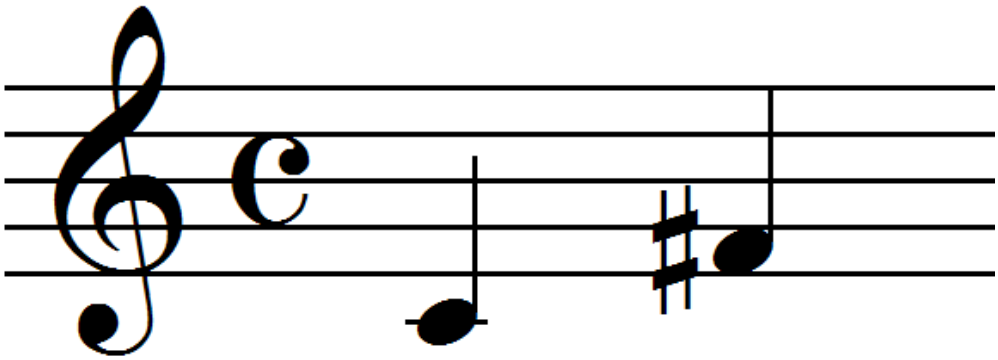
Proceedings of

DiSS 2017

The 8th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
18–19 August 2017

TMH-QPSR
Volume 58(1)



Edited by
Robert Eklund & Ralph Rose

< This page intentionally left blank >

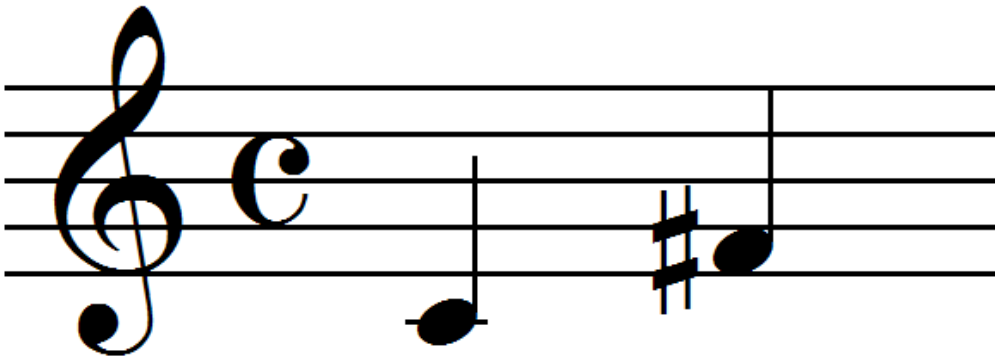
Proceedings of

DiSS 2017

The 8th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
18–19 August 2017

TMH-QPSR
Volume 58(1)



Edited by
Robert Eklund & Ralph Rose

Conference website: http://www.diss2017.org
Proceedings also available at: http://roberteklund.info/conferences/diss2017

Cover design by Robert Eklund
Graphics and photographs by Robert Eklund (except ISCA and KTH logotypes)

Proceedings of DiSS 2017, Disfluency in Spontaneous Speech
Workshop held at the Royal Institute of Technology (KTH), Stockholm, Sweden, 18–19 August 2017
TMH-QPSR volume 58(1)
Editors: Robert Eklund & Ralph Rose
Department of Speech, Music and Hearing
Royal Institute of Technology (KTH)
Lindstedtsvägen 24
SE-100 44 Stockholm, Sweden

ISSN 1104-5787
ISRN KTH/CSC/TMH–17/01-SE

Preface

Following the successes of the previously organized Disfluency in Spontaneous Speech (DiSS) workshops held in Berkeley (1999), Edinburgh (2001), Göteborg (2003), Aix-en-Provence (2005), Tokyo (2010), Stockholm (2013) and Edinburgh (2015), the organizers are proud to present DiSS 2017, held at the Royal Institute of Technology (KTH), Stockholm, Sweden, in August 2017.

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

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

Thanks to ISCA for administrative and financial support. Special thanks to Anders Eriksson, Olof Engwall, Gerard Bailly and Martin Cooke.

Stockholm, August 2017

Robert Eklund
Robin Lickley
Jens Edlund
Joakim Gustafson

Committees

Program and organization

Robert Eklund
Linköping University, Sweden

Robert Lickley
Queen Margaret University, Scotland

Jens Edlund
KTH Royal Institute of Technology, Sweden

Joakim Gustafson
KTH Royal Institute of Technology, Sweden

Scientific committee

Jens Allwood
Gothenburg University, Sweden

Liesbeth Degand
Université Catholique de Louvain, Belgium

Yasuharu Den
Chiba University, Japan

Danielle Duez
CNRS, Aix-en-Provence, France

Mária Gósy
Hungarian Academy of Sciences, Research Institute for Linguistics, Hungary

Robert Eklund
Linköping University, Sweden

Peter Heeman
Oregon Health and Science University, USA

Robert Hartsuiker
Ghent University, Belgium

Robin Lickley
Queen Margaret University, Scotland

Kikuo Maekawa
National Institute for Japanese Language and Linguistics, Japan

Hugo Quené
Utrecht University, The Netherlands

Ralph Rose
Waseda University, Japan

Vered Silber-Varod
The Open University of Israel

Elizabeth Shriberg
SRI International, Menlo Park, USA

Marc Swerts
Tilburg University, The Netherlands

Shu-Chuan Tseng
Academica Sinica, Taiwan

Michiko Watanabe
National Institute for Japanese Language and Linguistics, Japan

Åsa Wengelin
Gothenburg University, Sweden

Table of contents

Plenary talk

Presented papers

< This page intentionally left blank >

Plenary Talk

Fluency or disfluency?

Jens Allwood
SCCIIL Interdisciplinary Center, University of Gothenburg, Gothenburg, Sweden

Abstract

In this paper, I investigate the concepts of "fluency" and "disfluency" and argue that the application of the two concepts must be relativized to type of communicative activity. It is not clear that there is a generic sense of fluency or disfluency, rather what contributes to fluency and disfluency depends on what type of communication we are dealing with. The paper then turns to a brief investigation of what makes interactive face-to-face communication fluent or disfluent and argues that many of the features that have been labeled as disfluent, in fact, contribute to the fluency of interactive communication. Finally, I suggest that maybe it is time for a change of terminology and abandon the term "disfluent" for more positive or neutral terminology.

Why interesting?

The phenomena that sometimes go under the name of "disfluencies" are a pervasive feature of human communication. In written communication, they are to some extent edited out on the basis of normative criteria. In spoken and gestural communication, they are, however, a regular part of the ongoing flow. It seems unlikely that such a common, regular phenomenon is only "dysfunctional" or "dis-functional" or has no function at all. Rather, it seems to have functions that are interesting in themselves and deserve further study.

An ancillary reason for an interest in "disfluencies" is the question of whether artificial dialog systems in virtual agents or robots should be devoid of this feature. This is related to the more general question of what features a dialog system should have. Are "disfluent" features desirable or not desirable in a dialog system? Is what is "disfluent" constant across different human types of communication? All these questions lead back to the question of the nature of fluent and disfluent communication.

What is fluent varies with type of communication

A first observation we can make is that the ideals of fluency vary with type of communication. Fluency in written language involves writing in a manner, which is easy to read, making use of full sentences and judicious punctuation, while fluency in spoken language involves clear pronunciation, audibility and clear relevant gestures. In addition, spoken language ideals of fluency are different in different social activities. Fluency in public speaking involves such things as not presupposing context not shared, making good use of what could possibly be shared, being clear, holding attention, evoking interest and positive emotions, being audible and visible, while fluency in interactive (small) talk, with friends, involves such things as making efficient use of the much larger amounts of shared background information available, as well as being flexible and open for interactive cooperation in co-constructing content which, in turn involves such things as being able to change one's mind and having time to think.

Usually, disfluency varies with fluency, so that what is seen as "disfluent" can be seen as the negation of what is seen as "fluent", that is, not being clear, audible, presupposing as shared what is not shared etc.

So, what is fluent or disfluent in written language is not necessarily fluent or disfluent in interactive face-to-face communication and vice versa. Nor is what is fluent or disfluent in public speaking necessarily fluent or disfluent in private friendly interactive face-to-face communication and vice versa.

Finally, we may note a related use of the term "fluent" in connection with learning a new language. We talk about "fluency in a foreign language", referring to the ability to find words and use grammar easily. For a discussion of other aspects of fluency and disfluency, see Lickley (2015).

Fluency and disfluency in interactive communication

A model of interactive communication

Let us now consider some of the features of fluency and disfluency in interactive communication. We will take as our point of departure the model of interactive embodied communication proposed in Allwood et al. (2006) (see Figure 1). The model shows how interactive communication involves at least two communicators (A and B), forming a dynamic system of co-activation involving several different levels of awareness.

Like in Kahneman (2011), the model distinguishes processes on a high level of awareness, that are slower and involve responses based on evaluation and deliberation, from processes on a low level of awareness, that are faster and involve reactions based on more automatic appraisal and cognition.

The processes on higher levels of awareness are related to the processes on lower levels of awareness through a gradient, the specific nature of which needs to further investigated.

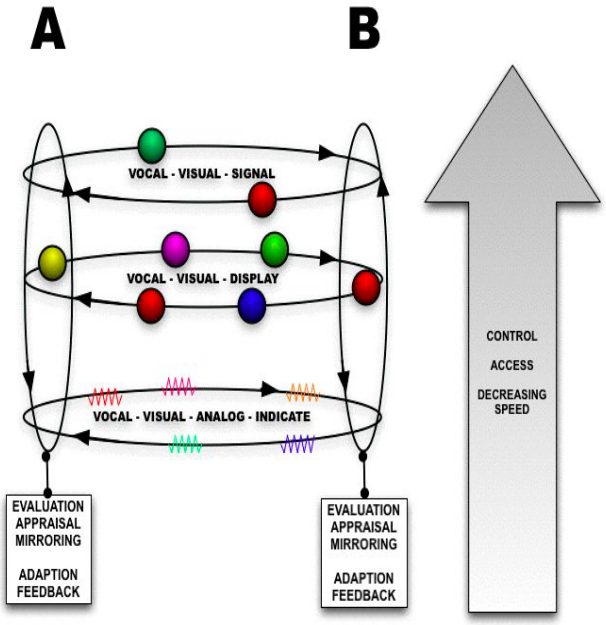


Figure 1. General model of embodied communication.

As a start of such an investigation, three levels of awareness are distinguished in production:

(i) *indicate* – the lowest level of awareness and conscious control. This involves being informative to an interlocutor without any communicative intention, e.g. through vocal features that indicate age, gender or dialect,

(ii) *display* – an intermediate stage of awareness and conscious control, involves intentionally expressing information (for an interlocutor), e.g. an emotion like joy or sorrow,

(iii) *signal* – the highest stage of awareness and control, involves expressing information for an interlocutor in such a way that the interlocutor should notice that the information is being expressed for him/her.

The three levels of awareness are connected and interact so that a feeling of joy, initially automatically "indicated" in intonation or facial gestures can become more aware and then more intentionally "displayed" and finally also intentionally "signaled" through a verbal utterance like "Great that you could come". Other processes go the other way and connect impulses on a high level of awareness with more automatic reactions on lower levels

Also, on the recipient side, higher levels of awareness and control are integrated with lower levels of awareness and control. Automatic fast processes of perception, reaction and appraisal are connected with and can influence slower processes of evaluation, deliberation, planning and response and going the other way slower processes can influence the faster less aware processes.

Both in production and reception, the processes can be sequential and simultaneous.

In interactive communication, vertical processes connecting higher levels of awareness and control with lower levels of awareness and control, interact with horizontal processes, connecting interlocutors with each other, on different levels of awareness, so that we both influence and are influenced by others on several levels of awareness. Interactive communication, in this way, forms a partly self-organizing system with vertical and horizontal subsystems.

The horizontal system (interactive communication management (ICM, Allwood, 2013) involves many interactive communication components, the most important being the feedback system, whereby interlocutors give each other multimodal feedback (mostly visual and auditory) concerning perception, understanding, emotional and other attitudinal reactions. The information given in the feedback system can be indicated, displayed or signaled. This also means that the means of expression can range from more or less conventionalized vocal verbal expressions, like *yes*, *no*, *mm*, (Lindblad & Allwood, 2013) or gestured verbal expressions, like head nods or head shakes, to less conventionalized, so called "conversational grunts" (Ward, 2006).

The vertical system (own communication management (OCM, Allwood, 2013), similarly, involves many components, two of the most important being:

(i) mechanisms for planning and selection of expressions and their combination (lexicon and grammar), for short "choice mechanisms" and

(ii) mechanisms for on-line modification and change of ongoing production, for short "change mechanisms" (see Allwood, Nivre & Ahlsén, 1990).

Like in ICM, OCM processes can be indicated, displayed or signaled, leading to means of expression that can be more or less conventionalized, ranging from fully

conventionalized hesitation words like *eh* and facial gestures (to gain time) to displayed and indicated such means, including also processes allowing on-line change management, ranging from signaled explicit negation to more ad hoc indicated means.

The two systems are integrated, so that many expressions can function both in vertical and horizontal processes, e.g. a hesitation expression can give feedback to an interlocutor (ICM), while also gaining time for a speaker to plan and select appropriate means of expression (OCM).

The important thing in all cases is that all processes, (both OCM and ICM) should be means of joint sharing of content and sometimes also explicit co-construction of content.

Fluency in interactive face-to-face communication

Achieving fluency in interactive face-to-face communication involves achieving at least the following goals:

(i) Being able to communicate while taking context and your interlocutor(s) into account, i.e., not belaboring what is given by context and being sensitive to simultaneously indicated, displayed and signaled vocal and gestural feedback, which is conventionalized to varying extents.

(ii) Being able to hold the floor in order to plan and select what you want to express.

(iii) Being able to manage, e.g. change what you are communicating in such a way that your interlocutor can follow you.

(iv) Being able to keep, yield, give, assign, take and accept turns.

(v) Being able to actively listen, react and respond by giving vocal and gestural feedback regarding perception, understanding, emotional and other attitudes.

(vi) Being able to co-construct content with your interlocutor, often using short and relevant utterances and gestures.

What is disfluency?

Let us now define "communicative disfluency" in the following manner:

"Communicative disfluency = Something in the communicative performance that disturbs the flow of communication". For a discussion of different definitions and characterizations of "disfluency", see Eklund (2004: 1548–160).

Some examples of what has been proposed as "communicative disfluencies" include:

(i) Mechanisms for hesitation or clarification, like *eh* or *I mean*, lengthening, pausing or self-repetition, which all have the effect of holding the floor.

(ii) Mechanisms for changing the expression or content of what you are communicating.

(iii) Short words, phrases to give feedback.

(iv) Stammering.

With the possible exception of stammering, we can now raise the question: are these really examples of disfluencies? Are they not rather examples of phenomena that are needed to make interactive communication fluent? Even for stammering, we might wonder if this phenomenon for a particular individual in a particular state might not be what is required to communicate.

Another way of approaching the "disfluent" phenomena exemplified above is to ask if they are fluent or disfluent in all types of communicative activity. It seems fairly clear that most of them would be "disfluent" in written language, if we are not trying to capture authentic speech in writing. It also seems clear that many of them might be disfluent in many types of public speaking. But this does not mean that they are disfluent in interactive (small) talk, where it is important that you are able to hesitate, change your mind, repeat for clarity, be flexible and non-categorical and give continuous unobtrusive feedback. It seems fairly clear, that many of the functional means for achieving these goals have been labelled as "disfluencies", since they have no role in the kind of fluency required in written language or public speaking, but are concerned with the "communication management" (both ICM and OCM) required in fluent interactive communication.

My claim is thus that many "disfluencies" really are examples of mechanisms that are required for rational, efficient interactive communication, especially making use of processes on lower levels of awareness.

This justifies the question: Is the term disfluency (dysfluency) never appropriate? Two cases may be distinguished:

(i) Looking at one type of communication from the point of view of another, e.g. looking at interactive face-to-face communication from the point of view of written language (this is seldom, if ever, appropriate).

(ii) Comparing the ideal-normative function and goals of a particular communicative activity with actual performance, e.g. mistakes in spelling or grammar in written language or exaggerated stammering or overlong pauses in interaction where a faster tempo was expected (this can be appropriate and be the basis for attempts at change).

Can terminology be changed in science?

Sometimes terminology changes in science. Usually, this signals a change of perspective or that an earlier view is seen as inappropriate or incorrect. "Phlogiston" disappeared and "oxygen" took over, when we changed our views of how what we now think of as oxidation, takes place. "Alchemy" became "chemistry", as part of an attempt to purge the field of practices considered to be less scientific. Charles Sanders Peirce changed the name of his philosophy from "pragmatism" to "pragmaticism" – "a name so ghastly that nobody will use it", when he was dissatisfied with some of the uses made of his philosophy. There are many other examples. Change of terminology is not uncommon.

Maybe it is time to change the terminology; abandon the term "disfluent" for more positive or neutral terminology, except in a few, well defined cases where really the goals of a particular communicative activity are not being met. For these cases, perhaps the word "dysfluency" could be used.

Conclusion

I have tried to argue that the notions of fluency and disfluency need to be relativized to type of communication. I have also argued that some interactive communicative practices that might seem "disfluent" from the perspective of public speaking or written language, in fact, in interactive communication, in most cases, are the opposite, i.e. features that help interactive communication become more fluent and efficient.

Finally, I have also suggested that it might be good if our common terminology for the phenomena discussed, reflected this.

References

Allwood, J. 2013. A multidimensional activity based approach to communication. In I. Wachsmuth, J. de Ruiter, P. Jaecks, P. & S. Kopp (eds.): *Alignment in Communication*. Amsterdam: John Benjamins, 33–55.

Allwood, J., K. Grammer, S. Kopp & E. Ahlsén. 2006. A framework for analyzing embodied communicative feedback in multimodal corpora. In *Proceedings of Workshop on Multimodal Corpora – From Multimodal Behaviour Theories to Usable Models*, Saturday 28 May 2006, Bielefeld, Germany, 43–47.

Allwood, J., J. Nivre & E. Ahlsén. 1990. Speech Management: on the Non-Written Life of Speech. *Nordic Journal of Linguistics* 13(1):3–48

Eklund, R. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, *Linköping Studies in Science and Technology*, Dissertation No. 882, Department of Computer and Information Science, Linköping University, Sweden.

Kahneman, D. 2011. *Thinking Fast and Slow.* New York: Farrar, Straus and Giroux.

Lickley, R. 2015. Fluency and disfluency. In M. Redford (ed.): *The Handbook of Speech Production*. Chichester, UK: John Wiley & Sons, 445–469.

Lindblad, G. & J. Allwood. 2013. Prosodic expressions of emotions and attitudes in communicative feedback. In J. Allwood, E. Ahlsén, P. Paggio, C. Navaretta & K. Jokinen (eds.): Proceedings of the 4th Nordic Symposium on Multimodal Communication, 15–16 November 2012. Gothenburg, Sweden. *NEALT Proceedings* 21:70–76. Linköping University Electronic Press 093, 2013.

Ward, N. 2006. Non-lexical conversational sounds in American English. *Pragmatics and Cognition* 14(1):129–182.

Glottal filled pauses in German

Malte Belz
Department of German Studies and Linguistics, Humboldt-Universität zu Berlin, Berlin, Germany

Abstract

For German, filled pauses are traditionally described with a vocalic form *äh* and a vocalic-nasal form *ähm*. A corpus-based approach and a closer phonetic inspection is used here to argue for an additional form, namely glottal filled pauses. In the data analysed for this study, the glottal form is produced by all seven speakers and amounts to 21% of all filled pauses. Contexts and durations of occurrences are discussed and compared to earlier studies on traditional filled pauses. It is suggested that the glottal variant should be considered in future studies on filled pauses and disfluencies.

Acoustic forms of filled pauses

Filled pauses (FPs) are defined in many ways. In this paper, I will use as a working hypothesis the notion of non-lexical entities, without considering extra-linguistic events (laughing, coughing, etc.). FPs are used as hesitation devices, but also serve other functions (Lickley 2015: 463). They are often exemplified with graphemic or phonetic realizations of the most frequent forms in a respective language. Lounsbury (1954) transliterates FPs in English as *hem* and *haw*, while Maclay and Osgood (1959) give more phonetic detail by listing the transcripts [ɛ æ r ə m]. It is largely agreed on that FPs often exhibit "both a prolonged vowel sound and a vowel (usually) followed by a nasal" (Lickley 2015: 458). Other forms that may be subsumed under FPs are clicks (Trouvain, Fauth & Möbius, 2015) and breath pauses (Trouvain et al. 2016). For German, the most cited forms are probably *äh* and *ähm*, with possible phonetic transcriptions varying between [ɛː] or [əː] and [ɛːm] or [əːm], although other forms are mentioned as well (cf. Schönle & Conrad 1985 for *ah* and *mh*).

In this paper, I will explore whether an additional form – a glottal filled pause – can be assumed for German spontaneous speech. This research question is part of a PhD project, in which I am currently investigating the link between form and function of FPs. In the process of annotation (cf. Section 2), I noticed sequences of glottal pulses and creak phonation without coarticulated vowels that seem to be used in a similar way to other FPs.

Example 1 gives a broad transcription of a speaker's utterance in a dialogue[i] of the GECO corpus (cf. Data and annotation). Durations are given within angle brackets. The speaker produces a glottalized sequence of approximately 16 glottal pulses before uttering another *yes* (cf. Figure 1a for a depiction of the signal). For a first description, a wildcard notation of a creaky sonorant is used.

(1) <[jaː] 580 ms> <*exhalation* 373 ms> <*inhalation* 718 ms> <[S̰ː][a] 492 ms> < [jaː] 630 ms> <[ç ˈvaɪs nɪ ˈalzoː] 706 ms>

a. S ≙ sonorant

Transliterations:
‚ja, ja ich weiß nicht also'
'yes, yes I don't know, well'

This glottalized sequence seems to be different from the rule-governed [ʔ]-epenthesis as predicted in German phonology:

(2) $\emptyset \rightarrow [ʔ] / \left\{ \begin{matrix} V \_ \ 'V \\ \# \_ \ V \end{matrix} \right\}$ (Hall 2011: 66)

Data and annotation

To investigate the forms of filled pauses, a multi-layer annotation scheme was added to the corpus GECO (GErman COnvergence) (Schweitzer & Lewandowski 2013). Six dialogues of the multi-modal condition are annotated as to now. In this condition, interlocutors are visible to each other, while speaking freely about any subject, separated by a transparent window and connected via headphones. Five of seven participants (A, C, K, M, D) participate in two dialogues. Each dialogue lasts 25 minutes. All speakers are female students, some with a noticeable Southern German (Swabian) accent.

Filled pauses are marked on a hesitation tier in Praat (Boersma 2001) with *fv* (vocalic or vocalic-nasal or nasal filler), *fg* (glottal filler) or *fc* (click filler), based on the perceptual categorization of the annotator. For the annotation of *fg*, additional cues were used such as irregular voicing periods (oscillogram) or clearly visible glottal stops (spectrogram). A further, yet preliminary approach is that no prominent vowel quality can be perceived.

The last speech segment to the immediate left and the first to the right of an FP are marked on the hesitation layer with *as* (antecedent segment), *ap* (antecedent silent pause), *ah* (antecedent breath pause), *ac* (antecedent click), *at* (antecedent turn), and postcedent *ps, pp, ph, pc* and *pt*, respectively. Further transcriptions of *a, f* and *p* categories and pause specifics (inhalation, exhalation) are annotated on a segmentation layer.

All files are converted to an EMU speech database (Winkelmann, Harrington & Jänsch, 2017) with help of the *emuR* package 0.2.1 (Winkelmann et al. 2016) in R (R Core Team 2016).

Results

Filled pause types

Table 1 shows the distribution of glottal, vocalic and click filled pauses per speaker.

Table 1. Filled pause type, word count and total frequency of FP per speaker.

	fc		fg		fv		Σ	Words	FP
	N	%	N	%	N	%	N	N	%
A	3	9.1	13	39.4	17	51.5	33	4214	.78
C	27	25.7	23	21.9	55	52.4	105	4766	2.2
K	9	7.6	31	26.1	79	66.4	119	8193	1.5
D	0	0	15	18.5	66	81.5	81	4987	1.6
F	0	0	1	16.7	5	83.3	6	1401	.43
M	0	0	8	11.0	65	89.0	73	5880	1.2
J	0	0	1	3.8	25	96.2	26	3092	.84
Σ	39	8.8	92	20.7	312	70.4	443	32533	1.4

Vocalic FPs are the most frequent form (70.4%), followed by glottal FPs (20.7%) and click FPs (8.8%). Some speakers do not utter any click FPs at all, whereas glottal FPs do occur at least once per speaker. The glottal FP with antecedent inhalation pause and postcedent *ja* (cf. Example 1) is shown in Figure 1a. An example with antecedent segmental and postcedent silent context is given in Figure 1b.

Immediate context

Figure 2 shows the distribution of antecedent, FP, and postcedent per FP type. The right tail of the distribution is cut off at five instances or less for plotting purposes, omitting 9.4% of the data. In the first bar of Figure 2 (*as_FP_ps*), 23.2% of all contexts with adjacent speech segments are glottal FPs. In the second and third bar, contexts with antecedent silent pauses and segmental postcedents (*ap_FP_ps*) exhibit more glottal FPs than those with antecedent breath pauses (*ah_FP_ps*).

This difference between silent and breath pauses is furthered when the contexts are reversed. In the fourth and fifth bar of Figure 2, contexts with segmental antecedents and postcedent silent pauses (*as_FP_pp*) exhibit glottal FPs, whereas those with postcedent breath pauses (*as_FP_ph*) show no glottal FPs at all. Most of the breath pauses considered here are inhalation breath pauses (50 of 52 in *ah_FP_*ps and 27 of 36 in *as_FP_ph*).

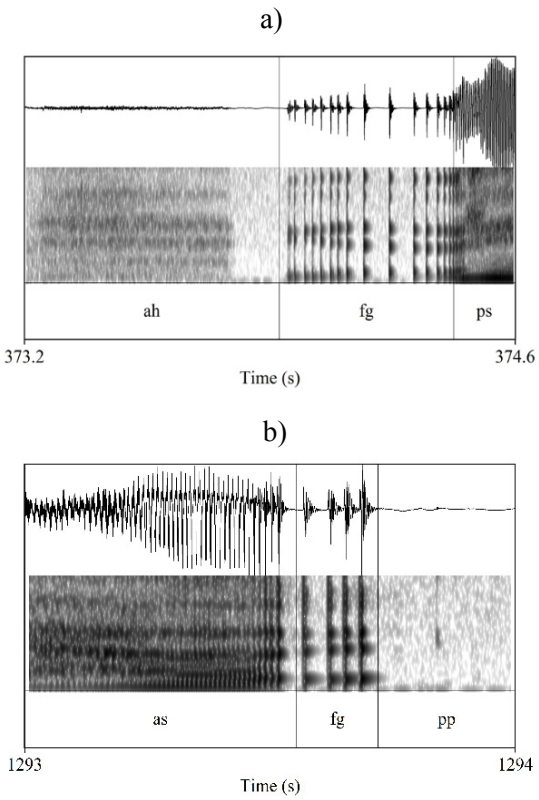


Figure 1. ***a)*** *Glottal filled pause (492 ms) labelled* fg*, preceded by breath intake (718 ms) and followed by [j] (173 ms) in* ja 'yes' *of subject C speaking with D.* ***b)*** *Glottal filled pause (88 ms) labelled* fg*, preceded by [a:] (296 ms) in* ja 'yes' *and followed by a silent pause (149 ms) of subject C speaking with D.*

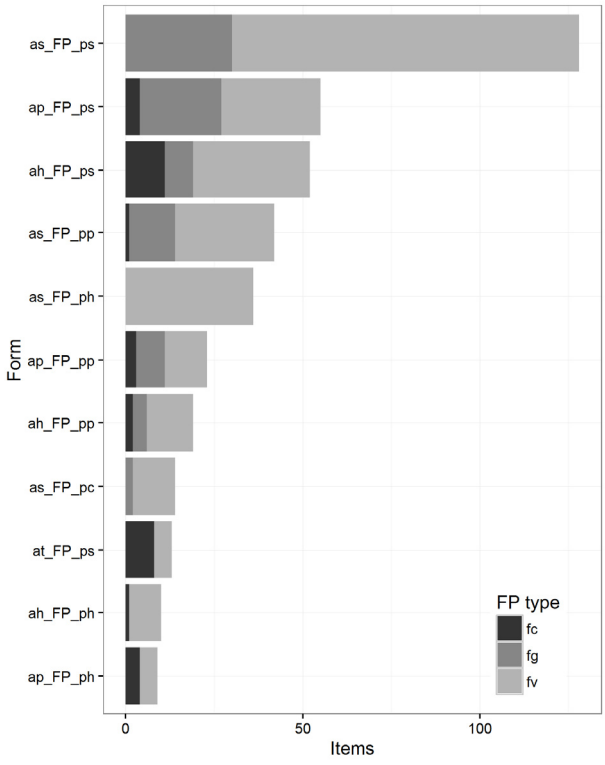


Figure 2. Distribution of filled pause forms per type. 'FP' *is a substitution for any type of* fc*,* fg*, or* fv*.*

Durational features

Vocalic and vocalic-nasal form show the longest durations, click FPs the shortest. This is plausible, given the articulatory features of clicks. The durations of glottal FPs are in between. Table 2 presents the mean lengths and standard deviations of FP types. Each paired comparison is significant (*fc* vs. *fg*: $t = 5.4$, $df = 106$, $p < .001$; *fc* vs. *fv*: $t = 14$, $df = 61$, $p < .001$; *fg* vs. *fv*: $t = 6.2$, $df = 140$, $p < .001$). The lengths of the vocalic and vocalic-nasal forms of *fv* (*äh* and *ähm*, without a closer inspection with respect to their vowel quality) also differ significantly ($t = -7.8$, $df = 233.3$, $p < .001$). However, the difference between glottal FPs (*fg*) and the *fv* variant *äh* is not significant ($t = -1.6$, $df = 163$, $p = .1$).

