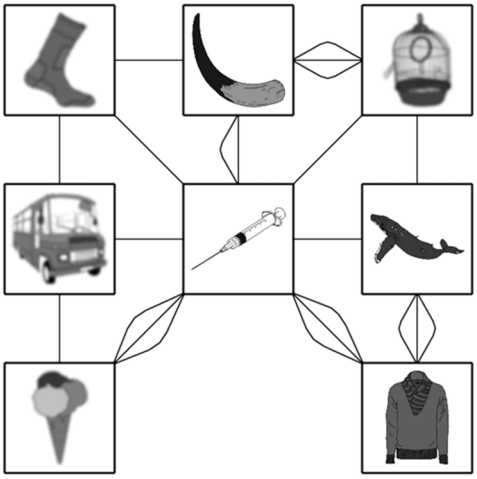

Figure 1. Example of a network.

Material and Methods

Twenty bachelor students, all native speakers of Dutch, participated in the experiment in exchange for course credit (18 Females and 2 Males, mean age was 18.6 ± 0.6 years old). The samples have been calculated using guidelines for mixed models in designs with repeated measures (Brysbaert & Stevens, 2018).

Material

We constructed 20 networks using a program written in Psychopy (Peirce, 2007). We used 160

https://doi.org/10.18463/diss-2021-017-pistono-hartsuiker

pictures from the Multipic database (Duñabeitia et al., 2018) and split it into two sets, matched for name agreement, age of acquisition, and visual complexity (Table 1).

Table 1. Mean (±SD), age of acquisition (AoA), visual complexity, name agreement (H-statistic), in isolation for each set of pictures (i.e. each set was the control condition for half the participants).

	Set1	Set2	*p* value
AoA	6.2±1.1	6.2±1.3	0.96
Visual complexity	3.0±0.6	3.0±0.5	0.62
Name agreement (H-statistic)	0.8±0.2	0.8±0.2	0.83

These sets were counterbalanced across participants, so that each part was the control condition for half the participants. Each network consisted of eight interconnected black-and-white line pictures: four blurred (4 pixels radial blur) and four control pictures (Figure 1). Within each network, pictures were either connected by one, two, or three straight lines or curves. Lines were either horizontal, vertical, or diagonal. Curves could be horizontal, vertical, or diagonal. The type and number of lines connecting the pictures, as well as the order and location of appearance of the 160 pictures were randomized across participants. The route through the network was indicated by a moving red dot that traversed the network in 42 seconds.

Procedure

Participants took place in front of a computer screen which displayed an example network. Instructions were given to provide an accurate description of the network while staying synchronized with the dot that moved through the network. Subsequently, three practice networks were run. The first network was described by the experimenter and the next two networks were described by the participant. During the experiment, each network was preceded by a fixation cross in the upper center of the computer screen and started with a two seconds period for visual inspection after which the movement of the dot started.

Scoring

A native Dutch speaker independently transcribed and scored all networks. In a subsequent phase, a second transcriber listened to all the productions and checked the transcriptions. They disagreed on 17.8% of trials. Disagreements were solved by a third person.

Only disfluencies preceding pictures names were analyzed. They were grouped into broad categories, to ensure a sufficient amount of data within each category: repetitions (of a sound, syllable, word, or phrase), filled pauses, silent pauses, prolongations, and self-corrections (substitutions, additions, or deletions). However, repetitions were not analyzed because there were only 10 observations in total in this category.

In a second step, we focused on the location of disfluency. We distinguished "early disfluencies" as disfluencies occurring before the determiner. Because self-corrections related to a picture name could hardly happen before determiner selection, we only focused on pauses. Early prolongations were prolongations occurring on the preposition "naar" (e.g. "naar de tafel"; "to the table"); early silent pause occurred between the preposition and the determiner (e.g. "naar (.) de tafel"; "to (.) the table"); so as early filled pause ("naar hm de tafel"; "to hm the table").

Results

We analyzed phrases corresponding to 1280 pictures (40% were excluded because the wrong target was produced or the gender-marked determiner—"de" or "het"—was omitted). There was at least one disfluency on 36% of these observations: 7.8% of pictures included at least one self-correction, 10.15% a silent pause, 5.5% a filled pause, and 10.54% a prolongation.

All disfluencies

The effect of blurriness on disfluency was tested using linear mixed effects (lme4 package in *R*, Bates et al., 2015). For the random part of the model, the maximal random effects structure was included. We then chose a backward-selection heuristic by reducing the model complexity until a further reduction would imply a significant loss in the goodness-of-fit (Matuschek et al., 2017). For the analysis of all phenomena together, the model resulted in a random intercept for subjects, network order, and image order, and a random slope for blurriness over subjects. There was no significant effect of blurriness ($\chi^2(1) = 1.49$, $p = .2$). For the analysis of each phenomenon separately, generalized linear mixed effects model were used, following the same method. There was no significant effect of blurriness when each disfluency was analyzed individually (Table 2).

Location of disfluencies

In a further set of analyses, we tested the effect of blurriness on disfluency location (early vs. late

Table 2. Summary of results for disfluency production.

Variable	Random structure	Effect of blurriness
Silent pauses	random intercept for item, subject, network order	($\chi^2(1) = 0.67$, $p = 0.41$)
Filled pauses	random intercept for item and subject	($\chi^2(1) = 2.6$, $p = 0.11$)
Self-corrections	random intercept for item and subject	($\chi^2(1) = 1.35$, $p = 0.24$)
Prolongations	Random slope for blurriness over items, random intercept for subject, network order and image order	($\chi^2(1) = 0.12$, $p = 0.73$)

disfluency). For that purpose, we focused on pictures for which prolongations, filled pauses, or silent pauses were produced. We conducted generalized linear mixed effects models, using the same methods as described above (Matuschek et al., 2017). There was no significant effect of blurriness on disfluency location. Contrary to what was expected, disfluency related to visual identification difficulty did not occur more often before than after the determiner: silent pauses ($\chi^2(1) = 0.48$, $p = 0.49$); filled pauses ($\chi^2(1) = 0.92$ $p = 0.34$); prolongations ($\chi^2(1) = 0.42$, $p = 0.52$).

However, descriptive analyses showed that 58.2% of silent pauses were "early silent pauses"; 58.9% of prolongations were "early prolongations"; and 93% of filled pauses were "early filled pauses". It therefore seems that disfluency type depends on the location of disfluency. In particular filled pauses seemed most often produced before determiner selection, regardless of current manipulation (i.e. blurriness). To investigate whether this finding is replicable, we analyzed disfluency location on another set of data, in which we tested the effect of lexical selection difficulty and grammatical selection difficulty using network tasks (Pistono & Hartsuiker, 2021).

Exploratory analysis

In the second dataset (Pistono & Hartsuiker, 2021), 61.2% of silent pauses were "early silent pauses"; 65.7% of prolongations were "early prolongations"; and 80.4% of filled pauses were "early filled pauses". Although not as distinct as the blurriness manipulation, filled pauses were most often produced before determiner selection.

To confirm this effect, we compared the proportion of early disfluency produced by each participant in each experiment, using *t*-tests. Inter-group differences were not significant: proportion early prolongations in each experiment: $t(38) = 0.68$, $p = 0.5$; proportion early silent pause in each experiment: $t(36) = 0.95$, $p = 0.35$; proportion early filled pause in each experiment: $t(33) = -0.77$, $p = 0.45$. These results reinforce the hypothesis that disfluency type and disfluency location are closely related.

Discussion

Contrary to what was expected, impeding conceptual access of object representations did not elicit more disfluency. However, the rate of disfluency was quite substantial (26% of trials had at least one disfluency) compared to studies manipulating lexical selection for example (Hartsuiker & Notebaert, 2010; Pistono & Hartsuiker, 2021). It is therefore possible that, because the manipulation was visually salient, the complexity of the whole task increased, leading to a high rate of disfluency. The current results also differ from those of Schnadt and Corley (2006), who found an effect of blurriness on prolongations. However, their method was quite different: different set of pictures (Snodgrass & Vanderwart, 1980), different analyses (ANOVAs), on a different population (English speakers).

The current study also tested whether blurriness was associated with disfluencies occurring earlier than the ones associated with clear pictures. Because gender is marked in Dutch, we differentiated "early disfluencies", occurring before the determiner, from disfluency occurring afterwards. Contrary to what was expected, we did not find any effect of blurriness on disfluency location. However, filled pauses were mostly produced before the determiner, which suggests that their underlying function differs from silent pauses or prolongations. This pattern of disfluency was found for blurred and control pictures. We replicated this exploratory finding by re-analyzing a previous set of data, in which we found a similar pattern of disfluency: filled pauses occurred most often before the determiner, while silent pauses and prolongations occurred either before or after the determiner, regardless of the manipulated difficulty. This finding suggests that filled pauses may be related to difficulties occurring at a lemma level, while silent pauses and prolongations may reflect a delay occurring at either a lemma level or at a lexeme level. The distinction between lemma and lexeme is crucial in most of the speech production theories (e.g. Levelt, 1989) for which lexical access consists of two major steps. During a first step (lemma retrieval), a word's syntactic properties are retrieved. During a second

step (lexeme retrieval), the word's morphological and phonological properties are recovered. Filled pauses could therefore reflect at delay occurring during the first step (e.g. related to the meanings of words or their syntactic properties), but they do not seem related to word-form encoding difficulties. On the contrary, silent pauses and prolongations could indicate a delay at both levels. Further work is required to analyze more specifically the role of each disfluency phenomenon in the language production system.

Acknowledgements

The authors thank the research assistants (Amber Van Goethem and Steffi Van Goethem) who transcribed participants' productions and coded disfluencies. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Individual fellowship, grant agreement No 832298.

References

Bates, D. M., M. Maechler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 1 (67), 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Brysbaert, M. & M. Stevens. 2018. Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition* 1(1), 1–20. https://doi.org/10.5334/joc.10.

Dhooge, E., W. De Baene, & R. J. Hartsuiker. 2016. The Mechanisms of Determiner Selection and Its Relation to Lexical Selection: An ERP Study. *Journal of Memory and Language* 88, 28–38. https://doi.org/10.1016/j.jml.2015.12.004

Duñabeitia, J. A., D. Crepaldi, A. S. Meyer, B. New, C. Pliatsikas, E. Smolka, & M. Brysbaert. 2018. MultiPic: A Standardized Set of 750 Drawings with Norms for Six European Languages. *Quarterly Journal of Experimental Psychology* 71(4), 808–16. https://doi.org/10.1080/17470218.2017.1310261.

Fox Tree, J. E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, 34(6), 709–738. https://doi.org/10.1006/jmla.1995.1032

Hartsuiker, R. J. & L. Notebaert. 2010. Lexical Access Problems Lead to Disfluencies in Speech. *Experimental Psychology*, 57(3), 169–177. https://doi.org/10.1027/1618-3169/a000021.

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

Matuschek, H., R. Kliegl, S. Vasishth, H. Baayen, & D. Bates. 2017. Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language* 94, 305–15. https://doi.org/10.1016/j.jml.2017.01.001

Oomen, C. C. & A Postma. 2001. Effects of Time Pressure on Mechanisms of Speech Production and Self-Monitoring. *Journal of Psycholinguist Research* 30(2), 163–84. https://doi.org/10.1023/A:1010377828778

Peirce, J. W. 2007. PsychoPy—Psychophysics Software in Python. *Journal of Neuroscience Methods* 162(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pistono, A. & R. Hartsuiker. 2021. Eye-Movements Can Help Disentangle Mechanisms Underlying Disfluency. *Language, Cognition and Neuroscience*. Published online. https://doi.org/10.31234/osf.io/6mx2y.

Schnadt, M. J. & M. Corley. 2006. The Influence of Lexical, Conceptual and Planning Based Factors on Disfluency Production. In: *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 26–29 July, 2006, Vancouver, Canada, 8–13.

Snodgrass, J. G. & M. Vanderwart. 1980. A Standardized Set of 260 Pictures: Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity. *Journal of Experimental Psychology: Human Learning and Memory* 6(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Disfluencies in spontaneous speech: The effect of age, sex and speech task

Judit Bóna
ELTE Eötvös Loránd University, Budapest, Hungary

Abstract

The main question of this study is if there are differences in the occurrence of disfluencies of young and old males and females depending on speech task. Frequency and types of disfluencies of 20 young and 20 old speakers were analyzed in three different speech tasks. Results show that speakers' age has significant effect on the frequency of disfluencies only in males' speech. There are disfluencies which are more characteristic of old speakers' speech, and others of young speakers' speech. Speech task has significant effect on the analyzed parameters in both ages, while sex has the least impact on frequency.

Introduction

Speech planning processes are closely related to cognitive processes like perception, memory and attention, which change with ageing (Holland & Rabbit, 1990; Humes, 1996; Schneider, Daneman, & Pichora-Fuller, 2002; Hnath-Chisolm, Willot, & Lister, 2003). In the elderly, hearing, speech understanding and memory capacity decrease (Holland & Rabbit, 1990; Humes, 1996; Schneider et al., 2002; Hnath-Chisolm et al., 2003; Duboisdindien, 2019), and the attentional processes change. This can lead to several speech planning problems.

The most characteristic feature of elderly speech is the word activation problem which refers to the deterioration of memory (Burke et al., 1991; Kemper, 1992); but there might also be problems on other speech planning levels. There are relatively few studies about disfluencies in elderly speech, and consensus hasn't been reached so far by the authors concerning frequency and occurences. Analyzing American English, some authors found that there were no differences between the speech of young and old speakers (Duchin & Mysak, 1987; Leeper & Culatta, 1995); while other authors (also in native American English speakers) found that elderly speakers produced more disfluencies than younger speakers did (Yairi & Clifton, 1972). Analyzing the speech of seven mentally intact 100–103-year-old American English speakers, it was found that disfluencies occurred with the same frequency in their speech as in the speech of 70–80–90-year-old speakers (Searl, Gabel, & Fulks, 2002). Similar results were found in the study of Andrade and Martins (2010) who investigated the speech of Brazilian elderly speakers. In their research there was no significant difference between the speech of 60, 70 and 80+ year-old people, although there was an increasing tendency of the disruption rates along the decades. In the background of the different results, there could be individual differences of speakers, different lengths of speech samples, counting syllables or words (only intended syllables or words, or every syllable and word), and the different speech tasks: Duchin and Mysak (1987) analyzed oral reading (Rainbow Passage), conversation (favorite summertime activities, jobs, family) and picture descripition tasks; Yairi and Clifton (1972) story-telling based on three picture-cards; Searl et al. (2002) investigated interviews where subjects were asked open-ended questions about a variety of topics; while Andrade and Martins (2010) gathered speech samples according to the speech Fluency Assessment Protocol (Andrade, 2000).

Taking both aspects into consideration, analyzing spontaneous speech in different speech tasks of various age groups (speakers between the ages 21 and 91) Duchin and Mysak (1987) found no differences in disfluencies between the age groups, but they found significant differences between the different speech tasks. According to their results, the frequency of disfluencies was higher in conversations than in picture description tasks.

In the study of Duchin and Mysak (1987) speech samples of only male speakers were analyzed, although it is well known that there might be sex related differences in the occurrence of disfluencies. It was found, that male speakers produce disfluencies more frequently than female speakers do (Bortfeld et al., 2001; Shriberg, 2001). Shriberg (2001) found that the main difference between males and females was in the frequency of hesitations; in her corpus male speakers produced more filled pauses than female speakers. Furthermore, she found that if in conversation the listener was male, speakers produced more disfluencies than if the listener was female (independently from the speakers' sex). In a research (in which gender was not confounded with occupation and education level, like in Shriberg, 2001), Bortfeld et al. (2001) obtained similar results as Shriberg (2001): male speakers produced more

https://doi.org/10.18463/diss-2021-018-bona

disfluencies than female speakers did. The frequency of hesitations and repetitions was significantly higher in males' speech than in females' speech, and there was a tendency-like (not significant) difference in the frequency of part-word repetitions between the two groups (men produced part-word repetitions more frequently). In another study (Branigan, Lickley, & McKelvie, 1999) it was proven that female speakers produced disfluencies less frequently than male speakers if they could see their speaking partner. The sex of the speaking partner didn't influence the occurrences of disfluencies.

The novelty of the study is that it unlike previous work which specifically targeted age or task, or sex alone, it triangulates evidence from all these effects within the same investigation. The aim of this study is to define differences in the frequency of disfluencies of young and old male and female speakers in three different speech tasks: in spontaneous narratives, in narrative recalls, and in conversations. In addition, another question is if there are not only quantitative but also qualitative differences in the occurrence of disfluencies depending on speakers' age and sex, and speech task.

The hypotheses were the following: 1) the age will influence the frequency of all disfluencies in every speech tasks; 2) the speech task have high effect on the frequency of disfluencies in both age groups; 3) there will be significant differences between males and females in both age groups and in all speech tasks.

Procedure

Subjects

Recordings of 40 subjects from the BEA Hungarian Spoken Language Database (Gósy, 2012) were selected for the study, all of them volunteered for the tasks. Twenty subjects were between 66 and 90 years of age (mean = 76.9 years), and twenty between 21 and 32 (mean = 25.3 years). In both groups there were 10 males and 10 females. All of them were native Hungarian speakers with normal hearing, and without any mental problem or speech disorder. All of them spoke standard Hungarian, and the subjects of the different (age and sex) groups were matched in education (all subjects had at least 12 years of education). Their speech rate and articulation rate were significantly different, young speakers spoke significantly faster than old speakers (their speech tempo was measured in Bóna 2014, see Table 1).

Table 1. Speech and articulation rates of the speakers (mean and Standard deviation) (Bóna 2014)

	Young speakers	Old speakers
Speech rate		
Narrative	4.3 (0.6)	3.6 (0.5)
Recall	3.6 (0.7)	3.2 (0.6)
Conversation	4.7 (0.6)	3.9 (0.5)
Articulation rate		
Narrative	5.8 (0.6)	4.8 (0.4)
Recall	5.4 (0.7)	4.6 (0.5)
Conversation	5.8 (0.6)	4.9 (0.5)

Material

Since recordings were selected from the same speech database, they were gathered in the same manner: the circumstances, the interviewer and the topics were the same in case of each speaker. Recordings were made with each subject in three situations which represented different speech tasks: 1) spontaneous narrative, 2) narrative recall, and 3) a three-participant conversation. These speech tasks require various cognitive skills with various levels of difficulty. 1. In spontaneous narratives participants spoke about their own lives, families and hobbies, and the interviewer only took the turn when they couldn't continue speaking. 2. In narrative recalls the task was to retell two heard text as accurately as possible. One of them was a popular science text (the duration was 97 s and contained 174 words), the other one was a historical anecdote (125 s and 270 words). 3. In conversations participants had to talk with two interviewers about everyday topics. One of the interviewers was always the same person who recorded the other two tasks, too. She was a young woman. The other interviewer was another young person, in some cases a male, in other cases a female one. These situations were really conversations, which means that compared to the first two tasks which were rather monologic, this shared close characteristics with face-to-face interactions. All participants (the subject and the two interviewers) wanted to speak, and a "competition" developed among them for speaking. These conversations needed rapid reaction time, good speech perception (because participants had to respond to each other), and fast speech planning processes, however, speakers could plan their speech while the others spoke. Conversation is also a joint activity of the participants, whose common aim is to maintain it.

Because the frequency of disfluencies might depend on the sample length (McLaughin & Cullinan, 1989; Shriberg, 1996), for each speaker the same length of speech material was selected. Taking the shortest duration of recalls into consideration, 300 syllables were analyzed from each speech task. This is longer than the recommended 200 syllables, and longer than 100 words which is usual in other

literature (Andrade & Martins, 2010; Roberts, Meltzer, & Wilding, 2009). Only intended syllables were calculated as conventional in the literature (Andrade & Martins, 2010; Roberts et al., 2009). The 300 syllables were selected from the recordings after the first 30 seconds, so the speaker had time to become comfortable in the speaking situation.

Data analysis

For comparability with data of previous studies, the number of disfluencies per 100 syllables (Roberts et al., 2009) were calculated, although ratio per hundred word is also commonly used in the literature. This means that the number of disfluencies occurred in the 300-syllable-long speech sample was divided into three. (It was only important for the comparability with other researches.) Each occurrence and type of disfluencies were identified and coded by the author. For reliability, two weeks after the first encoding, the author repeated the encoding in all speech material. The rate of agreement was 99.2% between the two coding. The cases which were not uniformly identified were excluded from the study.

Disfluencies were categorized in the following types (Roberts et al., 2009): interjections, revisions, word- or phrase-repetitions, part-word repetitions, and lengthenings. The data were compared across the two age groups, sex, and three speech tasks.

Statistical analyses (in case of normal distribution repeated-measure ANOVA, UNIANOVA, Tukey post hoc test, in case of non-normal distribution Mann–Whitney test and Wilcoxon-test) were performed by the SPSS 13.0 software at the 95% confidence level.

Results

Altogether 2097 disfluencies were analyzed. Not taking account of sex, there were very few significant differences between young and old speakers in the frequency of disfluencies. Young speakers produced more word- or phrase-repetitions in narratives ($Z = -2.658$; $p = 0.008$); and more lengthenings ($Z = -2.514$; $p = 0.012$) and less revisions ($Z = -3.386$; $p = 0.001$) in conversations than old speakers.

The frequency of types of disfluencies depending on age, sex and speech tasks is summarized in Table 2. The most frequent occurrence characterized recalls in both age and gender. Disfluencies were more frequent in males' speech by young speakers, and in females' speech in old speakers (except conversation).

According to the statistical analysis, significant differences occurred only in some types of disfluencies between young and old speakers in all speech tasks. As regards all occurrences of disfluencies, there were significant differences between the age groups only in males' narratives [$F(1, 19) = 8.704$; $p = 0.009$; $\eta^2 = 0.326$], and recalls [$F(1, 19) = 13.094$; $p = 0.002$; $\eta^2 = 0.421$], while in conversation, and in females' speech there were no significant differences between young and old speakers. In narratives, there were significant differences between the two age groups in the frequency of interjections [$F(1, 19) = 8.322$; $p = 0.010$; $\eta^2 = 0.316$] and word- or phrase-repetitions [$F(1, 19) = 8.203$; $p = 0.010$; $\eta^2 = 0.313$] of male speakers. In recalls, there were significant differences between young and old male speakers in interjections [$F(1, 19) = 12.153$; $p = 0.003$; $\eta^2 = 0.403$]. In conversations, there were significant differences in lengthenings [$F(1, 19) = 11.782$;

Table 2. Frequency of types of disfluencies per 100 syllables (mean and standard deviation). YW = Young women, YM = Young men, OW = Old women, OM = Old men

	Narrative	Recall	Conversation
All disfluencies per 100 syllables			
YW	4.5 (1.5)	8.6 (2.9)	4.2 (2.0)
YM	7.2 (3.1)	9.6 (4.1)	5.8 (3.5)
OW	4.5 (2.3)	10.0 (4.2)	3.8 (1.6)
OM	3.7 (2.0)	4.2 (2.3)	3.9 (1.6)
Interjections per 100 syllables			
YW	2.7 (1.0)	5.2 (2.2)	2.7 (2.0)
YM	4.4 (1.9)	6.2 (3.1)	2.9 (2.3)
OW	3.1 (2.0)	5.8 (3.1)	2.0 (1.0)
OM	2.2 (1.4)	2.3 (1.6)	1.6 (1.1)
Word- or phrase-repetitions per 100 syll.			
YW	0.6 (0.5)	0.6 (1.0)	0.9 (0.8)
YM	1.0 (0.8)	0.9 (0.9)	1.5 (0.9)
OW	0.4 (0.4)	1.0 (0.8)	0.9 (0.6)
OM	0.1 (0.4)	0.4 (0.4)	0.9 (0.9)
Part-word repetitions per 100 syllables			
YW	0.3 (0.3)	0.4 (0.4)	0.1 (0.1)
YM	0.2 (0.2)	0.4 (0.5)	0.6 (0.5)
OW	0.1 (0.2)	0.4 (0.4)	0.3 (0.5)
OM	0.1 (0.2)	0.3 (0.2)	0.5 (0.4)
Lengthenings per 100 syllables			
YW	0.8 (0.8)	1.7 (1.2)	0.4 (0.3)
YM	1.4 (1.1)	1.9 (1.3)	0.6 (0.6)
OW	0.7 (0.6)	1.7 (1.4)	0.0 (0.1)
OM	0.8 (0.5)	0.9 (0.8)	0.4 (0.5)
Revisions per 100 syllables			
YW	0.2 (0.2)	0.6 (0.7)	0.1 (0.2)
YM	0.2 (0.3)	0.3 (0.4)	0.1 (0.3)
OW	0.2 (0.2)	1.1 (0.9)	0.5 (0.5)
OM	0.4 (0.2)	0.3 (0.4)	0.5 (0.4)

Table 3: Results of the statistical analysis (the comparison of speech tasks in the two age groups)

Type of disfluency	Disfluencies per 100 syllables		
	F	***p***	η^2
Young women			
All	14.793	0.001	0.622
Interjection	7.464	0.008	0.453
Part-word repetition	4.504	0.043	0.334
Lengthening	9.967	0.006	0.525
Revision	4.348	0.043	0.326
Young men			
All	5.602	0.028	0.384
Interjection	6.958	0.010	0.436
Word- or phrase-repetition	3.809	0.047	0.297
Lengthening	7.947	0.007	0.469
Old women			
All	18.054	0.001	0.667
Interjection	8.262	0.008	0.479
Word- or phrase-repetition	4.625	0.028	0.339
Lengthening	7.701	0.014	0.461
Revision	4.878	0.033	0.352
Old men			
Word- or phrase-repetition	6.257	0.019	0.410
Part-word repetition	6.142	0.014	0.406

Table 4: Results of the Tukey post hoc test (the comparison of the speech tasks, p value). N = spontaneous narratives, R = Narrative recall, C = Conversation

Type of disfluency	N&R	N&C	R&C
Young women			
All	0.002	–	0.001
Interjection	0.010	–	–
Part-word repetition	–	–	0.033
Lengthening	0.003	–	0.023
Revision	0.031	–	0.016
Young men			
All	–	–	–
Interjection	–	–	0.039
Lengthening	–	0.030	0.029
Old women			
All	0.006	–	0.001
Interjection	–	–	0.012
Lengthening	–	0.030	0.029
Revision	0.047	–	–
Old men			
Word- or phrase-repetition	–	0.034	–
Part-word repetition	–	0.025	–

$p = 0.003$; $\eta^2 = 0.396$] and revisions [$F(1, 19) = 5.439$; $p = 0.032$; $\eta^2 = 0.232$] of female speakers, and in revisions of male speakers [$F(1, 19) = 5.702$; $p = 0.028$; $\eta^2 = 0.241$], too.

In both age groups, there were disfluencies the frequency of which was significantly different in the three speech tasks (for statistical data see Table 3). The most differences were between narrative recalls and conversations, while the least were between spontaneous narratives and conversations (Table 4).

Regarding the occurrence of all types of disfluencies (Table 2. and Table 3.), there were significant differences in the speech of young males, in the speech of young females, and in the speech of old females, while in the speech of old males there were no significant differences between narratives, recalls, and conversations. Analyzing the types of disfluencies, in the speech of young females, there were significant differences between the speech tasks in interjections, part-word repetitions, lengthenings, and revisions. In the speech of young males, there were significant differences between the speech tasks in interjections, repetitions, and lengthenings. In the speech of old females, there were significant differences between the speech tasks in interjections, repetitions, lengthenings, and revisions. In the speech of old males, there were significant differences between the speech tasks in word- or phrase-repetitions and part-word repetitions.

Differences between males and females were analyzed, too. Regarding to all disfluencies, between males and females there were significant differences only in the narratives of young speakers [$F(1, 19) = 5.490$; $p = 0.031$; $\eta^2 = 0.234$] and recalls of old speakers [$F(1, 19) = 14.653$; $p = 0.001$; $\eta^2 = 0.449$]. In the speech of young speakers, interjections of narratives [$F(1, 19) = 6.532$; $p = 0.020$; $\eta^2 = 0.266$], and part-word repetitions of conversations [$F(1, 19) = 10.913$; $p = 0.004$; $\eta^2 = 0.377$] showed significant differences. There were no significant sex differences in any type of disfluencies in recalls. In the speech of old speakers, there were significant differences between males and females in interjections [$F(1, 19) = 9.862$; $p = 0.006$; $\eta^2 = 0.354$] and revisions [$F(1, 19) = 7.525$; $p = 0.013$; $\eta^2 = 0.295$] of recalls, and in lengthenings of conversations ($Z = -2.715$; $p = 0.007$).

Discussion and Conclusion

In this paper frequency and types of disfluencies were analyzed depending on speakers' age, sex and speech task. The hypothesis relating to age (the first hypotheses) is partially confirmed. Although there were some differences in the frequency of disfluencies of both sex and in all speech tasks between young and old speakers, significant differences occurred only in the speech of male speakers. The occurrences of disfluencies in the speech of old males were less frequent than in young ones'. Analyzing the types of disfluencies revealed that there are disfluencies which are more characteristic of the speech of old speakers, or of the speech of young speakers, respectively, but the frequency depends on the speakers' sex and speech task, too. In the speech of young speakers, interjections, word- or phrase-repetitions, and lengthenings were more frequent, while in the speech of old speakers, revisions. The latter might be due to that in old age lexical access becomes more difficult (Schmitter-Edgecombe, Vesneski, & Jones, 2000), which is often associated with false start or false-word activation and their revisions. In the background of the significantly less disfluencies of old males, there might be slower speech rate, because slowing-down of speech rate can give more time for speech planning. The fact that there were no differences in female's speech between young and old speakers might have two reasons. One reason might be the individual differences (and standard deviation), and another reason might be the extremely low frequencies of certain types of disfluencies.

The second hypothesis regarding the effect of speech task has been confirmed in this study: disfluencies occurred more frequently in recalls than in narratives in both age groups, and the frequency of disfluencies decreased in conversations (which contained shorter turns, and gave more time for speech planning) compared to the speech tasks. In both age groups there were disfluencies which were more characteristic of the particular speech task. Interjection was the type of disfluency which showed significant difference in most speakers' groups. The frequency of this type greatly increased in recalls.