Table 2. Measures of central tendencies for vocalic fillers (fv), *glottal fillers* (fg), *and click fillers* (fc) *in milliseconds. Vocalic fillers are further split in vocalic-only and vocalic-nasal fillers.*

		x̄	SD
fc		109.7	86.1
fg		232.9	159.2
fv		352.9	151.2
fv	*äh*	267.1	125.9
fv	*ähm*	390.7	122.4

Glottal FP vs. laryngealization

Utterance-final or word-final glottalization due to a declined fundamental frequency (f_0) and the nearing minimum of air capacity in the lungs is sometimes called laryngealization (Kohler, Peters & Wesener, 2005: 189). How are instances of glottal FPs as in *as_fg_pp* different, then, from laryngealized speech? Figure 3 and Figure 1b give some qualitative evidence by comparing two within-speaker instances of the lemma *ja* 'yes'. In Figure 3, the vowel of *ja* is laryngealized towards the end. The glottal sequence in Figure 1b, however, is made out of four clearly perceivable single pulses and an interrupted voice bar.

Discussion and conclusion

Forms of FPs are language-specific (Clark & Fox Tree 2002: 92; Leeuw 2007; Wieling et al. 2016). However, two forms of FPs are ubiquitously mentioned in the literature for various vernaculars – a vocalic-only form (*uh* in English, *äh* in German), and a vocalic-nasal form (*uhm* and *ähm*, respectively). Perceivable breath pauses and clicks are sometimes also considered FPs. This paper argues for another type of FP – a glottal variant.

In spontaneous speech, the rule of [ʔ]-epenthesis for German (cf. Example 2) is not always met.

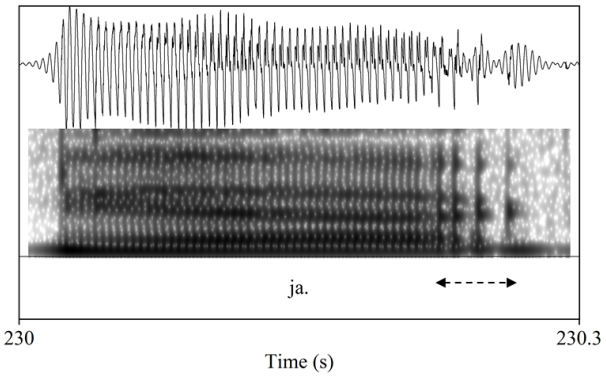


Figure 3. Example of laryngalized [jaː] ja 'yes' *(303 ms) uttered by subject C speaking with D. The laryngalized part lasts 75 ms and is indicated with the dashed arrows.*

Glottal marking of vowel-initial words in German also depends on speaking rate: the faster the rate, the less glottalization is observed (Pompino-Marschall & Żygis 2010). Nevertheless, the tendency for vowel-initial glottal stop insertion in German might add to the emergence of a glottal FP. The use of glottal stops as a functional marker is not a new phenomenon. After all, glottalization in German is also used to mark truncations (Kohler, Peters & Wesener, 2005).

Individual variation between speakers is a known challenge in FP research, and Table 1 gives ample evidence for that. For example, click FPs are only uttered by three speakers. However, although the speaker sample up to now is small and criticizable, the argumentation in favor of a new glottal FP category is strengthened by the fact that each of the speakers produces a glottal FP at least once.

Glottal FPs are found in many, but not all of the contexts where traditional vocalic and vocalic-nasal FPs occur. Even though their communicative function is yet unclear, speakers use them frequently. The non-occurrence of sequences consisting of a segmental antecedent, a glottal FP and a breath pause (*as_fg_ph*) might be related to the beginning inhalation process, in which the vocal folds are in abducted position, thus physiologically inhibiting the production of a continuing sequence of adduction gestures. A tentative implication of this physiological restriction is that speakers avoid glottal FPs in this context and produce vocalic or vocalic-nasal FPs before they run out of breath.

Strikingly, the durational distributions of glottal FPs and vocalic-only FPs overlap (cf. Table 2). One explanation is that glottal FPs consist of either one to three clearly perceivable glottal stops, or a larger sequence of creak phonation on top of an underspecified sonorant (in lack of a better description). This creaky sonorant can then be lengthened. The glottal FP, therefore, is either used as an allo-FP to *äh*, or speakers ascribe another function to it.

It seems that glottal FPs differ from non-FP word-final laryngealization and from coarticulated vowel-internal glottalization as in [ʔɛ̰ː]. Glottal FPs are, impressionistically, auditorily more prominent than word-final laryngealization and show higher glottal pulse energy. From a comparison of the sound pressure level of the laryngealized part of *ja* (Figure 3) with the glottal FP after (a different) *ja* (Figure 4) it seems that the glottal FP is uttered with a higher articulatory effort. However, a clear distinction between glottalized [ʔɛ̰ː] and a glottal FP with a conjectural form [ʢ] remains dubious. Further research will show whether they are used in a distinct or interchangeable way.

What are the merits from yet another (along with breath pauses and clicks) attested form of FPs?

First, we have to account for its mere observance.

Second, it has been shown that a more specific phonetic description apart from the assumption of standardized graphematic FP forms is a fruitful approach to reflect the actual variability in FP type, context and timing distributions more clearly.

Third, the analysis of contexts and glottal FPs might add to the debate of FPs being an intentional signal vs. FPs being an epiphenomenal, cognitive-burden induced entity (cf. Nicholson (2007) for an overview). At least the non-occurrence of glottal FPs in certain contexts might be explained by respiratory limitations.

References

Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9). 341–345.

Clark, H. H. & J. E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84(1):73–111.

De Leeuw, E. 2007. Hesitation Markers in English, German, and Dutch. *Journal of Germanic Linguistics* 19(2):85–114.

Hall, T. A. 2011. *Phonologie*: *Eine Einführung*. Berlin/New York: de Gruyter.

Kohler, K. J., B. Peters & T. Wesener. 2005. Phonetic Exponents of Disfluency in German Spontaneous Speech. In *Prosodic Structures in German Spontaneous Speech* (Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel), 185–201.

Lickley, R. J. 2015. Fluency and Disfluency. In Melissa A. Redford (ed.), *The Handbook of Speech Production*, Hoboken, NJ: John Wiley & Sons, Inc., 445–469.

Lounsbury, F. G. 1954. Transitional Probability, Linguistic Structure, and Systems of Habit-family Hierarchies. In Charles E. Osgood & Thomas A. Sebeok (eds.), *Psycholinguistics*: *A survey of theory and research problems*. Baltimore: Waverly Press, 93–101.

Maclay, H. & C. E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 5:19–44.

Nicholson, H. B. M. 2007. *Disfluency in Dialogue: Attention, Structure and Function*. PhD Thesis, University of Edinburgh.

Pompino-Marschall, B. & M. Żygis. 2010. Glottal Marking of Vowel-Initial Words in German. *ZAS Papers in Linguistics*(52). 1–17.

R Core Team. 2016. *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

Schönle, P.-W. & B. Conrad. 1985. Hesitation vowels: a motor speech respiration hypothesis. *Neuroscience Letters* 55:293–296.

Schweitzer, A. & N. Lewandowski. 2013. Convergence of Articulation Rate in Spontaneous Speech. In *Proceedings of Interspeech*, 25–29 August 2013, Lyon, France, 525–529.

Trouvain, J.. 2015. On clicks in German. In A. Leemann, M.-J. Kolly, S. Schmid & V. Dellwo (eds.), *Trends in phonetics and phonology*: *Studies from German-speaking Europe*, 21–34. Bern: Peter Lang.

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 *Proceedings of Speech Prosody (SP8)*, 31 May – 3 June 2016, Boston, USA, 31–35.

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.

Winkelmann, R., J. Harrington & K. Jänsch. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*.

Winkelmann, R., K. Jaensch, S. Cassidy & J. Harrington. 2016. *emuR: Main Package of the EMU Speech Database Management System*.

[i] The example is taken from dialogue multi_C-D_left, 372.2 ms–375.7 ms.

Differences in production of disfluencies in children with typical language development and children with mixed receptive-expressive language disorder

Axel Bergström [1], Martin Johansson [1] and Robert Eklund [2]
[1] Institute of Clinical and Experimental Medicine, Linköping University, Sweden
[2] Department of Culture and Communication, Linköping University, Sweden

Abstract

There are several studies about non-fluency in people who stutter, but comparatively few regarding children with language impairment. The current research body regarding disfluencies in children with language impairment has been using different study-designs and definitions, making some results rather contradictory.

The purpose of the present study is to expand the knowledge about disfluencies in children with language impairment and compare the occurrence of disfluencies between children with language impairment and children with typical language development in the same age group.

A total of ten children with language impairment and six children with typical language development participated in this study. The subjects were recorded when talking freely about a thematic picture or toys and then analysed by calculating disfluencies per 50 words including frequency of different kinds of disfluencies according to Johnson and Associates' (1959) classic taxonomy.

Our results show that children with language impairment do produce statistically significant more disfluency in general, notably sound and syllable repetition, broken words and prolongations.

Background

There are several studies on about non-fluency on people who stutter, but comparatively few regarding children with language impairment. The current research body regarding disfluencies in children with language impairment has been using different study-designs and definitions, making some results rather contradictory.

The purpose of the present study is to expand the knowledge about disfluencies in children with language impairment and compare the occurrence of disfluencies between children with language impairment and children with typical language development in the same age group.

A total of ten children with language impairment and six children with typical language development participated in this study. The subjects were recorded when talking freely about a thematic picture or toys and then analysed by calculating disfluencies per 50 words including frequency of different kinds of disfluencies according to Johnson and Associates' (1959) classic taxonomy.

By extensive mapping of the disfluencies used by children with language disorder possible predicative factors concerning their continued language development might be found, as well as potential new connections between disfluencies and linguistic deficiencies. This also expands upon the previously limited amount of research on the subject and can possibly be of clinical value in assessment of language disorders.

Language production can be viewed as a series of interconnected modules as in the Levelt (1989) psycholinguistic model, which (in a simplified form) explains disfluencies as delays in retrieving certain linguistic components of an utterance, or as a way of revising errors found in the utterance.

Studies of disfluency production in children with typical development have shown that the total rate of disfluency per 100 words does not change significantly between four and eight years of age (Haynes & Hood, 1977). A certain increase of interjections and a decrease of word repetitions was found, which might be attributed to the pragmatic maturation seen between ages four and eight (Haynes & Hood, 1977). The overall rate of disfluency seems to decrease first between eight and eighteen years of age, which also is thought to be due to the further development of pragmatic skill that takes place during that time period (Yairi & Clifton, 1972).

Studies searching for possible differences in disfluency production between boys and girls have shown somewhat contradictory results, though those that did find a significantly higher disfluency rate in girls than in boys attributed this to mostly contextual factors during testing (Hedenqvist & Persson, 2014; Buaka, Ström & Lóránt, 2016). One study that found a higher disfluency rate in boys than in girls used adult participants and the significance was only found in interjections and repetitions (Bortfeld et al., 2009) while Haynes and Hood (1977) and Kools and Berryman (1971) found no significant difference between sexes.

A study by Yaruss, Newman and Flora (1999) showed that the disfluency rate in children increases

in longer utterances, which is supported by McLaughlin and Cullinan (1989) who found that disfluency rate increases when the relative linguistic complexity of an utterance increases.

Bishop (1997) describes children with language disorders as a very heterogeneous group defined by when one or more domains of language is impaired. Children with language disorders generally get lower results on tests for assessing language development than their age-matched peers with typical language development. For example, Westby (1974) found that children with language disorder score lower in naming test than children with typical development.

Ullman and Pierpoint (2005) suggest that a language disorder is a result of impairments in nerves which constitutes the procedural memory, basal ganglia and parts of the frontal lobe cortex including Broca's area and supplementary motor cortex, also known as the *Procedural Deficit Hypothesis*.

Alm (2005) focused on stuttering but did also suggest that fluency disruptions are the result of neurological perturbations but etiologically because of disturbances in the '*medial premotor system'*. This is defined as nerves which travels through the cerebral cortex, the basal ganglia and finally to the supplementary motor cortex. Many of these structures seem, according to Murdoch (2010), to be involved in regular language production.

Purpose

Our two main research questions were:

1. Does the frequency of disfluencies differ between children with typical language development and children with language difficulties in both receptive and expressive language domains?

2. Do the types of disfluencies used differ between children with typical language development and children with language difficulties in both receptive and expressive language domains?

Relevant aspects not covered in this study

Due to the limited size of this study, and with regard to amount and type of data collected, we have chosen not to include analyses of syntactic placement of disfluencies or frequency and type of disfluency in relation to word classes, utterance length or linguistic complexity of utterances.

We have also chosen not to include different types of language deficits but limited ourselves to mixed receptive-expressive language disorder.

Method

Data were collected by letting the participants talk freely about a thematic picture (Lindström & Werner, 1995) and various toys. The participants were recruited by e-mail communication with four pre-schools, two of which were specialized in children with speech and language difficulties. Inclusion criteria were that the participants were to be between three and five years of age and native speakers of Swedish. The participants had either a typical language development or a diagnosed language disorder affecting both expressive and receptive language processing. The participants' parents were informed of the study and signed a letter of consent.

The data were recorded with a dictaphone by the brand Olympus, model VN-8500PC.

The data were transcribed orthographically and disfluencies per words were calculated for every child by dividing the total number of uttered words by 50, and then multiplying this by the number of disfluencies uttered. This process was divided evenly between the first two authors. Every transcription and analysis was then checked for errors or uncertainties by the other writer. The two test groups were then analysed individually and statistically compared with Mann Whitney *U*-tests using SPSS version 24. The disfluencies were classified using Johnson and Associates' (1959) taxonomy, which divides disfluencies into the categories interjections, sound and syllable repetitions, word repetitions, phrase repetitions, revisions, incomplete phrases, broken words and prolonged sounds.

The groups were not compared by age or sex because of the previously mentioned studies by Haynes and Hood (1977) and Yairi and Clifton (1972) since the focus of this study was comparing the children with typical language development and children with mixed receptive-expressive language disorder.

Results

A total of 10 children with language impairment and 6 children with typical language development (N=16) were recorded. Type (and total amount) of disfluency for every 50 words in the language impairment group is illustrated in Table 1 and Table 2 for the typical language development group. In the language impairment group (n = 10) the mean for total uttered words was 103.8 (min = 48; max = 158) with a 95% confidence interval, upper value = 129.2 and lower value = 78.4 In the typical language development group (n = 6) the mean for total uttered words was 204.2 (min = 89; max = 397) with a 95% confidence interval, upper value = 320.1 and lower value = 8.22.

Comparison

Statistically significant differences were found between the groups in the disfluency types *sound and syllable repetitions* $U = 7.5$ $p = 0.014$, *broken words* $U = 6.0$; $p = 0.009$, *prolonged sounds* $U = 3.12$ $p = 0.007$ with one-tail significance measure since zero prolongations occurred in the typical development group. Total amount of disfluencies produced over all $U = 0$; $p = 0.0288$.

Table 1. Each disfluency type for each child in the language impairment group per 50 words. C = Child.

Type of disfluency	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	% total	m/sd
Interjections	0.66	1.72	3.25	1.26	0	1.84	0	1.72	1.03	5	20	1.65 1.44(sd)
Sound and syllable repetitions	0.66	1.72	1.3	0.31	1.0	0	1.13	0.34	1.03	0.83	11.8	0.832 0.49(sd)
Word repetitions	0.66	0.57	2.6	0.95	0	0	1.13	0.69	0.52	1.25	11.9	0.837 0.76(sd)
Phrase repetitions	0	0	0.65	0.31	0	0	0	0	0	0.42	2	0.138 0.22(sd)
Revisions	0.66	0	0	0.31	0	0.37	0.57	0.7	0	0.42	4.3	0.303 0.27(sd)
Incomplete phrases	0	0	1.3	0.95	1.0	0.73	0	1.72	1.55	0	10.3	0.725 0.65(sd)
Broken Words	1.97	2.82	1.94	1.9	2.1	0.37	1.13	1.0	1.03	0.42	20.9	1.468 0.76(sd)
Prolonged sounds	0	1.72	1.94	0.63	1.0	1.1	0.57	0.69	2.56	0	14.15	1.02 0.79(sd)
Total	5.2	8.6	13.0	6.7	5.2	4.4	4.5	6.9	7.7	7.9	100	7.02 2.43(sd)

Table 2. Each disfluency type for each child in the typical language development group per 50 words. C = Child.

Type of disfluency	C1	C2	C3	C4	C5	C6	% total	m/sd
Interjections	0.76	0.56	0.58	1.42	1.28	1.03	33.5	0.94 0.36(sd)
Sound and Syllable repetitions	0.13	0	0.19	0	0.77	0	6.5	0.18 0.29(sd)
Word repetitions	0.25	0	0.58	0.36	0.25	0.34	10.5	0.3 0.14(sd)
Phrase repetitions	0.38	0	0.58	0.36	0.25	0.69	13.4	0.38 0.24(sd)
Revisions	0.5	0.56	0.19	0	0.77	0.34	11	0.39 0.27(sd)
Incomplete phrases	0	0.56	0.19	0	0.25	0.34	8	0.22 0.21(sd)
Broken words	0.25	1.1	0.19	0.71	0.25	0.34	16.9	0.47 0.36(sd)
Prolonged sounds	0	0	0	0	0	0	0	0
Total	2.3	2.2	2.5	2.9	3.8	3.1	100	2.8 0.6(sd)

Discussion

As for research question 1, our results show a significant difference in the general frequency of disfluencies produced between children with language impairment that affects both receptive and expressive language domains and children with typically developed language.

The results regarding a generally higher disfluency production in children with language impairment compared to their peers with typical language development confirm what Befi-Lopes et al. (2014) and Guo, Tomblin and Samelson (2008) found.

One way to explain the difference is to look at McLaughlin and Cullinan (1989) who stated that one specific sentence or utterance can have different linguistic complexity for two different individuals if put in relation to their respective linguistic abilities. Since children with language impairment are on a lower level regarding linguistic abilities one could possibly assume that the same utterance for a child with language impairment and a child with typically developed language could differ in fluency.

Another question is why disfluencies appear at all. Alm (2005) explains fluctuations of fluency as an interruption anywhere in what he calls the *medial premotor system*. Since these neurological structures are largely the same as Ullman and Pierpoint (2005) point out as divergent in children with language disorders there might be an etiological link between language disorders and high disfluency rate. It would be interesting if future studies would compare children with language disorder and language matched children who stutter. However, there are other perspectives on disfluencies. Allwood, Nivre and Ahlsén (1990) for example, suggest that disfluencies rather are a communicative tool.

As for research question 2, our results show there are differences in sound and syllable repetitions, broken words and prolonged sounds. Regarding prolonged sounds and broken words one possible explanation might be that it is the result of either an incomplete or slow semantic retrieval of a word and therefore, in line with Levelt (1989) and Westby (1974), could be a consequence of an exceeded linguistic demand for the child in relation to its unique linguistic abilities.

Yairi and Clifton (1972) discussed a possible link between pragmatic development and disfluency production. This was studied further by Haynes and Hood (1977) who linked decreases in specific disfluency types to pragmatic maturation. Since the terminology of linguistic pragmatics is not completely clear-cut, we have refrained from making any strong assumptions in this study. It would, however, with no doubt be interesting to look closer at possible correlations between specific pragmatic abilities and disfluencies since disfluencies seem to be affected by pragmatic development.

The present study is clearly rather limited in size and it is consequently difficult to draw strong conclusions as to how and why specific phenomenon seem to appear. However, we argue that the present study might reinforce the conclusions reported in previous studies in the same area, and hopefully thoughts about further research that can be made in purpose to contribute to what possibly could give disfluency analysis a role of a diagnostic tool in assessing risk factors in children developing a language disorder.

Conclusions and future research

In this study we found that children with mixed receptive–expressive language disorder produced a significantly higher rate of total disfluency than children with typical language development. Furthermore we have found that children with

language disorders produce higher numbers of prolongations, sound and syllable repetitions and broken words. In relation to prior research this might be explained from pragmatic, neurologic and linguistic perspectives.

The present study focused on comparison between children with language disorder children with typical language development. To further examine the linguistic components of disfluencies it would be interesting to use language-matched children with typical language development. It would also be interesting to further explore the similarities and differences between disfluency behavior in children with a language disorder and children who stutter, both age-matched and language-matched.

Finally, as was mentioned in the discussion, further studies on similarities and differences in disfluency behavior in children with language disorder, typical development and pragmatic deficits such as autism spectrum disorder could be of great worth for exploring the possible pragmatic components of disfluency.

Acknowledgements

We would like to thank all our participants.

References

Allwood, J., J. Nivre & E. Ahlsén. 1990. Speech Management: on the Non-Written Life of Speech. *Nordic Journal of Linguistics* 13(1):3–48

Alm, A. P. 2005. *On the Causal Mechanisms of Stuttering*. Diss., Lund University.

Befi-Lopes, D. M., A. M. Cáceres-Assenço., S. F., Marques & M., Vieira. 2014. School-age children with specific language impairment produce more speech disfluencies than their peers. *CODAS* 26(6):439–443.

Bishop, D. 1997. *Uncommon Understanding: Development and Disorders of Language Comprehension In Children*. Hove: Psychology Press Ltd.

Bortfeld, H., S. Leon, J. Bloom, M. Schober & S. Brennan. 2009. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language And Speech* 44(2): 123–147.

Buaka, P., N. Ström. & B. Lóránt. 2016. *Förekomsten av disfluenser hos svenska 6-åriga barn med typisk utveckling.* BA thesis, of Clinical and Experimental Medicine, Linköping University.

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 & Hearing Research* 51(3):722–738.

Haynes, W. & S. Hood. 1977. Language and disfluency variables in normal speaking children from discrete chronological age groups. *Journal of Fluency Disorders.* 2(1):57–74.

Hedenqvist, C. & F. Persson. 2014. *Förekomsten av disfluenser hos svenska 6-åringar med typisk utveckling.* MA thesis, Institute of Clinical and Experimental Medicine, Linköping University.

Johnson, W. & Associates. 1959. *The Onset of Stuttering.* Minneapolis: University of Minnesota Press.

Kools, J. & J. Berryman. 1971. Differences in Disfluency Behaviour Between Male and Female Nonstuttering Children. *Journal of Speech and Hearing* 14(1):125–130.

Lindström, E. & C. Werner. 1995. *A-ning – neurolingvistisk undersökning.* Ersta högskola – Ersta utbildningsinstitut, Stockholm.

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

McLaughlin S. & W. Cullinan. 1989. Disfluencies, Utterance Length, and Linguistic Complexity in Nonstuttering Children. *Journal of Fluency Disorders* 14(1):17–36.

Murdoch, B. E. 2010 (2nd edition). *Acquired speech and language disorders: a neuroanatomical and functional neurological approach*. Chichester, West Sussex: Wiley-Blackwell

Ullman, M. & E. Pierpont. 2005. Specific Language Impairment is not Specific to Language: The Procedural Deficit Hypothesis. *Cortex* 41(3):399–433.

Westby C. 1974. Language Performance of Stuttering and Nonstuttering Children. *Journal of Communication Disorders* 12(2): 133–145.

Yairi, E. & N. F. Clifton. 1972. Disfluent Speech Behavior of Preschool Children, High School Seniors, and Geriatric Persons. *Journal of Fluency Disorders.* 15(4):714–219.

Yaruss J., R. Newman & T. Flora. 1999. Language and Disfluency in nonstuttering children's conversational speech. *Journal of Fluency Disorders* 24(3):185–207.

Prolongation in German

Simon Betz [1,2], Robert Eklund [3] and Petra Wagner [1,2]
[1]Phonetics and Phonology Workgroup, Bielefeld University, Bielefeld, Germany
[2]CITEC, Bielefeld University, Bielefeld, Germany
[3]Department of Culture and Communication, Linköping University, Sweden

Abstract

We investigate segment prolongation as a means of disfluent hesitation in spontaneous German speech. We describe phonetic and structural features of disfluent prolongation and compare it to data of other languages and to non-disfluent prolongations.

Introduction

We investigate segment prolongation as a means of disfluent hesitation in spontaneous German speech. Prolongation is a common feature of speech occurring near phrase boundaries as a correlate of speakers coming to a halt in articulation. This phenomenon, known as phrase-final lengthening (Turk & Shattuck-Hufnagel, 2007; Umeda, 1977), utterance-final lengthening (Kohler, 1983), prepausal lengthening (O'Shaughnessy, 1995), or boundary-related lengthening (Turk & Shattuck-Hufnagel, 2007) also signals the boundary to the listener (Peters, Kohler & Wesener, 2005).

Prolongation also occurs in disfluent contexts, often in connection with other disfluencies. Within disfluency research there are only a few studies that have dealt with prolongation as a disfluency in its own right, namely corpus studies by Eklund and colleagues (Eklund & Shriberg, 1998; Eklund, 2001, 2004; Den (2003) and Lee et al., 2004) and speech synthesis studies by Betz and Wagner (2016) and Betz et al. (2016, 2017).

In this study, we follow the strand of corpus studies by Eklund and colleagues and present phonetic and structural data on prolongation in German and compare the analyses and distributions to the data available on other languages. In addition, we compare disfluent prolongations to other types of prolongation, showing that there are disfluency-specific features such as syllable position and pitch contour. We use the term *prolongation* with optional extra specifications such as "phrase-final" for all phenomena.

Method and data

We used one part of the DUEL corpus (Hough et al., 2016), called "Dreamapartment". In this corpus, two speakers have the task to build and furnish the apartment of their dreams in their imagination, with a hypothetical budget of 500.000 € and 200 m² to spare. This results in highly engaged dialogue with frequent disfluency and laughter. Eighteen speakers were recorded in 9 sessions of 30 minutes each, resulting in 4.5 hours of speech. Speakers were seated next to each other and each speaker was recorded in a separate channel.

The corpus is annotated for disfluencies following an annotation scheme specifically designed for this task (Hough et al., 2015). As there are detection problems regarding prolongations, the corpus has an extra annotation tier for lengthening created semi-automatically (Betz et al., 2017).

In the following section, we present results of the disfluent prolongation corpus data analyses with regard to frequency of occurrence (rates), duration and direct adjacency to other disfluencies, position, morphological complexity, part of speech, segment type and phonological length, and compare it to accentuation lengthening, where appropriate.

Results

Prolongation rates

Prolongation occurrence varies depending on the speaker, between 0.9 and 3.5 per 100 words. On average, there are 1.9 prolonged segments per 100 words with sd=0.8. On the time domain, there are 1.6 prolongations per minute of speech (including pauses).

The rate of 1.9% per word is higher than that reported for Swedish (1.27%), Japanese (1.13%; Den, 2003) and American English (0.5%; Eklund & Shriberg, 1998), and lower compared to Mandarin (3.5%; Lee et al., 2004). It has to be considered, however, that comparisons between different kinds of corpora might be difficult, as the DUEL corpus is specifically designed to elicit disfluencies and might thus feature a higher rate of prolongations on average. On the other hand, as shown in Betz et al., (forthcoming), there might be undetected instances of prolongations left in corpora, which would lower the rate accordingly.

Prolongations and fillers

Prolongations are closely linked to other disfluencies. Eklund (2001) reasoned that they might behave similar to fillers as they both signal hesitation by means of vocalization and duration, distinguishing them from other disfluencies, such as silences and repetitions. Adell, Bonafonte and Escudero Mancebo (2008) found in their corpus data that all filled pauses are preceded by prolongations. Betz, Wagner and Voße (2016)

reasoned that this might be related to the phenomenon of phrase final lengthening, as hesitations insert an intonation phrase boundary, which requires prolongation.

Eklund (2001) found that prolongations and filled pauses differ significantly in duration. We can confirm this finding using German data. We compare the phone duration of hesitant prolongation with the duration of prolonged phones preceding a filler and with prolonged phones that a part of a filler. As is shown in Table 1, there is a significant difference in duration. Prolonged phones in fillers are significantly longer than other prolonged phones. Prolongations without contact to fillers are slightly longer than prolongations before fillers, but not significantly. Consequently, Eklund's (2001) conclusion that prolongations and fillers are not similar in function thus receives support from German.