From the three factors, sex had the least impact on frequency (third hypothesis). In young speakers' groups there were less frequent disfluencies in females' speech, while in old speakers' groups in males' speech, but the frequency values were strongly influenced by speech tasks and types of disfluencies, too. Results show that in the analyses of the frequency of disfluencies, speakers' age, sex and speech tasks always have to be considered.

Acknowledgements

This study was supported by the National Research, Development and Innovation Office of Hungary, project No. K-128810, and the Thematic Excellence Program.

References

Andrade, C. R. F. de 2000. Protocolo para avaliação da fluência da fala [Speech fluency assessment protocol]. *Pró-Fono Revista de Atualização Científica* 12(2), 131–134.

Andrade, C. R. F. de & V. de O. Martins. 2010. Speech fluency variation in elderly. *Pró-Fono Revista de Atualização Científica* 22(1), 13–18. doi.org/10.1590/S0104-56872010000100004

Bóna, J. 2014. Temporal characteristics of speech: The effect of age and speech style. *Journal of the Acoustical Society of America* 136 (2), EL116–EL121. https://doi.org/10.1121/1.4885482

Bortfeld, H., S. D. Leon, J. E. Bloom, M. F. Schober, & S. E. Brennan. 2001. Disfluency rates in conversation: efects of age, relationship, topic, role, and gender. *Language and Speech* 44(2), 123–147. https://doi.org/10.1177/00238309010440020101

Branigan, H., R. Lickley, & D. McKelvie. 1999. Non-linguistic influences on rates of disfluency in spontaneous speech. In: J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (eds.), *Proceedings of the 14th International Conference of Phonetic Sciences*, 1–7August, 1999, San Francisco, CA, USA, 387–390.

Burke, D. M., D. G. MacKay, J. S. Worthley, & E. Wade. 1991. On the tip of the tongue: What causes word finding failures in young and older adults. *Journal of Memory and Language* 30(5), 542–579. https://doi.org/10.1016/0749-596X(91)90026-G

Duboisdindien, G. 2019. *Analyse multimodale des marqueurs pragmatiques au sein du vieillissement langagier en situation de Trouble Cognitif Léger* [Multimodal analysis of pragmatic markers in language aging in a situation of Mild Cognitive Disorder]. Ph.D. dissertation, Université Paris-Nanterre.

Duchin, S. W. & E. D. Mysak. 1987. Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders* 20(3), 245–257. https://doi.org/10.1016/0021-9924(87)90022-0

Gósy, M. 2012. BEA – A multifunctional Hungarian spoken language database. *Phonetician* 105–106, 50–61.

Holland, C. A. & P. M. A. Rabbit. 1990. Autobiographical and text recall in the elderly: An investigation of a processing resource deficit. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology* 42(3), 441–470. https://doi.org/10.1080/14640749008401232

Hnath-Chisolm, T., J. F. Willott, & J. J. Lister. 2003. The aging auditory system: anatomic and physiologic changes and implications for rehabilitation. *International Journal of Audiology* 42(Supplement 2), 3–10.

Humes, L. E. 1996. Speech understanding in the elderly. *Journal of the American Academy of Audiology* 7(3), 161–167.

Kemper, S. 1992. Adults' sentence fragments: Who, what, when, where, and why. *Communication Research* 19(4), 445–458. https://doi.org/10.1177/009365092019004003

Leeper, L. H. & R. Culatta. 1995. Speech fluency: Effect of age, gender and context. *Folia Phoniatrica et Logopedia* 47(1), 1–14. https://doi.org/10.1159/000266337

McLaughin, S. F. & W. L. Cullinan. 1989. Disfluencies, utterance length, and linguistic complexity in nonstuttering children. *Journal of Fluency Disorders* 14(1), 17–36. https://doi.org/10.1016/0094-730X(89)90021-1

Roberts, P. M., A. Meltzer, & J. Wilding. 2009. Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of communication disorders* 42(6), 414–427. https://doi.org/10.1016/j.jcomdis.2009.06.001

Schneider, B. A., M. Daneman, & K. M. Pichora-Fuller. 2002. Listening in aging adults: From discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology* 56(3), 139–152. https://doi.org/10.1037/h0087392

Searl, J. P., R. M. Gabel, & J. S. Fulks. 2002. Speech disfluency in centenarians. *Journal of Communication Disorders* 35(5), 383–392. https://doi.org/10.1016/S0021-9924(02)00084-9

Shriberg, E. 1996. Disfluencies in Switchboard. In: *The 4th International Conference on Spoken Language Processing (Addendum)*, October 3–6, 1996, Philadelphia, PA, USA, 11–14.

Shriberg, E. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31(1), 153–169. https://doi.org/10.1017/S0025100301001128

Schmitter-Edgecombe, M., M. Vesneski, & D. Jones. 2000. Aging and word finding: A comparison of discourse and nondiscourse tests. *Archives of Clinical Neuropsychology* 15(6), 479–493. https://doi.org/10.1016/S0887-6177(99)00039-6

Yairi, E. & N. F. Clifton. 1972. Disfluent speech behavior of preschool children, high school seniors, and geriatric persons. *Journal of Speech and Hearing Research* 15(4), 714–719. https://doi.org/10.1044/jshr.1504.714

Dynamic changes of pausing in triadic conversations

Dorottya Gyarmathy, Valéria Krepsz, Anna Huszár and Viktória Horváth
Hungarian Research Centre for Linguistics, Budapest, Hungary

Abstract

Pausing in conversation has several roles from speech planning to managing turn-takings (TTs). However, less is known about the dynamic changes of pauses over time or with regard to the turn-taking system. The frequency and the duration of silent and filled pauses (SPs and FPs) as well as shared silences was analyzed in 20 triadic Hungarian conversations using dynamic frames (altogether more than 7700 items). Data showed that the frequency of silent and FPs decreased over time across conversations. As opposite, shared silences were found to be the most frequent in the last sections of conversations. However, the duration of the pauses did not change over time across conversation—it may be influenced by other factors. We found that the SPs containing audible breathing were longer than other SPs. The SPs were less frequent before turn-takings than in other positions. However, their duration was not affected by the turn-taking system.

Introduction

Pauses have many kinds of roles in communication (e.g. respiration, cognitive load, production problems), both in speech production and perception next to boundary marking. So not every pause necessarily behaves as TRP (transition relevance place, which is defined as timing when the current speaker's turn can be completed and other participants are able to take the turn, cf. Sacks, Schegloff, & Jefferson, 1974), several factors can affect its appearance, frequency and duration as well. Local and Kelly (1986) investigated two different kinds of pauses: 1. 'trail-off silences' (a possible point for switching the role of the speaker) and 2. 'holding silences' (the speaker keeps the floor, it serves as an inhalation point or as a rhetorical tool). In case of trail-off pauses they found open glottis, out-breath, vowel centralization, and diminished loudness and tempo, preceded the pause. In the case of holding silence, they found closed glottis and no final lengthening preceding the pause. Levelt (1989) also differentiated types of silences according to their position and function. The speakers' tempo increased in the vicinity of syntactic boundaries to keep the floor and the rights of speaking, however they slow down and take a pause in the next phrase (Schegloff, 1996, Eggins & Slade, 1997).

The analysis of breathing in dyadic conversations corroborated that the speakers coordinate breathing to turn-takings (TTs). Inhalations inside a turn were shorter than when starting a new turn, suggesting that participants also adapt their breathing to hold turns (Rochet-Capellan & Fuchs, 2014). Inbreaths were analyzed in question-answer sequences in Dutch conversations, and they were found to be more frequent preceding long answers than short answers (Torreira, Bögels, & Levinson, 2015).

Filled pauses (FPs) also have several functions in the organization of the TT system as well. FPs may have pragmatic functions as indicators of the Feeling-of-Another's-Knowing in a dialogue (Brennan & Williams, 1995), or as turn-holders (Stenström, 1994). Therefore, some works described FPs as an interactional phenomenon (Levinson, 1983, Clark, 1994). FPs mark for the listeners that the next utterance will be more complex and the speaker needs more time for speech planning. Swerts (1998) found that FPs after stronger breaks tend to occur phrase-initially, whereas the majority of the FPs after weak boundaries are in phrase-internal position. The type and the position of FPs showed connection: 'um' was found to be more frequent at turn-initial position than 'uh', while 'uh' occurred rather at turn-medial position. Another study corroborated that FPs are often used to initiate the speaker's turn. In addition, when a speaker is confronted with unsuccessful answers in the course of the dialog, hesitations may also stand for marking his/her embarrassment and wish to close the dialog (Vasilescu, Rosset, & Adda-Decker, 2010). Isolated FPs occurred more frequently within their host unit than between clauses in English and French as well (Crible, Degand, & Gilquin, 2017). The FPs were also analyzed with regard to TTs from the Columbia Games Corpus (Benus, 2009). 33% of all FPs were in turn-initial position; so, FPs are linked to TT because these peripheral positions suggest several floor-management functions. FPs in this pre-start function allows the speaker some time for planning and the listener for tuning in.

The aim of the present study was to analyze the silent and filled pauses with regard to their position in the conversations. The main question was, how does pausing change across conversations? Which part of the conversation does contain the least pause or the shortest shared silences (ShS)?

Our hypotheses are the following:

https://doi.org/10.18463/diss-2021-019-gyarmathy-etal

1. We assume that during the conversations, the frequency and duration of silent, filled, and shared pauses decreases due to the accustoming and synchronization of the speakers. At the end of the conversations, an increase would be observed, as the participants run out of the topic of conversation and intend to close the communication event.

2. Silent pauses (SPs) are less frequent and shorter near to TTs, while more frequent and longer further from TTs.

3. FPs would occur less frequently in the position near before to the TTs than further from them. In addition, FPs would be more frequent and longer after turn-taking in turn-initial position (cf. Swerts, 1998, Benus, 2009).

Material and method

20 conversations were selected for the present study from the Hungarian Spontaneous Speech Database (BEA, cf. Neuberger et al., 2014) prepared in the phonetic lab of the Hungarian Research Centre for Linguistics. The BEA database consists of 460 recordings, which contain 7 different speech tasks, for example reading sentences and text, narratives. The conversation task is the 5th task in the whole recording. Three people participate in each conversation: the fieldworker1 (Fw1), the experimental speaker (S) and fieldworker2 (Fw2). The conversations are seminatural: the participants have no time for preparation, the first topic is given by the Fw1, but further topics are not fixed—the speech planning processes and the organization of the conversation are spontaneous. The two fieldworkers were the same people in each conversation (two female speakers, linguists, colleagues, 27–38 years old during conducting the database), while S changes across conversations (aged between 20–45 years). The conversations are about 18 mins long on average (8.5–23.5 min), the 20 conversations took almost 6 hours. The annotation of the material was carried out manually using Praat (Boersma & Weenink, 2018) by two trained annotators. The value of the inner-annotator agreement was 95%. In the case of disagreement, a third senior annotator checked the problematic parts and helped to decide. The annotation includes the level of interpausal units of the 3 speakers, the SPs and the hesitations as well. Furthermore, TTs, overlapping speech, backchannel responses were annotated in additional tiers (Horváth et al., 2019).

The patterns of the (silent and filled) pauses as well as shared silences were analyzed: i) frequency, ii) duration iii) types iv) audible breathings in SPs. Silent pauses and shared silences were differentiated based on their position: SPs were defined within a speaker's utterance, while ShSs were defined between the different speaker's units, when no one was speaking. The analysis was carried out using a dynamic approach: how these patterns change i) over time across conversations ii) near and further from TTs? The changes over time were analyzed using the following method: each conversation was split into 5 equal parts based on their duration automatically by a Praat script. For example, a 15-minute long conversation was cut up into five 3-minute long subsections (0–20%, 21–40% etc.). With this method, we can eliminate the unequal durations of the conversations, and the occurences of the given parameter can be comparable. The connection between pauses and TTs was analyzed with the following method (see Figure 1). The distance between pauses and the nearest TT was extracted automatically using a Praat script. The pauses and in most of the cases the TTs are not a point extend phenomenons; therefore, we calculated with the centers of the intervals. Based on the distance values of the pauses from the nearest TTs, the pauses were split into four groups according to two parameters: 1) nearer or further from TTs 2) before or after TTs. The border of closure vicinity was determined at 5 s based on the length of turns and the context.

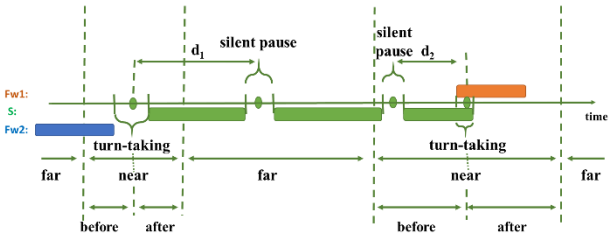

Figure 1. Method of the analysis of pauses near and further from TTs

The duration of the pauses was analyzed with linear mixed models in the R program (R Core Team, 2018) with the lme4 package (Bates et al., 2015), and the *p* values were obtained by Satterthwaite approximation (lmerTest package, ANOVA function, Kuznetsova, Brockhoff, & Christensen, 2015). The independent factors were the duration of the pauses, while the dependent factors were the five-partitions of the conversations. For each parameter, a random intercept and slope model were used (with the speaker as a random factor for each variable) and compared to the two models. There were no significant differences between the models, and because of the lower AIC (Akaike, 1973) values, the random intercept models were used during the analyzes. The frequency of the pauses were analysed

with Friedman Test (R Core Team, 2018): the dependent variable was the frequency of the pauses and the independent variable was the position of the pauses (near before, near after, further before, further after TT).

Results

Silent pauses

5881 SPs occurred in the 20 recordings. The mean frequency was 18.03 SPs/min ($SD = 4.76$). The frequency of SPs was analyzed in the 5 equal parts of the conversations. Results showed that SPs' occurrence was affected by their position in the conversation: $\chi^2(4) = 18.025$, $p = 0.001$ (Figure 2). They occurred least frequently in the first part of the conversation, then their frequency increased in the middle sections, while decreased again in the last section of the conversations.

The mean duration of the SPs was 431 ms ($SD = 336$ ms). The duration of SPs was also analyzed with regard to their position in the conversation. Data showed that there was no significant difference between the sections of the conversation in the duration of SPs. SPs were also analyzed with regard to their breathiness. 35% of the SPs contained audible breathing. The SPs with audible breathing were significantly longer than pauses without audible breathing [$F(1, 5781) = 248.625$, $p < 0.001$], irrespectively of the participant's role (Figure 3).

The mean duration of SPs with audible breathing were 577 ms ($SD = 305$ ms), without audible breathing were 354 ms ($SD = 326$ ms).

Filled pauses

A total of 1240 FPs occurred in the 20 recordings. The mean frequency of the FPs was 3.77 item/min ($SD = 2.31$). The dynamic change in the frequency data was analyzed in the 5 equal parts of the conversations. FPs occurred the least frequently in the last section (*mean*=3.5 item/min), while the most frequently in the 2nd section (*mean* = 3.87 item/min). However, the difference was not significant between the sections. The type of the FPs was analyzed. 57% of the FPs occurred as a monophthong schwa, while 35% realized as a nasal consonant. The ratio of diphthongs (like [əm] or [əh]) was altogether less than 10%. The duration of FPs significantly differed from their forms ($F(1239, 4) = 31.439$, $p < 0.001$): the more sounds the FP involved the longer duration it had (e.g. the duration of the swa form ([ə]) was on average 306 ms, while the average duration of [əhm] was 630 ms).