Table 1. Differences in duration.

	t-value(df)	p-value
PR vs. filler	t(63) = -4.6	< 0.001
PR vs. pre-filler	t(40) = 1.96	0.057
Filler vs. pre-filler	t(79) = 5.18	< 0.001
	mean duration (ms)	**sd**
Prolongation	293.9	130.2
Pre-filler	261.1	78.7
Filler	419.6	19.7

Word position & morphology/syllable structure

In the following, we investigate where hesitant prolongation in German is placed. For illustration, we compare it to prolongation that is due to accentuation from the same dataset.

Swedish is characterized by complex consonant clusters, created by additive affixation of grammatical morphemes, and the maximum allowed complexity of syllables in Swedish is C^3VC^9 (three syllable-initial consonants, and up to nine syllable-final consonants). Given that e.g. Japanese and Tok Pisin are far less permissive in this respect Eklund (2004:251) proposed that PR distribution might be the function of the syllable structure in the language, something Eklund (somewhat misleadingly) called the 'morphology matters hypothesis'. In this respect German is more similar to Swedish, of course.

First, we look at word position, distinguishing three levels: *initial* (first segment in a word), *medial* (a segment in a word that is neither first nor last) and *final* (last segment in a word). There are special cases of one-segmental variants of German words. Most common among these is the indefinite article *ein* which is frequently reduced to *n*. This would be labelled as "final" as the segment it is reduced to originally was word-final according to our definition. As shown in Table 2, disfluent prolongations have a strong tendency to fall on the last segment in a word. In this respect, it differs from accentuation where medial position is almost as frequent. The observed 7–15–78% distribution found in the word position is markedly different from the observed 30–20–50% distribution reported for American English and Swedish (Eklund & Shriberg, 1998; Eklund, 2001, 2004), two languages with a similar degree of morphological complexity.

Given that the reported distribution in morphologically less complex languages such as Tok Pisin (Eklund, 2001, 2004:251) where the figures are 15–0–85% and Mandarin (Lee et al., 2004), with 4–1–95% and Japanese (Den, 2003), with 0–5–95% suggest that syllable structure plays a vital role in what segment positions are subject to prolongation, but our finding here do not seem to lend support to at least a strong version of Eklund (2004:251).

However, recent findings from Hungarian (Gósy & Eklund, 2017) exhibit a distribution of prolongations similar to that of American English and Swedish, with the figures 18–19–63%. Compared to the very strong tendency found in Japanese and Tok Pisin to produce prolongations mainly on the final segment of words, Hungarian approaches English and Swedish in exhibiting prolongation on initial and medial segments.

Table 2. Word positions for disfluent and accentuation prolongations.

	disfluent	%	accentuation	%
initial	30	7.0	18	19.3
medial	65	15.1	34	36.6
final	336	78.0	41	44.1
Σ	431	100	93	100

Syllable position

We zoom in further and examine syllable position. As can be seen in Table 3, onsets correspond to word-initials and are dispreferred. Most disfluent prolongations are in a syllable's nucleus or coda, whereas accentuation has a strong tendency to fall on the nucleus. This supports the idea that the vocalic core of a syllable is the target for accentuation, whereas a continuant coda is as good for hesitation as a vocalic nucleus.

Table 3. Syllable positions for disfluent and accentuation prolongations.

	disfluent	%	accentuation	%
onset	30	7.0	17	18,2
nucleus	213	49,4	65	70
coda	188	43,6	11	11,8
Σ	431	100	93	100

Segment types, classes and lengths

As summarized in Table 4, sonorants like [m] and [n] outnumber the aggregate of diphthongs and long vowels in being the target of prolongations. Fricatives are more frequent than short vowels. Plosives are very rarely prolonged in German. The instances observed here are either word-initial suspensions of the occlusion (e.g. *das p:asst* ("that fits") or aspiration added to a word-final stop (e.g. *gut:*h* ("good"). Vowel length is distinctive in German, which is why it makes sense that speakers try to avoid short vowels for hesitant prolongation. Two of the most common words on which disfluent prolongation occurs in German are *und* ("and") and *dann* ("then") – both of which have a short vocalic nucleus and both are always prolonged in the final [n] instead of in the nucleus.

The segment type distribution found here exhibits a marked contrast with Swedish, where plosives are frequently prolonged, and where [t] makes the top five list in all corpora examined (Eklund, 2004:247). Once again, Hungarian (Gósy & Eklund, submitted) is similar to Swedish in that all kinds of segments are subject to prolongation.

Open vs. closed class words

Prolongation in German mainly occurs on function words/closed-class words. In the DUEL corpus, this is observed in 62.4% of all cases. In an earlier study on the GECO corpus (Schweitzer & Lewandowski, 2013), the rate is 77% (Betz, Wagner & Voße, 2016).

Table 4. Counts of most frequent phone classes and types. Percentage calculated on the total of 431 instances of disfluent prolongation.

Count	% of total	Phone class
160	37.1	sonorants
150	34.8	diphthongs + long vowels
62	14.4	fricatives
41	9.5	short vowels
10	2.3	plosives
Count		**Phone type**
98	22.7	n
50	11.6	m
30	7.0	o:
30	7.0	s
22	5.1	ə

While both rates exhibit a strong tendency towards closed-class words, the difference between the two corpora is striking. We can only speculate about the reasons for this. One reason might be the difference in corpus design, GECO being free dialogue and DUEL being highly engaged task-oriented dialogue, which might constrain speaker's freedom of prolongation placement.

Pitch contour

Research on disfluency pitch exist mainly with regard to fillers (e.g. Adell, Bonafonte & Escudero Mancebo, 2010; Belz & Reichel, 2015), clitical prolongations that resemble fillers in Japanese (Goto, Itou & Hayamizu, 1999) and Hebrew (Silber-Varod, 2010) or repetitions (Reddy & Hasegawa-Johnson, 2006), but there are no studies on pitch variations of disfluent prolongations.

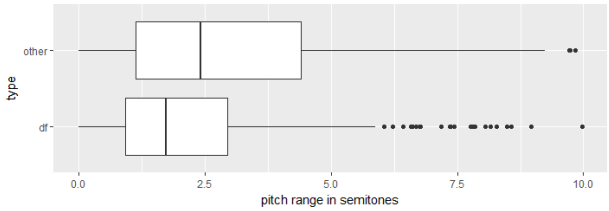


Figure 1. Pitch range differences in disfluent (df) and other prolongations.

We analysed the pitch variations of disfluent and non-disfluent prolongations. The examined data consists of the 431 disfluent prolongations extracted from the DUEL corpus and 250 other prolongations that are not disfluent, but prolonged for other reasons, such as accentuation. Our hypothesis is that pitch is one key feature to distinguish disfluent hesitation prolongations from other types of prolongation, in the sense that hesitations tend to have a flat pitch contour, whereas accentuations naturally exhibit more pitch movement, i.e. pitch accents. To investigate this, we obtained pitch values every 10ms for each instance of prolongation at hand. For each instance, we then calculated the pitch range (in semitones) by subtracting the minimum value from the maximum value. As automatic pitch extraction is known to be prone to errors, we discarded every pitch range value greater than 10 semitones. We then compared the pitch ranges of disfluent and other prolongations using a *t*-test.

As can be seen in Figure 1, non-disfluent prolongations exhibit a higher pitch range with higher variability compared to disfluent prolongations. This difference is significant, $t(543) = 4.07$, $p < 0.001$. This confirms the hypothesis that there is less pitch movement in disfluent prolongations.

Summary

German exhibits a higher rate of prolongations than most other languages tested in the series of previous corpus studies by Eklund and colleagues, although this might be due to the disfluency-specific design of the corpus at hand.

In terms of duration, German is comparable to Swedish, especially in the sense that fillers are significantly different in duration from prolongations. The preferred segmental targets for prolongations are long vocalic nuclei or sonorants codas. The nuclei will often be word-final, resulting

in a high percentage of word-final prolongations. This is markedly different from Swedish, where consonant prolongation is a common phenomenon, which can also occur word-initially.

In line with earlier studies, we observe a strong tendency for disfluency-related prolongations occurring in closed-class words, although with differences with regard to corpus type.

Pitch variation defines the type of the prolongation: Disfluent prolongations have a comparatively flatter pitch contour compared to other prolongations such as accentuation related ones. For future work, these analyses can be extended to the interaction of prolongations and fillers, for which studies on pitch are available.

Acknowledgements

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

References

Adell, J., A. Bonafonte & D. Escudero Mancebo. 2010. Modelling filled pauses prosody to synthesise disfluent speech. In *Proceedings of Speech Prosody* 2010, Chicago, USA.

Adell, J., A. Bonafonte & D. Escudero Mancebo. 2008. On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms. In *Proceedings of Interspeech*, 22–26 September 2008, Brisbane, Australia, 2278–2281.

Belz, M. & U. D. Reichel, 2015. Pitch characteristics of filled pauses. Presented at *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS)*, 8–9 August 2015, Edinburgh, UK (no page numbers).

Betz, S., J. Voße, S. Zarrieß & P. Wagner. 2017. Increasing recall of lengthening detection via semi-automatic classification. Accepted for Interspeech 2017.

Betz, S., P. Wagner & J. Voße. 2016. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. *Phonetik und Phonologie* 12:19–22

Betz, S. & P. Wagner. 2016. Disfluent Lengthening in Spontaneous Speech. *Studientexte zur Sprach-kommunikation (81). Elektronische Sprach-signalverarbeitung (ESSV) 2016*, 135–144.

Den, Y, 2003. Some strategies in prolonging speech segments in spontaneous Japanese. In R. Eklund (ed.), *Proceedings of DiSS'03, Disfluency in Spontaneous Speech*, 5–8 September 2003, Göteborg, Sweden. *Gothenburg Papers in Theoretical Linguistics 90*, ISSN 0349–1021, 87–90.

Eklund, R. 2001. Prolongations: A dark horse in the disfluency stable. In *Proceedings of DISS 2001, Disfluency in Spontaneous Speech*. 29–30 August 2001, Edinburgh, UK, 5–8.

Eklund, R. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping University, Sweden. ISBN 91-7373-966-9, ISSN 0345-7524

Eklund, R. & E. Shriberg. 1998. Crosslinguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. In *Proceedings of ICSLP 98*, 30 November – 5 December 1998, Sydney, Australia, 6:2631–2634.

Gósy, M. & R. Eklund. 2017. Segment Prolongation in Hungarian. In R. Eklund (ed.): *Proceedings of DiSS 2017*, 18–19 June 2017. Royal Institute of Technology, Stockholm, Sweden [this volume], 29–32.

Goto, M., K. Itou & S. Hayamizu. 1999. A real-time filled pause detection system for spontaneous speech recognition. In *Proceedings of Eurospeech,* 1999, Budapest, Hungary, 227–230.

Hough, J., L. de Ruiter, S. Betz & D. Schlangen. 2015. Disfluency and laughter annotation in a light-weight dialogue mark-up protocol. Presented at *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS)*, 8–9 August 2015, Edinburgh, UK (no page numbers).

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 *Proceedings of LREC 2016*, 23–28 May 2016, Portorož, Slovenia, 1784–1788.

Kohler, K. J. 1983. Prosodic boundary signals in German. *Phonetica* 40(2):89–134.

Lee, T.-L., Y.-F. He, Y.-J. Huang, S.-C. Tseng & R. Eklund. 2004. Prolongation in spontaneous Mandarin. In *Proceedings of Interspeech 2004*, 4–8 October 2004, Jeju Island, Korea, vol. III, 2181–2184.

O'Shaughnessy, D. 1995. Timing patterns in fluent and disfluent spontaneous speech. *Proceedings of ICASSP-95*, 9–12 1995, Detroit, Michigan, vol. 1, 600–603.

Peters, B., K. J. Kohler & T. Wesener. 2005. Phonetische Merkmale prosodischer Phrasierung in deutscher Spontansprache. In *K. J. Kohler, F. Kleber und B. Peters (eds.), Prosodic Structures in German Spontaneous Speech,* Kiel: IPDS, 143–184.

Reddy, R. M. & M. A. Hasegawa-Johnson. 2006. Analysis of Pitch Contours in Repetition-Disfluency using Stem-ML. *Midwest Computational Linguistics Colloquium, 2006.*

Schweitzer, A. & N. Lewandowski. 2013. Convergence of articulation rate in spontaneous speech. In *Proceedings of Interspeech 2013*, 25–29 August 2013, Lyon, France, 525–529.

Silber-Varod, V. 2010. Phonological aspects of hesitation disfluencies. In *Proceedings of Speech Prosody 2010*, 11–14 May 2010, Chicago, USA, 14–19.

Turk, A. E. & S. Shattuck-Hufnagel. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 35(4):445–472.

Umeda, N. 1977. Consonant duration in American English. *The Journal of the Acoustical Society of America* 61(3):846–858.

The effects of disfluent repetitions and speech rate on recall accuracy in a discourse listening task

Jillian Donahue, Christine Schoepfer and Robin Lickley
Clinical Speech Audiology and Language Research Centre, Queen Margaret University, Edinburgh, UK

Abstract

While many studies have examined the effects of disfluency on word recognition and local syntactic or semantic issues, fewer have addressed the impact on comprehension at a discourse level. In this work, we ask what effects features typical in the pathological condition of cluttering (essentially, rapid, disfluent and unintelligible speech) have on our ability to retain the information conveyed in speech. Specifically, we manipulate repetition disfluencies and speech rate in passages of running speech. Forty participants listened to four recordings of passages presented in four conditions: Control, Rapid, Disfluent, Rapid + Disfluent. They were asked to recall details of the passages and rate their speed, fluency and comprehensibility. Both repetition disfluencies and increased speech rate significantly reduced recall of information from discourse. Though no relationship was found between the working memory span of individuals and information recall, we argue that the cognitive load of these features of cluttered speech significantly affects intelligibility and thus recall of speech.

Introduction

There is increasing interest in the effects of disfluencies of various kinds on speech comprehension. Many studies focus on the effects of filled pauses (uh, um). While these hesitation markers may often prepare the listener for unexpected words (Arnold et al., 2004; Corley, MacGregor & Donaldson, 2007; Maxfield, Lyon & Silliman, 2009), others types, including repetitions, may not have the same effect (MacGregor, Corley & Donaldson, 2009).

Most studies on comprehension of disfluent speech focus on effects on word recognition (as above) or on local syntactic interpretations (e.g., Ferreira & Bailey, 2004), but few have examined the broader impact on discourse comprehension. An exception to this is Fraundorf and Watson (2011), who found that filled pauses inserted into spoken narratives enhanced recall of plot points in the narratives, while coughs did not. The effect of disfluent repetition on recall of discourse has not been tested.

Models of discourse comprehension emphasise the involvement of working memory in recall (Kintsch & Van Dijk, 1978), but it is unknown how this interacts with the effects of repetition disfluencies and speech rate.

Some studies have focused on the local effects of typical repetition disfluencies within single sentences. In some cases, listeners fail to recall brief repetitions in spontaneous speech (Bard & Lickley, 1998). In a word monitoring task, repetitions have been found to associate with faster monitoring latencies when they disrupt a phonological phrase, but not otherwise (Fox Tree, 1995).

A study measuring elicited neural responses to words in repetition disfluencies suggested that typical repetitions may have an impact on processing of subsequent words (MacGregor, Corley & Donaldson, 2009).

In this study, rather than examining effects within isolated sentences, we focus on the effect on recall of discourse of word-onset repetitions typical of stuttering and cluttering. One reason for this is that we were interested in the intelligibility of speech that exhibits characteristics of the disorder of fluency known as Cluttering. Cluttering is an under-researched disorder, characterised by disfluencies and rapid speech rate (St Louis et al., 2007).

Word onset (fragment) repetitions are frequent in typical speech, but also in cluttered (and stuttered) speech. They occur most often on stressed syllables of content words. Since content words are crucial in conveying meaning in speech, breakdown in fluency on these words may be expected to have an impact on a listener's comprehension.

The other major aspect of cluttered speech is an elevated speech rate. Faster speech presents the listener with a somewhat degraded signal and less time to process it, so it is only to be expected that there would be effects on comprehension. Such effects have been found in studies with both younger and older adults, where speech rate was systematically varied (Wingfield et al. 1999; Wingfield, Peelle & Grossman, 2003).

Finally, comprehension depends to some extent on the listener. In older people, a clear relationship has been found between an individual's working memory capacity and their recall of information in degraded speech (Ward et al., 2016).

We ask whether any effects of disfluency or speech rate on recall of speech will vary with working memory capacity in younger adults.

In our experiment, we follow the method used by Fraundorf and Watson (2011) whereby participants listen to retold versions of readings abridged from Lewis Carroll's *Alice's Adventures in Wonderland* (1865). We used 4 passages, one for each experimental condition.

There 2 main hypotheses concerning effects on recall: (1) that recall of the passages will be adversely affected by the presence of repetition disfluencies; (2) that recall will be adversely affected by an elevated rate of speech.

We also hypothesised that participants with poorer scores for working memory would recall less of the passages than other participants, with greater effects in degraded passages.

Method

Participants

Forty individuals participated in the study, all having learnt English before the age of six, and all high-school qualified and aged between 18 and 33 years (mean age: 26.4 years and median age: 27 years; male to female ratio 23:17). Participants were excluded if they reported having a history of hearing or language comprehension difficulties, dyslexia or fluency disorder. They were also excluded if they had significant experience of interacting with individuals with stuttered or cluttered speech.

Materials

Four passages (mean word count 302) were used. Three of the passages were those used by Fraundorf and Watson (2011), and a fourth one was selected (also from Carroll, 1865) to fit in with the design of our study. The stories were paraphrased from the original story, scripted and then retold by a 26 year old female with a North American accent, using as natural a delivery as possible, and a typical rate of speech. Each story contained 14 key plot points, and the speaker retold the story one plot point at a time. If any disfluencies occurred in the retelling, the section was re-recorded until it was fluent. The recordings of the passages were created using a TASCAM DR-100 digital audio recorder in the Queen Margaret University speech laboratory.
Each passage was then edited using Audacity® 2.1.0 so that there was a total of four versions. A control version contained no further edits.

A disfluent version of each passage was created by cutting and splicing word onsets on 10% of the content words in each passage.

A fast version of the control and the disfluent passages was created Audacity's "Change Tempo" tool, keeping fundamental frequency the same as the original. A percentage increase in speech rate (51%) was determined by calculating the difference in number of syllables per minute between three conversational samples of cluttered speech and two 'typical' speech samples, as well as figures for typical speech rate found in Tauroza and Allison (1990) and Wood (2001).

Working memory (WM) span was measured using the Forward Digit Span, Reverse Digit Span and Non-Word Repetition tasks (FDS, RDS, NWR), adapted from the *Comprehensive Test of Phonological Processing* (Wagner, Torgesen & Rashotte, 1999). Three tests were chosen since a clear definition of WM is still evasive and it was felt that a score based on a range of tests would provide a more representative measure of WM capacity (Conway et al., 2005).

Procedure

Participants were seated in a quiet room and equipped with headphones (Sennheiser HD201), with one of the experimenters (first two authors) present.

All tasks were preceded by short practice tests.

First, the Working Memory assessments, FDS, RDS, NWR were administered. The digit span tasks began with three strings of two digits and ended with three strings of nine digits. A maximum score of 24 was possible. The non-word repetition task consisted of 27 test items. In each case, the test was terminated after a participant had produced three errors or reached ceiling. The overall WM score had a range of 0–25.

Next, participants heard the four passages in conditions and orders determined by Latin Square, such that each participant was tested in each condition and each passage was presented in all four conditions across the experiment. After hearing each passage, the participant was asked to retell the story that they had just heard as accurately as possible. There was a maximum score of 28 for each passage, with points given on the basis of whether full (2 points) or partial (1 point) information had been provided for each plot point. After each passage, participants also gave subjective ratings of speed, fluency and intelligibility on a 1–5 Likert scale.

Ethical permission was granted by the QMU Ethics panel, and all participants gave informed consent.

Results

Since one passage was found to have lower recall scores across all conditions, recall scores were normalized and *z* scores used.

We expected that repetition disfluencies would have a negative impact on recall and they did. We also hypothesized that an increased rate of speech would result in poorer recall and it did. There was

also a combined effect of disfluency and increased rate, such that passages with both disfluency and increased speech rate attracted the lowest recall scores (Table 1).

Table 1. Means and Standard Deviations of Raw and Normalised Passage Recall Scores under each Condition. Maximum possible raw score is 28.

Condition	N	Raw Mean	Raw SD	Normalised Mean	Normalised SD
Control	40	15.65	5.53	0.33	0.84
Rapid	39	12.82	5.65	– 0.14	0.96
Disfluent	40	13.70	6.99	0.04	1.03
Combined	39	11.79	7.11	– 0.25	1.06

Scores for control passages were higher than all other conditions. There was a significant difference in the scores between Control and Rapid; $t(38) = 3.96$, $p < .001$, between Control and Disfluent; $t(39) = 2.53$, $p = .016$ and between Control and Combined; $t(38) = 5.82$, $p < .001$. Significant differences were also found between Disfluent and Combined; $t(38) = 2.58$, $p = .014$. The difference between the Rapid and Disfluent conditions approached significance; $t(38) = \_ 1.79$, $p = .082$, but Rapid and Combined conditions were not significantly different; $t(37) = 0.668$, $p = .508$. These results are illustrated in Figure 1.

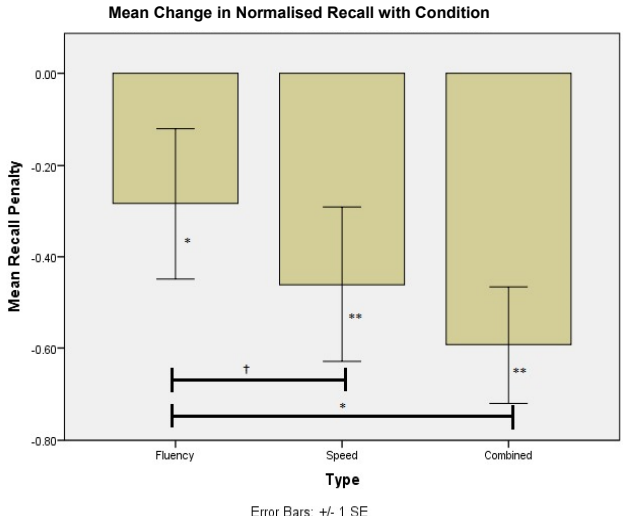


Figure 1. The Effect of Disfluencies, Rapid Speech Rate and Combined Conditions on Passage Recall Scores. Boxes show the difference between control recall scores and recall scores in each test condition.
** $p < .05$, ** $p < .01$*

Subjective judgements of disfluency, speed and intelligibility reflected the manipulations in the passages. Ratings for disfluency and speed are summarised in Table 2.

For intelligibility, the subjective ratings correlated strongly with recall scores (r_s between –41 and –.59) in the manipulated conditions, but no correlation was found in the control condition, since intelligibility scores were at ceiling.

Working memory scores are displayed in Table 3. Contrary to our expectation, there was no significant correlation between the composite WM score and recall scores in any condition. Similarly, no association was found between any of the components of the WM tests and recall performance in any conditions, except for the NWR assessment where a significant positive correlation was found with normalised Recall Scores during the Rapid condition ($r_s(39) = +.329$, $p = .041$).

Table 2. Means of Participant Subjective Ratings of Speed and Fluency for Each Condition (Scale 1–5, where 5 means fast/disfluent).

	Speed Mean	SD	Fluency Mean	SD
Control Condition	2.60	0.81	1.50	0.78
Rapid Condition	4.51	0.56	2.56	1.05
Disfluent Condition	2.58	0.81	3.75	0.70
Combined Condition	4.23	0.71	3.97	0.78

Table 3. Means and Standard Deviations (SD) of the individual WM assessments (FDS, RDS, NWR) and Comprehensive WM Score

	N	Mean	SD
Forward Digit Span	40	16.55	2.94
Non-Word Repetition	40	17.30	2.55
Reverse Digit Span	40	10.88	3.06
Comprehensive WM Score (Mean Score)	40	14.92	2.32

Discussion

In our experiment, both repetition disfluencies and increased rate of speech adversely affected recall of the content of what was heard. We did not find a clear relationship between listeners' working memory scores and their ability to recall details of what they had heard.

Our stuttering-like disfluency types and the overall methodology mean that it is hard to draw comparisons with previous work that examined the effects of typical repetitions in isolated sentences. The repetitions that we used were similar to stuttered repetitions, so we assume that they would be more disruptive to phonological phrases and have some effect on the processing of subsequent words (as per Fox Tree (1995) and MacGregor, Corley & Donaldson (2009)). It seems that this disruption has an effect on recall for the content of the speech.

Similarly, for speech rate, previous work has examined response latency to single sentences (Wingfield et al. 1999; Wingfield, Peelle & Grossman, 2003). Our findings suggest that the difficulty observed for single sentences translates

into negative effects on recall of the information in longer passages of speech.

Taken together, the findings confirm that the faster speech and stutter-like disfluencies that characterise pathologically cluttered speech are likely to have a significant effect on how much listeners can understand of what is being said.

The speech manipulations also affected subjective judgements of disfluency, speech rate and intelligibility of speech, lending some support to the use of such judgements in clinical assessment of cluttering (e.g., the *Cluttering Severity Instrument*, Bakker & Myers (2011)).

That the recall results did not vary significantly with individuals' working memory scores was unexpected. However, the participants were from a relatively homogeneous group, of young adults with a minimum of a full high school level of education, and the range of their working memory scores was fairly compact.

Future work in recall may take various directions. First, it is interesting to ask to what extent typical repetition disfluencies, involving single repetitions of short function words or word onsets, hinder recall of information, if at all and whether this varies with prosodic disturbance. It is also of interest to know at what point speech becomes too fast for full intelligibility. In both cases, there may be a habituation effect, which may have clinical implications for assessment of cluttered speech. Finally, despite our null findings, past work suggests that a listener's working memory capacity should have an impact on recall of speech affected by disfluency and increased rate, so work with populations with more varied working memory would be of interest.

Acknowledgements

This work was completed as part of a Masters dissertation project in Speech and Language Therapy at Queen Margaret University, Edinburgh, carried out by the first two authors, supervised by the third author.

References

Arnold, J. E., M. K. Tanenhaus, R. J. Altmann & M. Fagnano. 2004. The Old and Thee, uh, New Disfluency and Reference Resolution. *Psychological Science,* 15(9):578–582.

Bakker, K. & F. Myers. 2011. Cluttering Severity Instrument. *Computer software and manual]*. http://associations.missouristate.edu/ICA/

Bard, E. G. & R. Lickley. 1998. Disfluency Deafness: Graceful Failure in the Recognition of Running Speech. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, University of Wisconsin, Madison, Minnesota, USA, Lawrence Erlbaum, 108–113.

Carroll, L. 1865. *Alice's Adventures in Wonderland*. New York: MacMillan.

Conway, A. R. A., M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm & R. W. Engle. 2005. Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review* 12(5):769–786.

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.

Ferreira, F. & K. G. Bailey. 2004. Disfluencies and human language comprehension. *Trends in cognitive sciences 8*(5):231–237.

Fox Tree, Jean, E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34:709–738.

Fraundorf, S. H. & D. G. Watson. 2011. The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language* 65(2):161–175.

Kintsch, W. & T. A. Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review* 85(5):363–394.

MacGregor, L. J., M. Corley & D. I. Donaldson. 2009. Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language* 111(1):36–45.

Maxfield, N. D., J. M Lyon & E. R. Silliman. 2009. Disfluencies along the garden path: Brain electrophysiological evidence of disrupted sentence processing. *Brain and Language* 111(2):86–100.

St Louis, K.O., F. Myers, K. Bakker & R. Lawrence 2007. Understanding and Treating Cluttering. In E. G. Conture & R. F. Curlee (eds.), *Stuttering and Related Disorders of Fluency* (3rd ed). New York: Thieme, 297–325.

Tauroza, S. & D. Allison. 1990. Speech rates in British English. *Applied Linguistics* 11(1):90–105.

Wagner, R. K., J. K. Torgesen & C. A. Rashotte. 1999. *Comprehensive test of phonological processing.* Texas: Pro-Ed.

Ward, C. M., C. S. Rogers, K. J. Van Engen & J. E. Peelle. 2016. Effects of age, acoustic challenge, and verbal working memory on recall of narrative speech. *Experimental aging research* 42(1):97–111.

Wingfield, A., J. E. Peelle & M. Grossman. 2003. Speech rate and syntactic complexity as multiplicative factors in speech comprehension by young and older adults. *Aging, Neuropsychology, and Cognition,* 10(4):310–322.

Wingfield, A., P. A. Tun, C. K. Koh & M. J. Rosen. 1999. Regaining lost time: adult aging and the effect of time restoration on recall of time-compressed speech. *Psychology and Aging* 14(3):380

Wood, D. 2001. In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review* 57(4):573–589.