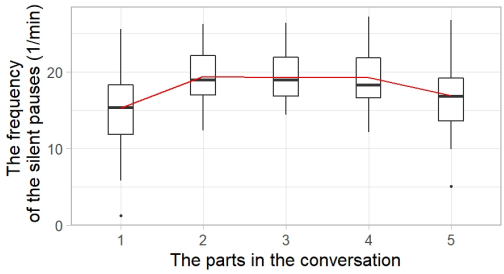

Figure 2. The frequency of the SPs in the 5 equal parts of the conversations (red line represents the means while black line on the boxes represents the medians).

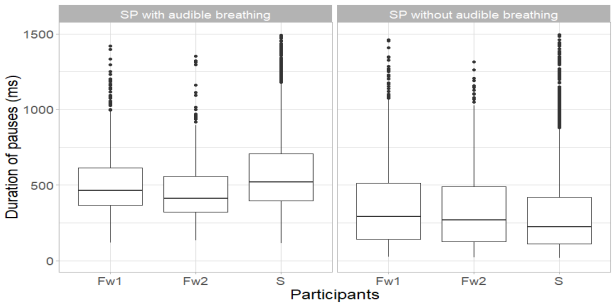

Figure 3. The duration of the SP with regard to breathiness and participant's role.

Shared silences

628 shared silences occurred in the 20 conversations with a mean duration of 510 ms ($SD = 513$ ms). The frequency of silences was 1.98 item/min on average ($SD = 1.38$). The frequency was also analyzed with regard to the changes over time across conversations. The conversations were split into 5 equal parts, and the occurrence of the silences was analyzed in these equal parts. Significant difference was found between the parts of the conversations in the number of silences per minute [$F(4, 76) = 3.684$, $p < 0.05$]. The least silences were found in the middle of the conversations (*mean* = 1.40 item/min), while the most of the silences occurred at the last two sections of the conversations (*mean* = 2.59 item/min, Figure 4).

The duration of the shared silences was analyzed with regard to their position of the conversation (in the 5 equal parts). The standard deviation of the values was huge and showed great overlaps; therefore data can not show any trend (Table 1).

Pauses and turn-taking

The dynamic changes of pausing was not only analyzed with regard to the equal parts of the conversation, but with regard to the TT system as well. The frequency of the SPs was analyzed with regard to their position to the TTs (near before, near

after, further before, further after TT), and significant differences were found among the positions ($\chi^2(3) = 19.599$, $p < 0.001$, Figure 5).

SPs occurred less frequently near the TTs than further from TTs. The frequency of the FPs were analyzed with regard to their position to the TTs, and the data showed significant differences among the positions ($\chi^2(3) = 17.65$, $p < 0.001$, Figure 6); FPs were the least frequent near before TTs.

The duration of the SPs and FPs was analyzed with regard to the distance from TTs. The duration of pauses did not differ significantly near TTs compared to further position from TT.

Discussion

Dynamic changes of pausing were analyzed in triadic conversations, firstly in Hungarian. The aim was to analyze how pausing changes in conversation over time as well as in the vicinity of TTs. Based on the analysis of more than 7700 items, results corroborated the first hypothesis: the frequency SPs and FPs changed over time across conversations. Pauses were the least frequent in the first and in the last sections. However, the duration of the pauses did not change over time across conversation—it may be influenced by other factors. One of these factors may be the breathiness: we found that the SPs containing audible breathing were longer than other SPs. The frequency of pauses with regard to turn-takings was analyzed as well. The SPs were less frequent in the vicinity of turn-takings than in other positions, according to our second hypothesis. Based on an earlier study for Hungarian on the same corpus (Horváth et al., 2021), the articulation rate was found to be increased in the vicinity of turn-takings. The increasing rate with the decreasing frequency of SPs signals that the current speaker is not yielding the floor yet ("rush-through", cf. Walker, 2010). FPs occurred the least frequently near before TTs, according to our hypothesis. The analysis of shared silences showed that their frequency changed over time, however, the difference was not significant. They occurred the least frequently in the middle section of the conversations. The silences were the most frequent in the last section of conversation marking that the participants were getting run out of the topic—the fieldworker should end the conversation. The duration of pauses was not affected by the TT system significantly, contrary to our hypotheses. Our results based on conversations add new information on the timing patterns as well as on the fluency patterns of speech, which was mainly analyzed previously in narrative speech style.

Table 1. Duration of shared silences in the 5 equal parts of the conversations.

Duration of silences in the 5 parts of the conversations (ms)		
parts of conversations	mean	SD
1	541	534
2	400	426
3	388	364
4	518	460
5	625	633

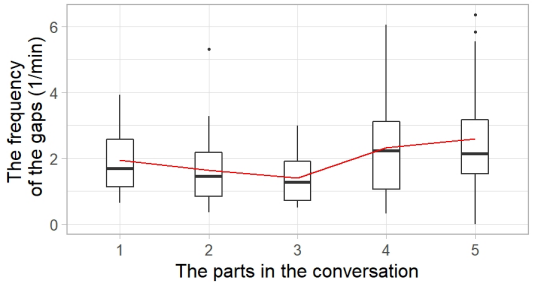

Figure 4. The frequency of shared silences in the 5 equal parts of the conversations (red line represents the means while black line on the boxes represents the medians).

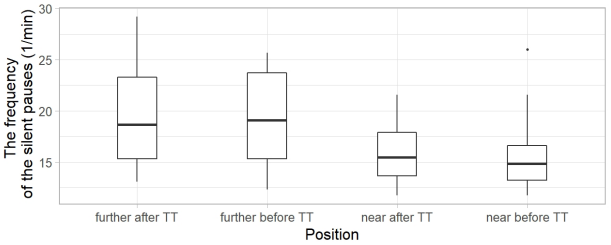

Figure 5. The frequency of the SPs according to their position to the TTs.

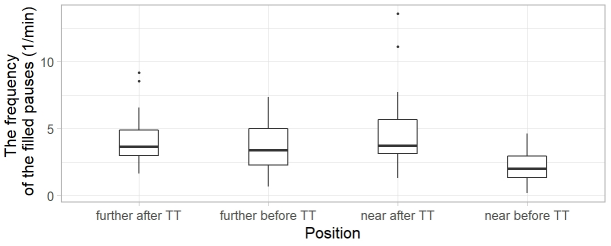

Figure 6. The frequency of the FPs according to their position to the TTs.

Acknowledgements

The research was supported by the Hungarian National Research, Development and Innovation Office of Hungary [projects No. K-128810] and the Bolyai János Research Scholarship.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov & F. Caski (eds.), *Proceedings of the Second international symposium on information theory*, Budapest: Akadémiai Kiadó. 267−281.

Bates, D., M. Mächler, B. Bolker, & S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Benus, S. 2009. Variability and stability in collaborative dialogues: Turn-taking and filled pauses. In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 6–10 September, 2009, Brighton, UK, 796–799.

Boersma, P. & D. Weenink. 2018. Praat: Doing phonetics by computer (version 6.1.38). https://www.praat.org/ (accessed 24 January 2021).

Brennan, S. E. & M. Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* 34(3), 383–398. https://doi.org/10.1006/jmla.1995.1017

Clark, H. H. 1994. Managing problems in speaking. *Speech Communication* 15(3–4), 243–250. https://doi.org/10.1016/0167-6393(94)90075-2

Crible, L., L. Degand, & G. Gilquin. 2017. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. Languages in Contrast 17(1), 69–95. https://doi.org/10.1075/lic.17.1.04cri

Eggins, S. & D. Slade. 1997. *Analysing casual conversation*. London, UK: Cassell.

Horváth, V., V. Krepsz, D. Gyarmathy, Á. Hámori, J. Bóna, C. I. Dér, & Z. Weidl. 2019. Háromfős társalgások annotálása a BEA-adatbázisban: elvek és kihívások [The principles and challenges of annotating the triadic conversations in the Hungarian Spontaneous Speech Database BEA]. *Nyelvtudományi Közlemények* 115, 255–274. https://doi.org/10.15776/NyK/2019.115.9

Horváth, V., V. Krepsz, D. Gyarmathy, A. Huszár, & Á. Hámori. 2021. Dynamic changes of speech patterns as cues of smooth turn-takings in Hungarian triadic conversations. Presentation at The Role of the Current Speaker in Conversational Turn Taking Workshop, 14–15 January, 2021, Berlin, Germany.

Kuznetsova, A., P. B. Brockhoff, & R. H. B. Christensen. 2015. LmerTest Package: Tests in linear mixed effects models. *Journal of statistical software* 82(13). http://dx.doi.org/10.18637/jss.v082.i13

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

Levinson, S. C. 1983. *Pragmatics*. Cambridge, UK: Cambridge University Press.

Local, J. & J. Kelly. 1986. Projection and 'silences': Notes on phonetic and conversational structure. Human Studies 9(2–3), 185–204. https://doi.org/10.1007/BF00148126

Neuberger T., D. Gyarmathy, T. E. Gráczi, V. Horváth, M. Gósy, & A. Beke. 2014. Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language. In: P. Sojka., A. Horák, I. Kopeček & K. Pala (eds.), *Text, Speech and Dialogue. TSD 2014, Lecture Notes in Computer Science*, Cham: Springer, 424–431. https://doi.org/10.1007/978-3-319-10816-2_51

R Core Team 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (version 4.0.5). https://www.R-project.org/

Rochet-Capellan A. & S. Fuchs. 2014. Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal. Society B* 369(1658), Article 20130399. https://doi.org/10.1098/rstb.2013.0399

Sacks, H., E. Schegloff, & G. Jefferson. 1974. A simplest systematic for the organization of Turn-Taking for conversation. *Language* 50(4/1), 696–735. https://doi.org/10.2307/412243

Schegloff, E. 1996. Turn organization: one intersection of grammar and interaction. In: E. Ochs, E. Schegloff, & S. Thompson (eds.), *Interaction and grammar*, Cambridge, UK: Cambridge University Press, 52–133. https://doi.org/10.1017/CBO9780511620874.002

Stenström, A.-B. 1994. *An Introduction to Spoken Language Interaction*, London, UK: Longman.

Swerts, M. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics* 30(4), 485–496. https://doi.org/10.1016/S0378-2166(98)00014-9

Torreira, F., S. Bögels, & S. C. Levinson. 2015. Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology* 6, Article 284. https://dx.doi.org/10.3389%2Ffpsyg.2015.00284

Vasilescu, I., S. Rosset, & M. Adda-Decker. 2010. On the Role of Discourse Markers in Interactive Spoken Question Answering Systems. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 17–23 May, 2010, Valletta, Malta, Paper 1333.

Walker, G. 2010. The phonetic constitution of a turn-holding practice: rush-throughs in English talk-in-interaction. In: D. Barth-Weingarten, E., Reber & M. Selting, (eds.), *Prosody in Interaction. Studies in discourse and grammar*, Amsterdam, The Netherlands: John Benjamins, 51–72.

Special day on (dis)fluency in speech and language disorders Preface

DiSS 2021 featured a co-located event on August 27th, 2021 which focused on (dis)fluency in speech and language disorders. Following are the abstracts from the various presentations given that day. Full paper versions of the research presented is expected to be published in 2022. The abstracts are provided here as a convenience to readers and as a record of the DiSS 2021 event. However, researchers are encouraged to read and cite the full versions of the following papers once they appear.

Disfluency characteristics predict stuttering persistency in preschool-aged children

Bridget Walsh
Michigan State University, Michigan, USA

Introduction

Early childhood stuttering is a neurodevelopmental speech disorder that typically emerges between the ages of two to five years, most often around a child's third birthday (Yairi & Ambrose, 2005). Approximately 5%–11% of children go through a period of stuttering (Andrews & Harris, 1964; Reilly et al., 2013; Yairi & Ambrose, 2005), yet most of these children, up to 80%, will recover from stuttering within a year or so of onset (Yairi & Ambrose, 2005). Children who persist are at risk for developing a chronic stuttering disorder that has negative implications for psychosocial development and academic and vocational achievement (Blumgart, Tran, & Craig, 2010; Craig, Blumgart, & Tran, 2009; Klein & Hood, 2004).

Given the relatively high incidence of stuttering in preschoolers, a significant concern is how to diagnostically differentiate children at risk for persisting from those children more likely to recover. Research that specifies factors associated with stuttering persistency is critical as it offers insight into the underpinnings of early childhood stuttering and helps clinicians identify children in need of immediate treatment. Prior research, from our laboratory and others has identified demographic factors associated with stuttering persistence, such as a positive family history of stuttering, being male, an older age at stuttering onset, and time since onset (Singer et al., 2020; Walsh and colleagues, 2018, 2021; Yairi & Ambrose, 2005).

We have also explored clinical factors associated with stuttering persistence. Results from this work reveals that phonological abilities assessed with a standardized assessment, accuracy on a nonword repetition test, and the frequency and nature of preschooler's stuttering-like disfluencies when children are 3–5 years of age differentiated children who eventually recovered or persisted in stuttering (Spencer & Weber-Fox, 2014; Walsh and colleagues, 2020, 2021).

All preschoolers produce developmentally appropriate disfluencies as a natural part of language/speech acquisition. These disfluencies include revisions, multisyllabic word and phrase repetitions, hesitation/pauses, and interjections. On the other hand, stuttering-like disfluencies (SLDs) such as sound or syllable repetitions (p…p…p…please), prolongations (sssssssssss nake), or blocks where no sound or air emerges occur less commonly in typical speakers. SLDs serve as a reliable diagnostic marker of stuttering in preschoolers (Ambrose & Yairi, 1999). The purpose of this study was to determine whether the type and frequency of disfluencies produced by children who stutter (CWS) during spontaneous speech predicted later stuttering persistence and recovery.

Method

We analyzed spontaneous speech samples from 47 preschool children aged 4–5 years diagnosed with early childhood stuttering using established criteria (Walsh et al., 2020). In this longitudinal study, children were reassessed in subsequent years to determine if their stuttering resolved or persisted. Based on these longitudinal diagnoses, we formed two groups of children: CWS who persisted (CWS-ePer = 18) and CWS who eventually recovered from stuttering (CWS-eRec = 29). We compared the frequency of subcategories of SLDs [part-word reps (PW), single-syllable whole word reps (SS), and Dysrhythmic Phonations (DP)—blocks, prolongations] between CWS-ePer and CWS-eRec. We also compared the frequency of typical disfluencies (TD): interjections, revisions, and multisyllabic word or phrase repetitions). Finally, we computed a weighted stuttering-like disfluency index (WSLD), a composite index of global stuttering disfluency severity that considers the frequency, type, and number of repetitions (Yairi & Ambrose, 1999) for each child. This index has been used to diagnose stuttering in young children, but we were interested to learn whether it was also sensitive to stuttering outcomes—persistence or recovery.

Results

Bonferroni-corrected Mann Whitney U tests indicated that there was not a significant difference between CWS-Per/Rec in the frequency of SS disfluencies, although PW and DP disfluencies were significantly higher in CWS-ePer. We found no differences in the occurrence of typical disfluencies (interjections, revisions, multisyllabic word repetitions, and phrase repetitions) between CWS-eRec and CWS-ePer. Bivariate regression results

revealed that the WSLD significantly predicted stuttering persistence. As the WSLD increased (denoting more severe stuttering) the odds of persisting also increased. See Walsh et al., 2020 for specific statistical results.

Conclusion

By ages 4 and 5 years, SLD characteristics of CWS-ePer and CWS-eRec have diverged and can be captured with the WSLD. Stuttering severity measured by the WSLD contributes to a sparse list of risk factors available to clinicians to evaluate a child's risk for persistence. We suggest that a 4–5-year-old's WSLD score be considered when assessing their risk for stuttering persistency keeping in mind that a comprehensive stuttering assessment should also consider reactions from the child/parents and feelings/attitudes emerging in a child who is stuttering.

References

Ambrose, N. G. & E. Yairi. 1999. Normative disfluency data for early childhood stuttering. *Journal of Speech, Language, and Hearing Research* 42(4), 895–909.

Andrews, G. & M. Harris. 1964. *The Syndrome of Stuttering*. Spastics Society Medical Education and Information Unit.

Blumgart, E., Y. Tran, & A. Craig. 2010. An investigation into the personal financial costs associated with stuttering. *Journal of Fluency Disorders* 35(3), 203–215. https://doi.org/10.1016/j.jfludis.2010.03.002

Craig, A., E. Blumgart, & Y. Tran. 2009. The impact of stuttering on the quality of life in adults who stutter. *Journal of Fluency Disorders* 34(2), 61–71. https://doi.org/10.1016/j.jfludis.2009.05.002

Klein, J. F. & S. B. Hood. 2004. The impact of stuttering on employment opportunities and job performance. *Journal of Fluency Disorders* 29(4), 255–273. https://doi.org/10.1016/j.jfludis.2004.08.001

Reilly, S., M. Onslow, A. Packman, E. Cini, L. Conway, O. C. Ukoumunne, E. L. Bavin, M. Prior, P. Eadie, S. Block, & M. Wake. 2013. Natural history of stuttering to 4 years of age: A prospective community-based study. *Pediatrics* 132(3), 460–467. https://doi.org/10.1542/peds.2012-3067

Singer, C. M., A. Hessling, E. M. Kelly, L. Singer, & R. M. Jones. 2020. Clinical Characteristics Associated With Stuttering Persistence: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research: JSLHR* 63(9), 2995–3018. https://doi.org/10.1044/2020_JSLHR-20-00096

Spencer, C. & C. Weber-Fox. 2014. Preschool speech articulation and nonword repetition abilities may help predict eventual recovery or persistence of stuttering. *Journal of Fluency Disorders* 41, 32–46. https://doi.org/10.1016/j.jfludis.2014.06.001

Walsh, B., A. Bostian, S. E. Tichenor, B. Brown, & C. Weber. 2020. Disfluency characteristics of 4- and 5-year-old children who stutter and their relationship to stuttering persistence and recovery. *Journal of Speech, Language, and Hearing Research* 63(8), 2555–2566. https://doi.org/10.1044/2020_JSLHR-19-00395

Walsh, B., S. Christ, & C. Weber. 2021. Exploring Relationships Among Risk Factors for Persistence in Early Childhood Stuttering. *Journal of Speech, Language, and Hearing Research* 64(8), 2909–2927. https://doi.org/10.1044/2021_JSLHR-21-00034

Walsh, B., E. Usler, A. Bostian, R. Mohan, K. L. Gerwin, B. Brown, C. Weber, & A. Smith. 2018. What are predictors for persistence in childhood stuttering? *Seminars in Speech and Language* 39(4), 299–312. https://doi.org/10.1055/s-0038-1667159

Yairi, E. & N. G. Ambrose. 1999. Early childhood stuttering I: Persistency and recovery rates. *Journal of Speech, Language, and Hearing Research* 42(5), 1097–1112.

Yairi, E. & N. G. Ambrose. 2005. *Early Childhood Stuttering for Clinicians by Clinicians*. Pro-Ed.

Speech rhythm abnormality in Japanese: Analysis of mora duration, pause, and non-segmented mora of dysarthric speech

Fumie Namba [1], Ryoko Hayashi [2] and Jun Tanemura [3]
[1]Kawasaki University of Medical Welfare, Okayama, Japan
[2]Kobe University, Hyogo, Japan
[3]Kawasaki University of Medical Welfare, Okayama, Japan

The purpose of this study is to determine the temporal factors that convey an abnormal impression of the Japanese speech rhythm to a listener. Abnormality of speech rhythm is observed especially in dysarthria and stuttering, but the criteria for evaluation may not be objective. For example, Fukusako et al. (1983) proposed an evaluation scale for paralytic speech in dysarthria and described the characteristics as "the impression of being scattered," "the speech rhythm broken down irregularly," and "unnaturally interrupted." Meanwhile, the Standardized Test for Stuttering (Ozawa et al., 2016) adopts the classification of "unnaturally prolonged" and "heard as unnatural in the fluency of speech." Most evaluations and classifications are based on subjective auditory impressions of speech pathologists. In this study, we aimed to provide objective indices for fluency of dysarthric speech, so that a patient's speech can be described more concretely and compared with other cases.

The subjects included six normal speakers (Nf1, Nf2, Nf3, Nm1, Nm2, Nm3; three women and three men; mean age 45) and six dysarthria patients with unnatural speech rhythm (Pm1–Pm6: Pm1–Pm3: Parkinson's disease; Pm4: cerebral infarction; Pm5 and Pm6: spinocerebellar degeneration; six men; mean age 61). The subjects were asked to read "The North Wind and the Sun" in Japanese. The duration of each mora and pause in the first four sentences of the text were measured, and the position and frequency of the pauses investigated.

In the utterances of the patients, the duration of a mora was widely distributed from 35 ms to 533 ms, while in normal speakers, it ranged from 50 ms to 350 ms. Japanese is a mora-timed language, and each mora is said to be of similar duration (Warner & Arai, 2001). However, morae with a very long duration (more than 400 ms) are found even in the middle of clauses in the utterance of patients. The difference in the duration between two adjacent morae was distributed from 0 to 235 ms for normal speakers, and from 0 to 314 ms for patients. Furthermore, approximately 15% of data points of patients were outliers, that is, more than Q3 + 1.5*IQR, whereas only 4% of those of normal speakers were outliers. Four of the six patients showed that the duration of one mora was three times longer than that of the adjacent mora.

Pauses at inappropriate positions in a sentence and very long pauses of more than 1,800 ms were also frequently found only in patients. If the pause was longer than 1,800 ms, there would no longer be the perception of a chain of patterns, but only isolated patterns (Fraisse, 1982).

Second, morae with no clear acoustic boundary ("non-segmented mora" (NSM)) were investigated. We defined the morae that were not segmented clearly on the spectrogram, including the vowel or semi-vowel transition as NSM. In the case of normal speakers, such morae are often observed in the vowel sequence, as shown in /taijoo/ of Figure 1. The patients were categorized based on the two types of utterances with more and less NSM than normal speakers. The former case was caused by unclear articulation of consonants (Figure 2), and the latter by pauses at inappropriate positions (a long pause after /ta/ in Figure 3). The speech rate of each sentence varied between patients (2.7–15.0 morae per second) and normal speakers (4.5–12.5 morae

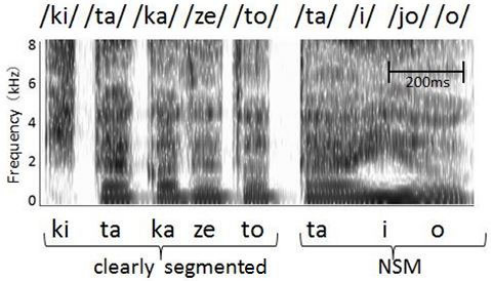

Figure 1. /ki ta ka ze to ta i jo o/ by Nf1

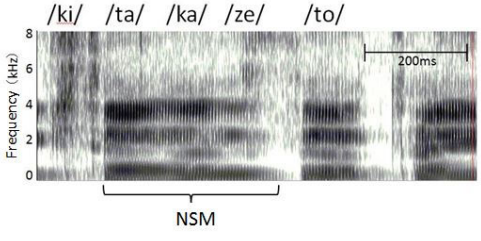

Figure 2. /ki ta ka ze to/ by Pm2

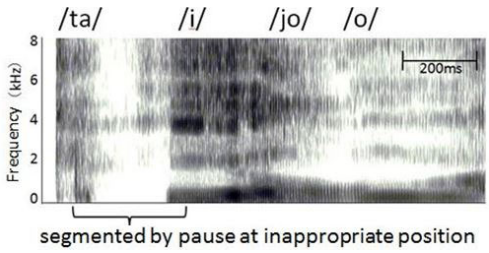

Figure 3. /ta i jo/ by Pm5

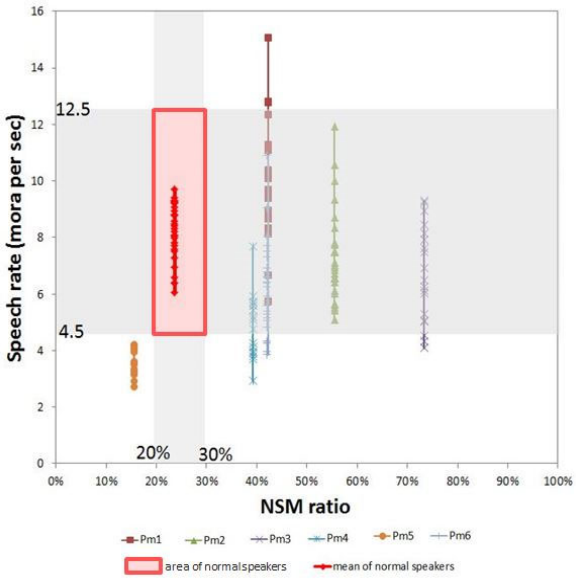

Figure 4. NSM ratio and speech rate of patients

per second). Figure 4 shows the relationship between the NSM ratio and the speech rate. The results showed that the NSM ratio ranged from 20% to 30% regardless of the speech rate in normal speakers, whereas it varied from case to case in patients.

In conclusion, the present study suggests the following necessary conditions for speech to not be classified as having abnormal rhythm: (1) The difference in the duration between two adjacent morae should be less than 235 ms, and one mora should not be longer than three times that of the adjacent mora. (2) The pauses should be placed in a syntactically correct position and not exceed 1,800 ms.

Collectively, NSM, speech rate, and isochrony of mora can be useful measures for determining speech rhythm in Japanese language for dysarthric speech.

References

Fraisse, P. 1982. Rhythm and tempo, In: D. Deutsch (ed.), *Psychology of Music*, New York, NY, USA: Academic Press, 149–180.

Fukusako, Y., H. Monoi, T. Itaru, K. Kumai, N. Hijikata, & H. Hirose. (1983). Mahi-sei (undō shōgai-sei) kōonshōgai no hanashikotoba no tokuchō — chōkaku inshō ni yoru hyōka — [Analysis of characteristics of dysarthric speech based on auditory impressions]. *Nihon onsei gengo igaku* [Journal of Japanese Society of Logopedics and Phoniatrics] 24(2), 149–164. https://doi.org/10.5112/jjlp.24.149

Ozawa, E., Y. Hara, N. Suzuki, H. Moriyama, Y. Oohashi, A. Iida, Y. Sakata, & N. Sakai. 2016. *Kitsuon kensahō* [Standardized Test for Stuttering 2nd edition], Tokyo: Gakuensha.

Warner, N. & T. Arai. 2001. The role of the mora in the timing of spontaneous Japanese speech, *The Journal of the Acoustical Society of America* 109(3), 1144–1156. https://doi.org/10.1121/1.1344156

Pauses and disfluencies in speech of patients with Multiple Sclerosis

Judit Bóna[1], Veronika Svindt[2] and Ildikó Hoffmann[2,3]
[1] ELTE Eötvös Loránd University, Budapest, Hungary
[2] Research Institute for Linguistics, Eötvös Loránd Research Network, Budapest, Hungary
[3] University of Szeged, Szeged, Hungary

Multiple Sclerosis (MS) causes various symptoms in speech production. About 60% of patients can be affected by these symptoms. The most frequent speech- and language symptoms in MS are the following: (1) disorders in articulation and voicing (a) dysarthria, (b) dysphonia, (c) deceleration of articulation rate and speech rate; (2) language disorders occurring in both speech production and perception (d) difficulties in word retrieval, (e) naming disorder, (f) decrease of verbal fluency, (g) semantic paraphrases, (h) disorders in speech comprehension (e.g. Renauld, Mohamed-Saïd, & Macoir, 2016; Sonkaya & Bayazit, 2018). According to the literature (Noffs et al., 2018; Svindt, Bóna, & Hoffmann, 2020; Feenaughty et al., 2021), the deceleration of speech rate is accompanied by more frequent pauses and an increase in the duration of pauses in the speech of MS patients.

There is almost no data about that whether the disease affects speech fluency in other ways, such as whether the frequency of disfluencies in speech changes. There is very little information on speech fluency in the literature regarding MS, and these studies do not show a difference between MS patients and controls in the occurrences of disfluencies. The main question of this study is the following: if we examine speech fluency in speech tasks requiring different cognitive load, will there be a difference between patients and controls?

15 MS patients and 15 age-, sex- and education-matched control speakers participated in the study (there were 14 females and 1 male in both groups). All of them were native Hungarian speaker without any hearing disorders and dementia. They were recruited by a neurologist for the study. All patients had speech symptoms according to their own subjective judgment. They disease began 6–33 years ago. There were two patients with secondary progressive multiple sclerosis, one with primary progressive multiple sclerosis, and twelve with relapsing-remitting multiple sclerosis among them.

Speech samples were recorded with all participants in three speech tasks: 1) spontaneous narratives about their own lives, 2) narratives about their day before, and 3) narrative recalls on the basis of a heard text. These three speech tasks require different cognitive loads during speech planning and production. 1) During spontaneous narratives about their own lives, the speakers can speak freely, they can plan both the content and the linguistic form. 2) During narratives about the day before, the speakers have to recall the events of the previous day, as accurately as possible. This task requires more cognitive load than the first task. 3) The narrative recall task requires the greatest cognitive load. The success of this task depends on the speech processing, attentional and working memory mechanisms, and narrative competence (Juncos-Rabadán & Pereiro, 1999).

Pauses and disfluencies were annotated in the speech samples by Praat, and duration of pauses were measured automatically. After that the frequency of pauses and disfluencies were calculated, and types of disfluencies were examined. The analyzed types of disfluencies were the following: interjections, revisions, word- or phrase-repetitions, part-word repetitions, and lengthenings (Roberts, Meltzer, & Wilding, 2009).

Results show that there are differences between MS patients and controls in the frequency and duration of pauses. However, the frequency of disfluencies showed differences between the two groups only in the speech tasks which required more cognitive load. The same types of disfluencies occurred in the same proportion in both speakers' group.

Results help to better understand speech production processes in MS.

Acknowledgments

This study was supported by the National Research, Development and Innovation Office of Hungary, project No. K-132460, and the Thematic Excellence Program.