A psycholinguistic exploration of disfluency behaviour during the tip-of-the-tongue phenomenon

Megan Drevets and Robin Lickley
Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, Edinburgh, UK

Abstract

A tip-of-the-tongue state (TOT) occurs when a speaker knows a word but cannot retrieve its phonological form from memory. While previous studies have found that disfluencies are related to lexical retrieval difficulties, the literature lacks studies which have specifically investigated the impact of TOTs on disfluency. This study explores the relationship between TOTs and such disfluency behaviours as hesitations and target approximations (i.e. incorrect attempts to produce targets). TOTs were induced using the TOTimal method (Smith, Brown & Balfour, 1991), where participants memorised and retrieved the names of imaginary animals. Speech samples were analysed for TOTs and disfluencies. Disfluency rates increased with retrieval times during resolved TOTs. Additionally, target approximation rates correlated with the rates of both TOTs and "Don't Know" responses, suggesting that target approximations are not unique to TOTs but are indicative of general uncertainty during lexical retrieval.

Introduction

The tip-of-the-tongue state (TOT) is the feeling that accompanies an attempt to retrieve a word that is known but temporarily inaccessible from memory. TOTs reflect a failure in the process of lexical retrieval, as speakers in TOTs know the meaning of the word they are searching for yet are unable to retrieve the phonological representation of that word (Levelt, 1989). Two competing theoretical accounts to explain TOTs are the inhibition hypothesis and the transmission deficit hypothesis. The inhibition hypothesis holds that word retrieval is prevented or delayed by the retrieval of phonologically similar targets, and that the presence of these so-called "blockers" leads to a TOT (Jones, 1989). The transmission deficit hypothesis, on the other hand, maintains that a TOT exists when the target is of insufficient memory strength to be retrieved, despite the partial information recalled (Burke et al., 1991). Both theories centre around the observation that speakers in TOT frequently generate phonologically and semantically related words in lieu of the target (known in this study as "target approximations"). The critical difference between the two accounts is that the inhibition theory holds that these related forms are inhibitory to retrieval, while the transmission deficit theory states that they facilitate TOT resolution.

TOTs have been an area of extensive study since Brown and McNeill (1966) induced TOTs in a laboratory setting for the first time. The bulk of TOT research since has focused on the cognitive processes of speech production during TOTs, with less research investigating the outward processes of speech that occur during this phenomenon. One of these outward processes is disfluency.

Disfluencies are breaks in the continuous production of speech and are associated with both normal and abnormal motor speech and linguistic functioning (Lickley 2015). There is some evidence that specific types of disfluencies can be associated with the different levels that make up traditional psycholinguistic models of speech production. For example, filled pauses have been associated with difficulties at the early level of semantic message planning (e.g., Clark & Fox Tree, 2002; Fraundorf & Watson, 2014), while some repetitions are thought to be associated with the late level of speech error correction when an error has been detected after articulation has already begun (e.g., Postma 2000; Fraundorf & Watson, 2014). There is less available research investigating how difficulties at the intermediary stages of speech processing (e.g. lexical retrieval) can affect disfluency production. Schnadt (2009), for example, found that lexical retrieval difficulties (as demonstrated by picture naming latencies) are closely related to the likelihood of producing associated disfluencies. The literature lacks studies, however, which have specifically looked at the incidence of disfluencies during the most extreme example of lexical retrieval – namely, the TOT.

Research questions and paradigm

The main question of this study is whether there is a relationship between TOTs and both lexical retrieval times and disfluency behaviours. It was hypothesised that the number of TOTs experienced would correlate positively with the number of disfluency behaviours (e.g., target approximations) produced. Additionally, it was hypothesised that the time taken to both retrieve the correct targets and to resolve TOTs would correlate positively with the number of disfluency behaviours produced.

To this end, an experiment was designed that would induce TOTs in participants and allow the disfluency behaviours produced during lexical retrieval to be recorded and analysed. This elicitation method was based on the TOTimal method, in which participants memorise and then try to recall the names of imaginary animals (Smith, Brown & Balfour, 1991). While this elicitation method utilises newly learnt pseudowords as targets, most studies which induce TOTs experimentally employ obscure real words as targets. These semantic paradigm studies have historically induced low TOT rates (13%), however, and have had to rely heavily on participants' prior experience with certain words – a variable which is difficult to control for (Brown, 1991). Pseudoword paradigms like the TOTimal method, on the other hand, induce comparatively high TOT rates (40%) and have a greater ability to control for participants' exposure to targets, as they have only just been learnt (Smith, Brown & Balfour, 1991). One of the main assumptions of the TOTimal method is that the TOTs experienced for newly learnt targets are comparable with those experienced for real words that are already known but temporarily inaccessible. Smith, Brown and Balfour (1991) found that participants experiencing TOTs elicited using the TOTimal method were more likely to report "feeling of knowing" states and to recall partial phonological information about the targets. These studies support the idea that the TOTs elicited using the TOTimal method are akin to naturalistic stimuli and are therefore appropriate to use in TOT experiments. Lexical retrieval comparisons between male and female participants were made post-hoc after perceptually significant gender differences were apparent after the initial analyses.

Methodology

A sample of 28 participants (16 women and 12 men) took part in the study. All participants were aged between 18 and 40, proficient in English, and had either completed or were currently enrolled in tertiary education. Additionally, participants did not have a hearing or visual impairment that could not be corrected by a hearing aid or glasses, nor a communication disorder such as stammering or dyslexia. Informed consent was provided in accordance with Queen Margaret University ethics procedures.

The 20 TOTimal names used as stimuli in the present study were generated from the ARC Nonword Database (Rastle, Harrington & Coltheart, 2002) with respect to the rules of Standard Southern British English phonology, phonotactics and orthography.

To control for specific pseudoword features, all TOTimal names were monosyllabic and had CVC structure, three phonemes, four letters and a different initial phoneme each. Each TOTimal name was paired with a randomly allocated diet and a drawing resembling a real animal, in order to facilitate learning (see Figure 1).

PAKE
Diet: Seeds

Figure 1: Example of TOTimal stimuli: Illustration by Daniela Barreto (2016)

Participants were asked to memorise the names of the 20 TOTimals in two sets of 10 stimuli. An audio clip of the TOTimal's name played while participants viewed each TOTimal picture and its corresponding written name and diet in a PowerPoint presentation. Participants viewed each TOTimal four times in total, with each slide presented for 15 seconds. Participants then took part in a naming exercise based on the newly memorised words. The TOTimal pictures and diets were presented in a second PowerPoint presentation without the audio and written names. Participants had 30 seconds per slide to attempt to retrieve and produce the name of each presented TOTimal. The naming phase of the experiment was audio recorded and then transcribed orthographically.

These transcriptions were then perceptually analysed for filled pauses (uh, um), prolongations, repetitions, repairs and target approximations (i.e., incorrect attempts to produce the target). If participants did not retrieve a target after 30 seconds, they were asked whether they were experiencing a TOT (as defined prior to starting the experiment as being the state "when you feel you know the name and that you might recall it any minute, but you cannot think of the name at the moment"). If participants responded affirmatively, these responses were coded as unresolved TOTs. If they responded that they were not experiencing a TOT but did not know the word, this response was coded as a "Don't Know" response. As in Beattie and Coughlin (1999), resolved TOTs were

coded when participants exhibited word-finding verbal behaviours, facial expressions, or gestures (e.g. wincing, head in hands, etc.) prior to retrieving the target.

Spearman's rank-order correlations were used to test for associations between rates of disfluency behaviours and rates of TOTs and "Don't Know" responses, as well as between disfluency rates and retrieval times during resolved TOTs. A series of t-tests was also used to investigate perceived differences in lexical retrieval between male and female participants.

Results

Audio recordings of participant responses were analysed for disfluencies and TOTs.

TOTs and disfluencies

Overall, for successful retrievals, participants who took longer to retrieve words also produced more disfluencies. A strong positive correlation was found between average retrieval times per participant and disfluency rates (Spearman's $r_s = .7$, $N = 28$, $p < .001$) and between average retrieval times per TOTimal stimulus and disfluency rates ($r_s = .84$, $N = 20$, $p < .001$). In addition, an expected positive correlation was found between retrieval times during resolved TOTs and disfluency rates ($r_s = .37$, $p < .001$).

It was hypothesised that TOT rates would correlate positively with disfluency rates and they did ($r_s = .84$, $N = 28$ $p < .001$). However, a positive correlation was also found between the rate of "Don't Know" responses and disfluency rates ($r_s = .46$, N – 28, $p = .015$).

More specifically, TOT rates correlated positively with the rates of target approximations ($r_s = .60$, $N = 28$, $p = .001$, Figure 2). As before, however, the number of "Don't Know" responses was also correlated with the number of target approximations ($r_s = .65$, $p < .001$, Figure 3).

Gender differences in lexical retrieval

A series of post-hoc independent-samples t-tests revealed significant differences in lexical retrieval between male and female participants. Male participants experienced more TOTs and had longer and more disfluent retrieval times than female participants (Table 1).

Discussion

This study supports the notion that there is a relationship between retrieval time and disfluencies, as the longer it took participants to remember the name of a TOTimal, the more disfluent they came. This correlation also applies to TOTs, as the longer it took participants to resolve TOTs, the more disfluency behaviours they produced. The study also provides evidence that uncertainty regarding targets during lexical retrieval is associated with an increase in disfluency behaviours.

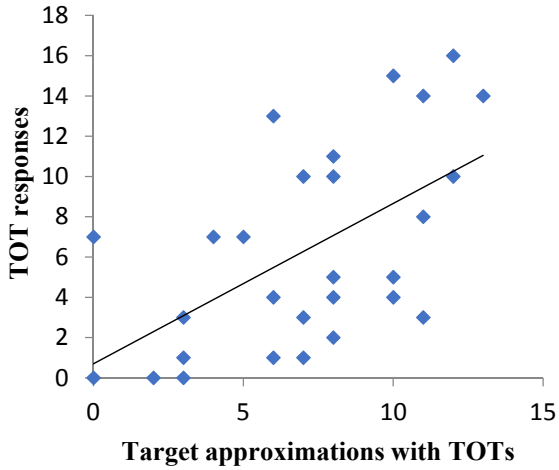


Figure 2: Number of target approximation disfluencies correlates positively with TOT rates by participants (N = 28)

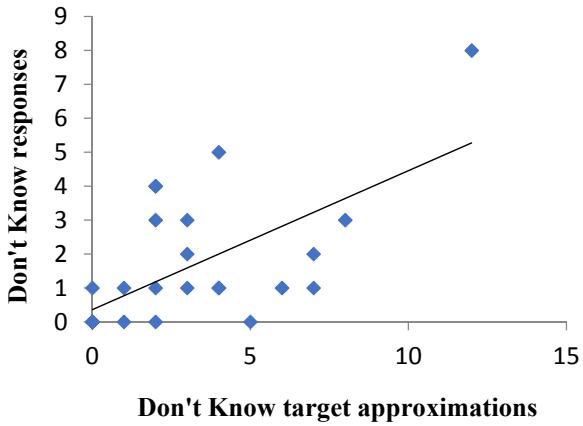


Figure 3: Number of target approximation disfluencies with "Don't Know" responses correlates positively with rate of "Don't Know" responses.

Table 1: Male vs female participants for retrieval time (RT), disfluency and TOTs.

Gender	N	Mean	St. Deviation
RT* Male	12	17.83	7.14
Female	16	9.52	6.11
Disfluency Male	12	37.67	15.11
Female	16	18.31	15.26
TOT Male	12	9.42	2.64
Female	16	6.44	3.58

* Retrieval time

This increase in disfluencies existed regardless as to whether the retrieval uncertainty was due to the participant experiencing a TOT or the participant simply not knowing the target.

Additionally, target approximations were a disfluency behaviour of interest due to their resemblance to the "blockers" held responsible for TOTs in the inhibition theory (Jones, 1989). Retrieval times were longer when participants spontaneously produced target approximations during retrieval. Additionally, a significant positive correlation was found between the rates of target approximations and TOTs, suggesting that target approximations are either the cause or consequence of TOTs. However, a significant positive correlation was also found between the rates of target approximations and "Don't Know" responses to TOTimal stimuli. These combined results indicate that target approximations (i.e. blockers) are more likely to be the cause or consequence of lexical retrieval difficulty in general, and are not — contrary to popular belief — exclusive to the TOT phenomenon. The present study acknowledges that retrieval was more difficult when participants produced target approximations. It does not provide evidence, however, that target approximations (i.e. blockers) cause TOTs, thus adding to the increasing number of studies that have not found evidence to support the widely-accepted inhibition hypothesis (e.g., Kornell & Metcalfe, 2006).

Finally, this study has identified significant differences in the way male and female participants experience lexical retrieval. As these findings were discovered post-hoc, they were not the main focus of analysis. Therefore, future research would be required to further investigate the incidence and cause of these gender differences.

Note

This work was completed as part of the first author's Master's Dissertation, contributing to her MSc degree in Speech and Language Therapy at Queen Margaret University, Edinburgh, UK. The dissertation was supervised by the second author. Address for correspondence is rlickley@qmu.ac.uk

References

Barreto, D. 2016. *Set of fantastical animals* [online]. [viewed 21 October 2016]. Available from: https://www.shutterstock.com/g/Daniela+Barreto?searchterm=fantastical+animals&sort=popular&search_source=base_gallery

Beattie, G. & J. Coughlin. 1999. An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology* 90(1):35–56.

Brown, A. S. 1991. A review of the tip-of-the-tongue experience. *Psychological Bulletin* 109(2):204–223.

Brown, R. & McNeill, D. 1966. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior* 5(4):325–337.

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.

Clark, H. H. & J. E. Fox Tree,. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1):73–111.

Fraundorf, S. H. & D. G. Watson. 2014. Alice's adventures in um-derland: psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience* 29(9):1083–1096.

Jones, G. V. 1989. Back to Woodworth: role of interlopers in the tip-of-the-tongue phenomenon. *Memory and Cognition* 17(1):69–76.

Kornell, N. & J. Metcalfe. 2006. "Blockers" do not block recall during tip-of-the-tongue states. *Metacognition and Learning* 1(3):248–261.

Levelt, W. J. 1989. *Speaking: From intention to articulation*. London: MIT Press.

Lickley, R. J. 2015. Fluency and disfluency. In M.A. Redford (ed.), *The Handbook of Speech Production,* Chichester, UK: John Wiley & Sons, 445–469.

Postma, A. 2000. Detection of errors during speech production: a review of speech monitoring models. *Cognition* 77(2):97–132.

Rastle, K. G., J. Harrington & M. Coltheart.. 2002. 358,534 nonwords: the ARC nonword database. *The Quarterly Journal of Experimental Psychology: Section A* 55(4):1339–1362.

Schnadt, M. J. 2009. *Lexical influences on disfluency production*. PhD thesis, University of Edinburgh.

Smith, S. M., J. M. Brown & S. P. Balfour 1991. TOTimals: a controlled experimental method for studying tip-of-the-tongue states. *Bulletin of the Psychonomic Society* 29(5):445–447.

Disfluency in chat and chunk phases of multiparty casual talk

Emer Gilmartin, Carl Vogel and Nick Campbell
Trinity College, Dublin

Abstract

Multiparty casual conversation lasting more than a few minutes can be viewed as a series of phases of chat and chunk type interaction, where chat is interactive conversation with several participants taking turns, and chunk refers to phases where one participant dominates the conversation, often by telling a story or giving an opinion. We investigate the distribution of disfluency in these phases in a 70-minute 5-party conversation where participants had no practical task to perform. This pilot study shows differences in the distribution of disfluency types and frequency in the two phases.

Introduction

Our understanding of task-based or instrumental talk has been greatly enhanced by corpora collected in natural and artificial settings including meetings as in the ICSI and AMI corpora (Janin et al., 2003; McCowan et al., 2005) and information gap activities such as the Map Task or Diapix. Collections of casual talk where there is no clear short term task are rarer, as are corpus-based studies of this speech activity. Interest in casual talk has resulted in the emergence of corpora of short, often dyadic social talk encounters, such as the Spontal, NOMCO and CCDb corpora, which are proving very useful in research on first encounters or short social chats (Edlund et al., 2010, ; Paggio et al., 2010; Aubrey et al,. 2013). However, much social talk involves more than two participants and often extends longer than the time allotted to such recordings. Casual talk seems to proceed in phases of chat and chunk interaction. We are currently investigating the characteristics of multiparty (3+) casual (without a clear short-term task) conversation, and have found differences in the distribution of laughter and silence in the two phases, and in the length of phases (Gilmartin et al., 2017). We aim to better understand this fundamental speech exchange system, and also to inform the design of spoken dialogue systems capable of interacting socially with users in companionship, educational and entertainment applications. Below we give a short overview of the structure of longer casual conversation, and report on a pilot study of the incidence of disfluency in one such conversation.

Multiparty casual talk

Social talk, rather than simply following Gricean maxims of efficient communication of information, is also bound by avoidance of silence and engagement in unthreatening but entertaining verbal display and interaction (Schneider, 1988). Participants can contribute at any time, unlike the more restricted roles found in more formal situations (Cheepen, 1988; Wilson, 1989). Casual conversation has been described as occurring in stages - chat and chunk (Eggins & Slade, 2004). In chat phases, participants contribute utterances more or less equally with questions and short comments. Chat is often used to 'break the ice' among strangers involved in casual talk (Laver, 1975). As conversation progresses, chat phases are interspersed with chunk phases – longer contributions from one participant – often in the form of narratives – anecdotes and recounts, opinion or discussion. The 'ownership' of chunks seems to pass around the participants in the talk (Eggins & Slade, 2004). The structure of casual conversation has also been described as a more detailed sequence of structural elements which may include Greeting, Address, Leave-taking and Goodbye sequences at the extremities, with Approach and Centring stages, somewhat similar to chat and chunk, forming the body of longer talk (Ventola, 1979). Figure 1 shows examples drawn from our data of typical chat and chunk phases in 5-party conversation. We are curious as to whether the distribution of disfluencies, including pauses, hesitations and repairs, and phenomena such as recycled restarts and abandoned utterances, will vary between different phases or subgenres of the same interaction, and indeed between different types of speech exchange system. We have prepared a 70-minute sample of extended casual conversation data, on which we are currently experimenting.

Disfluency and casual conversation

Disfluency has become a term for a range of phenomena in speech, where the speaker does not produce a full sentence. Disfluencies are usually defined around an interruption point, where the sentence flow is interrupted. Shriberg mentions that disfluencies account for up to 10% of words and over a third of utterances in natural conversation (Shriberg, 2001).

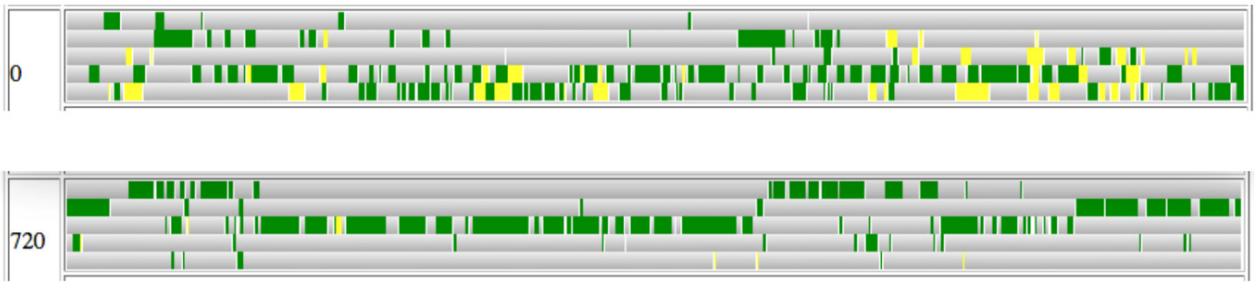


Figure 1. Examples of chat (top) and chunk (bottom) phases in two stretches from the 5-party conversation used in this study. Each row denotes the activity of one speaker across 120 seconds. Speech is green, and laughter is yellow on a grey background (silence).

Broen and Siegel's (1972) work on rate of disfluency and the speaker's perception of need to speak clearly showed fewest disfluencies occurring when the speakers spoke in front of an imaginary audience and most in casual conversation.

Of course, it should be borne in mind that much non-clinical work on disfluency has been based on adjusting 'ungrammatical' sentences to fit a text-based language model for Automatic Speech Recognition purposes. Casual conversation is not composed of spoken renditions of standard written text, but rather follows its own conventions, with an abundance of fragments or 'non-clausal units' (Biber et al., 1999). In addition, elements of spoken production labelled disfluencies, such as hesitations and filled pauses, have been shown to effect communicative goals in turn-taking management and highlighting of salient content.

We are interested in exploring the influence of conversational phase on the distribution of disfluency types in casual multiparty conversation.

Data and annotation

For this preliminary study, we use a 70-minute 5-party extract from D64, a multimodal corpus of informal conversational English recorded in an apartment living room (Oertel et al., 2010). There were no instructions to participants on topic, and participants were free to talk or not as the mood took them. The design, collection and processing of the corpus is fully described in Gilmartin and Campbell (2016). The audio recordings were found to be unsuitable for automatic segmentation due to overlap and bleed-over, and, after synchronisation, were segmented manually into speech and silence intervals using Praat (Boersma & Weenink, 2010) on 10 and 4-second or smaller windows, and doubts were settled with reference to the video using Elan (Wittenburg et al., 2006). Humans listening to speech can miss or imagine the existence of objectively measured short silences (Martin, 1970) and have difficulty recalling disfluencies from audio they have heard (Deese, 1980). However, during annotation speech could be slowed down and replayed and annotators could clearly see silences and differences in amplitude on the speech waveform and spectrogram, making it more likely that disfluencies would be noticed. After segmentation the data were manually transcribed, using a scheme largely derived from the TRAINS transcription scheme (Heeman & Allen, 1995). Words, filled and unfilled pauses, unfinished words, laughs and coughs were transcribed and marked.

The data were annotated for chat and chunk phases as described in (Gilmartin et al., 2017). The transcriptions were then automatically aligned by running the Penn Aligner (Yuan & Liberman, 2008) over a sound file and transcription for each intonation phrase annotated. Sections which could not be automatically aligned due to overlap or unfinished words were manually aligned. The word level transcription was then used with the sound files to manually annotate disfluencies. The scheme and procedures used were based largely on those in Shriberg's and Eklund's PhD theses (Shriberg, 1994; Eklund, 2004) and Lickley's Maptask corpus (Lickley, 1998), with extra labels and conventions for recycled turn beginnings (Schegloff, 1987), abandoned utterances, and disfluency during overlap. Complex, or nested, disfluencies, defined as having more than one insertion point, were labelled following Shriberg, and no indexing was used for substitutions or repetitions.

The annotated sample comprised 42 chat and 73 chunk phases, with 14,778 word tokens distributed across 2005 types. Interestingly, 'UM' and 'UH' were the 11th'and 15th most common words, reflecting the prominence of such tokens in casual talk. The most common word was 'YEAH'.

There were 1586 marked disfluencies, of which 101 were complex. Complex disfluencies were counted as single disfluency for statistical purposes. Silent pauses were disregarded for the study.

In the remaining dataset of 827 disfluencies, 726 were simple – with one interruption point. Of the 101 complex disfluencies, 82 had two interruption points, 15 had 3, and 2 each involved 4 and 5 interruption points.

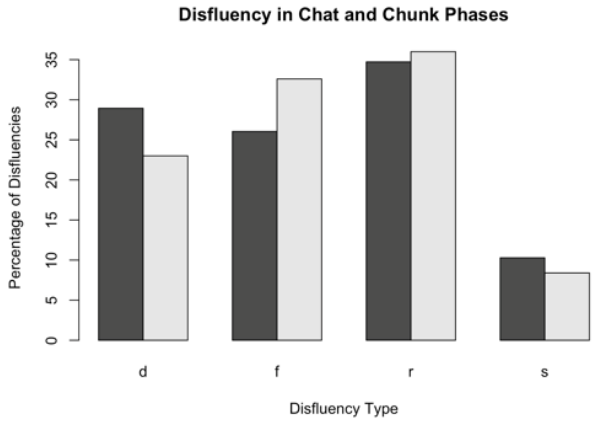


Figure 2. Distributions of disfluencies (deletions, filled pauses, repetitions, and substitutions) in chat (left) and chunk phases – as proportion in %.

There were a total of 280 distinct disfluency labels, of which 15 had 10 or more exemplars. In order to categorise this disparate group, Shriberg's Type Classification Algorithm was used to reduce label types to eight – ART (articulation), HYB (hybrid), SUB (substitution), INS (insertion), DEL (deletion), REP (repetition), CON (conjunction) and FP (filled pause). This resulted in 244 instances of filled pauses, 288 of repetition, 205 of deletions, 74 of substitutions, 8 insertions and 8 hybrid cases. For analysis, hybrids and insertions were disregarded.

Results

The distribution of disfluency types was contrasted depending on the phase of the conversation in which they occurred. Table 1 shows counts for deletions (d), filled pauses (f), repetitions (r) and substitutions (s) produced during chunk vs chat phases.

Table 1. Counts of disfluencies in chat and chunk phases.

	Chat	Chunk
d	90	115
f	81	163
r	108	180
s	32	42

The barplot for Experiment A in Figure 2 shows the proportional distribution of disfluency types (deletion, filled pause, repetition, substitution) occurring in chat (left) and chunk phases, regardless of speaker. It can be seen that filled pauses are more frequent in chunk speech than in chat phases – 33% vs 26%, while deletions are lower in chunk than in chat phases – 23% vs 29%. The proportions of repetition and substitution are quite similar in both conditions.

Discussion and conclusions

In a chunk phase, one speaker dominates the conversation, telling a story or giving an extended opinion. The reduction in the frequency of deletions in chunk vs chat in the conversation is interesting. A reduction in deletions in this modality probably reflects the lack of completion for turns – there could be fewer false starts than would occur when more than one speaker is trying to take a turn. The increase in filled pauses invites further study – the position of these pauses relative to utterance start could help distinguish whether they are the result of hesitation before semantically heavy items mid-speech or whether they are related to turn holding.

It should be noted that this is a small case study, but it makes a good case for investment of effort in collection of long form conversations, and annotation and analysis of disfluency in this genre.

Acknowledgements

This work is supported by the ERA-NET (CHISTERA) JOKER project, JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot, and by the Speech Communication Lab, Trinity College Dublin.

References

Aubrey, A. J., D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham & C. Wallraven. 2013. Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 23–28 June 2013, Portland, Oregon, USA, 277–82.

Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan & R. Quirk. 1999. *Longman Grammar of Spoken and Written English*, volume 2. London: Longman.

Boersma, P. & D. Weenink. 2010. *Praat: Doing Phonetics by Computer [Computer Program], Version 5.1. 44*. http://www.praat.org

Broen, P. A. & M. Siegel. 1972. Variations in Normal Speech Disfluencies. *Language and Speech* 15(3):219–231.

Cheepen, C. 1988. *The Predictability of Informal Conversation*. London: Pinter.

Deese, J. 1980. Pauses, Prosody, and the Demands of Production in Language. In H. W. Dechert & M. Raupach (eds.) *Temporal Variables in Speech, Studies in Honour of Frieda Goldman-Eisler*, The Hague: Mouton Publishers, 69–84.

Edlund, J., J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson & D. House. 2010. Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010),* 17–23 May 2010, Valetta, Malta, 2992–2995).

Eggins, S. & D. Slade. 2004. Analysing Casual Conversation. London: Equinox Publishing Ltd.

Eklund, R. 2004. *Disfluency in Swedish Human-Human and Human–Machine Travel Booking Dialogues*. Linköping, Sweden: Department of Computer and Information Science, Linköping Studies in Science and Technology.

Gilmartin, E. & N. Campbell. 2016. Capturing Chat: Annotation and Tools for Multiparty Casual Conversation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 25–27 May *2016*, Portoroz, Slovenia, 4453–4457.

Gilmartin, E., N. Campbell, B. R. Cowan & C. Vogel. 2017. Chunks in Multiparty Conversation - Building Blocks for Extended Social Talk. In *Proceedings of IWSDS 2017*, 6–9 June 2017, Farmington, PA.

Heeman, P. A. & J. F. Allen. 1995. *The TRAINS 93 Dialogues*. TRAINS-TN-94-2, Dept. of Computer Science, Rochester University, New York.

Janin, A., D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg & A. Stolcke. 2003. The ICSI Meeting Corpus. In 2003 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6–10 April 20003, Hongkong, China, I:364–367.

Laver, J. 1975. *Communicative Functions of Phatic Communion. Organization of Behavior in Face-to-Face Interaction*, 215–238. The Hague: Mouton

Lickley, R. J. 1998. HCRC Disfluency Coding Manual. Human Communication Research Centre, University of Edinburgh.

Martin, J. G. 1970. On Judging Pauses in Spontaneous Speech. *Journal of Verbal Learning and Verbal Behavior* 9(1):75–78.

McCowan, I., J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec & V. Karaiskos. 2005. The AMI Meeting Corpus. In L. P. J. J. Noldus, F. Grieco, L. W. S. Loijens & P. H. Zimmerman (eds.), *Proceedings of Measuring Behavior 2005*, *the 5th International Conference on Methods and Techniques in Behavioral Research*, 30 August – 2 September 2005, Wageningen, The Netherlands, ISNB 90-74821-71-5.

Oertel, C., F. Cummins, J. Edlund, P. Wagner & N. Campbell. 2010. D64: A Corpus of Richly Recorded Conversational Interaction. *Journal on Multimodal User Interfaces* 7:19-28.