References

Feenaughty, L., L. Guo, B. Weinstock-Guttman, M. Ray, R. Benedict, & K. Tjaden. 2021. Impact of Cognitive Impairment and Dysarthria on Spoken Language in Multiple Sclerosis. *Journal of the International Neuropsychological Society* 27(5), 450–460. https://doi.org/10.1017/s1355617720001113

Juncos-Rabadán, O., & A. X. Pereiro. 1999. Telling stories in the elderly. Influence of attentional and working memory proceses (preliminary study). In: M. G. L. C. Pinto, J. Veloso, & B. Maia (eds.), *Psycholinguistics on the threshold of the year 2000. Proceedings of the 5th International Congress of the International Society of Applied Psycholinguistics*, 25–27 June, 1997, Porto, Portugal, 155–159.

Noffs, G., T. Perera, S. C. Kolbe, C. J. Shanahan, F. M. Boonstra, A. Evans, H. Butzkueven, A. van der Walt, & A. P. Vogel. 2018. What speech can tell us: A systematic review of dysarthria characteristics in Multiple Sclerosis. *Autoimmunity reviews* 17(12), 1202–1209. https://doi.org/10.1016/j.autrev.2018.06.010

Renauld, S., L. Mohamed-Saïd, & J. Macoir. 2016. Language disorders in multiple sclerosis: A systematic review. *Multiple sclerosis and related disorders* 10, 103–111. https://doi.org/10.1016/j.msard.2016.09.005

Roberts, P. M., A. Meltzer, & J. Wilding. 2009. Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of communication disorders* 42(6), 414–427. https://doi.org/10.1016/j.jcomdis.2009.06.001

Sonkaya, A. R. & Z. Z. Bayazit. 2018. Language aspects of patients with multiple sclerosis. *European Journal of Medical Investigation*, 2(3), 133–138. https://doi.org/10.14744/ejmi.2018.96158

Svindt, V., J. Bóna, & I. Hoffmann. 2020. Changes in temporal features of speech in secondary progressive multiple sclerosis (SPMS)–case studies. *Clinical Linguistics & Phonetics*, 34(4), 339–356. https://doi.org/10.1080/02699206.2019.1645885

Towards an inclusive system for the annotation of (dis)fluency in typical and atypical speech

Ivana Didirková [1], Ludivine Crible [2], Christelle Dodane [3], Loulou Kosmala [4] and Aliyah Morgenstern [4], Berthille Pallaud [5], Marie-Claude Monfrais-Pfauwadel [3] and Fabrice Hirsch [3]

[1] UR 1569 TransCrit, Université Paris 8 Vincennes - Saint-Denis, France

[2] F.R.S. - FNRS & Université catholique de Louvain, Belgium

[3] UMR 5267 Praxiling, Université Paul Valéry Montpellier 3 & CNRS, France

[4] UR 4398 PRISMES, Université Sorbonne Nouvelle - Paris 3, France

[5] UMR 7309 Laboratoire parole et langage, Aix-Marseille Université & CNRS, France

All speakers experience hesitations and production difficulties in their everyday language use. Such episodes are often termed "disfluencies" and can be defined as disruptions of the speech flow, as opposed to the "fluent" state idealized as "smooth, rapid, effortless" (Crystal, 1987). Most frameworks consider filled and silent pauses, repeats, restarts, lengthening and word fragments as typical disfluencies (Shriberg, 1994). However, there is considerable variation in the conceptual approach, extent and format of the different typologies, reflecting the diversity of disciplines, languages and target populations (adults vs. children, native vs. non-native, disordered vs. non-disordered). Crucially, while the majority of annotation models focus on disfluencies as removable errors, recent approaches (Crible et al., 2019) strive to account for the ambivalence of elements that can be used either "fluently" or "disfluently" depending on the context. This paper presents a new annotation model that adopts such a functional view of "(dis)fluencies" and extends it to pathological speech.

While the presence of disfluencies is normal in oral discourse, severe alterations of the speech flow can also be the symptoms of a speech disorder such as stuttering. Contrary to non-pathological disfluencies (mainly due to speech planning issues), stuttering-like disfluencies (SLDs) have motor origins (Monfrais-Pfauwadel, 2014). Studies have shown that stutterers produce disfluencies shared with typical speakers, along with SLDs (e.g. blocks) and disfluencies that exhibit particularities (e.g. phoneme repetition). Studies on stuttering tend to focus on a restricted set of phenomena (often prolongations, repetitions, and blocks; see Lickley, 2017). As a result, the field lacks an integrated view of disfluencies in typical and atypical speech.

The present proposal addresses this gap and provides an operational annotation system for (dis)fluencies in pathological and non-pathological speech that overcomes the technical and conceptual limitations of previous frameworks. We identified three main shortcomings in the literature. Firstly, existing typologies often exclude several phenomena because they are "intentional" or "fluent" (e.g. discourse markers (DM) are often excluded, Meteer, 1995). By contrast, we cover any element that is *potentially* disfluent at all linguistic levels. Secondly, most annotation schemes are specific to pathological or non-pathological speech, with the exception of the FLUCALC project (Bernstein Ratner & MacWhinney, 2018), which concerns child data and is bound to specific conventions and format of the CLAN package (MacWhinney, 2000). Thirdly, we favor multi-layered annotations (instead of enriched transcriptions) and a user-friendly notation system compatible with natural language processing applications. Our system includes eight verbal and three paraverbal (dis)fluency categories, some of which can be subdivided for finer distinctions if required. This flexibility allows us to strike a balance between operationality and granularity. The verbal categories are summarized in Table 1 and defined below. Non-verbal categories include laughter (Ginzburg et al., 2014), mouth noise (including clicks, Ogden, 2018) and glottal stops.

Silent pauses correspond to the absence of vocalization, and have no predefined threshold. Unlike blocks, silent pauses are not accompanied by audible and visible tension and rarely occur within a word. Filled pauses are non-lexical vocalizations such as *uhm* or *euh* in French. Lengthenings are annotated perceptively and can be complemented if necessary, by automatic analyses of the syllabic duration (from the acoustic level). DMs include any syntactically optional, grammaticalized expression that performs a pragmatic function, such as *so*, *well*, or *actually* in English. Modified repetitions differ from identical ones by the removal, addition or substitution of an element. Fragments correspond to utterances left incomplete and not taken up in the next segment or word-truncations. Subcategories distinguish between positions or extents, depending

Table 1. Verbal (dis)fluencies. Betw - between; utt - utterance, phr - phrase, mult - multiple words, w - word, syl - syllable, C - consonant, V - vowel. Due to lack of space, annotations are represented within the text; screenshots will be presented during the conference.

Main category	*Symbol*	*Subcategories (if any) and position*	*Example*
Silent pause	PS	Position: betw-utt, betw-phr, mid-w, mid-syl	Can you say (PS:betw-phr) a little bit more?
Filled pause	PF		Uhm (PF), yes, I agree.
Lengthening	LG	Sound category: C, V	I'd like to hear more (LG:V).
Block	BL	Position: betw-w, mid-w, mid-syl Sound category: C, V	That is a (BL:C) bowl of popcorn.
Discourse marker	DM		And, well (DM), there's a little soda in there.
Identical repetition	RI	Extent: utt, mult, w, syl, pho	I doubt (RI:mult) I doubt it.
Modified repetition	RM	Extent: utt, mult, w, syl, pho	She wasn't paying (RM:mult) I don' think she was paying attention.
Fragment	FR		Bl(ue) (FR) purple.

on whether a (dis)fluency occurs within or between an utterance, phrase, word, syllable, or phoneme.

Annotation tags could be neither aligned with the segment (word(s), syllable, phone) affected by the disfluency. The hierarchical system allows us to extract annotations of various degrees of granularity (e.g. all repetitions "R", all identical repetitions "RI", only repetitions of single words "RI-w"). Disfluencies that are identified as pathological will be further labeled "path" in a separate tier. This model can be used on any multi-layered annotation software such as Praat (Boersma & Weenink, 2021). Beyond creating comparable annotated corpora, which will lead to new empirical findings, the project is intended to be fully inclusive. It will also feed NLP and health applications (in particular for diagnostic and therapy) and can be easily extended to other pathologies such as dysarthria, aphasia, and an autism spectrum disorder. We believe such an inclusive annotation system is crucial for the field of disfluency studies in that it will allow a more direct comparison of results obtained in different research fields and a better reproductivity.

References

Bernstein Ratner, N. & B. MacWhinney. 2018. Fluency Bank: A new resource for fluency research and practice. *Journal of Fluency Disorders* 56, 69–80. https://doi.org/10.1016/j.jfludisr.2018.03.002

Boersma, P. & D. Weenink. 2021. Praat: doing phonetics by computer (version 6.1.42). http://www.praat.org/ (accessed 15 April 2021).

Crible, L., A. Dumont, I. Grosman, & I. Notarrigo. 2019. (Dis)fluency across spoken and signed languages: Application of an interoperable annotation scheme. In: L. Degand, G. Gilquin, L. Meurant, A. C. Simon (eds.), *Fluency and Disfluency across Languages and Language Varieties*, Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain, 17–40.

Crystal, D. 1987. *The Cambridge Encyclopedia of Language (2nd edition)*. Cambridge, UK: Cambridge University Press.

Ginzburg, J., Y. Tian, P. Amsili, C. Beyssade, B. Hemforth, Y. Mathieu, C. Saillard, J. Hough, S. Kousidis, & D. Schlangen. 2014. The disfluency, exclamation, and laughter in dialogue (DUEL) project. In: V. Rieser & P. Muller (eds.), *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue*, 1–3 September, 2014, Edinburgh, Scotland, UK, 176–178

Lickley, R. 2017. Disfluency in typical and stuttered speech. In: C. Bertini, C. Celata, G. Lenoci, C. Meluzzi, I. Ricci, (eds.), *Fattori Socali e Biologici Nella Variazione Fonetica* [Social and Biological Factors in Speech Variation] (Studi AISV), Milano, Italy: Associazione Italiana Scienze della Voce, 373–387. https://doi.org/10.17469/O2103AISV000019

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd Edition. Mahwah, NJ, USA: Lawrence Erlbaum Associates. https://doi.org/10.21415/3mhn-0z89

Meteer, M. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus [Ms]. https://www.cs.brandeis.edu/~cs140b/CS140b_docs/DysfluencyGuide.pdf (accessed 27 November 2020).

Monfrais-Pfauwadel, M.-C. 2014. *Bégaiement, bégaiements: Un manuel clinique et thérapeutique* [Stuttering, Stutterings: A Clinical and Therapeutic Manual]. Brussels, Belgium: De Boeck.

Ogden, R. A. 2018. The actions of peripheral linguistic objects: Clicks. In: J. Ginzburg & C. Pelachaud (eds.), Proceedings of Laughter Workshop 2018, 27–28 September, 2018, Paris, France, 2–5.

Shriberg, E. E. 1994. *Preliminaries to a theory of speech disfluencies*, Ph.D. dissertation, University of California at Berkeley.

Silences and disfluencies in a corpus of patients with Alzheimer's Disease (CIPP-ma)

Francesca M. Dovetto[1], Alessia Guida[1], Anna Chiara Pagliaro[1] and Raffaele Guarasci[2]
[1] University of Naples Federico II, Naples, Italy
[2] ICAR-CNR, Naples, Italy

The work is aimed at disfluencies and in particular at silent pauses in speech produced by patients with Alzheimer's Disease (AD).

Starting from Hieke's (1981) classification of disfluencies, based on the speaker's intentions in speech planning, articulated in hesitation phenomena (in particular filled and silent pauses, lengthenings and repetitions) and in repairs (referred to the levels of phonology, syntax, text), considering also phenomena that interrupt the flow of speech (false starts, interrupted words, substitutions, repetitions, slips of the tongue and omissions, Lickley, 2015), we focused our attention on silent pauses, one of the most significant disfluency phenomena for AD. In fact, between the recognised symptoms of AD, disorders affecting the lexical-semantic sphere are observed, especially in the word-retrieval process (Huff, Corkin, & Growdon, 1986; Kempler, 1995). Subsequently, the progression of the disease also involves other aspects of verbal planning and production, following a hierarchy that allows the more automated mechanisms to be preserved longer (Emery, 2000) until mutism ensues. Although the manifestations of the disease, even with respect to language disorders, may differ among subjects as a consequence of the correlation with different clinical causes (such as the impairment of brain areas or comorbidities, Dubois et al., 2014), what scholars agree on is the relevant increase in silence and the difficulty in planning (Mazzon et al., 2019). In light of the above, silent pauses and disfluencies seem to play a central role in the analysis of AD patients' speech.

Data analysis was therefore mainly focused on the duration and position of silent pauses in the turn, in order to identify recurrent patterns of pauses, also in co-occurrence with other disfluency phenomena—in particular hesitations—related to their length and frequency.

This work was carried out in the framework of a project promoted at the LiSa (Linguaggio e Salute, 'Language and Health') Laboratory of the University of Naples Federico II. For the project (CIPP-ma Corpus), still in progress, 20 patients diagnosed with AD and 11 controls were recruited, balanced by age, schooling, gender, and economic status. Patients' speech was recorded at the Alzheimer's Evaluation Unit of the Second Division of Neurology of the University of Campania, subject to informed consent. Speech acquisition, both of patients and controls, was based on two different speech samples: a complex picture description task (Capasso & Miceli, 2001, 4) and a spontaneous speech task, obtained through a semi-structured interview organized around topics widely employed in the literature: family, home, organization of the day.

The speech of patients and controls was transcribed orthographically using PRAAT software (Boersma & Weenink, 2021), and annotated in XML-TEI formalism (Text Encoding Initiative, ver. 4.2.2), with attention to verbal and non-verbal disfluency phenomena. In addition to silent pauses and filled pauses (semi-lexical), we annotated verbal disfluencies including word truncations, false starts, self-repetitions and non-word realisations. Among the non-verbal disfluency phenomena (non-lexical) we annotated tongue clicks, inspirations, throat clearing, coughing and laughter, for their potential additional value within the communicative flow.

We hypothesized a higher incidence of both silence and other disfluencies in the patients' productions, in both types of speech recorded. Starting from this hypothesis, data were extracted regarding both the number and duration of silent pauses within the patients' turns, and the other disfluency phenomena, with particular attention to their co-occurrence. Chains extraction was carried out using NLTK library for text analysis (Bird, Klein, & Loper, 2009). From a computational point of view this extraction is configured as a pattern matching task aimed at extracting all permutations of length 3 comprising at least one silent pause and the possible co-occurrences of semi- and non-lexical items. The extractions of the chains were evaluated both quantitatively, using frequency distributions to rank most recurrent patterns, and qualitatively.

Although a higher number—both in frequency of occurrence and type—of chains as potential indicators of greater hesitation and difficulty in speech planning was assumed, data did not show such an obvious trend. Even if results did not confirm the higher incidence of disfluencies or combinations of disfluencies in patients than in controls, we noticed that pauses at the edges of the chain produced

by healthy subjects are always shorter, both in the interview and in the picture description task. Moreover, the average percentage of absolute silence produced by patients (25,26%) in both tasks is always higher than in controls (14,85%). This confirms first of all the general hypothesis that the most significant disfluency of the pathology is the silent pause.