Paggio, P., J. Allwood, E. Ahlsén & K. Jokinen. 2010. The NOMCO Multimodal Nordic Resource – Goals and Characteristics. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, 19–21 May 17–22 May 2010, Valetta, Malta, 2968–2973.

Schegloff, E., A. 1987. Recycled Turn Beginnings: A Precise Repair Mechanism in Conversation's Turn-Taking Organization. Talk and Social Organization, 70–85.

Schneider, K. P. 1988. *Small Talk: Analysing Phatic Discourse*. Marburg: Hitzeroth.

Shriberg, E. 2001. To "Errrr" Is Human: Ecology and Acoustics of Speech Disfluencies. Journal of the International Phonetic Association 31(01):153–169.

Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley.

Ventola, E. 1979. The Structure of Casual Conversation in English. *Journal of Pragmatics* 3(3):267–298.

Wilson, J. 1989. *On the Boundaries of Conversation*. Oxford & New York: Pergamon.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann, & H. Sloetjes. 2006. ELAN: A Professional Framework for Multimodality Research. I*n Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 22–28 May, Genoa, Italy, 1556–1559.

Yuan, J. & M. Liberman. 2008. Speaker Identification on the SCOTUS Corpus. *Journal of the Acoustical Society of America* 123(5):3878.

Segment prolongation in Hungarian

Mária Gósy [1] and Robert Eklund [2]
[1]Dept. of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[2]Department of Culture and Communication, Linköping University, Sweden

Abstract

Segment prolongation (PR) has been shown to be one of the most common forms of non-pathological speech disfluencies (Eklund, 2001). The distribution of PRs in the word (initial–medial–final segment) seems to vary between languages of different syllable-structure complexity, making it interesting to study segment prolongation in languages that exhibit different syllable structure characteristics. Previous studies have studied languages with complex syllable structure, such as English and Swedish (Eklund & Shriberg, 1998; Eklund, 2001, 2004) where affixation creates complex consonant clusters, and languages with very simple syllable, such as Japanese (Den, 2003) or Tok Pisin (Eklund, 2001, 2004), as well as Mandarin Chinese (Lee et al., 2004). In this paper we study PRs in Hungarian. Our results indicate that PRs in Hungarian are more similar to English and Swedish than it is to Japanese, Tok Pisin or Mandarin Chinese, which lends support to the notion that underlying morphology plays a role in how PRs is realised.

Introduction

Research on non-pathological disfluency has been carried out for very long, but although formal studies began already in the 1930s it was during the 1950s when extensive and formal studies saw the light. (For an overview of disfluency research the reader is referred to Eklund, 2004:51–171.)

From the very start of this research classification and terminology were at the core, but although 70 years have now passed there is still no general agreement on how to classify the different types of disfluencies in existence. In addition, even the term 'disfluency' itself is not generally agreed upon (although it is likely the most commonly used term for the phenomenon discussed here).

One type of disfluency that was recognized early on was *segment prolongation* (PR) – although the terminology varies; see Eklund (2004:163) – i.e. when a speech segment in a word is produced unusually long. Although this is similar to (what is perhaps most commonly called) filled pauses (FPs) in that both are durational and voiced, PRs have been shown to differ from FPs in some respects (e.g. Eklund, 2001).

However, one issue that has been discussed in the literature is *what segment* in the word that tends to get prolonged. A first categorization into different classes (used by Eklund & Shriberg, 1998) was to analyse PRs in three different positions: word initial (the first segment of a word), word final (the last segment of a word) and word medial (any position that is not initial or final). Eklund and Shriberg (1998) reported almost identical distribution for American English and Swedish, with a 30–20–50% distribution, for initial–medial–final position, respectively. What made these figures interesting, however, was the appearance of studies of other languages. Eklund (2001; 2004:251) reported that the corresponding figures for distribution in Tok Pisin were 15–0–85%. Den (2003) reported 10–5–85% for Japanese and Lee et al. (2004) reported 4–1–95% for Mandarin Chinese.

Swedish is characterized by complex consonant clusters, created by additive affixation of grammatical morphemes, and the maximum allowed complexity of syllables in Swedish is C^3VC^9 (three syllable-initial consonants, and up to nine syllable-final consonants). Given that e.g. Japanese and Tok Pisin are far less permissive in this respect Eklund (2004:251) proposed that PR distribution might be the function of the morphology in the language in which they appear, something Eklund (somewhat misleadingly) called the 'morphology matters hypothesis'.

Grammar and syntax, too, differ between those languages, so there might be other factors at play, and the "acid test" would then, of course, be to study languages that expand on both the grammar/syntax and the morphology scales.

The goal of this study

In this study we have set out to investigate segment prolongation in Hungarian, a language that is different from all the languages mentioned above. Hungarian is an agglutinative language that belongs to the Finno-Ugric language family with an extremely rich morphology and an extensive system of affixation. The syntactic and semantic functions of noun phrases are primarily expressed via suffixes and postpositions. Case markings are used extensively with Hungarian nouns, but pronouns, adjectives and numerals also take case and number markings. Verbs also have a considerable number of affixes (Kenesei, Vago & Fenyvesi, 2012). Hungarian words are relatively long due to the rich morphology. The number of syllables of words is 3.7 syllables on average in spontaneous speech.

Words can easily consist of 9 or more syllables. The vowel inventory of Hungarian contains 14 vowels and 36 consonants; there are short–long phonemic pairs both in vowels and consonants. Hungarian is a 'syllable-timed' language where word stress invariably falls on the initial syllable although in connected speech not all words are stressed (Siptár & Törkenczy, 2000). The goal of the study was to analyse Hungarian PRs to see to what degree that morphology and syllable structure might influence the distribution of prolonged segments in spontaneous speech of the language.

Method

Thirty-six speakers (aged between 22 and 32 years, mean age: 27 years; half of the speakers were females) participated in this study who were randomly selected from the BEA Hungarian Spontaneous Speech Database (Gósy, 2012).

All subjects were native monolingual speakers of Hungarian living in Budapest, and had a similar socio-economic status. Half of both females and males had mid-level education while the other halves had university degrees. There were no indications of language or speech disorders for any of the participants.

Recordings were made in a sound-attenuated room (the same for all), under identical technical conditions using an AT4040 microphone connected directly to a computer using GoldWave to record samples at 44.1 kHz, 16 bits, monaurally. In all recordings the interviewer was the same young female phonetician.

Various types of spontaneous speech materials were used in the analysis including narratives, storytelling and a three-member conversation with each participant. One of the narratives was about the participant's life, family, job and hobbies, while the participants talked about a topic of current interest in the other narrative and in conversations.

The duration of the analyzed spontaneous narratives was about 24 hours (ca. 40 minutes/speaker).

Target segments

All prolongations were considered occurring in the 24-hour speech material both concerning vowels and consonants. Prolongations were identified by one of the authors and was checked by another phonetician, also a native Hungarian. 0.3% of disagreement was found in the identification of prolongations between the two phoneticians; these cases were excluded from further analysis.

Prolongations were categorized according to their occurrence in the word.

Annotation was done manually using Praat software (Boersma & Weenink, 2015) according to criteria determined in advance. Vowel boundaries were marked between the onset and offset of the second formants of the vowels. Consonants were identified depending on their acoustic structures considering their voicing part (if any), burst, release, second formant information and the neighbourhood context, as appropriate. Duration measurements were carried out automatically using a specific script. A total of 948 prolongations were found which is 0.66 PRs per minute.

Examples (prolonged segment is marked bold; the English equivalent of the target word containing the prolonged segment, is given right after the Hungarian word): *olyan szülő**k**et* 'parents' *ismerek meg* 'I get acquainted with parents that', *huszonöt nagycsoporto**s*** 'preschool' *óvodás* 'twenty five preschool children', *egy tanító**́*** 'teacher' *a faluban* 'a teacher in the village', *tudod **m**ert* 'because' *nagyon elfáradtam* 'you know because I got very tired', *dolgoztam **é**s* 'and' *jól éreztem magam* 'I worked and felt well', *busszal ut**a**ztam* 'traveled' *tegnap* 'I traveled with bus yesterday', *ez az elektronikus könyvt**á**r* 'library' 'this is the electronic library', *hogya**n*** 'how' *lehet elérni* 'how can it be reached'.

Six factors were considered for analysis: 1: Position of the target segment in the word (initial, medial, final); 2: Type of segment (vowel vs. consonant); 3: Word type (content word vs. function word); 4: Number of syllables of the word containing the prolonged segment (from 1 to 7); 5: Duration of the target segment; and 6: Gender.

For statistical analysis, a Kruskal–Wallis test was performed. The confidence level was set at the conventional 95%.

Results

Position

Beginning with distributional patterns (see above), our results are shown in Table 1.

The general distribution observed (when the one-syllable word "a" is excluded from the analysis) is approximately 18–19–63%, i.e. a distribution which is quite similar to that of American English and Swedish, especially compared to the figures reported from Tok Pisin, Japanese and Mandarin Chinese.

Table 1. PR distribution in words. The total number of PRs = 779. Note that the one-syllable word, a definite article, "a", which arguably falls in all three categories (initial, medial, final) is reported separately.

Position	Number of occurrences	Percentage of total number
Initial	138	17.7%
Medial	148	19.0%
Final	493	63.3%
"a"	169	21.7%

Segments

What type of segments were subject to prolongation is shown in Table 2.

Table 2. Segments subject to prolongation, given in orthography and IPA and relative frequency given as percentages.

Vowels (N=628) (orthography)	IPA	Occurrence (%)
a	ɔ	37.1
e	ɛ	21.9
é	e:	13.0
i	i	10.3
á	a:	8.1
o	o	2.5
ó	o:	2.3
ő	ø:	2.3
í	i:	0.9
ö	ø	0.3
ü	y	0.1
ú	u:	0.1
Consonants (N=320) (orthography)	IPA	Occurrence (%)
s	ʃ	42.8
m	m	19.1
n	n	18.1
z	z	8.1
sz	s	3.7
h	h	1.8
gy	ɟ	1.2
k	k	1.2
f	f	0.9
l	l	0.9
ty	c	0.3
v	v	0.3
tt	t:	0.3
ny	ɲ	0.3
p	p	0.3
cs	tʃ	0.3

As is seen, prolongation affects all possible kinds of segments, similar to what has been reported for English and Swedish.

Word type

In Figure 1 we report how prolongation occurred as a function of whether the words affected occurred on content words or function words.

As is seen in Figure 1, prolongation on content words is, on average, shorter than it is on function words. This sits well with proposed theories that hesitation occurs whenever important choices are made in speech production, sometimes referred to as the "many-options hypothesis" (see e.g. Eklund & Wirén, 2010:24).

Number of syllables in words

We also set out to find out whether the number of syllables in the affected words played a role in segment prolongation. Our results are shown in Table 3. As can be seen there is a strong linear fall-off as a function of number of syllables in the affected words: the fewer the number of syllables, the more likely the word is to exhibit prolongation.

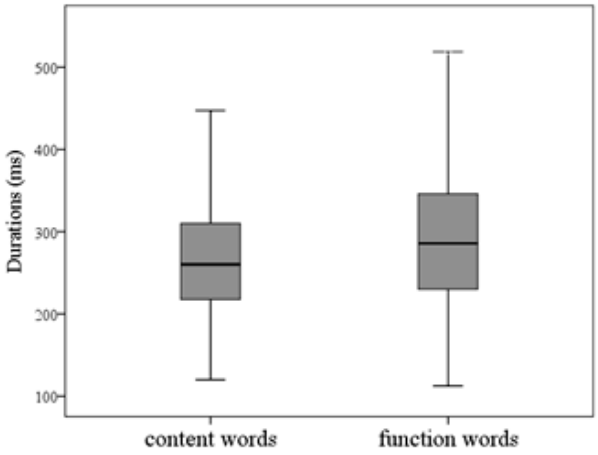


Figure 1. Prolongation as a function word type. Total number of content words = 371. Total number of function words = 577. The difference is significant, chi-square (two-tailed) at $p < 0.001$.

Table 3. Prolongation as a function of number of syllables in the affected word, given both as actual number of occurrences and as relative frequency, as well as the relative frequency of words in spontaneous speech. The total number of words analysed = 948.

Number of syllables of words	Occurrences of the words	Relative frequency (%)	Relative frequency of words in spontaneous speech
1	589	62.1	44.7
2	181	19.1	28.8
3	101	10.6	15.2
4	56	6.0	7.6
5	16	1.7	2.7
6	2	0.2	0.7
7	3	0.3	0.2

Duration of the prolonged segments

In Figure 2 below we show the results of our durational analysis, broken down for vowels and consonants. As is shown in Figure 2, prolongation is generally longer on vowels than on consonants.

Gender

Finally, we observed that there is a small, but significant, tendency for men to produce longer prolongations than females, chi-square (two–tailed); $p = 0.012$.

Discussion and conclusions

Starting with the *Distribution*, there is a remarkable similarity between our results from Hungarian and previous reports on American English and Swedish, especially when compared with the reported figures from Tok Pisin, Japanese and Mandarin Chinese. So, at a first glance it would seem as the proposed 'morphology matters hypothesis' is given some support in the present study.

However, recent results from German seem to point in another direction, and suggest that at least a strong version of the morphology matters hypothesis is not supported.

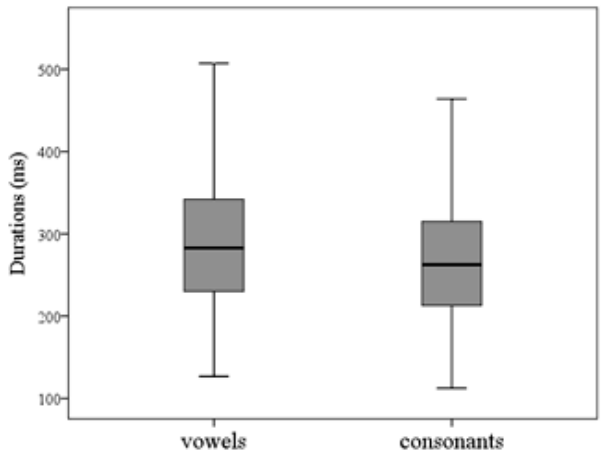


Figure 2. Durations of PRs broken down for vowels and consonants. The difference is significant, chi-square (two-tailed) at $p < 0.001$.

Evidence against such a strong interpretation comes from German, where the distribution 7–15–78% was found (Betz, Eklund & Wagner, 2017). Since German and Swedish have very similar morphology – more similar than that of Hungarian and Swedish – and both exhibit phenomena like frequent and creative compounding it would seem that morphology alone cannot explain the observed differences in distribution.

As for *Segments*, the most striking observation is that tantamount to American English and Swedish, all types of segments are subject to prolongation.

As for *Word Type*, the tendency is to prolong function words more than content words, something that sits well with the "many-options hypothesis" of the roll hesitation plays in speech production.

As for *Duration*, vowels are, on the whole and perhaps not surprisingly, more prolonged than consonants in our data.

As for *Sex*, there is a small tendency for male speakers to produce longer PRs than female speakers, supporting the proposed hypothesis that men are less prone to yielding the floor in dialog (see Eklund & Wirén, 2010:23). Females produced more PRs (500 items) than males (448 items) which can also explain their shorter lengthened segments.

We think that our paper not only sheds light on previous research on speech prolongation but also reveals many new details about this sometimes neglected disfluency. It is our hope that future studies will provide even more insights into segment prolongation in non-pathological speech.

Finally, it must be pointed out that the reported figures might be indicative of what *kind* of data was used. For example, the American English and Swedish data used by Eklund and Shriberg (1998) were all telephone data, and disfluency in dialog over a telephone line, where interlocutors cannot make use of visual cues, might be different from disfluency in face-to-face dialog.

Acknowledgements

The research was supported by OTKA Project, #108762. Thanks to Beáta Megyesi for comments on Hungarian morphology.

References

Betz, S., R. Eklund & P. Wagner. 2017. Prolongation in German. In R. Eklund (ed.): *Proceedings of DiSS 2017*, 18–19 August, Royal Institute of Technology, Stockholm, Sweden [this volume], 5–8.

Boersma, P. & D. Weenink. 2015. *Praat: doing phonetics by computer.* http://www.praat.org (Accessed 2014).

Den, Y, 2003. Some strategies in prolonging speech segments in spontaneous Japanese. In R. Eklund (ed.), *Proceedings of DiSS'03, Disfluency in Spontaneous Speech*, 5–8 September 2003, Göteborg, Sweden. *Gothenburg Papers in Theoretical Linguistics 90*, ISSN 0349–1021, 87–90.

Eklund, R. 2001. Prolongations: A dark horse in the disfluency stable. In *Proceedings of DISS 2001, Disfluency in Spontaneous Speech*. 29–30 August 2001, Edinburgh, Scotland, 5–8.

Eklund, R. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping University, Sweden. ISBN 91-7373-966-9, ISSN 0345-7524

Eklund, R. & E. Shriberg. 1998. Crosslinguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. *Proceedings of ICSLP 98*, 30 November – 5 December 1998, Sydney, Australia, 6:2631–2634.

Eklund, R. & M. Wirén. 2010. Effects of open and directed prompts on filled pauses and utterance production. In: *Proceedings of Fonetik 2010*, 2–4 June 2010, Lund, Sweden, 23–28.

Gósy, M. 2012. BEA – A multifunctional Hungarian spoken language database. *The Phonetician* 105/106, 50–61.

Kenesei, I., R. Vago. & A. Fenyvesi. 2012. *Hungarian*. New York: Routledge.

Lee, T.-L., Y.-F. He, Y.-J. Huang, S.-C. Tseng & R. Eklund. 2004. Prolongation in spontaneous Mandarin. In *Proceedings of Interspeech 2004*, 4–8 October 2004, Jeju Island, Korea, vol. III, 2181–2184.

Siptár, P. & M. Törkenczy. 2000. *The phonology of Hungarian*. Oxford: Oxford University Press.

Intervention for word-finding difficulty for children starting school who have diverse language backgrounds

Peter Howell [1], Kaho Yoshikawa[1], Kevin Tang[1, 2], John Harris[1] and Clarissa Sorger[1]
[1]Division of Psychology and Language Sciences, University College London, London, UK
[2]Department of Linguistics, Yale University, New Haven, US

Abstract

Children who have word-finding difficulty can be identified by the pattern of disfluencies in their spontaneous speech; in particular whole-word repetition of prior words often occurs when they cannot retrieve the subsequent word. Work is reviewed that shows whole-word repetitions can be used to identify children from diverse language backgrounds who have word-finding difficulty. The symptom-based identification procedure was validated using a non-word repetition task. Children who were identified as having word-finding difficulty were given phonological training that taught them features of English that they lacked (this depended on their language background). Then they received semantic training. In the cases of children whose first language was not English, the children were primed to use English and then presented with material where there was interference in meanings across the languages (English names had to be produced). It was found that this training improved a range of outcome measures related to education.

Introduction

Various disfluencies occur in the speech of young children attending UK schools. Some of these (e.g. stuttering) can require intervention by speech and language therapists whereas others (e.g. word-finding difficulty, WFD) can be addressed in school. WFD can happen irrespective of language background but is frequently seen in children who use English as an Additional Language (CwEAL) if their vocabulary is not well developed. One question that arises is whether speech difficulties and WFD lead to different patterns of disfluency that allow them to be distinguished. Once children with either form of difficulty have been distinguished, appropriate interventions can be given. A procedure that separates children with WFD (in CwEAL and children with English Only, CwEO) from both children who are fluent and also from those who have serious speech difficulty is reviewed below. Then research on an in-school intervention that addresses the best way to improve the English vocabulary acquisition of children with WFD is reported.

Identification of children with WFD

Fluent children have, by definition, low rates of all types of disfluency in their speech (Howell & Davis, 2011; Howell, 2013; Howell et al., in press; Mirawdeli & Howell, 2016). Children with serious speech difficulty have high rates of symptoms that fragment words (prolongations, word breaks and part-word repetitions) which are used to assess stuttering severity in Riley's (2009) test. Campbell (2014) showed that this set of symptoms also identified pediatric speech problems such as dyspraxia. Whole-word repetitions are not considered 'disfluent' by Riley (2009). However, whole-word repetitions are used as a diagnostic sign in the standard test for WFD (German, 1991) because children who are affected often repeat the words prior to the one they are trying to retrieve. Consequently, as whole-word repetitions are not used in Riley (2009), high rates of whole-word repetitions can be used as a selective indicator of WFD. However, whilst considerable evidence shows that whole-word repetitions are not indicative of serious fluency problems (see Howell, 2010 for review), DSM-5 maintains that whole-word repetitions are a sign of stuttering. Hence, use of whole-word repetitions as an indication of WFD, not speech difficulty, requires external validation. Also, since WFD is common in CwEAL, an appropriate validation procedure needs to work equitably whatever language a child speaks.

Non-word repetition occurs frequently in the speech of children who have serious speech difficulties (Bakhtiar, Ali & Sadegh, 2007), but happens less often in the speech of children who are fluent (Gathercole et al., 1994). Non-word repetition should also be unaffected in CwEAL who have WFD because they often have to produce unfamiliar English phoneme sequences which makes similar demands to non-word repetition (Bialystok, Majumder & Martin, 2003). CwEO who have WFD should not have non-word repetition problems if the problem is due to limitations of vocabulary, rather than articulation (Ellis Weismer et al., 2000). In summary, for children showing disfluencies, only those with high rates of Riley's (2009) symptoms should show non-word repetition problems which provides an alternative way of

distinguishing them from children with WFD (who should not have non-word repetition problems).

A further issue to address before this prediction could be tested arose because non-word repetition tasks designed for one language may not be appropriate for other languages. For instance, Masoura and Gathercole (1999) reported that Greek children who learned English at school performed better on a Greek than an English non-word repetition test. Therefore, a task that applies equitably across languages was required for assessing non-word repetition ability in heterogeneous language samples. The Universal Non-Word Repetition test was developed for this purpose (Howell et al., in press). The Universal Non-Word Repetition test employs a common core of syllabic phonotactic constraints. Non-words generated according to these constraints are phonologically well-formed for all of these languages. Howell et al. (in press) administered the Universal Non-Word Repetition test to 96 4-5 year old children attending UK schools (20.83% of whom had English as an additional language). The children's speech samples in English were assessed separately for disfluency and whole-word repetitions. Regression models showed Riley's (2009) symptom score predicted Universal Non-Word Repetition test performance, which indicated that the Universal Non-Word Repetition test scores depended on whether there was speech difficulty. Whole-word repetitions did not predict Universal Non-Word Repetition test scores, which showed that Universal Non-Word Repetition test scores did not depend on whether there was WFD. The results for both speech and WFD applied whatever language was spoken; there were no effects of language group in the analyses.

Intervention for WFD

Wing (1990) determined that both phonological and semantic training were necessary when treating people with WFD whereas Ebbels et al., (2012) reported success with semantic training alone; both studies used participants who spoke English only. CwEAL do not use all phonological constructions employed in English in their native language and can encounter lexical interference when individual sound patterns result in different meanings in English and the alternative language To design appropriate phonological training material, note that the Universal Non-Word Repetition test involves material that is easy to produce as the constraints apply to many languages. Conversely, the material excluded from the Universal Non-Word Repetition test stimuli is 'difficult' (forms that are idiosyncratic to particular languages). The phonological material that is challenging depends on the language(s) children speak. For example, Polish does not use h and θ and Urdu does not have s-consonant and s-consonant-consonant clusters. Hence material that included these forms would pose problems when children speaking these languages learn English. Improving performance on unfamiliar phonological constructions that are used in English should improve access to English vocabulary items that apply these constraints (Wing, 1990). To date, appropriate material has been developed for Polish, Urdu, Lithuanian and Mandarin.

A problem specifically faced by CwEAL is that sound forms may occur in both their languages and activate either the same semantic representation (true friends, e.g. 'cat' has the same meaning in both English and Polish), or different semantic representations (false friends, e.g. 'pet' means 'cigarette stub' in Polish but 'tame household animal' in English). False friends cause semantic interference.

Phonological and semantic training material was developed for English, Polish and Urdu. These were used to assess and then train the children. The impact of the training on a battery of educational outcomes was established to see whether training for WFD improved their school performance in general.

Method

Participants

The phonological and semantic training was delivered in group sessions over three weeks (one session per week). (The intervention takes five weeks in total. including baseline and end of intervention assessments.) Groups consisted of 38 English seven Polish and six Urdu children.

Intervention

Phonological training involved repeating a non-word that corresponded in phoneme type to an actual English word that had voicing and place on consonants changed. Participants repeated the non-words (which served as primes), and in subsequent tests saw pictures of the corresponding word items which they were required to name.

In the semantic procedure true friends and items that were only words in English were presented first to prime the children in employing their English lexicon. Then false friend materials were presented. For true friends and words with meaning in English alone, the English meaning would be primed, and it was hypothesized that this training would induce children to continue producing English words when faced with false friend material.

Assessment battery made at baseline and follow-ups

Educational impact of the training was measured before and after the intervention period, as well as at follow-up. First, children with speech difficulties other than WFD were identified (Mirawdeli & Howell, 2016). These children may have WFD *as well as* speech difficulties, but the remaining children have WFD only. After children with speech difficulties were excluded, children who have WFD in the remaining sample were identified based on rate of whole-word repetition symptoms.

The children with WFD underwent a battery of assessments. The assessments collected specific measures expected to be affected by the intervention (word-finding and fluency) and a selection from the Get Ready for Learning assessment (Bowyer-Crane et al., 2008). Inclusion of Get Ready for Learning assessment tasks ensured that a comprehensive set of language and literacy outcomes were obtained and allowed comparison with other studies. All children received all baseline, post-treatment and follow-up assessments individually.

Procedure

Children's entire performance was audio-recorded, allowing the appropriate parts to be selected for analysis. At least one test from the three skill areas (language, literacy, and phonological skills) examined in Get Ready for Learning assessment was included, with modifications made to achieve efficiency and to ensure CwEAL were assessed equitably. The assessments chosen from Get Ready for Learning assessment were: language: narrative comprehension; literacy: letter-sound knowledge and early word recognition; phonological assessment: was made using the Universal Non-Word Repetition test (Howell et al., in press).

Additional language tests were: conducted For fluency, Riley (2009) was used; two measures of WFD were employed (whole-word repetition rate and tests from German's Test of Word Finding Difficulty (German, 1991), standardised from age 4;6). The phonological and semantic components of the intervention were also assessed.

Intervention

A picture-naming task was given after both phonological and semantic training had taken place. The picture-naming tasks that were given after phonological (non-word training) and semantic training (training to produce English words) were conducted using picture material not seen during intervention. These picture-naming tasks allowed the effects of long-term changes due to training and any retention over time to be assessed.

Results and discussion

The following table gives the effect sizes (differences pre- and post-treatment) for the three language groups (English, Polish, Urdu). Effect sizes are given for four assessment categories with several measures within each category. These are: 1) Language (narrative comprehension, disfluency rate, whole-word repetitions rate, test of word-finding difficulty score (%T-units)); 2) Literacy (letter-sound knowledge, early word recognition); 3) Phonological (scores on Universal Non-Word Repetition test); 4) Picture-naming tasks corresponding to those used in the intervention (phonological and semantic).

Table 1 Results of assessment for the language groups.

Lang.	Assessment category	Measure	Cohen's D*
English	Language	Narrative comp.	0.155
		Disfluency rate	0.409
		WWR r whole-word repetitions ate	0.452
		Test of word-finding difficulty score	0.502
	Literacy	Letter-sound	0.295
		Early word	0.407
	Phonological	Universal Non-Word Repetition test score	0.417
	Picture naming	Phonological	0.552
		Semantic	0.496
Polish	Language	Narrative comp.	0.127
		Disfluency rate	0.455
		WWR whole-word repetitions rate	0.858
		Test of word-finding difficulty score	0.749
	Literacy	Letter-sound	0.633
		Early word	0.590
	Phonological	Universal Non-Word Repetition test score	0.398
	Picture naming	Phonological	0.497
		Semantic	0.502
Urdu	Language	Narrative comp.	0.135
		Disfluency rate	0.448
		WWR whole-word repetitions rate	0.508
		Test of word-finding difficulty score	0.489
	Literacy	Letter-sound	0.462
		Early word	0.324
	Phonological	Universal Non-Word Repetition test score	0.563
	Picture naming	Phonological	0.525
		Semantic	0.511

* Modulus is given as gains due to training are indicated by positive and negative values.

The results generally show moderate to good effect sizes for all measures and language groups with the exception of narrative comprehension (mean without narrative comprehension is approximately 0.5).

Whilst effects on literacy could be because of schooling or the intervention (there was no control group at present – a delayed treatment group is being run), the lack of effects on narrative comprehension suggests otherwise (this should have improved over this period too if the effects were due to schooling). Literacy showed greater effects than in Get Ready for Learning assessment for letter sound and positive effect on word reading arose (they found a negative effect).

Even CwEO may have difficulty with advanced English phonology which is a property of the material designed for training for the individual languages. Hence performance differences between CwEO and CwEAL language groups would be reduced and new ways of generating material are being developed that should increase these differences.

Acknowledgement

Supported by the Dominic Barker Trust (whose support is gratefully acknowledged).