However, it should be noted that in the descriptive task the difference in the length of the pauses at the edges of the chains is less substantial than the difference in the length of the pauses at the edges of the chains in the interview task. In our opinion, this finding can be attributed to the cognitive complexity of the picture description task also for the control subjects, who are recruited in the age range 64–84 (young-old and middle-old). In general, in fact, senile age often correlates with visual difficulties and a greater slowness in the cognitive processing of stimuli (Favilla & Iagulli, 2014). On the other hand, this is also reinforced by the fact that the few long pauses found in the control subjects are detected almost exclusively in this task. Vice versa in the interview the controls present only 3 pauses longer than 2 s, of which 1 longer than 3 s, while the Alzheimer's subjects present 44 pauses longer than 2 s, 21 of which are longer than 3 s (with maximum value longer than 9 s).

In conclusion, the data extracted from the analyses show that in the observed pathology the phenomenon that is undoubtedly more significant is silence with respect to other disfluencies, whether single or in chain, as confirmed by a longitudinal case study in which, as the disease progresses, there is an increase in the number and duration of silent pauses against an unexpected decrease in the number of other hesitation phenomena (Dovetto et al., 2020).

References

Bird, S., E. Klein, & E. Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. Sebastopol, CA, USA: O 'Reilly Media.

Boersma, P. & D. Weenink. 2021, Praat: doing phonetics by computer (version 6.1.40) http://www.praat.org/ (accessed 15 March 2021).

Capasso, R. & G. Miceli. 2001. *Esame Neuropsicologico per l'Afasia: ENPA* [Neuropsychological Examination for Aphasia: ENPA], Berlin, Germany: Springer Science & Business Media.

CIPP-ma: *Corpus di Italiano Parlato Patologico della Malattia di Alzheimer* (corpus being implemented; the transcribed and annotated corpus will be archived at the LiSa Laboratory of the LUPT Research Center https://www.lupt.it/attivita/lisa.html

Dovetto, F. M., S. Schiattarella, A. Guida, C. Coppola, & M. Melone. 2020. Relationships between cognition, emotion and language in Dementia Syndrome, from interdisciplinary to transdisciplinary research: A case study. Poster Presentation at Alzheimer's Disease International (ADI) Conference, 10–12 December, 2020, Virtual.

Dubois, B. et al. 2014. Advancing research diagnostic criteria for AD: the IWG-2 criteria. *The Lancet. Neurology* 13(6), 614–629. https://doi.org/10.1016/s1474-4422(14)70090-0

Emery, V. O. B. 2000. Language impairment in dementia of the Alzheimer type: a hierarchical decline? *The International Journal of Psychiatry in Medicine* 30(2), 145–164. https://doi.org/10.2190/x09p-n7au-ucha-vw08

Favilla M. E. & A. Iagulli. 2014. Le abilità linguistiche e comunicative negli anziani. In: L. Cerrocchi & E. Gilberti (eds), *Educare «nella e alla» età senile. Processi e pratiche di alfabetizzazione digitale e di socializzazione intra- ed inter-generazionale* [Educating "in and at" senile age. Processes and practices of digital literacy and intra- and inter-generational socialization], Reggio Emilia, Italy: Edizioni Junior, 59–72.

Hieke, A. E. 1981. A content-processing view of hesitation phenomena. *Language and Speech* 24(2), 1981, 147–160. https://doi.org/10.1177%2F002383098102400203

Huff, F. J., S. Corkin, & J. H. Growdon. 1986. Semantic impairment and anomia in AD. *Brain and language* 28(2), 235–249. https://doi.org/10.1016/0093-934x(86)90103-3

Kempler, D. 1995. Language changes in dementia of the Alzheimer type. In: R. Lubinski (ed.), *Dementia and communication*, Philadelphia, PA, USA: Decker, 98–114.

Lickley, R. 2015. Fluency and Disfluency. In: M. A. Redford (ed.), *The Handbook of Speech Production*, Hoboken, NJ, USA: John Wiley, 445–469. https://doi.org/10.1002/9781118584156.ch20

Mazzon, A. et al. *2019.* Connected Speech Deficit as an Early Hallmark of CSF-defined AD and Correlation with Cerebral Hypoperfusion Pattern. *Current Alzheimer Research* 16(6), 483–494. https://doi.org/10.2174/1567205016666190506141733

Text Encoding Initiative (TEI). Transcriptions of Speech (version 4.2.2) https://www.tei-c.org/release/doc/tei-p5-doc/it/html/TS.html (accessed 12 March 2021).

Disfluency patterns in Alzheimer's Disease and frontotemporal lobar degeneration

Aurélie Pistono [1], Marie Rafiq [2], Jérémie Pariente [2,3] and Mélanie Jucla [4]
[1] Ghent University, Ghent, Belgium
[2] Toulouse University, Inserm, UPS, France
[3] Toulouse University Hospital, Toulouse, France
[4] University of Toulouse II-Jean Jaurès, Toulouse, France.

Introduction

Patients with neurodegenerative diseases most often present with discourse impairment compared to healthy older adults (Ash, Avants, & Grossman, 2011). In the current study, we focus on two neurodegenerative diseases: Alzheimer's disease (AD) and the behavioral variant of Frontotemporal lobar degeneration (FTLD-bv). FTLD-bv involves a deficit of social comportment that often co-occurs with executive function disorders (Rascovsky et al., 2011). Although these patients are not aphasic, they present with a reduced speech rate, correlated with executive functioning limitations. Unlike FLTD-bv, AD is characterized by lexical-semantic impairment (Taler & Phillips, 2008). AD patients show an increase of disfluency production, attributed to word finding difficulties (Lira et al., 2011). The current study aims at analyzing disfluency in early AD and FTLD-bv compared to healthy older adults. If disfluency reflects word finding difficulties in AD, these patients will produce more disfluencies than healthy controls, and these markers will be correlated with poorer lexical-semantic abilities. If reduced fluency in FTLD-bv reflects executive function limitations, this group will have lower speech rate and more interrupted utterances (that could reflect their planning and inhibition difficulties). Contrary to the AD group, disfluency will not be correlated with language tasks.

Methods

Fifteen AD participants, 12 FTLD-bv participants and 15 healthy controls (HC) were recruited. The three groups were matched for age, gender and level of education. Language was assessed with the GREMOTs battery (Bézy, Renard, & Pariente, 2016), which includes a picture-based narrative task. For this task, the following variables were analyzed: speech rate (number of words/discourse duration in seconds); proportion of self-corrections per 100 words (i.e. when the speaker stops and resumes with a substitution for a word or a new utterance); proportion of repetitions (of sounds, syllables, words) per 100 words; proportion of filled pauses per 100 words; proportion of semantic shifts per 100 words (i.e. abrupt interruption of an utterance, after which a new concept begins (Marini et al., 2005)). Because of the small sample size, inter-groups comparisons were performed with permutation tests. We then conducted Kendall correlations between language tests and disfluency, for each group separately.

Results

Both AD and FTLD-bv groups had lower performance compared to HC during fluency and naming tasks. The three groups did not significantly differ in terms of discourse length (in number of words, FTLD-bv: 97.5±46.8; AD: 109.3±48.1; HC: 135.1±110.8), but the two groups of patients had lower speech rate than HC. Additionally, FTLD-bv participants produced significantly more semantic shifts than HC during their narrative (see Table 1). Correlations were performed with the most discriminant lexical-semantic tasks based on inter-group comparisons, i.e. semantic fluency, phonemic fluency, action naming and famous faces naming. In the HC and FTLD-bv group, language performance was not correlated with disfluency. In the AD group, self-corrections were negatively correlated with naming famous faces ($r = -0.52$, $p < 0.01$), and speech rate was positively correlated with semantic fluency ($r = 0.44$, $p < 0.05$).

Discussion

During discourse production, FTLD-bv participants had reduced speech rate, in line with Ash et al. (2011), but also more semantic shifts. While previous studies did not focus on disfluency in this population, current results show that interrupted utterances could be a more specific feature than speech rate for this group of patients. Similarly to FTLD-bv participants, AD participants had lower speech rate, which is coherent with previous studies (Pistono et al., 2016). However, they did not differ from HC on other measures. Despite this lack of significant differences, self-corrections and speech rate were correlated with lexical-sematic tasks in this

*Table 1. Inter-group differences. *significant results after Bonferroni-Holm corrections.*

		FTLD	AD	HC	p value	post-hoc tests
GREMOTs battery	Semantic fluency	9.2±3.6	14.3±6.7	19.9±4.7	<0.0001*	HC>AD>FTLD
	Phonemic fluency	7.6±5.9	16.3±7	18.8±7.5	<0.0001*	HC=AD>FTLD
	Object naming	29.1±6	32.1±2.9	34.3±1.3	0.02	HC>AD=FTLD
	Action naming	29.1±2.4	30.5±3.4	33.1±2.7	0.002*	HC>AD=FTLD
	Famous faces naming	5.5±2.3	4.7±2.7	8.3±2.2	<0.0001*	HC>AD=FTLD
	Syntactic comprehension	18.8±5.4	18.9±3.3	21.6±2.2	0.04	HC>AD=FTLD
	Word spelling	10±2.4	10.3±2.2	11.3±0.9	ns	ns
	Sentence spelling	24.4±2.2	24.8±1.8	25.1±2.2	ns	ns
Connected-speech task	Speech rate	1.7±0.5	1.8±0.4	2.4±0.8	<0.01*	HC>AD=FTLD
	Repetitions	4.7±4.5	2.8±3	2.3±3.3	ns	
	Filled pauses	3.5±3.4	4.6±4.2	5.1±4.8	ns	
	Self-corrections	1.8±1.8	2.7±1.8	2.2±1.4	ns	
	Semantic shifts	1.9±1.8	0.9±1.2	0.5±0.9	<0.05*	FTLD>HC

group, unlike HC. Self-corrections could be therefore related to word finding difficulties in the AD group, but not in the HC group. Current findings have implications regarding the classification of disfluency phenomena. Indeed, interrupted utterances are sometimes grouped with other self-corrections (e.g. Hartsuiker & Notebaert, 2010). On the contrary, current study shows that interrupted utterances are hallmarks of FTLD-bv while other self-corrections are better features of AD and word-finding difficulties. Additionally, these results stress the influence of non-linguistic cognitive factors on disfluency production (Engelhardt, McMullon, & Corley, 2018).

References

Ash, S., B. Avants, & M. Grossman. 2011. Non-Fluent Speech in Frontotemporal Lobar Degeneration. *Journal of Neurolinguistics* 22(4), 370–383. https://doi.org/10.1016/j.jneuroling.2008.12.001

Bézy, C., A. Renard, & J. Pariente. 2016. *GREMOTS Batterie d'évaluation des troubles du langage dans les maladies neurodégénératives* [GREMOTS Battery for evaluating language disorders in neurodegenerative diseases]. Brussels, Belgium: De Boeck.

Engelhardt, P. E., M. E. G. McMullon, & M. Corley. 2018. Individual differences in the production of disfluency: a latent variable analysis of memory ability and verbal intelligence Paul. *Quarterly Journal of Experimental Psychology* 72(5), 1084–1101. https://doi.org/10.1177/1747021818778752

Hartsuiker, R. J., & L. Notebaert. 2010. Lexical access problems lead to disfluencies in speech. *Experimental Psychology* 57(3), 169–177. https://doi.org/10.1027/1618-3169/a000021

Lira, J. O. de, K. Z. Ortiz, A. C. Campanha, P. H. F. Bertolucci, & T. S. C. Minett. 2011. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics* 23(3), 404–412. https://doi.org/10.1017/S1041610210001092

Marini, A., A. Boewe, C. Caltagirone, & S. Carlomagno. 2005. Age-related differences in the production of textual descriptions. *Journal of Psycholinguistic Research* 34(5), 439–463. https://doi.org/10.1007/s10936-005-6203-z

Pistono, A., M. Jucla, E. J. Barbeau, L. Saint-Aubert, B. Köpke, M. Puel, & J. Pariente. 2016. Pauses during Autobiographical Discourse Reflect Episodic Memory Processes in Early Alzheimer's Disease. *Journal of Alzheimer's Disease* 50(3), 687–698. https://doi.org/10.3233/jad-150408

Rascovsky, K., et al. 2011. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, 134(9), 2456–2477. https://doi.org/10.1093/brain/awr179

Taler, V. & N. A. Phillips. 2008. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5), 501–556. https://doi.org/10.1080/13803390701550128

Linguistic disfluencies in Russian-speaking children with developmental language disorder

Alexandr Kornev [1] and Ingrida Balčiūnienė [1,2]
[1] Saint-Petersburg State Pediatric Medical University, Saint-Petersburg, Russia
2 Vytautas Magnus University, Kaunas, Lithuania

Introduction

Linguistic disfluency might be defined as various forms of hesitations, repetitions, revisions, false starts, incomplete utterances, etc. They are quite natural elements of spontaneous speech in both child and adult speech; however, a number of disfluencies tends to increase along the speaker's age (Evans, 1985) and growing linguistic skills (Starkweather, 1987; Fiestas et al., 2005). Results of the studies in various typical and atypical populations have highlighted that children with language and learning difficulties produced more linguistic disfluencies than did their typically developing peers (Guo, Tombline, & Samelson, 2008); also, linguistic disfluencies were more numerous in bilinguals' than in monolinguals' speech (Fiestas et al., 2005). Presumably, numerous linguistic disfluencies might be a symptom of atypical language development; on the other hand, production of linguistic disfluencies might be explained by a self-monitoring, as an obligatory skill for speech production (e.g. Bock & Levelt, 1994); in some other studies, a correlation between linguistic disfluencies, individual intelligence, and executive function has been revealed (Engelhardt et al., 2010). However, despite numerous studies, substantially a nature and mechanisms of linguistic disfluency remain unclear both in children and adults.

In this study, we aimed at highlighting linguistic disfluencies typical for Russian-speaking children with developmental language disorder (DLD). Whereas many previous studies have been based on particular types of linguistic disfluency (Evans, 1985; Levelt, 1983; 1984; Corley & Stewart, 2008), we analyzed linguistic disfluencies as the entity of various kinds of disruptions of linguistically fluent speech.

Research method

The subjects of the study were 12 clinically referred preschoolers (mean age 76 months) with DLD and 12 typically developing (TD) peers. The DLD children were recruited from those who attended remedial treatment unit for speech and language disordered kindergartens. Exclusion criterion was non-verbal IQ on Raven's matrix below 84. In all cases, morphosyntactic backwardness (below 5-year level) was coupled with articulation/phonological disorders. The TD children were recruited from day care center for kindergartens. All subjects were monolinguals living in Saint-Petersburg (the second largest city); for both the TD and DLD group; informed consent was obtained from parents before the experiment.

The subjects were assessed individually. Each subject performed storytelling according to wordless picture sequence (Balčiūnienė & Kornev, 2017). All the stories were video-recorded and transcribed according to the CHAT tools (MacWhinney, 2010). Linguistic disfluencies represented by hesitations, repeats, revisions, false starts, and incomplete utterances were encoded with special symbols according to internationally accepted principles of discursive annotation (Bernstein Ratner & MacWhinney, 2019). Then, individual measures (the number and distribution of each type of disfluencies) were estimated and submitted for statistical analysis.