References

Bakhtiar, M., D. A. A. Ali & S. P. M. Sadegh. 2007. Non-word repetition ability of children who do and do not stutter and covert repair hypothesis. *Indian Journal of Medical Sciences* 61:462–470.

Bialystok, E., Majumder, S. & Martin, M. M. 2003. Developing phonological awareness: Is there a bilingual advantage? *Applied Psycholinguistics,* 24(1):27–44. DOI: 10.1017/S014271640300002X

Bowyer-Crane, C., M. J. Snowling, F. J. Duff, E. Fieldsend, J. M. Carroll, J. Miles, K. Gotz & C. Hulme. 2005. Improving early language and literacy skills. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *49* (4):422–432.

Campbell, A. 2014. *The separation of spoken communication disorders based on symptoms in speech of four and five year old children.* Unpublished Master's thesis. University College London, United Kingdom.

Ebbels, S. H., H. Nicoll, B. Clark, B. Eachus, A. L. Gallagher, K. Horniman & G. Turner. 2012. Effectiveness of semantic therapy for word-finding difficulties in pupils with persistent language impairments: a randomized control trial. International *Journal of Language & Communication Disorders*, 47(1), 35–51. DOI:10.1111/j.1460-6984.2011.00073.x

Ellis Weismer, S., B. Tomblin, X. Zhang, P. Buckwalter, J. Chynoweth & M. Jones. 2000. Non-word repetition in school-aged children with and without language impairment. *Journal of Speech, Language and Hearing Research* 43(4):867–878.

Gathercole, S. E., C. S. Willis, A. D. Baddeley & H. Emslie. 1994. The Children's Test of Nonword Repetition: A test of phonological working memory. *Memory* 2:103–127.

German, D. J. 1991. Test of Word Finding in Discourse. Austin, TX: PRO-ED.

Howell, P. 2010. *Recovery from Stuttering*. New York: Psychology Press. ISBN-13: 978-1-84872-916-2

Howell, P. 2013. Screening school-aged children for risk of stuttering. *Journal of Fluency Disorders* 38:102–123.

Howell, P. & S. Davis. 2011. Predicting persistence of and recovery from stuttering at teenage based on information gathered at age eight. *Journal of Developmental and Behavioral Pediatrics* 32:196–205.

Howell, P., K. Tang, O. Tuomainen, K. Chan, K. Beltran, A. Mirawdeli & J. Harris. In press. Identification of fluency and word-finding difficulty in samples of children with diverse language backgrounds. *International Journal of Language & Communication Disorders.*

Masoura, E. & S. E. Gathercole. 1999. Phonological short-term memory and foreign vocabulary learning. *International Journal of Psychology* 34:383–388.

Mirawdeli, A. & P. Howell. 2016. Is it necessary to assess fluent symptoms, duration of dysfluent events and physical concomitants when identifying children who are at risk of speech difficulties? *Clinical Linguistics & Phonetics* 30:696–719.

Riley, G. 2009. *The Stuttering Severity Instrument for Adults and Children (SSI-4)* (4th ed.). Austin, TX: PRO-ED.

Wing, C. S. 1990. A preliminary investigation of generalization to untrained words following two treatments of children's word-finding problems. *Language, Speech, and Hearing Services in Schools* 21(3), 151–156.

A preliminary study of hesitation phenomena in L1 and L2 productions: a multimodal approach

Loulou Kosmala and Aliyah Morgenstern
Sorbonne Nouvelle University

Abstract

This paper presents a preliminary study of vocal hesitations in L1 and L2 productions using a multimodal perspective. It investigates the use of vocal hesitations of French learners of English interacting in tandem with American speakers in semi-spontaneous speech. Several hesitation markers were analyzed (filled pauses, unfilled pauses, prolongations and non-lexical sounds) based on formal and functional features as well as their relation to gesture. Results do not show great differences in the frequency of vocal hesitations between L1 and L2 productions overall; however, we find differences in duration and combination complexity. Our study indicated that vocal hesitations mainly served planning functions and were very often accompanied with gaze aversion both in L1 and L2 productions. Moreover, speakers did not tend to gesture while hesitating. We conclude that hesitations mainly served planning strategies both in L1 and L2 speech, but with some differences in duration and complexity.

Introduction

One aspect of hesitation disfluency is that it is sequential (Crible, Degand & Gilquin 2017:71) – disfluency and hesitation markers do not only appear in isolation, but in combination. Merlo and Mansur (2004:496) showed that disfluencies were correlated (P-value of 0.01), and Rieger (2003:42) found combinations of filled pauses as well as combinations with other markers (such as 'a:and uh'). Therefore, some disfluencies can be 'complex' (Shriberg, 1994). Another aspect of hesitation disfluency is its duration—hesitation is manifested acoustically as speakers generally hesitate for a certain period of time (Barr 2003:461).

Indeed, time is required in conversation as speech flow can never be fully fluent. Planning discourse can be a demanding cognitive activity as speakers need to make processing decisions such as selecting the appropriate sentence, phrase, or lexical item. Hesitation markers, such as filled pauses, can be used by speakers to buy time in order to plan their utterances. (Jucker 2015; Holmes 1988; Fehringer & Fry 2007; Tottie 2014). Lengthenings can also serve as a cue to indicate that speakers are still currently planning their speech (Betz & Wagner 2016). In this sense, the main function of vocal hesitations is to provide thinking time to introduce a 'thought unit' (Kjellmer 2003:174) and to maintain contact with the interlocutor (Guaïtella 1993).

In L2 acquisition, such processing decisions are even more challenging and can lead to additional difficulties (Watanabe and Rose 2012). L2 speech is usually characterized by slower pace and frequent use of filled and unfilled pauses (Tavakoli 2011:71). Some studies show that speakers produce longer pauses in their L2 than in their L1 (Riazantseva 2001; Tavakoli 2011), and there is a significant correlation between the number of hesitation phenomena produced in L1 and L2 speech. (Fehringer & Fry, 2007:57).

Hesitation can also be expressed non-verbally. A number of studies have shown that speakers tend to avert their gaze while hesitating (Schober et al. 2012; Glenberg, Schroeder & Robertson 1998; Swerts & Krahmer 2005) and typically produce a 'thinking face' when they are searching for a word (Goodwin & Goodwin 1996). Though gestures do not often co-occur with filled pauses (Christenfeld, Schachter & Bilous 1991); some studies have shown that gesture suspensions (such as hold and retraction) tend to co-occur with disfluency (see Seyfeddinipur 2006).

Few studies have looked at the use of 'vocal hesitations' in relation to gesture. Vocal hesitations (Guaïtella 1993:128) include unfilled pauses (silence), filled pauses ('uh' and 'um') and lengthening (word or syllable prolongation). In this preliminary study, we focus on the use of vocal hesitations in L1–L1 and L1–L2 settings. They are analyzed based on duration and complexity, as well as their non-verbal features expressed in visual modalities.

Corpus and methods

The materials for the study were provided by the SITAF corpus (Spécificité des Interactions Verbales dans le Cadre de Tandems Anglais-Français) collected at Sorbonne Nouvelle University between 2012 and 2014 (Horgues & Scheuer 2015). The data include a 25-hour video-recorded corpus, comprising 21 pairs of undergraduate students, 21 native French speakers and 21 native English speakers. For this study, two tasks were focused on: (1) "Liar, Liar"—a storytelling task in which one participant has to tell a story and insert three lies

that the partner will have to identify; (2) "Like Minds" – a question-answering task in which both participants have to answer a controversial question and decide on their degree of agreement. Eight recordings were selected, four comprising L1-L1 pairings and four L1-L2. They involved four subjects: A03 and A07 (American speakers): F03 and F07 (French speakers). All participants were female, aged 18 to 21. The duration of our selected corpus is approximately 30 minutes.

Three types of hesitation markers were analyzed, focusing on vocal productions: (1) filled pauses—non-lexical autonomous vocalic fillers, i.e. 'uh', 'um', 'euh', 'eum'; (2) unfilled pauses–perceptible pauses exceeding 400ms; (3) prolongations–syllable, vowel or consonant lengthening that does not signal the end of an intonation unit. The duration and combination of these markers were also analyzed, and coded in three distinct categories: (1) brief hesitation—hesitation made of one marker that does not exceed 600 ms; (2) elongated hesitation—hesitation made of one marker which exceeds 600 ms; (3) complex hesitation (following Shriberg, 1994)—hesitation made of several hesitation markers. Our analysis of prolongations and the distinction between brief and elongated hesitations was done based on perception.

Table 1. Examples of hesitation categories

Category	Utterance	Duration	Marker(s)
Brief hesitation	**&euh** then we came back because we were really tired.	540ms	filled pause (FP)
Elongated hesitation	**a:and** yeah that was real long and we were really tired.	1.859ms	prolongation
Complex hesitation	**&um a:and &um** it was fun.	2.555ms	FP+prolong+ filler

We also coded the functions of hesitations, making two distinctions: (1) planning function–hesitation used to plan at the macro- or micro-level, i.e. plan a new utterance, continue planning, or plan a specific lexical item; (2) reformulating function–hesitations used to reformulate parts of the utterance, either by repeating, repairing or starting a new constituent; here the hesitation typically co-occurs with repairs, restarts and repeats.

Table 2. Examples of functions

Function	Utterance
Planning	**a:and** you know they're still part of society even if they're not living in it
Reformulating	**the:e** [/] the to the flower garden

The goal was also to see what happens non-verbally when speakers produce a hesitation (aligning verbal and non-verbal modalities) when there was a change in their non-verbal behavior. We observed the following features: (1) gaze direction–towards or away from the interlocutor (2) head movement–tilts, head shakes, head nods and downward head movement; (3) facial expressions–frowning, smiling or eyes closed (4) gesture phases (Seyfeddinipur 2006: 106)–rest position or return to rest position, preparation phase, hold, gesture unit, interruption of the gesture.

Results

A total of 330 hesitations were found in the data. Table 4 indicates that on average, speakers produced 11 hesitations per minute. It seems that French learners produced more hesitations in their L1 (16.2 per minute) than in their L2 (12.3 per minute), which does not support the view that L2 learners are more hesitant in their L2 (Fehringer & Fry, 2007:57). However, our results show great individual differences between the two French speakers, which is in favour of the idea that hesitation phenomena vary from speaker to speaker (Fehringer & Fry, 2007:57), and so no conclusions can be made at that point. No significant differences were found in the functions; our results show that hesitations were mostly used for planning.

The difference between L1 and L2 seem to be at the level of the structure of hesitations. As shown in Table 3, hesitations produced by the French speakers in their L2 are longer than in their L1. They are also longer than the ones produced by the English speakers in their L1. French speakers tend to use slightly more elongated hesitations and slightly fewer brief hesitations when speaking their L2. This could indicate that French learners need more time to plan their utterances in their L2 than in their L1. Their elongated hesitations are longer when using their L1, but only represent 17% of the hesitations. However, complex hesitations produced in their L2 are much longer than those produced in their L1 and those produced by the English speakers. Hesitations comprising a single form (brief and elongated) represent 65% of the hesitations (218 out of 330 in total). French speakers use more brief filled pauses in their L1 (48 out of 55) than in their L2 (30 out of 43), while English speakers use a higher number of brief prolongations. Our results suggest differences in duration in L1 and L2 speech.

Even though no striking differences were found in the number of complex hesitations in L1 and L2 speech (Table 3), we find differences in the complexity of hesitations (Table 4). We looked at all the different combinations in the complex hesitations produced by the French learners, and we found 14 combinations in total for L1 speech: 11 composed of two forms; two composed of three forms, and one composed of four forms.

Table 3. Overall results.

	Task 1 L1-L1						Task 1 & 2 L1-L2						Task 1 L1-L1 & 2 L1-L2						
	L1 French						L2 English						L1 English						Total
	F03		F07		Total		F03		F07		Total		A03		A07		Total		
Speaking time	1.74 mn		5.28 mn		7.02 mn		4.17 mn		4.26 mn		8.43 mn		7.14 mn		6.27 mn		13.43 mn		30 mn
No. hesitations	12		102		114		28		76		104		68		44		112		330
No. hesitations per minute	6.8		19.3		16.2		6.7		17.8		12.3		9.5		7		8.3		11
Average duration	878ms		786ms		796ms		663ms		1142ms		1014ms		969ms		917ms		948ms		916ms
No. Brief Hesitations	3	25%	52	50%	55	48%	14	50%	29	38%	43	41%	31	45%	13	29%	44	39%	142
No. Elongated Hesitations	3	25%	17	16%	20	17%	8	28%	20	26%	28	26%	10	14%	18	40%	28	25%	76
No. Complex Hesitations	6	50%	33	32%	39	34%	6	22%	27	35%	33	31%	27	39%	13	29%	40	36%	112
No. Planning functions	10	83%	79	78%	88	78%	20	71%	63	82%	82	79%	56	82%	42	95%	98	88%	268
No.Reformulating functions	2	7%	23	22%	25	22%	8	29%	13	18%	21	21%	12	12%	2	5%	14	12%	60

Two recurrent combinations were found: "FP+pause" and "prolongation+FP", which are the most frequent ones (produced 11 and 8 times). In L2 speech, however, our results include 18 combinations in total: 6 combinations of 2 forms, 7 combinations of 3 forms, 4 combinations of 4 forms and 1 combination of 5 forms. No recurrent combinations were found. Hesitations produced in the L2 have greater complexity. Table 5 shows that in 63% of cases, hands tend to be in rest position while the speaker produces the hesitation (209 out of 330). This suggests that speakers tend not to gesture when they hesitate. This is consistent with previous studies (Christenfeld, Schachter & Bilous, 1991).

However, in cases when they do gesture, they sometimes produce a gesture unit (13%) or their gesture tends to be held, interrupted or return to rest position at the same time as the hesitation (20%). Such interruptions indicate a suspension from the speaker both in verbal and non-verbal modalities. Speakers momentarily retreat from the interaction to gaze away (82% of the time) and think.

Figure 1 shows that the speaker produces a complex hesitation characterized by the prolongation of the vowel 'al' in the adjective 'traditional' and a non-lexical sound (a click). While she hesitates, her hands simultaneously return to rest position and she looks down. When she retrieves the noun phrase 'Christmas dinner', she opens her palms and gazes back at her interlocutor.

Table 4. Examples of complex combinations in L1 and L2

	Example	Combination
L1 French	e:et &euh	prolong+FP
	&eum (...)	FP+pause
L2 English	a:aand yea:ah &m (...) [click]	prolong+prolong+nl-sound+pause+nl-sound
	a:and the:e &um	prolong+prolong+FP

Table 5. Visual features accompanying hesitation.

	L1 Fr.	L2 En.	L1 En.	Total
gesture unit	7	19	18	44
hold	3	10	18	31
rest position	102	59	48	209
return to rest position	2	8	24	34
preparation phase	0	7	4	11
interrupted	0	1	0	1
head shake	2	1	2	5
tilt	1	2	2	5
head nod	0	1	0	1
head down	0	1	0	1
eyes closed	2	4	5	11
frowns	1	7	4	12
winces	2	0	5	7
Gaze away	90	82	101	273
Gaze toward interlocutor	24	22	11	57

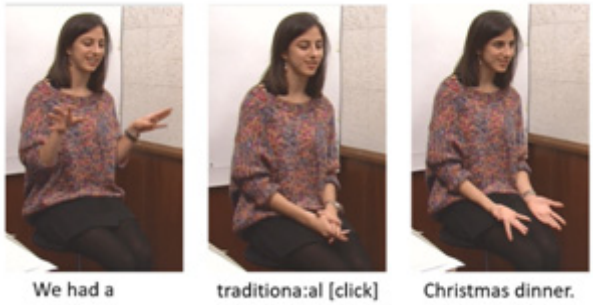


Figure 1. Multimodal activity during hesitation

Conclusions

This preliminary study was conducted in order to test new methods for analyzing vocal hesitations in L1 and L2 speech. Our results indicated that the frequency of hesitations did not differ greatly between L1 and L2 productions.

However, L2 learners seemed to require more time for planning in their L2 than in their L1, as their hesitations tended to be longer and showed greater complexity. This could suggest that hesitations did not result from speech processing difficulty, but rather helped speakers to structure their speech. Another interpretation would be that French speakers aligned their hesitations with American speakers as they produced longer hesitation markers in their L2 (see Finlayson, Lickley & Corley, 2010).

Speakers also often disengaged from interaction when hesitating, as they tended to avert their gaze from their interlocutor as a way to suppress the environment's control over cognition (Glenberg, Schroder & Robertson, 1998). They also did not tend to gesture often while hesitating, although a temporal connection between gesture suspension and speech suspension was found.

Our preliminary results are promising for conducting further analyses on vocal hesitations in L1 and L2 speech and their relation to gesture. Further research will be carried out on more subjects from the whole corpus in order to extend this analysis.

Acknowledgements

Many thanks to Céline Horgues, Maria Candea and Gaëtanelle Gilquin for their interest in our study and for reading a draft of this paper, and to our anonymous reviewers for their constructive input.

References

Barr, D. J. 2003. Paralinguistic Correlates of Conceptual Structure. *Psychonomic Bulletin & Review* 10 (2): 462–467.

Betz, S. & P. Wagner. 2016. Disfluent Lengthening in Spontaneous Speech. *Elektronische Sprachsignal-verarbeitung (ESSV) 2016*.

Christenfeld, N., S. Schachter & F. Bilous. 1991. Filled Pauses and Gestures: It's Not Coincidence. *Journal of Psycholinguistic Research* 20(1):1–10.

Crible, L., L. Degand & G. Gilquin. 2017. The Clustering of Discourse Markers and Filled Pauses. *Languages in Contrast* 17(1):69–95.

Fehringer, C. & C. Fry. 2007. Hesitation Phenomena in the Language Production of Bilingual Speakers: The Role of Working Memory. *Folia Linguistica* 41(1–2):37–72.

Finlayson, I., R. J. Lickley & M Corley. 2010. The influence of articulation rate, and the disfluency of others, on one's own speech. In *DiSS-LPSS Joint Workshop*, 25–26 September 2010, Tokyo, Japan, 119–122.

Glenberg, A. M., J. L. Schroeder & D. A. Robertson. 1998. Averting the Gaze Disengages the Environment and Facilitates Remembering. *Memory & Cognition* 26(4):651–658.

Goodwin, C. & M. H. Goodwin. 1996. Seeing as a Situated Activity: Formulating Planes. In D. Middleton & Y. Engestrom (eds.), *Cognition and Communication at Work*, edited by. Cambridge: Cambridge University Press.

Guaïtella, I. 1993. Functional, Acoustical and Perceptual Analysis of Vocal Hesitations in Spontaneous Speech. In *ESCA Workshop on Prosody*.

Holmes, V. M. 1988. Hesitations and Sentence Planning. *Language and Cognitive Processes* 3(4):323–361.

Horgues, C. & S. Scheuer, S. 2015. Why some things are better done in tandem? In J. A. Mompeán & J. Fouz-González (eds.), *Investigating English Pronunciation: Current Trends and Directions*. Basingstoke and NY: Palgrave Macmillan, 47–82.

Jucker, A. H. 2015. Uh and Um as Planners in the Corpus of Historical American English. *Developments in English: Expanding Electronic Evidence*, 162–77.

Kjellmer, G. 2003. Hesitation. In Defence of Er and Erm. *English Studies* 84(2):170–198.

Merlo, S. & L. L. Mansur. 2004. Descriptive Discourse: Topic Familiarity and Disfluencies. *Journal of Communication Disorders* 37(6):489–503.

Riazantseva, A. 2001. Second Language Proficiency and Pausing – A Study of Russian Speakers of English. *Studies in Second Language Acquisition* 23(04):497–526.

Rieger, C. L. 2003. Disfluencies and Hesitation Strategies in Oral L2 Tests. In R. Eklund (ed.), *Proceedings of DiSS 2003*, 5–8 September 2003, Göteborg University, Sweden, Gothenburg Papers in Theoretical Linguistics 90, 41–44.

Schober, M. F., F. G. Conrad, W. Dijkstra & Y. P. Ongena. 2012. Disfluencies and Gaze Aversion in Unreliable Responses to Survey Questions. *Journal of Official Statistics* 28(4):555.

Seyfeddinipur, M. 2006. *Disfluency: Interrupting Speech and Gesture*. MPI-Series in Psycholinguistics.

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

Swerts, M. & E. Krahmer, E. 2005. Audiovisual Prosody and Feeling of Knowing. *Journal of Memory and Language* 53(1):81–94.

Tavakoli, P. 2011. Pausing Patterns: Differences between L2 Learners and Native Speakers. *ELT Journal* 65(1):71–79.

Tottie, G. 2014. On the Use of Uh and Um in American English. *Functions of Language* 21(1):6–29.

Watanabe, M. & R. Rose. 2012. Pausology and Hesitation Phenomena in Second Language Acquisition. *The Routledge Encyclopedia of Second Language Acquisition*, 480–483.

Phonetic characteristics of filled pauses: a preliminary comparison between Japanese and Chinese

Kikuo Maekawa [1], Ken'ya Nishikawa [1] and Shu-Chuan Tseng [2]
[1]National Institute for Japanese Language and Linguistics, Tokyo, Japan
[2]Institute of Linguistics, Academia Sinica, Taipei, Taiwan

Abstract

Filled pauses in spontaneous Chinese and Japanese were analyzed to examine if there is systematic phonetic difference between the vowels of filled pauses and those occurred in ordinary lexical items. Also, the effect of the category of filled pauses (simple vocalic fillers versus fillers derived from demonstratives) was examined in both languages. Random forests analysis revealed that it was possible to construct automatic classifiers that achieved F-measure values of .7-.9. It turned out also that, in both languages, vowels in simple vocalic filled pauses showed higher F-values than the filled pauses derived from demonstratives. Lastly, it turned out that acoustic features distinguishing filled pauses from ordinary lexical items differ depending on both the category of filled pauses and languages.

Aim of the study

Filled pauses (FP hereafter) transmit various pragmatic/para-linguistic information, but at the same time, there are strong phonological constraints in the form of FP. It is accordingly expected that sub-phonemic phonetic details and voice quality play important roles in the information transmission by FP. It was reported recently for Japanese that there is systematic difference of voice quality between the vowels in FP and those in ordinary lexical items like nouns and verbs (LX hereafter), and it was possible to construct automatic classifiers of significant performance for the task of FP-LX vowel classification (Maekawa & Mori, 2015, 2016). The aim of the current paper is two-folds. First, to examine if similar conclusion could be obtained in language other than the Japanese, and second, to examine if the conclusion differs depending on the category of filled pauses.

Data

Two corpora of spontaneous speech were used for analysis. The core part of the Corpus of Spontaneous Japanese (CSJ-Core, 505455 words and 984092 morae spoken by 139 speakers; see Maekawa (2003) for details) and Mandarin Conversational Dialogue Corpus (MCDC8, 93533 words and 136229 syllables spoken by 16 speakers, see Tseng (2014) for details). An important difference between these corpora is that CSJ is a corpus of monologue, while MCDC8 is a corpus of dialogue. See conclusion on this issue.

In both Japanese and Chinese, the simplest and typical FP consists of a single vowel. In Japanese, all five vowels are used as FP but with considerably different frequencies. In Chinese, schwa vowel is used exclusively for simple fillers. Moreover, in both Chinese and Japanese, there are FPs that are morphologically derived from demonstratives. In Japanese, /ano/ ('that') and /sono/ ('this') are such FP. In Chinese /nà nà ge/ ('that', 'that+classifier') and /zhè zhège/ ('this', 'this+classifier') are used as FP (Zao & Jurafsky 2005). In the rest of this paper, these two FP categories will be referred to as simple-fillers (SF) and demo-fillers respectively.

In the Chinese corpus, there were 104 simple-fillers vowel (/ə/), 244 demo-filler vowels of /ə/, and 475 demo-filler vowels of /a/. Note only monophthong (excluding diphthongs and filler-like particles) were analysed in this study for the sake of comparison with the Japanese vowel.

Japanese corpus had 3450 simple-filler vowels of /eH/ (long /e/ vowel), 136 simple-filler vowels of /aH/, 1519 demo-filler vowels of /a/ (derived from /ano/) and 252 demo-filler vowels of /o/ (derived from the first syllable of /sono/).

Acoustic analysis

Acoustic features listed in Table 1 were computed for each vowel using Praat (Boersma & Weenink, 2013). Some features were Z-transformed using the mean and SD of each speaker. H1, H2 and A3 values were corrected by the method of Hanson (1997). Note only those vowels having at least ten cycles were selected for analysis. Vowels having one or more missing values in acoustic features were also excluded.

Tables 2 and 3 summarizes the results of a series of t-tests applied to the acoustic features of Table 1, the null hypothesis being no difference of means between the LX and FP vowels.

Note the test was applied to a data set consisting of 100 samples of LX vowels and 100 samples of FP vowels chosen randomly from the corpus. These data sets were prepared for random forests analysis presented in the next section.

Table 1: List of acoustic features.

FEATURE	GLOSS
TL	Spectral tile [dB] estimated from cepstrum
Jitter	Mean Jitter (PPQ5) [%]
Shimmer	Mean shimmer (APQ5) [%]
AutoCorr	Mean autocorrelation
Harm2noise	Mean harmonics to noise ratio
FOLZ	Mean Z-value of log10 F0 in Hz
F1LZ	Mean Z-value of log10 F1 in Hz
F2LZ	Mean Z-value of log10 F2 in Hz
F3LZ	Mean Z-value of log10 F3 in Hz
IntensityZ	Mean Z-value of intensity in dB
Duration	Duration [sec]
H1*-H2*	Difference of the first and second harmonics [dB]
H1*-A1	Difference of the fisrt harmonics and level of F1
H1*-A2	Difference of the fisrt harmonics and level of F2
H1*-A3*	Difference of the fisrt harmonics and level of F3

Table 2: T*-test of acoustic features, Japanese vowels.*

FEATURE	SF/aH/	SF/eH/	Demo/a/	Demo/o/
TL	n.s.	- LX<FP	* LX<FP	** LX<FP
Jitter	* LX<FP	n.s.	n.s.	n.s.
Shimmer	n.s.	n.s.	n.s.	n.s.
AutoCorr	- LX>FP	n.s.	*** LX<FP	* LX>FP
Harm2noise	n.s.	n.s.	*** LX<FP	* LX>FP
FOLZ	*** LX>FP	n.s.	n.s.	*** LX>FP
F1LZ	* LX<FP	** LX<FP	n.s.	n.s.
F2LZ	** LX<FP	n.s.	n.s.	n.s.
F3LZ	n.s.	n.s.	n.s.	n.s.
IntensityLZ	*** LX>FP	*** LX>FP	*** LX>FP	*** LX>FP
Duration	*** LX<FP	*** LX<FP	*** LX>FP	* LX>FP
H1*-H2*	n.s.	* LX<FP	- LX<FP	* LX<FP
H1*-A1	n.s.	n.s.	*** LX<FP	n.s.
H1*-A2	** LX<FP	n.s.	** LX<FP	n.s.
H1*-A3*	n.s.	- LX<FP	*** LX<FP	** LX<FP

*** p<.001; ** p<0.01; * p<0.05; - p<0.1

Tables 2 and 3 show that FP and LX vowels showed significant differences in many acoustic features, but the set of significant features differ depending on the category of FP. Moreover, it is interesting that the magnitude relationship of a given significant feature can be in opposite direction depending on the FP category. For example, in Table 2, FP vowels are significantly longer than LX vowels in simple fillers, while LX vowels are longer than FP vowels in demonstrative FPs. Inverted relationships can be found in each table and across two tables. We will return to this issue in the discussion section.

Table 3: T*-test of acoustic features, Chinese vowels.*

FEATURE	SF/ə/	Demo/ə/	Demo/a/
TL	*** LX<FP	n.s.	n.s.
Jitter	n.s.	** LX>FP	n.s.
Shimmer	- LX>FP	*** LX>FP	* LX>FP
AutoCorr	n.s.	** LX<FP	* LX<FP
Harm2noise	- LX<FP	*** LX<FP	* LX<FP
FOLZ	n.s.	*** LX<FP	*** LX<FP
F1LZ	*** LX<FP	** LX>FP	** LX>FP
F2LZ	*** LX>FP	** LX<FP	n.s.
F3LZ	n.s.	* LX>FP	* LX<FP
IntensityLZ	* LX>FP	- LX<FP	** LX<FP
Duration	*** LX<FP	n.s.	n.s.
H1*-H2*	n.s.	** LX<FP	** LX<FP
H1*-A1	** LX<FP	* LX<FP	n.s.
H1*-A2	n.s.	n.s.	* LX<FP
H1*-A3*	** LX<FP	n.s.	n.s.

Random forests analysis

The results of t-tests in Tables 2 and 3 do not provide direct evaluation on the effectiveness of the features for the classification of FP and LX vowels. Random forests analysis was used to examine this issue. Random forests is a machine learning technique to construct statistical classifier like the support vector machines. A crucial difference from the support vector machines is that random forests provides information on the contribution of features used for learning. The RandomForest package (Ver. 4.6-12) of the R language (Ver. 3.3.1) was used for computation. The data sets were the same as the ones used in the previous section.