Results

Comparative analysis revealed that the total number of linguistic disfluencies per utterance was very similar between the DLD and TD groups. However, some (sub-)types of disfluencies discriminated the groups (see Table 1):

1) incomplete utterances and fillers were significantly more numerous in the DLD than in the TD children;
2) among all hesitations, the filled hesitations (fillers) were more dominant (72%) in the TD children, whereas the unfilled hesitations (pauses) were more numerous (54%) in the DLD peers;
3) repeated parts of word were significantly prevalent among all repeats in the DLD children, while repeated words were the more frequent in the TD peers;
4) although the total number of revisions did not differ between the groups, the phonological revisions were significantly more numerous in the TD than in the DLD children.

Conclusions and discussion

In our paper, we address some issues permanently debated in many previous publications (e.g. Levelt, 1983, 1984; Evans, 1985; MacLachlan & Chapman, 1988; Leadholm & Miller, 1992; Wagner et al., 2000; Madon, 2007; Guo et al., 2008; Corley & Stewart,

Table 1. Distribution of linguistic disfluencies within the DLD vs. TD group

Measures	DLD (*N* = 12)		TD (*N* = 12)		*df*	*F*	*Sig.*
	M	*σ*	*M*	*σ*			
A number of false starts per utterance	0.008	0.021	0.064	0.086	1	3.189	0.092
A number of incomplete utterances per utterance	0.089	0.073	0.027	0.054	1	4.448	0.050
% of filled hesitations among all hesitations	0.340	0.334	0.719	0.202	1	9.343	0.007
% of unfilled hesitations among all hesitations	0.536	0.379	0.281	0.202	1	3.633	0.074
% of repeated parts of word among all repeats	0.405	0.457	0.012	0.039	1	8.241	0.011
% of repeated words among all repeats	0.309	0.442	0.716	0.461	1	3.743	0.070
% of phonetical revisions among all revisions	0.104	0.197	0.478	0.435	1	5.095	0.037

2008; Engelhardt et al., 2010). The given studies have focused on the nature of linguistic disfluency, as a manifestation of language immaturity *vs.* a complex of individual strategies of discourse production. Results of our study, instead, raise a question about the role of cognitive resource (and its deficit) in the linguistic dysfluency. The DLD children obviously were not able to produce complete utterances more often than their TD peers. Both TD and DLD children tried to build the proposition of the utterance and to perform its linguistic program in parallel, but mainly the DLD children struggled with combining these two processes. Unsuccessful attempts (and, especially, series of attempts) to find the proper grammatical (morphological or syntactic) forms and/or lexical items overloaded cognitive resources of the DLD children and, thus, prevented them to complete the most utterances. This twofold process was not only resource- but also time-consuming and lead to a high number of hesitations in the DLD children. Unfortunately, we did not measure the maturity of executive functions and cognitive recourses in the subjects, and this should be considered as a limitation of our study. In further researchers, we plan to add psychological evaluation of the given variables.

References

Balčiūnienė, I. & A. N. Kornev. 2017. Evaluation of narrative skills in language-impaired children. Advantages of a dynamic approach. In: E. Aguilar-Mediavilla, L. Buil-Legaz, R. López-Penadés, V. A. Sanchez-Azanza, & D. Adrover-Roig (eds.), *Atypical Language Development in Romance Languages*. Amsterdam, The Netherlands: John Benjamins, 127–141. https://doi.org/10.1075/z.223.08bal

Bock K. & W. J. M. Levelt. 1994. Language production. Grammatical encoding. In: M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*. New York, NY, USA: Academic Press, 741–779.

Corley M. & O. W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of *um*. *Language and Linguistics Compass* 2(4), 589–602. https://doi.org/10.1111/j.1749-818X.2008.00068.x

Engelhardt P. E., M. Corley, J. T. Nigg, & F. Ferreira. 2010. The role of inhibition in the production of disfluencies. *Memory & Cognition*, 38(5), 617–628. https://dx.doi.org/10.3758%2FMC.38.5.617

Evans, M. A. 1985. Self-initiated speech repairs: A reflection of communicative monitoring in young children. *Developmental Psychology*, 21(2), 365–371. https://psycnet.apa.org/doi/10.1037/0012-1649.21.2.365

Fiestas C. E., L. M. Bedore, E. D. Peña, & V. J. Nagy. 2005. Use of mazes in the narrative language samples of bilingual and monolingual 4- to 7-year-old children. In: J. Cohen, K.T. McAlister, K. Rolstad, J. MacSwan (eds.), *Proceedings of the 4th International Symposium on Bilingualism*. 30 April–3 May, 2003, Tempe, AZ, USA, 730–740.

Guo, L., J. B. Tomblin, & V. Samelson. 2008. Speech disruptions in the narratives of English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 51(3), 722–738. https://dx.doi.org/10.1044/1092-4388(2008/051)

Leadholm, B. J. & J. F. Miller. 1992. *Language Sample Analysis: The Wisconsin Guide*. Madison. WI: Wisconsin Department of Public Instruction.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. 1984. Spontaneous self-repairs in speech: Processes and representations. In: M. P. R. Van den Broecke & A. Cohen (eds.) *Proceedings of the Tenth International Congress of Phonetic Sciences*, 1–6 August, 1983, Utrecht, The Netherlands, 105–118.

MacLachlan, B. G. & R. S. Chapman. 1988. Communication breakdowns in normal and language learning-disabled children's conversation and narration. *Journal of Speech and Hearing Disorders*, 53(1), 2–7. https://doi.org/10.1044/jshd.5301.02

Bernstein Ratner, N. & B. MacWhinney. 2019. TalkBank resources for psycholinguistic analysis and clinical practice. In: A. Pareja-Lora, M. Blume, & B. Lust (eds.), *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*, Cambridge, MA, USA: MIT Press, 131–150.

Madon, Z. 2007. *Investigation of Maze Production in Children with Specific Language Impairment*. MA thesis, McGill University.

Starkweather, C. W. 1987. *Fluency and Stuttering*. Englewood Cliffs, NJ, USA: Prentice Hall.

Wagner, C. R., U. Nettelbladt, U. Sahlén, & C. Nilholm. 2000. Conversation versus narration in preschool children with language impairment. *International Journal of Language and Communication Disorders*, 35(1), 83–93. https://doi.org/10.1080/136828200247269

Jaw and lip amplitude and velocity in stuttered disfluencies. A preliminary study

Ivana Didirková [1], Shakeel Ahmad Sheikh [2], Slim Ouni [2], Anais Vallé [3] and Fabrice Hirsch [3]
[1] UR1569 TransCrit, Université Paris 8 Vincennes - Saint-Denis, France
[2] UMR7503 LORIA, Université de Lorraine, INRIA & CNRS, France
[3] UMR5267 Praxiling, Université Paul-Valéry Montpellier 3, France

Stuttering is a neurodevelopmental speech disorder affecting about 1% of the general population (Ward, 2018). Although several factors, including genetics, have been identified as favoring stuttering (Frigerio-Domingues et al., 2019), the disorder's etiology is still discussed. Moreover, several neurolo-gical specificities (for an exhaustive review, see Etchell et al., 2018) have also been observed in persons who stutter (henceforth PWS).

From a perceptual point of view, stuttering is characterized by an increased number of involuntary prolongations, repetitions, and silent blocks often accompanied by perceptible tensions and muscular spasmodic movements. Aside from these disfluencies, speech in PWS remains perceptually fluent, i.e., with no perceptible marks of a speech disorder. Nevertheless, speech production studies have identified several markers typical for such perceptually fluent speech in PWS compared to speech produced by persons who do not stutter (PWNS). For instance, they suggested specific behavior related to a greater vowel centralization (Blomgren, Robb, & Chen, 1998). Other authors used articulatory data (such as electromagnetic articulography (EMA) or ultrasound tongue imaging (UTI)). These studies conclude to a different articulatory sequencing in PWS (De Nil, 1995) but fail to find any difference in duration or peak velocity in onset gestures and coarticulation (Frisch, Maxfield, & Belmont, 2016; Heyde et al., 2016).

In contrast, however, comparatively few studies have been explicitly led on disfluencies up to date. Most of the time, research on stuttered disfluencies focuses on identifying stuttered structures. For instance, consonant clusters, phonetically complex sounds, and voiceless consonants would be more often concerned by a stuttered disfluency (SLD) than their singleton, simple, and voiced counterparts (Blomgren, 2012). Moreover, to our knowledge, articulatory studies describing disfluent speech sequences in PWS are relatively scarce up to date. Yet, using direct observations (i.e., not based on acoustic data interpretation) seems crucial for stuttered speech understanding. Furthermore, some of these techniques (such as UTI) are already used for real-time visual feedback in speech and language pathology or for speech modeling.

For all these reasons, this research aims at providing additional knowledge about stuttered speech using articulatory data. We aimed to investigate the motor events occurring during SLD and non-pathological disfluencies from a spatial and temporal perspective. More specifically, our purpose was to investigate amplitude and velocity of visible speech articulators (lips, mandible). We chose to concentrate on articulators which are visible to the interlocutor, but also to the subjects themselves, with no use of imaging techniques.

EMA data were collected using an electromagnetic articulograph Carstens AG501 3D at the Lorraine Research Laboratory in Computer Science and its Applications with a sampling rate of 250 Hz and spatial accuracy of 0.3 mm. Data were stocked in a *.pos* file and synchronized with a sound recording (44,1 kHz, 16 bits, .wav). 9 sensors (2×3 mm) per subject were used: two were fixed on each subject's lips (1 in the middle of the upper lip and another one in the middle of the lower lip). Another sensor was placed on the subjects' jaw to track the mandible's movements. Two coils were situated on each subject's tongue, one on the tongue tip, one on the tongue body; however, the tongue's movement was not investigated in this study. The palate's form was indicated using a ninth coil. Other sensors were used to control the head's movements. Four adults who stutter with no therapy for the past two years and four control subjects were recruited for this study. PWS were evaluated as presenting mild to severe stuttering on the SSI scale (Riley, 2009). All subjects were matched according to gender, age, and socio-professional category. They all were native speakers of French. Participants were recorded during several tasks; only semi-spontaneous speech will be presented in this paper. No disfluency elicitation technique was used.

All SLDs (i.e., sound prolongations, single sound repetitions, and blocks) were identified in the production of PWS by a speech-language pathologist. Further, non-pathological disfluencies (sound prolongations, single sound repetitions) and silent pauses were annotated both in PWNS and

PWS. After this classification, the VisArtico software (Ouni, Mangeonjean, & Steiner, 2012) was used to visualize the vertical and horizontal movements of the upper and lower lip and the mandible in segments that included the disfluent phone and its preceding and subsequent syllables. The movement extent, velocity, distance, and trajectory were analyzed in 80 disfluent sequences and 40 silent pauses.

Results reveal differences between PWS and PWNS in terms of movement amplitude and velocity. Modifications are observed in both pathological and non-pathological disfluencies produced by PWS compared to PWNS. More specifically, movement amplitudes were lower in PWNS than in PWS in prolongations, silent pauses, and single sound repetitions. This behavior applies to horizontal and vertical movements of the lower lip and for vertical movements of the upper lip and of the jaw. Relatedly, articulatory velocity was higher in PWNS compared to PWS for horizontal and vertical lip movements and vertical movements of the jaw. Most interestingly, these observations are observed for both pathological (SLD) and non-pathological disfluencies compared to non-pathological disfluencies produced by PWNS. In other terms, PWNS have less ample but faster movements than PWS. Thus, our observations seem to support statements on particular articulatory behavior in both fluent (e.g. De Nil, 1995) and stuttered (e.g. Didirková, Le Maguer, & Hirsch, 2021) speech in PWS. However, they are to be confirmed on a larger number of participants and using tongue movements.

The results presented in the abstract have a dual purpose: a better description of stuttered speech during disfluencies and a possible contribution to automatic stuttering detection, especially from a deep learning perspective. Indeed, the existing techniques for stuttering detection use spectral features as an input, such as spectrograms, mel-frequency cepstral coefficients (MFCCs), and linear prediction cepstral coefficients (LPCCs) or its variants which capture stuttering-related information (Khara, Singhr, & Vir, 2018). However, to date, there are few studies using EMA as features in stuttering detection. The findings described in the previous paragraph impact the properties of the speech that can be exploited and used as input features solely or in combination with other modalities like audio, video, or text in the detection of stuttering and its types using deep learning models.

References

Blomgren, M. 2012. Do speech sound characteristics really influence stuttering frequency? In: *Proceedings of the 7th World Congress of Fluency Disorders*, 2–5 July, 2012, Tours, France.

Blomgren, M., M. Robb, & Y. Chen. 1998. A Note on Vowel Centralization in Stuttering and Nonstuttering Individuals. *Journal of Speech, Language, and Hearing Research* 41(5), 1042–1051. https://doi.org/10.1044/jslhr.4105.1042

De Nil, L. F. 1995. The influence of phonetic context on temporal sequencing of upper lip, lower lip, and jaw peak velocity and movement onset during bilabial consonants in stuttering and nonstuttering adults. *Journal of Fluency Disorders* 20(2), 127–144. https://doi.org/10.1016/0094-730X(94)00024-N

Didirková, I., S. Le Maguer, & F. Hirsch. 2021. An articulatory study of differences and similarities between stuttered disfluencies and non-pathological disfluencies, *Clinical Linguistics & Phonetics* 35(3), 201–221. https://doi.org/10.1080/02699206.2020.1752803

Etchell, A. C., O. Civier, K. J. Ballard, & P. F. Sowman. 2018. A systematic literature review of neuroimaging research on developmental stuttering between 1995 and 2016. *Journal of Fluency Disorders* 55, 6–45. https://doi.org/10.1016/j.jfludis.2017.03.007

Frigerio-Domingues, C., Z. Gkalitsiou, A. Zezinka, E. Sainz, J. Gutierrez, C. Byrd, R. Webster, & D. Drayna. 2019. Genetic factors and therapy outcomes in persistent developmental stuttering. *Journal of Communication Disorders* 80, 11–17. https://doi.org/10.1016/j.jcomdis.2019.03.007

Frisch, S. A., N. Maxfield, & A. Belmont. 2016. Anticipatory coarticulation and stability of speech in typically fluent speakers and people who stutter. *Clinical linguistics & phonetics* 30(3–5), 277–291. https://doi.org/10.3109/02699206.2015.1137632

Heyde, C. J., J. M. Scobbie, R. Lickley, & E. K. E. Drake. 2016. How fluent is the fluent speech of people who stutter? A new approach to measuring kinematics with ultrasound. *Clinical Linguistics & Phonetics* 30(3–5), 292–312. https://doi.org/10.3109/02699206.2015.1100684

Khara, S., S. Singhr, & D. Vir. 2018. A comparative study of the techniques for feature extraction and classification in stuttering. In: *Proceedings of the Second International Conference on Inventive Communication and Computational Technologies*, 2021 April, 2018, Coimbatore, India, 887–893.

Ouni, S., L. Mangeonjean, & I. Steiner. 2012. VisArtico: A visualization tool for articulatory data. In: *Proceedings of 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 9–13 September, 2012, Portland, OR, USA, 1878–1881.

Riley, G. D. 2009. *The stuttering severity instrument – Fourth Edition*. PRO-ED.

Ward, D. 2018. *Stuttering and cluttering. Frameworks for understanding and treatment (2nd Edition)*. Milton Park, England, UK: Routledge.

Author index

<This page intentionally left blank.>

<This page intentionally left blank.>

ISBN: xxx-xxx-xxx-xxx-x