Table 4 and 5 summarize the performances of random forests analyses. Note these are the performance of cross-validation. In cross validation, data set was randomly split into two subsets; one of them contained 90% of data (i.e., 180 samples) and was used for training of classifier, and the other set containing the resulting 10% of data (20 samples) was used as a test set. The performance of the classifier was evaluated by the success rate of the classification when it is applied for the test set. This process was repeated 10 times for each class of FP and LX vowels. Numbers in Table 4 and 5 are the means of such repeated cross-validations.

The top 15 rows of Table 4 and 5 stand for the relative contributions of acoustic features. The numbers shown in each cell of the table is

called MDG (Mean Decrease in Gini). MDG shows the decrease in the value of Gini index caused by the exclusion of the predictor variable in question. The greater the MDG, the greater the contribution of the variable. In each FP category, features of top three MDG are shown by shading. Note that these are the mean values of MDGs over 10 repetitions of cross-validation.

Table 4: Results of random forests analysis. Cross-validation of the Japanese vowels. Mead MDG values.

FEATURE	SF/aH/	SF/eH/	Demo/ano/	Demo/sono/
TL	2.25	2.62	3.68	5.60
jitt_ppq5	3.51	4.26	4.35	3.77
shim_apq5	2.37	2.75	3.98	4.40
autoCorr	2.15	3.82	5.44	4.67
harm2noise	1.97	3.63	9.08	4.10
F0LZ	6.28	4.78	7.50	8.25
F1LZ	3.44	5.51	3.94	4.53
F2LZ	3.28	4.06	6.52	9.83
F3LZ	2.70	2.93	5.30	4.08
IntensityLZ	14.14	19.21	8.83	7.05
DurLZ	37.25	24.81	8.37	9.36
H1*-H2*	2.49	3.17	5.48	6.99
H1*-A1	2.07	2.58	6.50	7.08
H1*-A2	3.21	2.98	3.84	4.30
H1*-A3*	2.35	2.31	6.62	5.45
F-measure	0.87	0.89	0.73	0.76

Table 5: Results of random forests analysis. Cross-validation of the Chinese vowels. Mead MDG values.

FEATURE	SF/ə/	Demo/ə/	Demo/a/
TL	2.62	4.20	4.45
jitt_ppq5	4.26	4.46	6.58
shim_apq5	2.75	4.12	4.73
autoCorr	3.82	4.53	3.91
harm2noise	3.63	5.11	5.22
F0LZ	4.78	11.87	13.89
F1LZ	5.51	6.73	5.10
F2LZ	4.06	6.16	3.48
F3LZ	2.93	4.48	6.50
IntensityLZ	19.21	9.78	7.44
DurLZ	24.81	5.35	3.87
H1*-H2*	3.17	7.76	9.87
H1*-A1	2.58	6.88	5.22
H1*-A2	2.98	4.33	5.96
H1*-A3*	2.31	3.63	3.24
F-measure	0.89	0.69	0.72

The last rows of Tables 4 and 5 show F-measure, a commonly used measure of classification performance. This is also the mean over the ten repetitions.

Discussion

In tables 4 and 5, F-measure distributes in the range .69-.89, suggesting the effectiveness of the 15 acoustic features as the predictor variables. There is, however, difference of F-measure between the simple-filler and demo-filler categories in both languages. Vowels belonging to the simple-filler category showed higher F-values (.87-.89) than those belonging to demo-fillers (.69-.76) in both Japanese and Chinese.

In Table 4, as far as the vowels of simple-filler category are concerned, intensity and duration are among the most important. The importance of these prosodic features coincides with the conclusion reported in Maekawa and Mori (2016). In the category of demo-fillers, on the other hand, spectral features like F2 and harmonics to noise ratio make certain contribution to the classification; as the result, the contributions of prosodic features are smaller compared to simple-fillers.

In Chinese (Table 5), the situation is different. Prosodic features of duration is not as important as in Japanese in demo-fillers. F0 made large contribution in demo-fillers as in Japanese, but the influence of the variable is not the same as in Japanese. Table 2 shows that, in Japanese and where F0 showed significant difference, mean F0 is always lower in FP than in LX vowels, while in Chinese F0 is always higher in FP than in LX vowels (see Table 3).

Inter-language dissimilarities like this can be found in other acoustic features as well. When we compare the contribution of intensity in two languages, we found that in Japanese (See Table 2), mean intensity is always lower in FP than in LX vowels, while in the demo-fillers of Chinese, FP showed higher intensity than in LX vowels (see Table 3). In the same vein, autocorrelation and harmonics to noise ratio are lower in FP than in LX in Japanese, but in Chinese, they are higher in FP than in LX. Moreover, features concerning the spectral tilt of voice source (H1*-H2*) made certain contribution in Chinese, while their role in Japanese is limited.

Lastly, changes in formant frequencies between the FP and LX vowels is of some interest. Figure 1 compares the mean first (F1LZ) and second (F2LZ) formant frequencies of seven classes of vowels analysed so far. Rectangles and circles stand for the Japanese and Chinese vowels respectively. It can be seen in Figure 1 that while Japanese vowels do not show large displacement between the LX and corresponding FP vowels, Chinese vowels show larger displacements, especially in the /ə/ vowel.

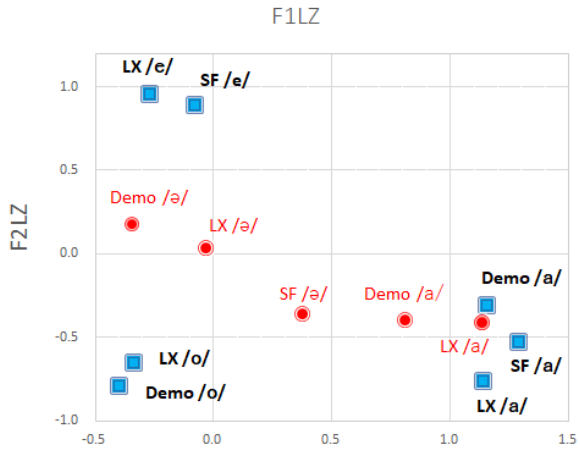


Figure 1: Mean formant frequency of seven vowel classes. Unit is standard deviation.

This difference is due probably to the typological difference of prosody between the two languages.

As some researchers believe, Chinese polysyllabic words have the specification of stress in addition to that of lexical tones (Lee, Tseng & Ouh-Young 1989), while Japanese is a pure pitch-accent language lacking any kind of stress.

To sum up, simple-fillers of Japanese and Chinese, as well as the demo-fillers of Japanese are similar in their behaviour. They are characterized by longer duration, lower intensity and lower F0. On the other hand, demo-fillers of Chinese are very different from all other fillers in that they are characterized by higher F0, higher intensity and they are not longer than LX vowels.

Also, they are characterized by lower shimmer, higher autocorrelation and higher harmonics to noise ratio. These features are characteristic of so-called clear and crisp speech. Our speculation is that this can be partly a result caused by the high-falling tone the lexical counterparts of the demo-fillers originally carry.

Conclusion

The study reported here revealed three new findings.

First, as reported in previous studies dealing with Japanese, phonetic characteristics of the vowels in FP are systematically different from those in LX in Chinese as well. It is possible to automatically classify the vowels by means of random forests classifiers with the mean F-measure of about 0.8.

Second, vowels in simple-fillers are much easier to classify than the vowels in demo-fillers. This tendency is found in both languages.

Third, the substantial phonetic difference between the FP and LX items can be different dependent on both the category of FP and the languages. Especially, the Chinese demo-fillers make a phonetic class that is drastically different from all other filled pauses across two languages.

To conclude, we found both language-independent and language-dependent aspects of the phonetics of FP in the present study. More analysis is needed, however, for the fuller understanding of the issue. Especially, the analyses of Japanese dialogue data and Chinese monologue data are badly needed, although unfortunately, it is currently impossible due to the lack of suitable (i.e. phonetically annotated, large-scale, spontaneous speech) corpora other than the CSJ-Core and MCDC8.

Acknowledgements

This work is supported by the JSPS Kakenhi grants to the first author (26284062). It was also supported by the MOST project grant to the third author (105-2410-H-001-084).

References

Boersma, P. & D. Weenink. 2013. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.21, retrieved from http://www.praat.org/

Hanson, H. 1997. Glottal characteristics of female speakers: Acoustic correlates". *Journal of the Acoustical Society of America*, 101 (1):466–481.

Lee. L.-S, C.-Y. Tseng & M. Ouh-Young. 1989. The synthesis rules in a Chinese text-to-speech system. *IEEE Transactions. Acoustics, Speech, and Signal Processing*, 37(9):1309–1320.

Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its Design and Evaluation. *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 13–16 April 2003, Tokyo, Japan, 7–12.

Maekawa, K. & H. Mori. 2015. Voice-quality analysis of Japanese filled pauses: A preliminary report. Paper presented at DiSS 2015, Edinburgh, Scotland (no page numbers).

Maekawa, K. & H. Mori. 2016. Voice-quality difference between the vowels of filled pauses and ordinary lexical items. *Proceedings of Interspeech 2016,* 8–12 September 2016, San Francisco, USA, 3171–3175.

Tseng, S-Ch. 2014. Chinese disyllabic words in conversation. *Chinese Language and Discourse*, 5(2):23–51.

Zao, Y. & D. Jurafsky. 2005. A preliminary study of Mandarin filled pauses. *Proceedings of DiSS 2005*, 10–12 September 2005, Aix-en-Provence, France, 179–182.

The time course of self-monitoring within words and utterances

Sieb Nooteboom and Hugo Quené
Utrecht Institute of Linguistics OTS, Utrecht University

Abstract

The within-word and within-utterance time course of internal and external self-monitoring is investigated in a four-word tongue twister experiment eliciting interactional word initial and word medial segmental errors and their repairs. It is found that detection rate for both internal and external self-monitoring decreases from early to late both within words and within utterances. Also, offset-to-repair times are more often of 0 ms in initial than in medial consonants.

Introduction

This paper is about the time course of both prearticulatory and postarticulatory self-monitoring within words and within utterances. We derive and test a few predictions on detecting and repairing segmental errors in different positions in the word and in the utterance. We make the following assumptions on the processes of internal and external detection of segmental errors. These assumptions are taken from the computational model described by Hartsuiker and Kolk (2001) plus a few modifications as suggested by Nooteboom and Quené (2017).

A list of phases in generating segmental speech errors and detecting these errors by self-monitoring may look as follows:

1) Lexical selection.
2) Phonological encoding.
3) Selection of motor plan: 100 ms per lexical item.
4) Execution of motor plan: 100 ms per syllable.
5) Parsing encoded form: 100 ms per lexical form.
6) Comparing error and target form 50 ms.
7) Error detected at least 150 ms after phonological encoding is completed.
8) After error detection in internal speech, both an interruption command is issued and executed (150 ms) and a command to repair is issued.

If a segmental error is not detected in internal speech, it may be detected later in overt speech. In that case parsing and comparing start about 50 ms after overt articulation has started. The time gap between internal and external detection of segmental errors is at least 350 ms according to Hartsuiker and Kolk (2001). Nooteboom and Quené, (2017) found this time gap to be roughly 500 ms.

Ad 1). During speech preparation successive lexical items are activated (cf. Levelt, Roelofs & Meyer, 1999). A lexical item sent to phonological encoding to become a lexical form remains active for a few hundreds of ms. This supports correct and not misspoken lexical forms during further processing.

Ad 2). For each lexical item a prosodic frame, specifying stress pattern and slots for segments, is selected, and segments are selected to fill these slots, leading to a lexical form (Levelt, Roelofs & Meyer, 1999). The buffer for phonological encoding may contain more than one lexical form. At this level segments in similar positions (cf. Nooteboom & Quené, 2015) may interact leading to one or more error forms. The moment when the phonological encoding is completed and the lexical form is forwarded will here be called T1.

Ad 3) and 4). These assumptions on timing are taken from Hartsuiker and Kolk (2001). Together they imply that articulation of a two-syllable word starts some 300 ms after T1.

Ad 5) and 6). These assumptions on timing too stem from Hartsuiker and Kolk (2001). Self-monitoring inspects the encoded forms by parsing each form (100 ms per lexical form) and comparing the parsed form with the still active correct target form (50 ms per form). We assume that error form and correct target form can be simultaneously active and in competition (cf. Nooteboom & Quené, 2017), potentially leading to interactions between correct and error segments in comparable positions and thus to segment replacements or articulatory blends as described by Goldstein et al. (2007) and McMillan and Corley (2010).

Ad 7). Because parsing and comparing together take some 150 ms, an error can be detected not sooner than 150 ms after T1. Let us call the moment of error detection T2. We assume that parsing and comparing in search of a speech error is similar to scanning a word form in a silent phoneme detection task as described by Wheeldon and Levelt (1995). They found that phoneme detection in internal speech takes progressively more time going from early to later phoneme positions. Likewise we assume that whereas segmental errors in initial

position can be detected at 150 ms after T1, detection of segmental errors in later positions takes more time. Assuming that the time needed for scanning a lexical form for speech errors is roughly equivalent to speaking time, T2 will fall substantially later after T1 for later segments than for initial segments, but the time interval between T2 and the moment the error segment is actually spoken will be the same for segments in different positions.

Ad 8). At the moment a segmental error is detected in internal speech, a command to interrupt the process of speaking the form containing the error is issued. According to Hartsuiker and Kolk (2001), execution of an interruption command takes 150 ms. As we have seen above, the process of starting articulation after phonological encoding takes some 300 ms. Now we see that the process of error detection (150 ms) plus speech interruption (150 ms) also takes some 300 ms. Assuming that there is noise in the timing of the various processes involved, it follows that after a segmental error is detected in internal speech, the process of speaking the form containing the error can be interrupted before or after articulation has started. For errors in initial position this implies that the distribution of error-to-offset times is incomplete: Cases in which speech is stopped before articulation has started are invisible. For errors later in the word form, this may lead to utterances for example of the form *ba..bakery* where the internal error may have been *bapery*. Here also the internal error is invisible.

During attempts to repair after internal error detection the correct target form is still active in many cases. If so, repairing can be very fast. After external error detection, however, which happens at least some 350 ms later, the chances are that the correct target has been de-activated. If so, planning a repair will be more time-consuming.

Because in the above view of self-monitoring the average moment of interruption is coupled to the position of the error segment in the word, potentially the proportion of observable errors would remain unaffected by the position in the word. But there is reason to suspect that the observed detection rate is indeed affected by the position in the word. Self-monitoring requires attention (Levelt, Roelofs & Meyer, 1999), and during speaking attention is divided over different processes, for example speech preparation, articulation, but also self-monitoring internal speech and self-monitoring external speech (Hartsuiker, Kolk & Martensen, 2005). In scanning a word form for errors, more and more attention will be needed for preparing and speaking the next word as the end of the current word comes nearer. Therefore one would not be surprised to find that less and less attention is paid to self-monitoring going from earlier to later in the word form. If so, this would lead for example to a predicted difference in detection rate between initial and medial consonant errors:

(1) Detection rate for segmental speech errors will be lower for medial than for initial consonant errors.

A similar argument can be made with respect to position of the word in the utterance. From the first word on, the number of possible interactions to be detected increases, decreasing the amount of attention for detecting later speech errors. From this we predict that:

(2) Detection rate for segmental speech errors will decrease with position of the word in the utterance from early to late.

According to Hartsuiker and Kolk (2001) both the interruption and the command to produce a repair are triggered by the error detection and executed in parallel. Therefore a difference between initial and medial consonants in the timing of a repair as measured in offset-to-repair times has already been compensated by the later moment of interruption. For this reason we predict that there is no difference between earlier and later segmental errors in offset-to-repair times:

(3) Offset-to-repair times are equal for initial and medial consonant errors.

A similar argument can be made for segmental errors in earlier and later words in the utterance:

(4) Offset-to-repair times do not depend on the position of the words in the utterance.

Experiment

We have elicited segmental speech errors in tongue twisters, each tongue twister consisting of 4 two-syllable CVCVC words (Shattuck-Hufnagel, 1992). There were 4 conditions, one with only strong-weak (Sw) stress patterns, one with only weak-strong (wS) stress patterns, one with the sequence Sw wS wS Sw, and one with the sequence wS Sw Sw wS. In each condition there were 48 tongue twisters, 24 meant to elicit interaction between initial or medial consonants by consonant repetition (as in *kennis gekkie gele kater* for initial position), and 24 without consonant repetition for that position. Initial and medial positions were controlled for the opportunities for interaction (Nooteboom & Quené, 2015). Thirty participants were asked to speak each tongue twister 6 times, 3 times reading from a screen and 3 times from memory (Shattuck-Hufnagel, 1992). All speech was transcribed and coded as to segmental speech errors, keeping the four word positions and within-word segmental

positions separate. For all repaired single segment word initial and word medial speech errors onset-to-cutoff intervals (from word onset to interruption), error-to-cutoff intervals (from onset of error segment to interruption), and offset-to-repair intervals (from interruption to repair onset) were measured.

We focus on consonant errors in initial and medial positions, because other positions were not controlled for the numbers of opportunity for interaction. We exclude all cases in which specific errors, due to hysteresis, were repetitions of the same error by the same speaker. In the current experiment we wished to elicit segmental errors in different positions in the word and in different words in the utterance. We were predictably punished by much variation in error-to-cutoff times, leading to considerable overlap between internally and externally detected repaired errors. Because of this, there is no possibility to estimate the form of the underlying distributions of internally and externally detected errors, and thus no possibility to separate between internal and external error detection, as was done in Nooteboom and Quené (2017).

The detection rates of speech errors are summarised in Figure 1. One may note that the total number of single segment errors is lower for medial than for initial position.

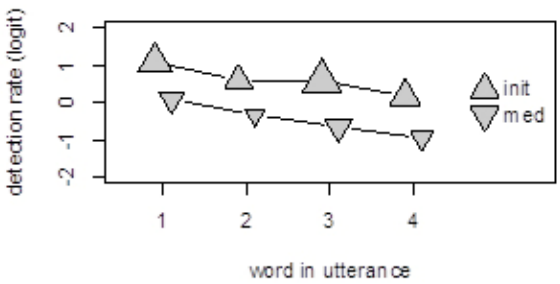


Figure 1: Detection rates of single-segment errors, broken down by word position and by within-word position of the error. Symbol size corresponds to numbers of errors in each cell.

Binomial detection status of each valid error was analysed by means of a GLMM (Quené & Van den Bergh, 2008), with position (initial vs medial) and word number (one to four) as two fixed factors, with response number (1,4,5,6 vs. 2,3) as an additional fixed factor, and with participants and items as random intercepts. Position and word number were also added as random slopes over participants. Interactions were dropped from the GLMM as these did not improve the model fit, according to Likelihood Ratio tests.

The GLMM shows a significant main effect of consonant position, with a significantly lower detection rate for medial than for initial consonant errors (beta=-1.169, Z=-7.1, p<.0001). We also found a significant main effect of word position (beta=−0.320, Z=−4.1, p<.0001): Detection rate decreases within utterances from earlier to later words. We explain the decrease of detection rate from initial to medial consonant error positions, and from earlier to later words in utterances as reflecting differences in attention available for self-monitoring.

We have also measured onset-to-offset times (from word onset to interruption) and error-to-offset times (from onset of error segment to interruption) for all initial and medial segmental errors. Of course, for initial segments onset-to-offset times and error-to-offset times are identical. For medial consonants the moment of interruption on average falls 155 ms later than for initial consonants. However, with Hartsuiker and Kolk (2001) we have assumed that both the command to interrupt speech and the command to plan a repair are triggered by the moment of error detection and executed in parallel. If the interruption after detecting a medial consonant error is 155 ms later than the interruption after detecting an initial consonant error, then also the initiation of generating the repair should be 155 ms later for medial than for initial consonant errors. As the offset-to-repair times were measured from the moments of interruption, this difference of 155 ms is already taken into account. Therefore a priori we expect for internally detected errors no difference in offset-to-repair times between initial and medial positions. Also for externally detected errors there is no reason to expect a difference in offset-to-repair times. The distributions of offset-to-repair times are shown in Figure 2.

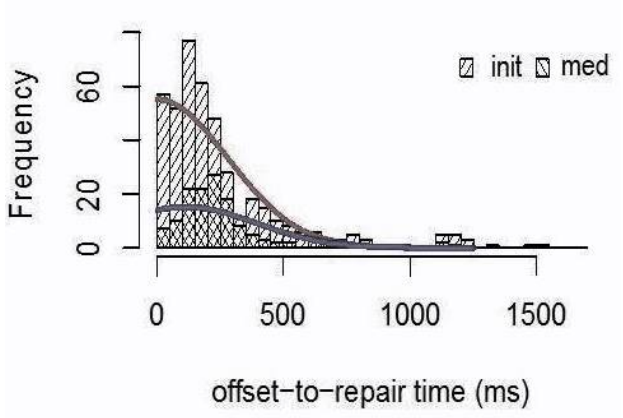


Figure 2: Histograms and fitted truncated gaussian distributions of offset-to-repair times, for initial and medial consonants (see text).

Figure 2 shows that offset-to-repair times are truncated at 0 ms, for errors involving initial (48/426 errors) and medial consonants (5/140 errors). The untransformed cutoff-to-repair times

were therefore analyzed using regression techniques for such truncated distributions (Croissant & Zeileis, 2016). Error times exceeding 1000 ms (3.7% of total) were ignored for this analysis. Since models including word position as a predictor could not be estimated reliably, we focus here on consonant position as the only predictor. Random intercepts and slopes of participants and of item sets were also ignored, as these could not be estimated.

The optimal model for the truncated cutoff-to-repair times showed a non-significant effect of consonant position, according to a Likelihood Ratio test (χ^2=2.72, df=1, p=.099) of the optimal model compared to a null model. For initial consonants, the estimated mean of cutoff-to-repair times is 0 ms [with bootstrapped 95% confidence interval (−210, +65) over 1000 iterations], whereas for medial consonants, the estimated mean cutoff-to-repair time is +113 ms [with 95% confidence interval of (−78,+181)]. It seems possible that there is a difference between initial and medial consonants, but that it cannot be demonstrated due to a lack of statistical power. This is confirmed by a separate analysis of the numbers of immediate (offset-to-repair time of 0 ms) vs non-immediate repair (offset-to-repair time > 0 ms). These were analyzed by means of a GLMM with the same fixed and random predictors as before. The odds of an immediate repair were significantly lower for medial consonants than for initial consonants (beta=-1.237, Z=−2.424, p=.0154), and the odds were significantly higher for words in the third position than for words in the baseline first position (beta=+0.918, Z=2.251, p=.0244). The interaction between consonant position and word position was not significant (p=.326, Likelihood Ratio Test).

The cases in which the offset-to-repair time is 0 ms demonstrate that a repair is ready to be spoken at the moment of interruption. Our results show that this happens significantly more often after word initial consonant errors than after medial consonant errors. This can hardly be explained from the small and insignificant difference in error-to-offset times between initial (217 ms) and medial (198 ms) consonant errors: The time for preparing a repair after error detection and before speech is interrupted is roughly similar for the two consonant positions. Therefore our results suggest that preparing a repair takes more time for medial than for initial consonant errors. Earlier, we have suggested that the amount of attention available for self-monitoring is less for medial than for initial consonants, causing a lower error detection rate for medial than for initial consonant errors. We now suggest that less attention not only lowers detection rate, but also leads to slower error repair.

Conclusion

Main findings are (1) Detection rate of errors is lower in medial than in initial consonants; (2) Detection rate decreases from the first to the last word in four-word utterances; (3) Relatively, there are many more offset-to-repair times of 0 ms for initial than for medial consonants.

Findings 1) and (2) suggest that attention for self-monitoring decreases from early to late both within words and within utterances.

Finding (3) shows that a repair is often available before speech is interrupted, and that repairing medial errors takes more time than repairing initial errors. These results confirm the computational model of Hartsuiker and Kolk (2001) and shed further some light on the time course of self-monitoring.

References

Croissant, Y. & A. Zeileis. 2016. *truncreg: Truncated Gaussian Regression Models*. R package version 0.2-4. Available at: https://CRAN.R-project.org/package=truncreg.

Goldstein, L., M. Pouplier, L. Chen, E. Saltzman & D. Byrd. 2007. Dynamic action units slip in speech production errors. *Cognition* 103:386–412.

Hartsuiker, R. J. & H. H. J. Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology* 42:113–157.

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 monitoring 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, 1–75.

McMillan, C. T. & M. Corley. 2010. Cascading influences on the production of speech: Evidence from articulation. *Cognition* 117:243–260.

Nooteboom, S.G. & H. Quené- 2015. Word onsets and speech errors. Explaining relative frequencies of segmental substitutions. *Journal of Memory and Language* 78:33–46.

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.

Quené, H. & H. Van den Bergh. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. Journal of Memory and Language 59:413–425.Shattuck-Hufnagel, S. 1992. The role of word structure in segmental serial ordering. *Cognition* 42:213–259.

Wheeldon, L. R. & W. J. M. Levelt. 1995 Monitoring the time course of phonological encoding. Journal of *Memory and Language* 34:311–334.

Silent and filled pauses and speech planning in first and second language production

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

Abstract

The present study looks at the relative association of silent and filled pauses to problems in discourse and syntactic planning via utterance and clause boundary phenomena, respectively, by focusing on crosslinguistic data. The occurrence of boundary pauses in a crosslinguistic corpus of speech suggests that silent pauses are more closely related to both discourse and syntactic planning than filled pauses, but more strongly so for discourse planning. These results were consistent across both first and second language production. However, clause boundary silent pauses in first language speech were more atypical (i.e., longer than average) than those in second language speech. This difference may be due to complexity differences in the first and second language speech samples.

Introduction

Two kinds of pauses are frequently examined in disfluency research (cf., Goldman-Eisler, 1961; Rochester, 1973; Duez, 1982; Campione & Véronis, 2005; inter alia). Silent pauses are periods of non-articulation by the speaker. However, not all silences are necessarily disfluencies since speakers may also take short pauses for the sake of breathing or other articulatory functions (Boomer & Dittmann, 1962). Thus, researchers frequently take some sort of lower-bound threshold as the definition of a silent pause, a typical threshold being around 250–300ms (cf., De Jong & Bosker, 2013). Filled pauses, on the other hand, are periods of articulation of non-propositional content and fitting a language-specific convention (Mahl, 1956; Clark & Fox Tree, 2002). In English, for instance, filled pauses are typically 'um' and 'uh' and similarly in other languages (e.g., German, French and Portuguese). But somewhat more phonemically complex forms can be observed in other languages such as Spanish 'este', Chinese 'nage'/'zhege' and Japanese 'e-to'. A comprehensive model of speech production needs to account for the occurrence of silent and filled pauses and observed differences between them.

One of the most well-known models of speech production (Levelt, 1983, 1989) partitions the production process into three domains or levels: conceptualization, formulation and articulation. At the lowest level, the speaker organizes the concepts to be communicated to the listener. Once organized, the speaker then formulates these concepts into a linguistic message according to the grammar and lexicon of the language. Finally, this message is articulated by the speaker for the listener's audition.

One assumption that could be made based on this tri-level model (and which will be made in this paper) is that problems with discourse-level planning are associated with problems at conceptualization while problems with syntactic-level planning are associated with problems at formulation. If utterance boundaries are assumed to be the domain of discourse planning and clause boundaries as the domain of syntactic planning, then disfluencies at these boundaries would reflect processing problems at these respective levels. Then, it becomes possible to explore what kind of disfluencies are more closely associated with discourse or syntactic processing problems. This is the approach taken in the present paper.

In particular, this paper looks at the occurrence of silent and filled pause disfluencies at utterance and clause boundaries in both first and second language speech production. Prior work (Watanabe, Kashiwigi & Maekawa, 2015) has shown that the duration of silent but not filled pauses is longer at major boundaries (i.e., comparable to utterance boundaries) than at minor boundaries (i.e., comparable to clause boundaries) in Japanese unscripted monologues. The present work seeks to build on this finding by comparing pausing patterns in first and second language speech. Those speaking in their second language—particularly those with lower proficiency in the second language—are surely more likely to have syntactic processing difficulties due to some inadequacies in their second language knowledge, but they should have less difficulty with discourse processing because the underlying conceptualization process should be similar to that in their first language. But the pausing strategies they use in their second language may differ from their first. The following study examines this possibility using a crosslinguistic speech corpus.

Methods

Materials

The present study makes use of a crosslinguistic corpus of speech which contains speech samples from native speakers of Japanese speaking in both

their first language (L1) as well as in English, their second language (L2). The corpus (see Rose (2013) for details) contains speech recordings in response to three elicitation tasks—reading aloud, picture description and topic narrative—in both L1 and L2 for each speaker. Full transcriptions and timing interval information (in Praat TextGrid format; Boersma & Weenink (2013)) was available. For the present research, though, only the spontaneous speech samples (i.e., excluding reading aloud) were used, comprising 4 hrs 34 mins in Japanese (L1) and 4 hrs 41 mins in English (L2). The corpus also contains a scalar estimate of each speaker's L2 proficiency. This was used to split the speakers into two broad proficiency groups: low and high.

Procedure

The timing information in the corpus was aligned with the transcripts. Utterance boundaries are already included in the corpus transcripts delimiting stretches of speech which are marked prosodically by definitive falling intonation at the end. Clause boundaries were marked as the initiation or completion of a syntactic clausal unit. All utterances were considered as consisting of at least one clause, even in the rare cases when an utterance consisted only of a noun phrase. A clause was still marked in those cases because some syntactic processing must have taken place in the formulation of the utterance for articulation.

Some complexities of natural speech necessitate some further explanation about clause boundary marking. Because of such things as repairs and restarts, there are clause start boundaries that do not have a corresponding end boundary, and alternatively, some clause starts which have multiple end boundaries. All of these clause starts are included as potential candidate sites for analysis because each start or restart represents some unique syntactic processing event or problem.

For each utterance or clause boundary, it was determined whether there was an immediately preceding silent or filled pause and the duration of this pause as well as its type (silent or filled) was recorded. For silent pauses, only pauses longer than 250 ms were considered (cf., De Jong & Bosker, 2013), while no lower threshold for filled pauses was used.

One simplification here is the case when a silent and a filled pause co-occur before a boundary—a not uncommon case: sometimes there is a filled pause followed by a silent pause, while other times there is a silent pause followed by a filled pause. In these cases, the pause that was more proximate to the boundary was selected for analysis. This was done under the assumption that the more proximate pause is a more accurate reflection of the processing difficulty being endured by the speaker at the utterance or clause boundary.

These data were analyzed using linear regression modeling (`lm` in `R`) and using an alpha level of 0.05.

Results and analysis

Although the corpus contains recordings for 35 speakers, one speaker had no L2 competence estimate and is therefore excluded from further study here. Of the remaining 34, three speakers used almost no filled pauses in one or more of their speech recordings. These three were also removed from the statistical analysis. The remaining N=31 speakers produced a total of 1291 utterance and 3018 clause starts in Japanese and 1024 utterance and 2313 clause starts in English. They also produced a total of 5391 silent and 1951 filled pauses in Japanese and 6068 silent and 1462 filled pauses in English. Table 1 shows detailed data on the relative proportion of these pauses at boundaries and Figure 1 shows detailed data on their duration.

Table 1. Proportion of utterance and clause boundaries with pauses by pause type, language and L2 (English) proficiency (all numbers percentages).

		Utterance boundaries		Clause boundaries	
		Silent pauses	Filled pauses	Silent pauses	Filled pauses
High	Japanese	46.8	40.8	29.8	30.9
	English	67.1	18.4	37.9	16.2
Low	Japanese	56.6	33.1	33.5	23.1
	English	66.0	23.3	47.8	19.4

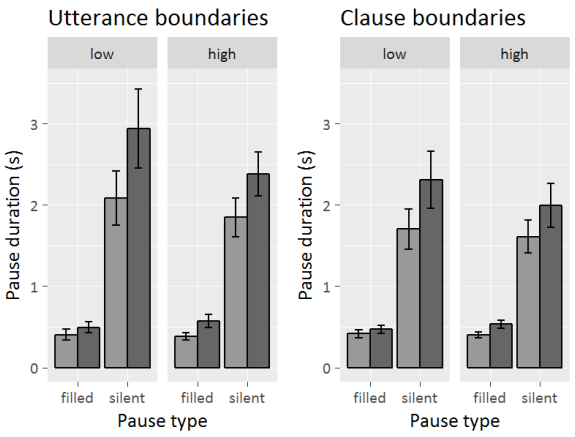


Figure 1. Pause duration at boundaries by L2 proficiency group and language (light gray = Japanese, dark gray = English). Error bars show standard error of the mean.

A linear regression using proportion of boundaries with pauses as the dependent variable and pause type (silent or filled), language (Japanese or English) and L2 proficiency level (low or high) as independent variables was performed for both the utterance boundary and clause boundary data. The optimal models for both boundary types showed that only pause type was a significant variable: for utterance boundaries, $F(3, 236) = 25.7$, $p<0.001$, Adj. $R^2 = 0.24$ and for clause boundaries, $F(3, 236) = 10.2$, $p<0.001$, Adj. $R^2 = 0.10$. In short, these speakers are more likely to produce silent pauses than filled pauses at *both* utterance and clause boundaries. However, the data as shown in Table 1 makes it clear that this trend is stronger at utterance boundaries than at clause boundaries: While well more than half of utterances have a silent pause at the start, rarely more than one third of clause starts do.

A similar analysis of the durational data yields several interesting results worth commenting on. Linear regressions of the two boundary types yield similar results, showing that both pause type and language are significant factors in determining the duration of pauses at boundaries: for utterance boundary starts, $F(3, 248) = 48.5$, $p<0.001$; Adj. $R^2 = 0.36$ and for clause boundary starts, $F(3, 258) = 53.6$, $p<0.001$; Adj. $R^2 = 0.38$. In terms of raw timing, pauses (both silent and filled) are longer in English than in Japanese, and silent pauses are much longer than filled pauses. However, the language factor may be modulated by general differences between first and second language production. For example, in this corpus, the speech rate in English is significantly slower than in Japanese [$F(1,34) = 161.0$, $p<0.001$]. In order to remove this possible influence, the pause durations were normalized against the overall pause durations in each speech sample. That is, the boundary silent or filled pause durations were normalized against the respective silent or filled pause durations in each speech sample as a whole. As a result, the normalized measure is a number ranging (in theory) between 0 to infinity: pauses with values less than 1 are pauses that are shorter in duration than the speech sample mean, and those with values greater than 1 are longer. Figure 2 shows a revised version of Figure 1 after this normalization.

Linear regressions on the utterance and clause boundary data show divergent results. For utterance boundaries, the optimal model has only pause type as a significant factor: $F(1, 250) = 156.7$, $p<0.001$; Adj. $R^2 = 0.38$. In other words, language drops out of the model and the only difference is between silent and filled pauses generally. Filled pauses have a duration that is close to the mean length of all filled pauses in speech samples while silent pauses are longer than average.

Figure 2. Normalized pause duration at boundaries by L2 proficiency group and language (light gray = Japanese, dark gray = English). Error bars show standard error of the mean.

For clause boundaries, on the other hand, both pause type and language remain as significant factors: $F(2, 259) = 40.9$, $p<0.001$; Adj. $R^2 = 0.23$. Filled pauses—as with those at utterance boundaries—are close to a typical length, but silent pauses are longer than typical. Furthermore, pauses are shorter (relative to their typical length in the speech sample) in English than in Japanese.

One final note is that proficiency level was never a significant factor in any of the linear regressions. Although it occasionally approached a marginal level, it never broke the alpha level threshold.

Discussion

This paper has reported on a corpus investigation to explore whether silent or filled pauses are more closely associated with problems at the discourse planning level (conceptualization) or syntactic planning level (formulation), operationalized here as utterance boundaries and clause boundaries, respectively. In particular, the investigation here has taken a crosslinguistic approach, looking at speech recordings of first and second language speech by native speakers of Japanese. Results show that both utterance and clause boundaries are more likely to have silent than filled pauses immediately preceding them, but also that this trend is stronger for utterance than clause boundaries. This behavior is consistent across first and second language speech production.

In terms of pause duration, the results are somewhat more complex. For both utterance and clause boundaries, silent pauses tend to be longer-than-average silent pauses, while filled pauses are of a typical length. Furthermore, when speaking in English (participants' L2), silent pauses at clause boundaries are closer to average duration than those when speaking in Japanese (participants' native language).

For silent pauses, these results suggest that they are strongly associated with both utterance and clause boundaries, and therefore with discourse planning and syntactic planning, respectively. But the observation that their L2 silent pause patterns are closer to average is somewhat surprising since nonnative speakers might be expected to have more difficulty with speech planning and therefore take more time on the average: That is, planning at utterance and clause boundaries might have attracted a greater share of longer pauses. One possible explanation could result from differences in how speakers approached the tasks in the different languages. If speakers actually planned simpler explanations in English than in Japanese, the syntactic planning task might be somewhat easier in English than in Japanese. This explanation is supported by a post-hoc analysis of the data which shows that the mean clause depth (i.e., the hierarchical depth of embedded subordinate clauses) is greater in Japanese than in English [$F(1,33) = 39.9$, $p<0.001$].

For filled pauses, the results suggest that they are not particularly associated with either utterance or clause boundaries. This result suggests that filled pauses are more generally associated with any sort of processing difficulty, at any level. Or, another possibility is that they are merely associated with very local processing difficulties such as immediate lexical retrieval (cf., Beattie & Butterworth, 1979).

Conclusion

Observations from the corpus in this study suggest that silent pauses are more closely associated with problems at both utterance and clause boundaries than are filled pauses, and thus, by extension, more closely related with discourse and syntactic planning. Further work is necessary to confirm these results, perhaps in a controlled experiment with a production paradigm in which speakers would need to produce speech of comparable complexity in both their first and second languages. Alternatively, assuming that speech perception would be dependent on expectations based on production patterns, a perception experiment in which the duration of silent and filled pauses are experimentally manipulated to be not well-formed would elicit different behavioral patterns from listeners (e.g., in a self-paced reading paradigm).

Acknowledgments

This research is partially supported by Japan Society for the Promotion of Sciences (JSPS) Grants-in-aid for Scientific Research (Project #15K02765).

References

Boomer, D. & A. Dittmann. 1962. Hesitation Pauses and Juncture Pauses in Speech. *Language and Speech* 5:215–220.

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.

Boersma, P. & D. Weenink. 2013. Praat: doing phonetics by computer [Computer program]. Version 5.4.03, retrieved 26 Dec 2014 from www.praat.org.

Campione, E. & J. Véronis. 2005. Pauses and hesitations in French spontaneous speech. *Proceedings of DiSS 2005*, 10–12 September 2005, Aix-en-Provence, France, 43–46.

Clark, H. & J. E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1):73–111.

De Jong, N. H. & H. R. Bosker. 2013. Choosing a threshold for silent pauses to measure second language fluency. *Proceedings of DiSS 2013, 21–23 August 2013*, Stockholm, Sweden, 17–20.

Duez, D. 1982. Silent and Non-Silent Pauses in Three Speech Styles. *Language and Speech* 25(1):11–28.

Goldman-Eisler, F. 1961. A Comparative Study of Two Hesitation Phenomena. *Language and Speech* 4(1):18–26.

Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition* 14(1):41–104.

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

Mahl, G. F. 1956. Disturbances and silences in the patient's speech in psychotherapy. *Journal of Abnormal and Social Psychology* 53(1):1–15.

Rochester, S. 1973. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research* 2(1):51–81.

Rose, R. 2013. Crosslinguistic Corpus of Hesitation Phenomena: A corpus for investigating first and second language speech performance. *Proceedings of the 14th Annual Conference of the International Speech Communication Association* (Interspeech 2013), Lyon, France, 992–996.

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. Paper presented at DiSS 2015, 8–9 August 2015, Edinburgh, Scotland.

Analysis of silences in unbalanced dialogues: the effect of genre and role

Vered Silber-Varod [1] and Anat Lerner [2]
[1]The Research Center for Innovation in Learning Technologies, The Open University of Israel
[2]Department of Mathematics and Computer Science, The Open University of Israel

Abstract

This study examines the diversity of silences in unbalanced dialogues, i.e. dialogues between speakers with different participation levels: responder and reporter. We examined two genres: therapeutic sessions and private dialogues that are based on this responder-reporter structure. When looking at silences versus speech ratios, we found no differences between the genres nor between the roles. However, when grouping the silences by their types: Pauses (intra-speaker silences), gaps (inter-speakers' silences) and silences that occur in the vicinity of speech overlaps, we found that the silence duration of pauses are role dependent in both genres, while the silence duration of gaps were found genre dependent, but not role dependent. Moreover, speech rate was not found genre dependent. It seems that although silences in unbalanced dialogues vary considerably, genre and speaker's role are influential.

Introduction

Silences are dealt here in the context of prosodic characteristics of speech genre and speakers' role in dialogues. We argue that studying silence types in a quantitative manner can shed light on the study of linguistic, paralinguistic and extra-linguistic aspects of silence. This approach of quantitative analyzes of silences goes back as early as 1939 (Ephratt, 2011:2294), and is carried out mainly in order to analyze the ratios of speech and non-speech, in isolation or in relation to personality variables. Indeed, such studies produced quantitative predictions, such as the "constant ratio" between vocalization and silence in spontaneous speech (Crown & Feldstein, 1991) and "standard maximum" silence in conversation (Jefferson, 1989). By taking this challenge, we declare that we will not consider the linguistic content that surrounds the silence (Wooffitt & Holt, 2010), nor will we try to understand other dimensions in silence research, such as the communicative functions of silence (Bruneau, 1973; Ephratt, 2008; 2011), or the interpretation of silence in the pragmatic sense (inter alia, Kurzon, 2007; Sacks, Schegloff & Jefferson, 1974). The examined two types of silences that this research is focused on are inter-speaker gaps and intra-speaker pauses. These two terms were mentioned in (Sacks, Schegloff & Jefferson, 1974), who defined pauses as "intra-turn silence (not at a transition-relevance place)", and gaps as "silence after a possible completion point". While pauses are relevant to cognitive processing and are related to effect, style and linguistic structures (Ramanarayanan et al., 2009; Zellner, 1994), gaps are interactive event and were found influential by different types of turns (Tannen, 2000). As Ephrat (2014) claims, "the most prominent case is silence as a discourse marker for turn taking". (Ephrat, 2014:127). Meaning, silence as a discourse marker for turn taking is a direct communicative act that the speaker operates in order to pass the turn from himself to his interlocutor(s).

In this study, we chose dialogues in which one of the speakers is a "reporter", the one who shares a story or a personal problem, while the other is a "responder", the one who listens and comments, hence – unbalanced dialogues.

The contribution of the present study is in the typology of silence types in dialogues with regard to the speakers who speak before and after the silence (reporter, responder or both together;AKA overlaps). We assume that this typology can shed light on broader research on dialogue styles and genres. For example, on the notion of "who owns the gap" – the preceding speaker, who initiated it, or the following speaker, who broke it? For example, in ANALOR tool (Lacheret-Dujour & Victorri, 2002), the silent pause is attached to the preceding speech turn. We further believe that the current study can augment automatic classification of dialogue styles and identifying speakers' role.

Corpus and method

The examined corpus consists of eight dialogues from two genres of speech. Both genres belong to the spontaneous spectrum: Four private face-to-face dialogues, taken from the Corpus of Spoken Israeli Hebrew (CoSIH) and four therapeutic sessions (for details see Lerner et al. 2016). All participants are fluent Hebrew speakers. All dialogues have the following common feature: in each of them there is one core speaker – the reporter who shares and tells a personal story, and the other participant – the responder, who listens and responds.

The spontaneous dialogues varied in their durations from 7 to 27 minutes. The therapeutic sessions consisted of 22–26.5 minutes.

The segmentation and annotation procedure was carried out manually by a phonetician using PRAAT textgrid tool (Boersma & Weenink, 2015). Each interval was labeled with one of the following four parameters:

Silence (non-speech communicative event): Acoustic silence (labeled by a hashtag - #) in the present study is when both speakers are not involved in any vocal production. According to this definition, inhales, exhales, sniffs, sighs, tutting (tsk sound articulation) and coughs were treated as silences; *Reporter*: The reporter's speech intervals (including hesitations and creaky voice); *Responder*: The responder's speech intervals (including hesitations and creaky voice); and *Overlap*: When the two speakers talk simultaneously.

Figure 1 illustrates the labeling scheme at the bottom tier, which is derived from the top and mid tiers, each of which consists of the annotations ascribed to one of the interlocutors in the dialogue.

We then distinguished between three classes of silences: *Pauses* – refer to intra-speaker silences. Pause's duration minimum threshold was set as 100 milliseconds, following Silber-Varod (2013). *Gaps* – refer to inter-speakers* silences (following Edlund, Heldner & Hirschberg (2009) and Heldner & Edlund (2010)), and *Pausal Interruption Silences* (henceforth, PIS) that interact with simultaneous speech (following Bruneau (1973:28–36) and Ephrat (2014:24)). Lapses ("extended silences at transition-relevance places" (Sacks, Schegloff & Jefferson, 1974:715, n. 26) above 10 seconds were omitted from the calculations. This was done since these cases do not reflect pauses and gaps, as argued in Bruneau (1973).

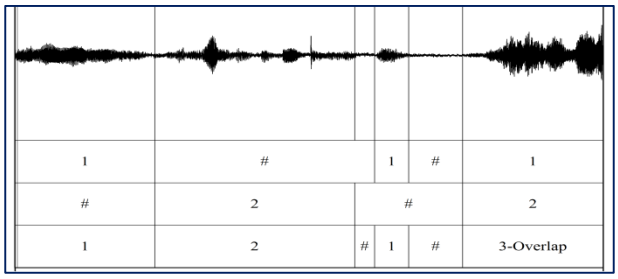


Figure 1. Annotation scheme via three tiers: Reporter-1 (top tier), responder-2 (middle tier) and the derived annotation (bottom tier).

Results

We first measured the relative distribution of the four major interactional parameters in each dialogue (Figure 2). On average, silences compose 35.8% of a dialogue. It is evident that although the therapeutic sessions were taken under specific psychological paradigm, they are varied in terms of the ratios of the relative amount of silence in each dialogue (range from 29% of a session to 47%, similar to the range in the private dialogues).

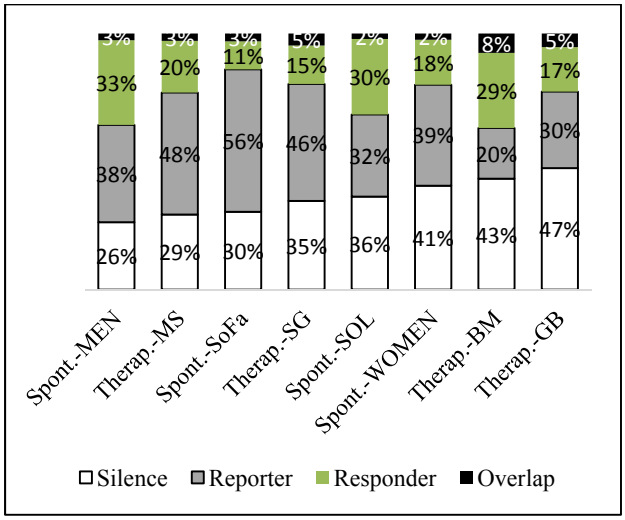


Figure 2: The durational distribution (%) of the four labels in each dialogue: Silences, Reporter's speech, Responder's speech and overlaps.

We then divided the pauses (1421 cases) and gaps (720 cases), and measured their duration. Due to the strongly skewed of the distributions to the left, for the duration of both pauses and gaps, the arithmetic values were transformed to geometric ones, as suggested by Campione and Veronis (2002). This suggests that all statistical calculations, that assume normal distributions, should be performed in the logarithmic domain. To this end, the input to all the t-tests is the natural logarithmic (ln) values of the duration rather than the duration (ms) values. Results show that the duration of most of the pauses is between 300 ms to 500 ms (299 cases). Average pause duration is 1114 ms; Most gaps are between 200 ms to 400 ms (164 cases). Average gap duration is 860 ms. The median pause value is 675 ms and the median gap value is 626 ms. In a *t*-test, the difference between pauses and gaps was found significant ($p = 0.012813$.).

We then tagged all the 2,287 silence intervals below 10 seconds according to the preceding and following participant. Reporter was tagged as 1; Responder as 2; an overlap as 3; and silences by the sign #. The number to the left of the hashtag # represents the participant who breaks his speech and the number to the right represents the one who follows the silence: one of the speakers or an overlap. For example, a silence interval "1#2" means that the reporter (1) spoke before the silence and the responder (2) after the silence (means, this is a gap). This process resulted in nine different

tags, as shown in Table 1. Finding shows that reporter's pauses are the most frequent (954 cases), followed by 467 cases of responder's pauses. Both 1#2 and 2#1 gaps have similar frequency (360 cases), while overlaps rarely occur. As to the distribution among the two genres; Pauses were found in similar distribution – 62%–63%, while gaps were found more in private dialogues (34%) than in therapeutic sessions (29%).

In Table 1 we present the average duration of each of the nine silence types, from the shortest "1#1" pause type (with an average of 1.086 s) to the longest "2#3" PIS type (with an average of 2.388 s). Several differences emerge from the average data: The average duration of reporters' pauses (1#1) is the shortest. The average responders' pause (2#2) is above 400 ms longer than the reporters' pause. The average duration of the 1#2 gap is 100 ms shorter than the 2#1 gap. Pausal interruptions (PIS) are the longest type of silences, except for 1#3 (i.e., reporter # overlap) type. All averages are above the 'tolerance interval' length "of approximately one second" (Jefferson, 1989:170), which is suggested to mark the normatively acceptable length of absence of talk in conversational interaction. Standard deviation values indicate the variance of silences in spontaneous speech. The general finding here is that, the silence length is affected by the event (speaker or overlap) prior to the silence more than by the event that follows the silence: the three cases of 1#? are the shortest; the three cases of 2#? are longer; and the three cases of 3#? are even longer, although the gap 2#3 (representing responder's break and both speakers initiate the next turn synchronically) is the longest.

Table 1. The average (and standard deviation) duration of the nine silence types.

Silence (#) types	Tags	Average duration (seconds) and standard deviation
Pause	1#1	1.086 (sd = 1.417)
Gap	1#2	1.202 (sd = 1.684)
PIS	1#3	1.261 (sd = 3.929)
Gap	2#1	1.368 (sd = 2.464)
Pause	2#2	1.523 (sd = 1.915)
PIS	3#2	1.579 (sd = 1.792)
PIS	3#1	1.720 (sd = 2.117)
PIS	3#3	1.795 (sd = 1.930)
PIS	2#3	2.388 (sd = 4.710)

Next, we compared the duration of the pauses between the two speakers, in the two genres. A *t*-test was carried out in order to measure the significance of the difference between intra-reporters' pauses ("1#1") and intra-responder pauses ("2#2"), and it was found that in both genres, the responder's pauses ("2#2") are significantly different from the reporter's ones ("1#1"): $t = 4.261$; $p = 2.2E\text{-}05$. For Private: $t = 2.657$; $p = 0.008$. For Therapeutic: $t = 3.027$; $p = 0.002$. Due to the unique setting of the therapeutic sessions, where each intern served once as a client and once as a therapist, we measured pause difference between "1#1" and "2#2" of the same speaker in *two different* sessions. For three speakers (S, G and M) results are statistically significant: for S: $t = 2.128$; $p = 0.034$; for G: $t = 5.478$; $p = 0.00001$; and for M: $t = 2.206$; $p = 0.028$. For speaker B, the difference was found not significant ($t = 0.128$; $p = 0.897$).

As to the gap duration between every turn taking – reporter-responder ("1#2") and responder-reporter ("2#1"), the difference was found not significant in both genres together: $t = 0.679$; $p = 0.497$. The result is not significant at $p < 0.05$ in Private: $t = 0.75$; $p = 0.451$, and in Therapy: $t = 1.493$; $p = 0.135$).

Last, we measured the differences of the average gaps and pauses durations for the two genres. It was found that the average durations of gaps are different, while pauses are almost identical in the two genres (Figure 3). The difference between the intra-speaker average pause durations in both genres was found statistically not significant for both roles: For 1#1: $t = 1.552$; $p = 0.120$; for 2#2: $t = 0.336$; $p = 0.736$. While the inter-speaker gaps were found significantly different: For "1#2": $t = 3.589$; $p = 0.0003$; for "2#1": $t = 5.652$; $p < 0.0001$.

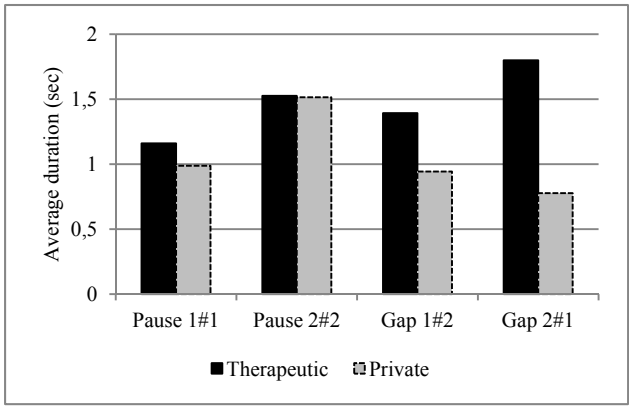


Figure 3. The average duration of gaps and pauses in the two genres.

We further tried to examine if the genre affects the speaking rate. We calculated the speech rates per dialogue in two steps. First, we created the graph of speech intervals' onset time and then ran linear regression to evaluate the speaking rate (as a function of speech intervals' onset rate). The speaking rate was then calculated as the slope of the linear regression evaluator (Table 2). Higher values indicate slower speaking rate (i.e., longer time laps between two speech intervals).

Table 2. Speaking rates of each dialogue (slope of the estimated linear regression).

Dialogue	Speaking rate
SOL	1.818
SG	2.002
MS	2.095
SOFA	2.314
MEN	2.591
WOMEN	2.628
GB	2.894
BM	3.304

Discussion

The present study suggests a basic method to learn about similarities and differences between genres and speaker's role, not only regarding silence types, but also to other parameters of the interaction.

As expected, there are more gaps (indication to turn-takings), in the private dialogues compared to therapeutic session: Gaps distribution ratios were found in private dialogues more than in the therapeutic sessions. On the other hand, pauses distribution ratios were found similar in both genre.

As to the durational distribution, we found the duration of pauses role dependent – reporters' pauses are shorter than responders', in both genres, while the duration of gaps was found genre dependent. These findings suggest that the socio-linguistic aspect of role affects not only speech acoustics (Lerner et al., 2016), but also silence duration. Moreover, speech rate was not found genre dependent. This calls for more inspection on the issue of silence variations, and the linguistics, as well as non-linguistics, parameters that affects silence behavior. As suggested by Jefferson (1989), it seems that although silences in unbalanced spontaneous dialogues vary considerably compared to silences in reading, this does not mean that in spontaneous speech sheer cognitive demands determine the silences. This complex form of silence variability should be further investigated by itself, and by its interface to other prosodic parameters and to speech interval size.

References

Boersma P. & D. Weenink, 2015. Praat: doing phonetics by computer [Computer program]. Version 5.4.06, retrieved 21 February 2015 from http://www.praat.org/

Bruneau, T. J. 1973. Communicative silences: Forms and functions. *Journal of Communication* 23(1):17–46.

Campione, E. & J. Véronis. 2002. A large-scale multilingual study of silent pause duration. *Proceedings of the first international conference on speech prosody (Speech Prosody 2002)*, 11–13 April 2002, Aix-en-Provence, France, 199–202.

The Corpus of Spoken Israeli Hebrew (CoSIH) http://humanities.tau.ac.il/~cosih

Crown, C. L. & S. Feldstein. 1991. The perception of speech rate from the sound-silence patterns of monologues. *Journal of Psycholinguistic Research* 20(1):47–63.

Edlund, J., M. Heldner & J. Hirschberg. 2009, January. Pause and gap length in face-to-face interaction. *Proceedings of Interspeech 2009, 2009*, 6–10 September 2009, Brighton, UK, 2779–2782.

Ephratt, M. 2014. *When Silence Speaks: Silence as Verbal Means of Expression from a Linguistic Point of View*, Magnes Press. (Hebrew)

Ephratt, M. 2011. Linguistic, paralinguistic and extralinguistic speech and silence. *Journal of pragmatics* 43(9):2286–2307.

Ephratt, M. 2008. The functions of silence. *Journal of Pragmatics* 40(11):1909–1938.

Heldner, M. & J. Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38(4):555–568.

Jefferson, G. 1989. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (eds.), *Conversation: An interdisciplinary perspective* (vol. 3), Multilingual matters.

Kurzon, D. 2007. Towards a typology of silence. *Journal of Pragmatics* 39(10):1673–1688.

Lacheret-Dujour, A. & B. Victorri. 2002. La période intonative comme unite d'analyse pour l'étude du français parlé, modélisation prosodiques et enjeux linguistiques. *Verbum* 24(1–2):55–72.

Lerner, A., V. Silber-Varod, F. Batista & H. Moniz. 2016. In search of the role's footprints in client-therapist dialogues. *Proceedings of Speech Prosody 2016 (SP2016)*, 31 May – 3 June 2016, Boston, USA, 400–404.

Ramanarayanan, V., E. Bresch, D. Byrd, L. Goldstein & S. S. Narayanan. 2009. Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation. *The Journal of the Acoustical Society of America* 126(5):EL160–EL165.

Sacks, H., E. A. Schegloff & G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 696–735.

Silber-Varod, V. 2013. *The SpeeCHain Perspective: Form and Function of Prosodic Boundary Tones in Spontaneous Spoken Hebrew*, Lambert Academic Publishing (LAP).

Tannen, D. 2000. "Don't Just Sit There-Interrupt!" Pacing and Pausing in Conversational Style. *American Speech* 75(4):393–395.

Wooffitt, R. & N. Holt. 2010. Silence and its organization in the pragmatics of introspection. *Discourse Studies* 12(3):379–406.

Zellner, B. 1994. Pauses and the temporal structure of speech. In E. Keller (ed.) *Fundamentals of speech synthesis and speech recognition*. Chichester: John Wiley, 41–62.

Author index

< This page intentionally left blank >

ISSN 1104-5787
ISRN KTH/CSC/TMH–17/01-SE